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 del 29 de noviembre de 1976.

Departamento de Electrónica, Sistemas e Informática

Maestría en Sistemas Computacionales

Predicción de egreso y abandono de una ingeniería en una universidad privada mexicana con modelos de curación

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN SISTEMAS COMPUTACIONALES

Presenta: **LEONARDO OMAR BOLAÑOS RIVERA**

Asesor **DRA. GISEL HERNÁNDEZ CHÁVEZ**

Tlaquepaque, Jalisco. mayo de 2025.

AGRADECIMIENTOS

Doy gracias a mi madre y a mi esposa por todo el apoyo y paciencia recibidos por parte de ambas a lo largo de esta nueva meta personal.

A la Dra. Gisel Hernández por la disposición, paciencia y conocimientos transmitidos que me han hecho reflexionar sobre lo mucho que me falta por aprender, pero sobre todo por lo que he aprendido que me servirá en mi vida profesional.

A todos los profesores que me enseñaron algo totalmente nuevo en cada sesión que tuve la oportunidad y privilegio de compartir.

Al ITESO por permitirme trabajar en un proyecto tan interesante de aplicación real con información de la institución.

DEDICATORIA

Dedico este trabajo a mis padres y a mi esposa que siempre han deseado lo mejor para mí y me motivan en mi día a día, aunque mi padre ya no está en este mundo sé que estaría feliz de ser parte de este proceso.

RESUMEN

Esta propuesta tiene como antecedente los modelos predictivos de eventos de historia de vida desarrollados por la Dra. Hernández que utilizan técnicas de regresión semi paramétrica de Cox, modelos de Cox extendidos y de bosques de supervivencia aleatorios (*RSF*). En este trabajo se hizo la comparación entre Modelos de Curación Mixtos (MCM) Paramétricos multivariados que utilizan desde 1 a 8 variables independientes como máximo, las funciones de distribución de los que abandonan que se utilizaron fueron Log Normal, Weibull y Exponencial, con funciones de enlace logit y loglog complementario. En este trabajo se pudieron comparar 1501 modelos de los cuales se seleccionaron 4 modelos con el mejor desempeño y para su comparación se hizo uso de las métricas de AUC (entre 0.725 y 0.734) y AIC (entre 713.874 y 716.671). Se identificaron como predictoras las variables ind_planea, PromPre, CatPAA y EdadIng. Los mejores 10 modelos MCM se compararon con Cox PH y en general los MCM tuvieron un mejor desempeño basado en la métrica AUC, esta comparación se hizo basado en las instrucciones dadas por la Dra. Hernández para utilizar la técnica de regresión semi paramétrica de Cox.

ABSTRACT

This proposal is based on the predictive models of life history events developed by Dr. Hernández, which use semi-parametric Cox regression techniques, extended Cox models, and random survival forests (RSF). This work compared multivariate parametric mixed curation models (MCM) that use from 1 to 8 independent variables at most. The distribution functions of those who dropped out were Log Normal, Weibull, and Exponential, with logit and complementary loglog link functions. This work compared 1,501 models, from which 4 models with the best performance were selected. The AUC (between 0.725 and 0.734) and AIC (between 713,874 and 716,671) metrics were used for their comparison. The variables ind_planea, PromPre, CatPAA, and EdadIng were identified as predictors. The top 10 MCM models were compared to Cox PH and overall, the MCMs performed better based on the AUC metric, this comparison was made based on the instructions given by Dr. Hernandez to use the semi-parametric Cox regression technique.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ACRÓNIMOS Y ABREVIATURAS

AIC	Criterio de Información de Akaike (Akaike Information Criterion)
AUC	Área Bajo la Curva (Area under the curve)
BIC	Criterio de información bayesiano (Bayesian information criterion)
C Index	Índice de Concordancia (Concordance Index)
Cox PH	Modelo de riesgos proporcionales de Cox (Cox Proportional Hazard)
EDA	Análisis de Datos Exploratorio
EPV	Eventos Por Variable (Events Per Variable)
KDD	Knowledge Discovery In Databases
LL	Verosimilitud (Likelihood)
MCM	Modelo de Curación Mixto (Mixture Cure Model)
NMCM	Modelo de Curación No Mixto (Non-Mixture Cure Model)
RMST	Tiempo medio de supervivencia restringido (Restricted Mean Survival Time)
RSF	Bosques de supervivencia aleatorios (Random Survival Forest)
TOG	Trabajo de obtención de grado.

1. INTRODUCCIÓN

Actualmente, se cuenta con modelos de estimación de abandono, cambio de carrera y egreso a nivel de la universidad y de grupos de alumnos (mujeres, hombres, por carreras, por rangos de edades, etc.), así como modelos predictivos a nivel individual. Para estos últimos se han empleado técnicas estadísticas como regresión semi paramétrica de Cox y modelos de Cox extendidos y modelos de aprendizaje de máquina como los bosques de supervivencia aleatorios (RSF). Los modelos se desarrollaron para predecir la deserción de alumnos de las licenciaturas del ITESO. Que incluye los eventos de cambio de carrera y de baja de la universidad. Un problema de esos modelos es que se han censurado a la derecha a los alumnos que egresan, cuando en realidad "se curan", o sea, no son susceptibles de abandonar una vez que han egresado.

Atendiendo al problema relacionado con la censura de los egresados es que se aborda este TOG para construir modelos en que los egresados se consideren "curados". Se empleará una metodología de Minería de Datos con elementos específicos del Análisis de Supervivencia propuesta por la Dra. Hernández en su tesis doctoral y mejorada con los aportes de este trabajo para el caso de los modelos de curación.

1.1. Antecedentes

Esta propuesta tiene como antecedente los modelos predictivos de eventos de historia de vida desarrollados por la Dra. Hernández los cuales pueden ser encontrados en su tesis doctoral “Aplicación de Bosques de Supervivencia Aleatorios a la Predicción de Abandono Universitario”[1] .

1.2. Justificación

Cuando se analizan datos de tiempo hasta el evento, a menudo sucede que una cierta fracción de los datos corresponde a sujetos que nunca experimentarán el evento de interés. Este es el caso de los alumnos que egresan cuando se estudia en abandono de una carrera. Estos tiempos de eventos se consideran infinitos y se dice que los sujetos están curados [2].

1.3. Problema

Los modelos desarrollados hasta ahora para el ITESO, en lugar de considerar a los egresados como curados, los censura por fin de observación, generando modelos con sobre estimación de la probabilidad de abandono y del riesgo.

1.4. Hipótesis

Nuestra hipótesis de investigación es que con los modelos de curación que se construirán en este TOG obtendríamos estimaciones más precisas haciendo uso de alguna métrica como AUC o el índice de concordancia C. Actualmente con Cox se logra un índice de concordancia C de aproximadamente 0.7, mismo que se espera superar.

1.5. Objetivos

1.5.1. Objetivo General:

Desarrollar modelos de supervivencia de curación (cure models) de una ingeniería y comparar la calidad de estos.

1.5.2. Objetivos Específicos:

1. Preparar los datos para poder utilizarlos en los diferentes modelos de curación.
2. Seleccionar mejores modelos de curación.

1.6. Novedad científica, tecnológica o aportación

Actualmente existen pocas implementaciones de modelos para abandono escolar y de los que existen son muy escasos los que utilizan modelos de curación. Por lo que hay mucho espacio para investigación en

este ámbito. Este trabajo aporta información sobre el uso de Modelos de Curación Mixtos Paramétricos multivariados para implementarse en el abandono escolar y el cual busca encontrar alternativas de uso en los lenguajes de programación Python [3] y R [4].

2. ESTADO DEL ARTE

En este capítulo se presenta un resumen de los trabajos relacionados con los modelos de supervivencia, entre los que se encuentran los modelos de curación. Además de trabajos relacionados al abandono escolar con modelos de curación.

2.1. Introducción a Modelos de Supervivencia

Con el siguiente término de búsqueda por modelo de curación "*cure model*" podemos encontrar información relacionada al análisis de supervivencia con aplicación en ciencias de la salud. Incluso haciendo una búsqueda excluyendo términos como: medicina y salud. Los resultados obtenidos son muy relacionados con áreas de la salud, como oncología, etc.

Al buscar información a través de "Google Scholar" en el periodo comprendido del 23 al 25 de noviembre de 2023 sobre trabajos relacionados a el abandono escolar en estudiantes universitarios en los últimos 10 años. Con términos como: "Cure models student drop out", "Cure models student graduation", "Cure models student performance", "Cure models student retention", "Cure models student prediction", "Cure models student prediction dropout", "Cure models student prediction withdraw", "Cure models to prevent student dropout", "Cure models student desertion prediction". Se pudo encontrar información relacionada con otros trabajos sobre estimación de abandono en alumnos de universidad. Pero que usan otros modelos de predicción.

Por ejemplo:

- "Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico". Fue desarrollado por Hernandez Gonzales y colegas [5]. El estudio comparativo está usando 4 algoritmos: regresión logística, agrupamiento, arboles de decisión y redes neuronales.
- "Student Dropout Model Based on Logistic Regression" desarrollado por Cuji Chacha y colegas[6] utilizando regresión logística.
- "Neural networks to predict dropout at the universities" desarrollado por Alban y Mauricio[7] el cual hace uso de redes neuronales con aplicación de algoritmos de perceptrón multicapas y base radial.

Los métodos de curación también pueden ser usados en otras áreas como: Economía y Biología Humana. Otros ejemplos comunes pueden ser encontrados en demografía (tiempo para casarse, tiempo para el primer hijo), mercadotecnia (tiempo para comprar un producto), educación (tiempo para aprender cómo hacer cierta tarea), etc. [8]

En el caso de Economía para el cálculo de riesgos en créditos bancarios o para evaluar costos en algún tratamiento.

Ejemplos:

- "MSc Dissertation An Application of Survival Analysis in Credit Risk Management" desarrollado por [9] el estudio hace uso de modelos de curación mixtos además de otros modelos de análisis de supervivencia para el manejo de crédito.
- "POSC322 The Impact of Treatment Coefficient Parameterization for Mixture Cure Models on Cost-Effectiveness Models" desarrollado por Al-Khayat y colegas [10] hace uso de modelos de curación mixtos para la rentabilidad.

- "Height and marital outcomes in the Netherlands, birth years 1841-1900" desarrollado por Thompson y colegas [11] haciendo uso de métodos de supervivencia y además de modelos de curación en el matrimonio.

Dentro del análisis de datos que involucran el tiempo para que ocurra un evento de interés, frecuentemente sucede que una parte de los datos corresponde a sujetos que jamás experimentaran este evento de interés. Este es el caso de estudiantes que se gradúan. A los modelos de supervivencia que toman en cuenta esta característica se les suele llamar modelos de curación. Es por ello por lo que a los sujetos de estudio se les considera como curados. Los métodos de curación separan la población de datos en dos grupos: los curados y los que son susceptibles del evento de interés [2].

A los modelos de curación también se les conoce como modelos de tasa de curación (The cure rate model) y otra forma de llamar a los sujetos curados es como sobrevivientes a largo plazo (long-term survivors) [12].

Existen 2 principales familias de modelos de regresión de curación: modelos de curación combinada (Mixture Cure Models) y modelos de mejora en el tiempo (Promotion Time Cure Models). Los modelos de curación combinada ofrecen un gran número de enfoques para implementar, desde modelos completamente paramétricos, semi paramétricos, a no paramétricos. Los modelos de mejora en el tiempo usan un enfoque más frecuentista y bayesiano, usando la medición de errores como parte importante de la información[12].

Las principales diferencias entre estas 2 familias de métodos de curación podrían ser descritas de la siguiente forma según Amico y colegas (2017):

- Los modelos de curación combinada no respetan el riesgo proporcional a diferencia de los modelos de mejora en el tiempo .
- Los modelos de mejora en el tiempo pueden ser interpretados en un contexto biológico (Sobre simplificación de un concepto con base en otro, por ejemplo, el crecimiento de un tumor respecto a la cinética tumoral). Además, tienen técnicas bayesianas que han sido desarrolladas para estimar la mejora en el tiempo de un sujeto curado.

Entiéndase por técnicas bayesianas al uso de técnicas estadísticas que usan el teorema de bayes. Thomas Bayes fue un clérigo del siglo XVII, que formuló un teorema matemático que lleva su nombre y que se sigue utilizando ampliamente [13].

En estadística bayesiana se permite y se defiende la utilidad de las probabilidades subjetivas mientras que la estadística tradicional denominada objetivista o frecuentista solo admiten probabilidades basadas en experimentos repetibles y que tengan confirmación empírica [13].

Los modelos de curación han sido ampliados unificando los modelos de mejora en el tiempo (*Promotion Time Cure Models*) y los modelos de curación combinada (*Mixture Cure Models*), ambos modelos están matemáticamente relacionados y en algunas situaciones son incluso equivalentes. A estos modelos se les

conoce como modelos unificados de curación (*Unifying Cure Models*) [2] La unión de ambos modelos curación en uno solo, evita por lo tanto la delicada tarea de elegir entre uno de los dos modelos. [14].

A primera vista los modelos de curación que usan regresión no son más que resultados binarios de curados contra no curados [15].

Patilea y Van (2020) proponen una serie de pasos que pueden conformar una idea general en cómo utilizar los métodos de curación y que podrían ser utilizados como un manual en la implementación de los métodos de curación y no solo verlos como modelos de regresión.

Otra forma común de identificar un modelo de curación es imponer la suposición de que se tiene un seguimiento suficiente (sufficient follow-up). Pero si el tiempo de seguimiento se cree lo suficientemente extenso de forma errónea, la tasa de curación seria sobreestimada y generando conclusiones con resultados posiblemente falsos. Para corregir este problema se puede utilizar la adición de un parámetro llamado índice de valor extremo (extreme value index) que tiene una distribución heavy-tailed y que ofrece mejores resultados [8].

Una distribución heavy-tailed tiene una cola más pesada que una distribución exponencial, esto quiere decir que la distribución heavy-tailed se aproxima más lentamente a cero que la distribución exponencial [16].

Al menos hasta la fecha de redacción de este documento no se encontraron muchos trabajos de investigación que usen métodos de curación para estimación de abandono en estudiantes universitarios. El único ejemplo de trabajo encontrado por el momento es una tesis doctoral que usa métodos de curación. "Modelling Time To Graduation of Durban University of Technology Students Using Event History Analysis" [17].

Bonginkosi [17] en su tesis doctoral, el evento de interés es la graduación y los sujetos que podrían nunca graduarse. Eligiendo a los sujetos curados como aquellos estudiantes que eventualmente no se gradúan. Los objetivos de su estudio los define en 5 puntos:

1. Modelar el tiempo de graduación de 2004 a 2008 en la Universidad Durban.
2. Investigar si hay una relación entre graduación y varias variables explicativas tales como raza, genero, facultad y edad.
3. Modelar un perfil de riesgo para graduarse (risk-to-graduate) en estudiantes, como una función de tiempo para identificar el periodo cuando un estudiante está en mayor riesgo.
4. Investigar la posibilidad de que podría existir una porción de estudiantes que podría eventualmente no graduarse incluso si el periodo fuera extendido.
5. Fraccionar los estudiantes censurados por la cercanía a su periodo de observación en el cual eventualmente se van a graduar y aquellos que eventualmente abandonarán la universidad.

Bonginkosi [17] en su estudio selecciono un modelo de curación mixto (*mixture cure model*), específicamente un modelo discreto: Regresión logística de riesgo (*logistic regression hazard*) y lo comparó contra un modelo de mezcla de riesgo competitivo (*mixture competing risks model*), en este caso extendiendo el modelo de cura a un modelo polinomial y con el cual obtuvo mejores resultados.

Los modelos polinomiales son aquellos que se utilizan para encontrar probabilidades en experimentos donde hay más de 2 resultados [18].

Esto nos deja ver que aún hay mucho campo para la investigación en el ámbito de abandono escolar a nivel universitario con el uso de modelos de curación.

2.2. Aplicación en Abandono Escolar

Existen diversos artículos que tratan de aplicar los métodos de supervivencia para el abandono escolar. Los siguientes son ejemplos de estudios en 2023 que usan métodos de supervivencia:

- "Study on Computer Science Undergraduate Students Dropout at the University of Brasilia" por Mathews y colegas [19] en el que hace uso de un modelo de regresión normal.
- "Stay or Leave? Exploring Student Factors Associated with Dropout Patterns in Massive Open Online Courses" por Shi and Zhou [20] en el que hace uso del modelo proporcional de riesgo Cox (The Cox proportional hazard model).
- "STUDENT DROPOUT IN A BRAZILIAN PUBLIC UNIVERSITY: A SURVIVAL ANALYSIS" por Klitzke y Carvelhaes [21] en el cual utilizó el modelo estadístico de supervivencia de tiempo discreto.

En consultas realizadas entre el 20 de abril y el 10 de mayo de 2024 utilizando los buscadores "Google Scholar", "IEEE Xplore", "Web of Science" y "ELSEVIER". Con frases como "cure model student drop out" entre otras frases que se usaron en una búsqueda anterior en el periodo comprendido del 23 al 25 de noviembre de 2023. Solo "google scholar" y "IEEE Xplore" arrojaron artículos relacionados. "google scholar" arrojo resultados con relación a abandono escolar para estudiantes de preparatoria y otros con relación a estudiantes en línea, pero en este caso se está buscando en alumnos de universidad y en específico con modelos de curación. En el caso de "IEEE Xplore" los resultados arrojados fueron positivos para abandono escolar en universidades, pero haciendo uso de otros modelos de supervivencia.

Hasta este momento el único artículo que contiene información sobre abandono escolar y que hace uso de un modelo de curación es el realizado por Bonginkosi [17] utilizando un modelo de curación mixto (*mixture cure model*).

En el caso de la utilización de métodos de curación en el abandono escolar la literatura es escasa por lo que hay espacio para investigación en este tema.

3. MARCO TEÓRICO/CONCEPTUAL

Aquí se presenta la metodología general de trabajo que se utilizó; se explican de manera general en qué consisten los métodos de análisis de supervivencia y se detallan en particular los modelos de curación aplicados. Además se explican formas de comparación de curvas de supervivencia y métricas para selección de modelos.

3.1. Metodología

La metodología que se utilizó fue KDD (*Knowledge Discovery In Databases*) [22] la cual se centra en el proceso de descubrimiento de conocimiento a partir de datos, cómo se almacenan, se acceden y como los algoritmos pueden escalar a conjuntos de datos masivos y aun así ejecutar eficientemente. Fayyand y colegas lo definen como un proceso no trivial de identificación de patrones de datos válidos, novedosos, potencialmente útiles y en última instancia comprensibles. Las etapas de los procesos de KDD se describen en la Tabla 3-1. En la figura Figura 3-1 se muestra la descripción general de los pasos que componen KDD que es una imagen basada en la figura 1 de Fayyad y colegas.

Fuente: Elaborado de la imagen original de Fayad y colegas [22]

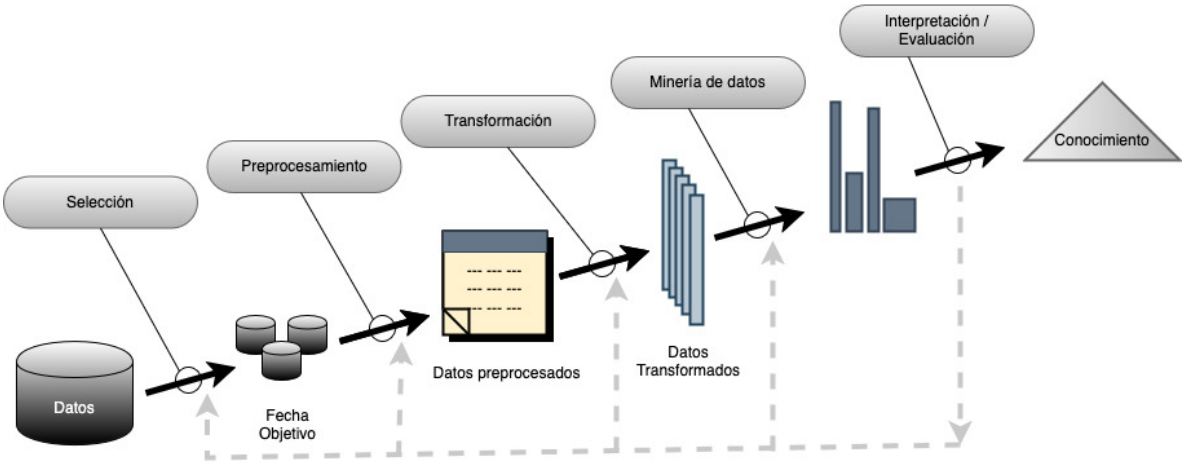


Figura 3-1 Descripción general de los pasos que componen KDD

Tabla 3-1 Etapas de KDD (Knowledge Discovery in Databases)

Etapas de proceso KDD	Descripción
Selección	Identificación de datos relevantes y selección para análisis. • Desarrollo de una comprensión del área para la cual se pretende desarrollar la aplicación, en la cual se identifican los datos relevantes. • Selección de un subconjunto de variables, ejemplos, así como otros subconjuntos para análisis.
Preprocesamiento	Limpieza de datos. Con la finalidad de tener datos confiables, consistentes y evitar errores. • Remoción de datos duplicados.

	• Identificación de datos faltantes y selección de método para manejarlos. • Reducción de ruido en los datos. Dentro de los cuales se encuentras los datos atípicos.
Transformación	Transformación y reducción de dimensionalidad. Conversión de datos a un formato más adecuado para su análisis. • Incluye operaciones sobre los datos como: Normalización, discretización, agregación, reducción de dimensionalidad, compresión de datos, entre otros. • En esta parte se incluye el Análisis Exploratorio de Datos (*EDA*).
Minería de Datos	Componente del proceso en el cual se hace uso de algoritmos para extraer y enumerar los patrones de predicción a partir de datos y la aplicación del análisis de datos. Que tienen las siguientes características clave: • **Validez**: Comprende la evaluación de la calidad de los datos y verificación de los patrones de predicción. • **Novedad**: Hace referencia a la originalidad de la contribución en la investigación. • **Utilidad**: Se refriere a si la información puede ser utilizada para fines prácticos. • **Simplicidad**: Referente a la importancia de mantener los métodos simples antes que algo demasiado complejo.
Interpretación y Evaluacion	Involucra el análisis manual de resultados y la interpretación de estos obtenidos a partir de los datos. • Análisis de los patrones encontrados para identificar lo que se puede considerar nuevo conocimiento. • Evaluación para garantizar que sea útil el conocimiento derivado de los datos.

3.2. Análisis de Supervivencia

En la etapa de Minería de Datos de KDD se aplicaron métodos de análisis de supervivencia. Shewhart y colegas definen al análisis de supervivencia como el análisis estadístico de datos de fallos en el tiempo [23]. Kleinbaum y Klein lo definen como la colección de procedimientos estadísticos para el análisis de datos en el cual la salida de la variable de interés es el tiempo hasta la ocurrencia de un evento [24]. Por evento se refieren a algún acontecimiento que le puede suceder a un individuo, como puede ser la muerte, incidencia de una enfermedad, su recuperación, etc. Klein y colegas señalan que este tipo de análisis también recibe el nombre de análisis de datos del tiempo de vida (*Analysis of Lifetime Data Analysis*) y lo definen como el área de la estadística que trata con el tiempo hasta la ocurrencia de un evento (*time-to-event*), el cual es complicado por la naturaleza dinámica de los eventos que ocurren en tiempo y por los datos censurados [25].

3.2.1. Censura de Datos

La censura es la ocurrencia de un evento de interés, pero del cual no se sabe el tiempo exacto. Las razones más comunes por la que la censura ocurre son 3, el estudio termina antes de que el individuo experimente el evento, se pierde el seguimiento de una persona durante el periodo del estudio, una persona se retira del estudio[25].

La censura por la izquierda ocurre si el individuo es observado que falla antes de algún tiempo determinado pero el tiempo actual de fallo por el contrario no se conoce [23].

La censura por la derecha es la más común en el análisis de supervivencia, cuando el fallo en el tiempo (evento) para un individuo esta más allá del tiempo de censura se dice que esta censurado por la derecha [25].

La censura de intervalo significa que el fallo en el tiempo de la variable de interés se observa o solo se conoce en algún intervalo o ventana en vez de observarse exactamente [25].

De acuerdo con Clark y colegas, los métodos estándar utilizados para analizar datos de supervivencia con observaciones censuradas son válidos solo si la censura es no informativa [26]. Esto ocurre cuando los individuos son censurados debido a una pérdida de seguimiento en un momento determinado, tienen la misma probabilidad de sufrir un evento posterior que aquellos individuos que permanecen en el estudio.

3.2.2. Función de Supervivencia

La función de supervivencia es fundamental para el análisis de supervivencia y es denotada por $S(t)$**.** Esta función da la probabilidad de que una persona sobreviva más allá que algún tiempo especificado t. Esto es que $S(t)$ da la probabilidad de que la variable aleatoria T excede el tiempo especificado $\boldsymbol{t}$ [24].

La función de supervivencia está definida para distribuciones discretas y continuas por la probabilidad de que $\boldsymbol{T}$ excede un valor de t en su rango [23].

3.2.3. Función de Riesgo

La función de riesgo (*Hazard Function*) es denotada por $h(t)$, la cual da el potencial instantáneo por unidad de tiempo para la ocurrencia de un evento, dado que el individuo ha sobrevivido a el tiempo t. La función de riesgo se enfoca en fallos que es el evento por ocurrir, al contrario que la función de supervivencia. Por lo que se le puede considerar el opuesto a la función de supervivencia [24].

3.2.4. Tipos de Modelos de Supervivencia

Un modelo paramétrico de supervivencia es el que su forma funcional se especifica, excepto los valores de los parámetros desconocidos, en el que la distribución de la salida (tiempo hasta el evento) se especifica según los parámetros desconocidos. Muchos modelos paramétricos son modelos de tiempo de fallo de aceleración, el cual provee una medida alternativa a la razón de riesgo (*hazard ratio*) llamado "factor de aceleración" [24].

Los modelos semi-paramétricos de supervivencia involucran una función de riesgo de referencia (línea de base) $h_0(t)$, que es una función no especifica [24]. Este modelo incorpora también un modelo paramétrico $h(X)$ de la relación entre la tasa de error y las covariables (parámetros) X especificadas [23] .

Los modelos no paramétricos son aquellos que no consideran ninguna covariable sino el tiempo hasta la ocurrencia del evento. Entre ellos se encuentran los modelos de Kaplan-Meier [25] y de Nelson Aalen [27].

Estimador "*Kaplan-Meier*" fue una versión moderna de las tablas de vida clásica usado en ciencia actuarial y desarrollado por Edmon Halley en 1693 [25]. Kaplan-Meier también se le conoce como estimador de límite de producto (*Product Limit Estimator*), la función empírica de distribución (*The empirical distribution function*). Es un estimado simple de la función de distribución $F(x) = P\ (X < x)$ y es una manera familiar y conveniente de resumir y mostrar datos. Un argumento de $F\boldsymbol{n}(x)$ contra X visualmente representa un ejemplo y provee información completa sobre los puntos percentiles, la dispersión y las características generales del ejemplo de distribución. Además de ser indispensable en el estudio de la forma de la distribución de la población de donde surge la muestra [23]. El estimado de "*Kaplan-Meier*"

es un estimador no paramétrico el cual puede ser usado para estimar la función de distribución de supervivencia a partir de datos censurados[28].

$$\hat{S}(t) = \prod_{t_j \leq t} (1 - \frac{d_j}{r_j}) \qquad \textbf{Ecuación 1}$$

Donde:

- t es el tiempo.
- d_j es el número de individuos que mueren al tiempo t_j.
- r_j es el número de individuos en riesgo. Por ejemplo vivos y no censurados.

Estimador "*Nelson-Aalen*" es un método no paramétrico que puede ser utilizado para estimar la tasa de riesgo acumulativo a partir de datos de supervivencia censurados [28].

$$\hat{A}(t) = \sum_{t_j \leq t} {}^{d_j}\!/_{r_j} \qquad \textbf{Ecuación 2}$$

Donde:

- t es el tiempo.
- d_j es el número de individuos que mueren al tiempo t_j.
- r_j es el número de individuos en riesgo.

3.2.5. Taxonomía de los Modelos de Análisis de Supervivencia

La Figura 3-2 Taxonomía de los modelos de curación, muestra en azul los métodos en los que este trabajo se enfoca dentro de los métodos de supervivencia. Está figura está basada en la Figura 3 "*Taxonomy of the methods developed for survival analysis*"[29] e información de los modelos de curación de [14].

Fuente: Basado en la imagen "*Taxonomy of the methods developed for survival analysis*" de [29]

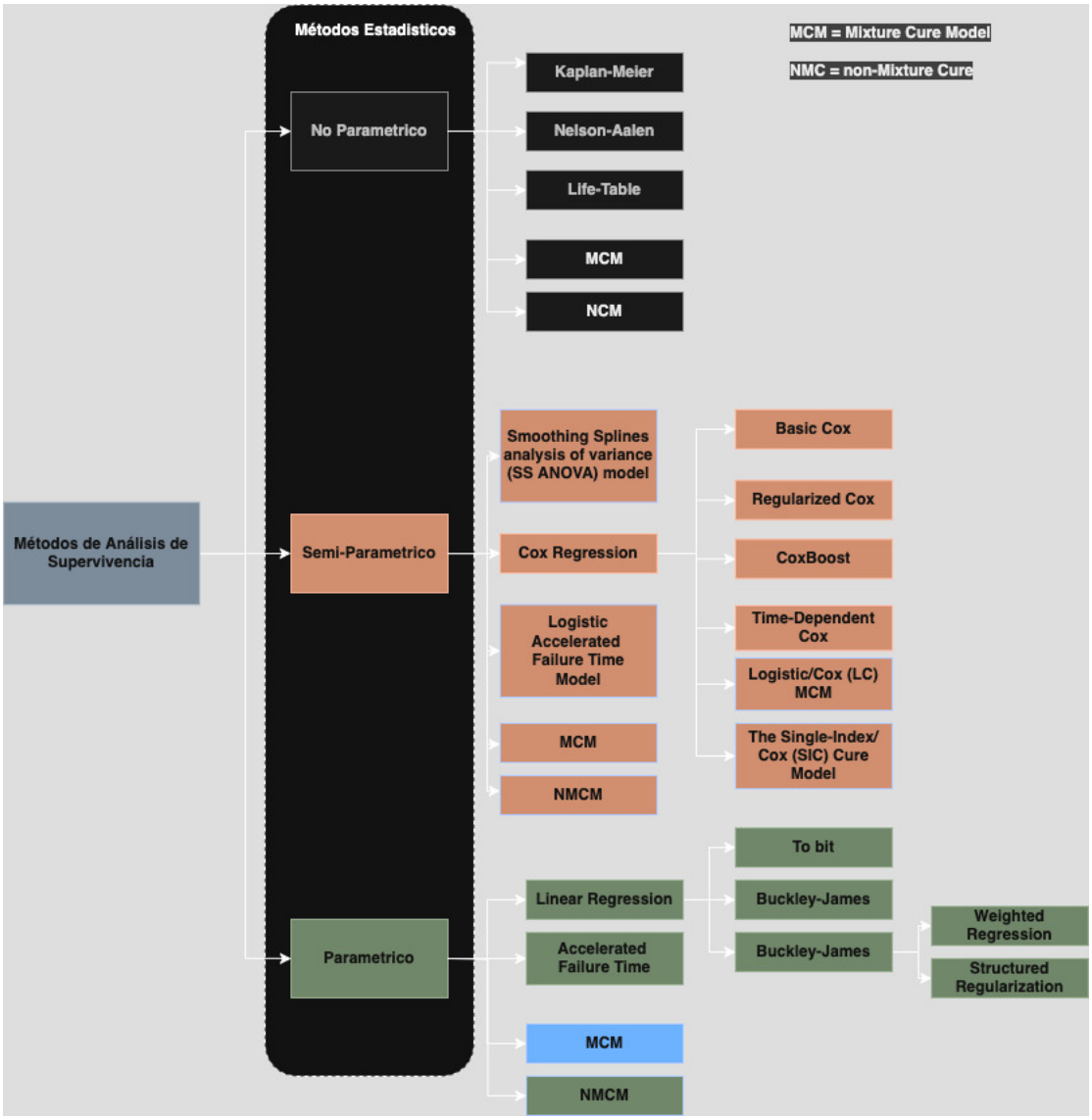


Figura 3-2 Taxonomía de los modelos de curación

3.2.6. Métodos Gráficos y Estadísticos para Comparar Curvas de Supervivencia

Una forma de comparar las curvas de supervivencia de varios grupos de individuos a analizar (por ejemplo mujeres y hombres) es graficando las mismas y analizando sus diferencias.

Otra forma de comparar curvas es utilizando la prueba *log-rank* (Rango logaritmico) es el método más popular para comparar grupos de supervivencia, el cual toma el seguimiento completo en cuenta. Tiene una considerable desventaja que es el no requerir saber nada acerca de la forma de la curva de supervivencia o de la distribución de tiempos de supervivencia. La prueba log-rank es usada para probar la hipótesis nula de si hay diferencia entre poblaciones en la probabilidad de un evento en cualquier punto en el tiempo [30].

3.2.7. Selección de Modelos Predictivos

De acuerdo con [31] hay tres términos importantes respecto a los modelos predictivos los cuales son predicción, modelo y modelado predictivo. Y los define de la siguiente manera. La predicción como el pronóstico de valores futuros en el tiempo que incluyen un punto, un intervalo, una región, una distribución, o una clasificación de nuevas observaciones. El modelo predictivo como cualquier método que produce una predicción sin importar el enfoque usado como paramétrico o no paramétrico etc. Y el modelado predictivo como el proceso de aplicar un modelo estadístico o algoritmo de minería de datos con el fin de predecir observaciones nuevas o futuras.

La selección de un modelo es el proceso que se sigue para decidir de entre varios candidatos disponibles, cuál es el mejor para solucionar un problema particular con base en diferentes criterios como precisión, complejidad, adaptación a los datos (Verosimilitud) etc. A continuación se explican las diferentes métricas aplicadas en este trabajo.

Eventos por Variable (EPV)
Usando una regla de pulgar se sugiere que para un modelo predictivo se requiere un mínimo de 10 eventos por variable (EPV) [32] , lo que significa que sí se tienen un numero de 5 variables se requerirían 50 observaciones con dicho evento. De acuerdo con Ogundimu y colegas un EPV ≥ 20 generalmente elimina sesgos en coeficientes de regresión con predictores de prevalencia baja incluidos en un modelo de regresión de Cox.

Es importante tomar en cuenta los EPV porque esto influye en la cantidad de muestras que se requerirán en nuestro conjunto de datos tanto para entrenamiento como pruebas. De acuerdo con algunos autores se requieren al menos un mínimo de 100 eventos e idealmente 200 o más eventos para el conjunto de pruebas [33].

Logaritmo de Verosimilitud

El concepto de verosimilitud fue introducido por Fisher en 1921 y es una metodología de las más usadas [34]. La verosimilitud (*Likelihood*) es una medida de qué tan bien un modelo, se adapta a los datos. Los logaritmos de verosimilitud hacen la misma actividad y son preferidos por las siguientes razones:

- En estimaciones de máxima verosimilitud es usualmente más sencilla.
- Son más fáciles de graficar.
- Usualmente es más fácil de optimizar.

Para comparar un modelo con logaritmo de verosimilitud, dado que lo que se desea maximizar es la log-verosimilitud el valor más alto es mejor[35].

Véase Ecuación 3 [36].

$$log\ L(\theta|x)\ =\ \sum_{i=1}^{n} log\ f(x_i|\theta)$$ **Ecuación 3**

Criterio de Información Akaike

Akaike Petrov descubrió una relación entre la máxima verosimilitud y la divergencia de Kullback-Leibler y basado en esta información definió el criterio de selección de modelos ahora conocido como Criterio de Información de Akaike (AIC) [37].

La idea detrás del AIC es penalizar la utilizacion de variables adicionales en un modelo [38]. Se define como la adición de 2 veces la cantidad de los parámetros k, al cual se le sustrae 2 veces el valor maximo del logaritmo de la verosimilitud del modelo.

$$\text{AIC}\ =\ 2\kappa\ -2ln(\hat{L})$$ **Ecuación 4**

Donde:

- κ = Numero de parámetros estimados en el modelo.
- $\hat{L}$ = valor máximo de la función de verosimilitud del modelo.

Para comparar modelos con AIC, el valor más bajo es mejor, si la diferencia es de más de 2 unidades de AIC se considera que es significativamente mejor [39].

Criterio de Información Bayesiano

El Criterio de Información Bayesiano (BIC) [40] es otro criterio para la selección de modelos que mide la compensación entre el ajuste y la complejidad del modelo [41]. Al igual que el AIC a menor valor del BIC, indica un mejor ajuste del modelo.

$$BIC = \kappa\, ln(N) - 2ln(\hat{L}) \quad \textbf{Ecuación 5}$$

Donde:

- κ = Numero de parámetros estimados en el modelo.
- $\hat{L}$ = Valor máximo de la función de verosimilitud del modelo.
- N =Numero de mediciones registradas

Área Bajo la Curva

Área Bajo la Curva (AUC) es una métrica que mide el rendimiento de un modelo, cuanto mayor sea la puntuación mejor es el rendimiento de un clasificador [42].

Las siguientes fórmulas fueron extraídas de Amico [14]:

$$Se(\kappa) = P(M > \kappa)|D = 1 \quad \textbf{Ecuación 6}$$

$$Sp(\kappa) = P(M \leq \kappa)|D = 0 \quad \textbf{Ecuación 7}$$

$$ROC(\mu) = Se\{(1 - Sp)^{-1}(\mu)\}, 0 < \mu < 1 \quad \textbf{Ecuación 8}$$

$$AUC = \int_0^1 ROC(\mu)d\mu \quad \textbf{Ecuación 9}$$

Donde:

- M = Clase M
- D = Clase D
- Se = Sensibilidad
- Sp = Especificidad
- ROC = Característica Operativa del Receptor (Receiver Operating Characteristic).

3.3. Modelo de Riesgos Proporcionales de Cox (Cox PH)

De acuerdo con Amico [14] el modelo de riesgos proporcionales fue introducido por Cox en 1972. El cual es un modelo semi-paramétrico muy popular que consiste en modelar la función de riesgo con la siguiente función. Una característica particular de este modelo es que los riesgos son proporcionales como la relación de riesgo para dos sujetos.

$$\lambda(t|z)\ \lambda_0(t)exp(\beta^t z) \qquad \textbf{Ecuación 10}$$

Donde:

- $\lambda_0(t)$: El riesgo base, que es la función de riego para $z\ = 0$
- β: Es el vector de parámetros asociados con z que no contiene un interceptor.
- z: Vector de covariables.

Este modelo en particular se utilizará para comparar los resultados obtenidos de Modelos de Curación Mixtos Paramétricos multivariados haciendo uso de la función coxph de la biblioteca survival [43] de R.

3.4. Modelos de Curación

En este trabajo se está interesado específicamente en los modelos de curación paramétricos, semi-paramétricos y no paramétricos.

Los modelos de curación son un tipo de análisis de supervivencia censurado donde algunos sujetos no van a experimentar el evento de interés, aunque se les siga por mucho tiempo. Aquellos quienes no van a desarrollar el evento de interés se les refiere como sujetos curados o sobrevivientes a largo plazo [25].

3.4.1. Distribuciones Utilizadas en Modelos de Curación Mixtos (MCM)

Los modelos de curación mixtos o Mixture Cure Models por sus siglas en ingles son mejor conocidos como MCM. En este trabajo se utilizaron diferentes distribuciones de acuerdo con el tipo de biblioteca utilizada. Para modelos paramétricos univariados se hizo uso de Python3 [3] así como de la biblioteca lifelines [44] de Python la cual utiliza la Ecuación 11.Y para los modelos paramétricos multivariados se hizo uso de una biblioteca de R [4] llamada *cuRe* [45]. En el caso del paquete de R al usar la opción "*mixture*" le indica que utilice la función indicada en la Ecuación 12, para utilizar MCM; dentro de este

paquete también se puede usar la opción “nmixture”, que hace referencia a los Modelos de Curación No Mixtos o Non-Cure Mixture Model por sus siglas en inglés o mejor conocidos como NMCM los cuales no son abordados en este trabajo.

$$S(t) = c + (1 - c)\, S_b(t), 1 > c > 0$$ **Ecuación 11**

Donde:

- c: Proporción de Curados
- $(1 - c)$: Proporción de no Curados
- $S_b(t)$: Función paramétrica de supervivencia. *Exponential, Log Normal, Log Logistic, Weibull, Generalized Gama, Log-Normal y Spline* que se explican más adelante.

$$\boldsymbol{R(t|z) = \pi(z) + (1 - \pi(z))S_u(t|z)}$$ **Ecuación 12**

Donde:

- $\boldsymbol{R(t|z)}$: Supervivencia Relativa.
- $\boldsymbol{\pi(z)}$: Proporción de curados.
- $\boldsymbol{(1 - \pi(z))}$: Proporción de no Curados
- $\boldsymbol{S_u(t|z)}$: Función de supervivencia relativa. *Exponential*, *Weibull* y *Log-Normal* que se explican a continuación.

La distribución Lognormal (Log Normal) es una modificación de la distribución normal. La cual trunca los valores de una variable aleatoria a un rango, permitiendo el uso de valores positivos[46]. El modelo paramétrico univariado de la biblioteca lifelines utiliza la Ecuación 13.El modelo multivariado hace uso de la Ecuación 14.

$$S(t) = 1 - \Phi\left(\frac{log(t) - \mu}{\sigma}\right), \sigma > 0$$ **Ecuación 13**

$$S_u(t) = 1 - \Phi\left(\frac{log(t) - \theta_1}{\theta_2}\right)$$ **Ecuación 14**

La distribución *Weibull* es una distribución de dos parámetros y es representada por un parámetro k de forma adimensional y un parámetro λ de escala[47]. El enfoque para el modelo paramétrico univariado de la biblioteca lifelines es la Ecuación 15. De forma similar el modelo paramétrico multivariado hace uso de la Ecuación 16.

$$S(t) = exp\left(-\left(\frac{t}{\lambda}\right)^{\rho}\right), \qquad \lambda > 0, \rho > 0$$ **Ecuación 15**

Donde:

- λ : Es el parámetro de escala que representa el momento en el que ha muerto un porcentaje de la población.
- ρ: Es el parámetro que influye en la forma del riesgo. Controla si el peligro acumulative es convexo o cóncavo. Lo que representa peligros de aceleración o desaceleración.

$$S_u(t) = \exp\left(-\theta_1 t^{\theta_2}\right)$$ **Ecuación 16**

La distribución *Generalized Gama* es una distribución más reciente que la distribución normal que se creó con la finalidad de combinar la distribución Gamma y Weibull en 1962, la cual es bastante flexible [48]. La fórmula utilizada por el modelo paramétrico univariado de la biblioteca lifelines es la Ecuación 17.

$$S(t) = \begin{cases} 1 - \Gamma_{RL}\left(\frac{1}{\lambda^2}; \frac{e^{\lambda\left(\frac{log(t)-\mu}{\sigma}\right)}}{\lambda^2}\right) \; if \; \lambda > 0 \\ \Gamma_{RL}\left(\frac{1}{\lambda^2}; \frac{e^{\lambda\left(\frac{log(t)-\mu}{\sigma}\right)}}{\lambda^2}\right) \; if \; \lambda \leq 0 \end{cases}$$ **Ecuación 17**

Donde:

- Γ_{RL}: Es la función Gamma incompleta inferior regularizada. Para que este modelo use distribución Gamma requiere que σ y λ sean iguales para usar esta configuración. Dado que este modelo en lifelines puede usar otros submodelos tales como Exponencial, Weibull, Log-Normal, entre otros.

La distribución *Log-logistic* es muy similar a en forma a la distribución *Lognormal*. Entre sus ventajas es que es más conveniente que la distribución normal debido a tener expresiones algebraicas simples[49]. La fórmula utilizada por el modelo univariado de la biblioteca lifelines es la Ecuación 18.

$$S(t) = \left(1 + \left(\frac{t}{\alpha}\right)^{\beta}\right)^{-1}, \alpha > 0, \beta > 0$$ **Ecuación 18**

Donde:

- α : Es el parámetro de escala que se interpreta como igual a la mediana de la vida de la población.
- β: Es el parámetro que influye en la forma del riesgo.

La distribución exponencial se usa para modelar el tiempo entre eventos que ocurren de manera aleatoria. La cual aparece en muchos contextos del mundo natural, la cual es la solución a una serie de ecuaciones relacionadas a procesos físicos que caracterizan el riesgo[50]. La fórmula utilizada por el modelo paramétrico univariado de la biblioteca lifelines se muestra en la Ecuación 19. El modelo paramétrico multivariado hace uso de la Ecuación 20.

$$S(t) = exp\left(\frac{-t}{\lambda}\right), \lambda > 0$$ **Ecuación 19**

$$S_u(t) = exp(-t\theta_1)$$ **Ecuación 20**

Dado que las distribuciones de los tiempos de supervivencia pueden no ser logística ni Weibull, se requieren modelos más flexibles como Splines [51]. El enfoque que toma lifelines es el de un modelo univariado que usa N cubic splines (spline cubicos o ranuras cubicas) el cual se muestra en la Ecuación 21.

$$H(t) = exp\left(\Phi_0 + \Phi_1 log(t) + \sum_{j=2}^{N} \Phi_j v_j(log(t))\right)$$ **Ecuación 21**

Donde:

- v_j: Funciones de base cubica en nudos predeterminados llamados knots.

3.4.2. Modelo de Curación Mixto (MCM)

El modelo de curación mixto *(mixture cure model*) o combinado está conformado de dos funciones de supervivencia relativa que se origina de las proporciones de los curados y no curados donde la proporción de curados es frecuentemente el interés primario [52]. Véase Ecuación 12.

Donde π es la proporción de los individuos curados y S_u es la función relativa de supervivencia que es multiplicada por la proporción de los no curados, las cuales pueden ser, entre otras: *Log Normal*, *Generalized Gama*, *Log Logistic*, Spline, Weibull y Exponential. Para π las covariables normalmente se modelan linealmente con el uso de funciones de enlace apropiadas, como *logit* y *cloglog* (log-log complementario) para asegurar valores entre 0 y 1.

3.4.3. Modelo de Curación No Mixto (NMCM)

Los Non-Mixture Cure Models o modelos de curación no mixtos, también conocidos como NMCM son modelos que pueden predecir la probabilidad de curación, pero sin separar la fracción de curados y no curados como lo hacen los modelos de curación mixtos, si no que toman a toda la población como un solo grupo, pero aun así estos modelos pueden estimar la fracción de cura y supervivencia [53].

En este trabajo se abordan únicamente MCM por lo que no se ahondara en este tipo de modelos.

4. DESARROLLO METODOLÓGICO

Aquí se presentan los resultados de la aplicación de las primeras 3 etapas de la metodología KDD que incluye el Análisis de Datos Exploratorio (EDA), en su primera etapa de Selección de los datos, así como varias tareas de preprocesamiento incluidas en las etapas de Limpieza, Transformación y Reducción de dimensionalidad.

El conjunto de datos de estudio contiene la información de 396 estudiantes de una Licenciatura del ITESO. La cual se separó en un conjunto de entrenamiento usando el 70 % y pruebas usando el 30 %. La decisión para seleccionar esta configuración está dada en tratar de dar un mayor número de eventos al conjunto de pruebas, debido a que se cuenta con un conjunto de datos demasiado pequeño, lo que podría afectar la validez del modelo. Para el conjunto de datos se seleccionó un EPV de 10 lo cual debería ser suficiente para la cantidad de variables con las cual se trabaja. Lo ideal sería tener 100 eventos para el conjunto de datos de prueba como lo sugieren Collins y colegas [33], pero aun con la configuración de 30% para datos de prueba solo se tienen 41 eventos.

4.1. Análisis Exploratorio Descriptivo Univariado

En la siguiente sección se muestra un Análisis Exploratorio Descriptivo de cada una de las variables haciendo uso de bibliotecas de Python como pandas [54] para manipulación de datos, pyplot [55] y seaborn [56] para generación de gráficos.

Las columnas para analizar el conjunto de datos, así como sus descripciones, tipos y rangos de valores se muestran en la Tabla 4-1.

Tabla 4-1 Especificación de columnas de archivo AlumISC_afterKM_.csv

Nombre del campo o columna	Descripción	Tipo csv	Tipo de dato según escala de medición estadística	Tipo sugerido en Python pandas	Valores válidos
T12	Total, de semestres transcurridos hasta que se da de baja o hasta que termina el seguimiento identificado con el numero 13	numérico	nominal	int64	Valores del 1 al 13
E12	Variable de salida que indica la ocurrencia o no del evento bajo estudio y está asociada con la variable T12 que contiene el tiempo en el cual ocurre el evento de estudio o el tiempo	numérico	nominal	int64	1: si cambió de carrera o no continuó en la universidad. 0: lo contrario
Sexo_M	Sexo del Alumno	numérico	nominal	category	1: Masculino 0:Femenino
EscolarizadaNacional	Preparatorias separadas en dos categorías. Sí son extranjeras o abiertas; o sí son públicas o privadas.	numérico	nominal	category	1: Extranjera, Abierta 0: Privada, Pública
Semestre_Primavera	Inició en Semestre Primavera u Otoño	numérico	nominal	category	1:Primavera 0: Otoño
EdadIng	Edad en el momento del ingreso.	numérico	razón	float64	Valores entre 17.0 y 35.0
CatPAA	Alumnos agrupados en 3 grupos dependiendo de la puntuación obtenida en la Prueba de Aptitud Académica o si tienen pase directo,	numérico	ordinal	float64	1: De 400 a 1224 puntos 2: De 1225 a 1600 puntos 3: Pase directo
CatFin	Identifica sí el alumno cuenta con financiamiento o no.	numérico	nominal	category	1: Tiene financiamiento 0: No tiene financiamiento

PromPre	Promedio obtenido por el alumno en la preparatoria.	numérico	razón	float64	Valores entre 65.0 y 100.0
Prepa_Encode_Prob_abandon	Probabilidad de abandono histórica de los alumnos provenientes de una preparatoria dada en la universidad.	numérico	razón	float64	Valores desde 0 a 1.0
CatPrepa_Encode_Prob_abandon	Prepa encode probability abandon usando cuartiles.	numérico	ordinal	float64	0: Desde 0 hasta 0.174 1: Entre 0.174 y 0.201 2: Entre 0.021 y 0.312 3: Entre 0.312 y 1.0
ind_planea [1]	Índice de calidad de una preparatoria en la prueba planea 2015.	numérico	razón	float	0 a 200

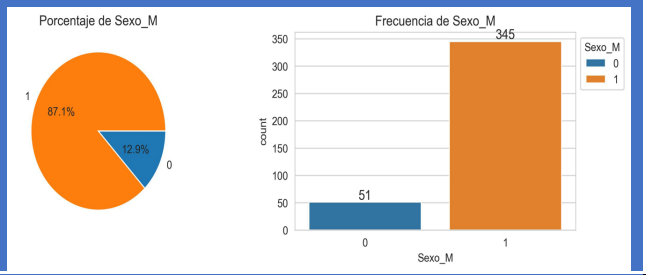
4.1.1. Análisis de Datos Nominales

En esta sección se hace el análisis de cada una de las variables de nominales que fueron previamente descritas en la tabla Tabla 4-1.

En la tabla Tabla 4-2 se puede observar que la mayoría de los estudiantes son del sexo masculino conformando el 87.1% con 345 estudiantes mientras que solo el 12.9% es conformado por mujeres es decir 51 estudiantes.

Tabla 4-2 Estadística descriptiva de Sexo_M

Valores	Frecuencia	%
1: Masculino	345	87.1
0: Femenino	51	12.9

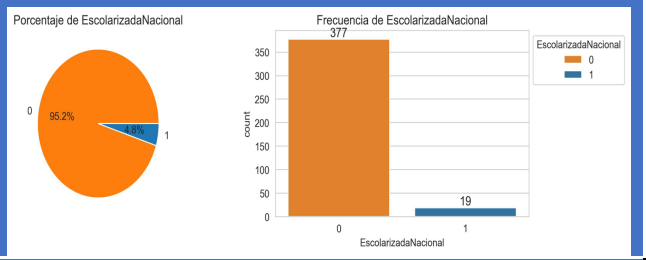


Columnas	conteo	valores únicos
Sexo_M	396	2

En la Tabla 4-3 se puede apreciar que la mayoría de los estudiantes en este caso 377 que conforman el 95.2% proviene de preparatorias privadas o públicas, y solo 19 estudiantes el 4.8 % proviene de escuelas abiertas o extranjeras.

Tabla 4-3 Estadística descriptiva de EscolarizadaNacional

Valores	Frecuencia	%
1: Preparatoria abierta o extranjera	19	4.8
0: Preparatoria pública o privada	377	95.2

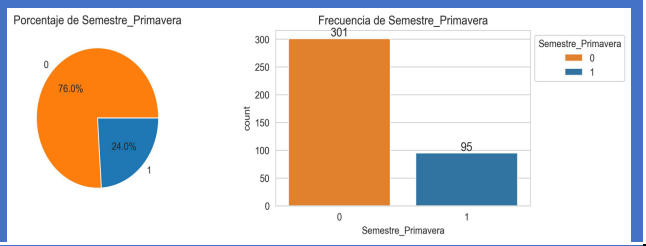


Columnas	conteo	valores únicos
EscolarizadaNacional	396	2

En la Tabla 4-4 se puede observar que la mayoría de los estudiantes, el 76.0% ingresaron en el semestre Otoño y que solo 24% de estudiantes ingresaron en el semestre Primavera. Esto se puede deber a que la mayoría de los estudiantes egresan a inicios del semestre Otoño y solo aquellos que se retrasan en sus estudios de preparatoria tienden a entrar en primavera.

Tabla 4-4 Estadística descriptiva de Semestre_Primavera

Valores	Frecuencia	%
1: Primavera	95	24.0
0: Otoño	301	76.0

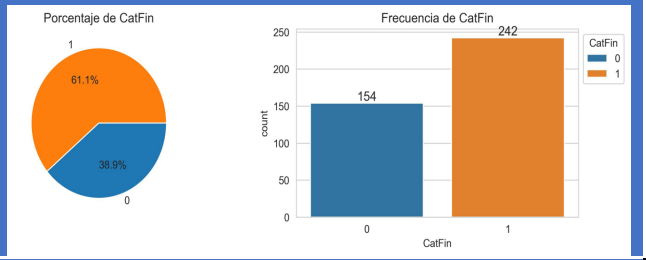


Columnas	conteo	valores únicos
Semestre_Primavera	396	2

En la tabla Tabla 4-5 de puede observar que el 61.1 % de los estudiantes cuentan con financiamiento y solo el 39.9 % no cuenta con financiamiento. Es decir más de la mitad de los estudiantes cuentan con financiamiento.

Tabla 4-5 Estadística descriptiva de CatFin

Valores	Frecuencia	%
1: Tiene financiamiento	242	61.1
0: No tiene financiamiento	154	39.9

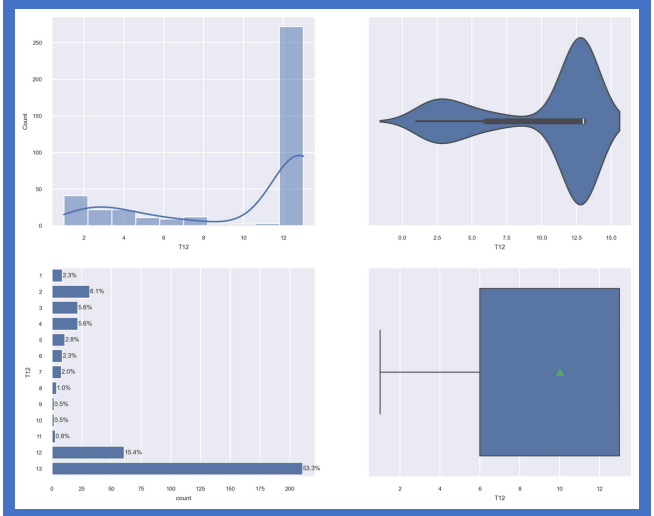


Columnas	conteo	valores únicos
CatFin	396	2

Como se observa en la Tabla 4-6 53.3 % de los estudiantes fueron censurados. Es decir que los alumnos no abandonaron la carrera antes de los 12 meses, pero están fuera del alcance del estudio. El 15.4 % de los estudiantes terminaron en el semestre 12, 8.1 % terminaron en el semestre 2, en los semestres 3 y 4 terminó el 5.6 %. Algo que es importante resaltar de la Tabla 4-6 es que tanto en la gráfica superior izquierda de densidad de kernel y en la gráfica superior derecha de violín se puede apreciar que hay una distribución bimodal, es decir tenemos 2 grupos de datos. En un primer grupo tenemos a los que tienen el valor de 13, es decir aquellos que fueron censurados y por otro lado tenemos a los alumnos que terminaron en un determinado semestre la carrera, los cuales pueden ser identificados con valores del 1 al 12 indicando el semestre de egreso.

Tabla 4-6 Estadística descriptiva de T12

Valores	Frecuencia	%
1	9	2.3
2	32	8.1
3	22	5.6
4	22	5.6
5	11	2.8
6	9	2.3
7	8	2.0
8	4	1.0
9	2	0.5
10	2	0.5
11	3	0.8
12	61	15.4
13: Estudiante censurado	211	53.3



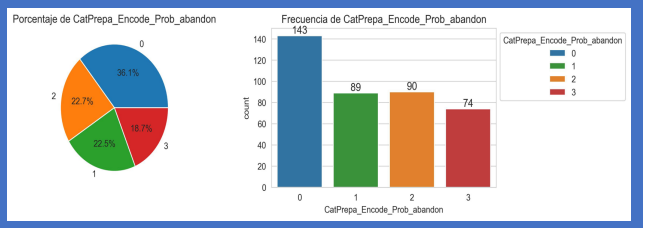
Columna	conteo	valores únicos	moda
T12	396	13	13: Estudiante censurado

4.1.2. Análisis de Datos Ordinales

En esta sección se hace el análisis de cada una de las variables nominales de tipo ordinal que fueron previamente descritas en la tabla Tabla 4-1.

En la tabla Tabla 4-7 se puede observar la probabilidad de abandono histórica de los alumnos provenientes de una preparatoria en la universidad la cual fue categorizada por cuartiles. Se puede observar que los categorizados como 0 son aquellos que tienen la menor probabilidad de abandonar conformando el 36.1 %, la categoría 1 que son el siguiente grupo de estudiantes con menor probabilidad de abandono conforman el 22.5 %, la categoría 2 incrementa la probabilidad de abandonar conformando el 22.7% y finalmente la categoría 3 son los alumnos con mayor riesgo de abandono son el 18.7 %.

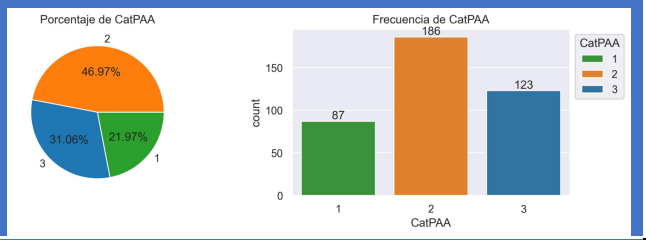
Tabla 4-7 Estadística descriptiva de CatPrepa_Encode_Prob_abandon

Valores	Frecuencia	%	
0: Desde 0 hasta 0.174	143	36.1	
1: Entre 0.174 y 0.201	89	22.5	
2: Entre 0.021 y 0.312	90	22.7	
3: Entre 0.312 y 1.0	74	18.7	

Columnas	conteo	valores únicos
CatPrepa_Encode_Prob_abandon	396	2

En la Tabla 4-8 se pueden observar a 3 grupos dependiendo la puntuación obtenida en la Prueba de Aptitud Académica o sí tienen paso directo. La mayoría de los alumnos el 46.97 % tienen una puntuación alta en la Prueba de Aptitud Académica, seguido del 31.06 % que cuentan con pase directo y solo el 21.97 % tienen una puntuación entre 400 y 1224 puntos.

Tabla 4-8 Estadística descriptiva de CatPAA

Valores	Frecuencia	%	
1: De 400 a 1224 puntos	87	21.97	
2: De 1225 a 1600 puntos	186	46.97	
3: Pase directo	123	31.06	

Columnas	conteo	valores únicos
CatPAA	396	2

4.1.3. Análisis Descriptivo Univariado de Datos de Razón

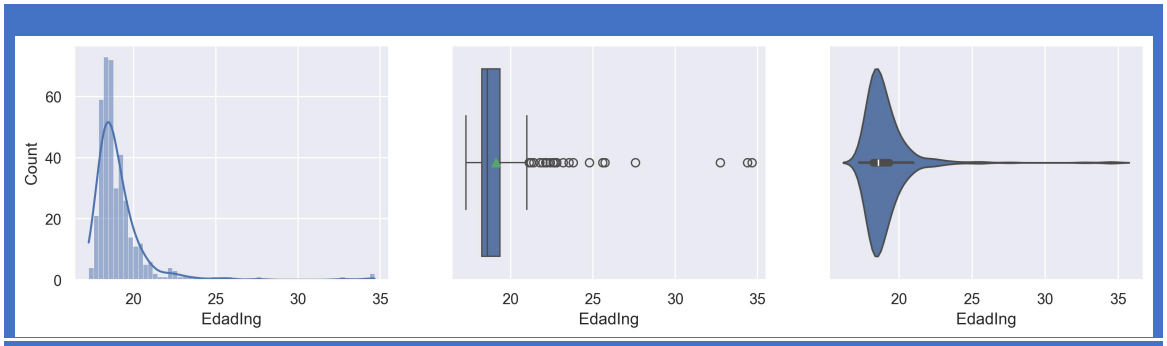
En esta sección se hace el análisis de cada una de las variables de razón que fueron previamente descritas en la tabla Tabla 4-1.

EdadIng: Edad en el Momento del Ingreso

Como se observa en la Tabla 4-9. Se observan valores atípicos arriba del cuarto cuartil. También se puede observar que la media es un poco mayor a la mediana. La grafica que sigue una distribución normal con cola hacia la derecha muestra que la gran mayoría de los estudiantes termina alrededor de los 19 años la preparatoria. Los casos atípicos que se observan pueden ser debido a estudiantes de preparatorias abiertas

que se retrasaron en sus estudios o a alumnos de escuelas extranjeras que revalidaron sus materias para continuar sus estudios.

Tabla 4-9 Estadística descriptiva de EdadIng

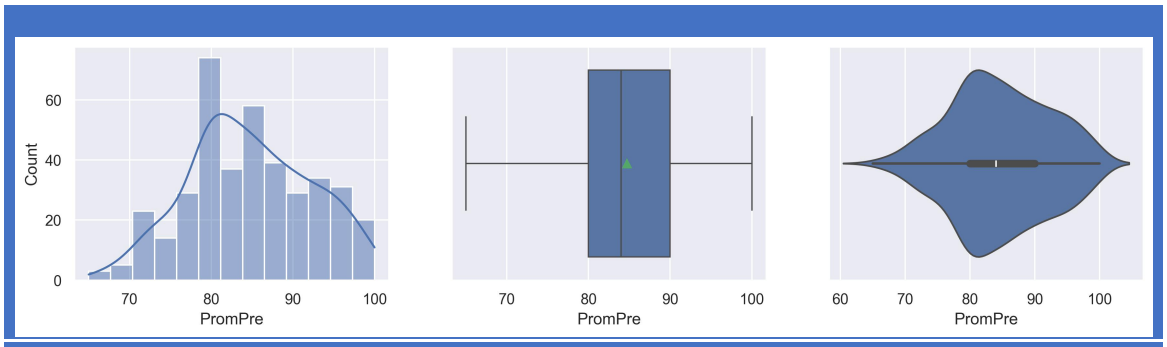


Columna	conteo	media	Desviación	min	25%	50%	75%	Max	curtosis	simetría
EdadIng	396.00	19.10	1.80	17.28	18.24	18.58	19.35	34.65	36.29	5.14

PromPre: Promedio Obtenido por el Alumno en la Preparatoria.

Como se observa en la Tabla 4-10. El grueso de los registros está entre los valores de 80 y 90. La media es ligeramente mayor a la mediana. En la gráfica de cajas y bigotes se puede apreciar que no tenemos valores atípicos. La distribución es una distribución normal con una ligera cola hacia la izquierda, donde los valores se concentran en un promedio de 80 con una desviación estándar de 7.51.

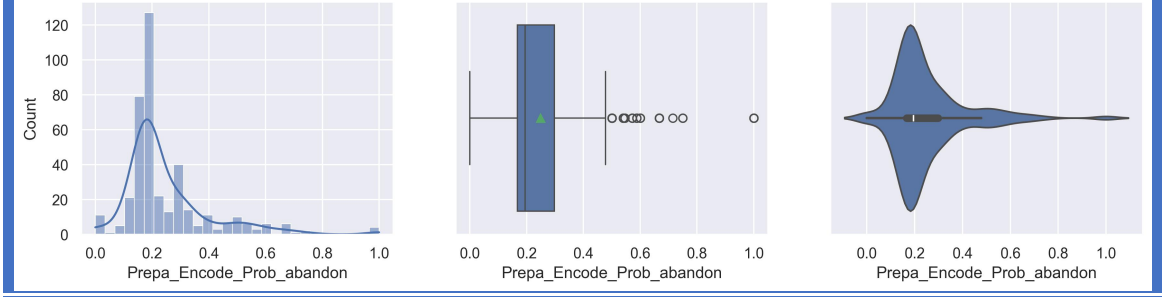
Tabla 4-10 Estadística descriptiva de PromPre



Columna	conteo	media	Desviación	min	25%	50%	75%	Max	curtosis	simetría
PromPre	396.00	84.71	7.51	65.00	80.00	84.00	90.00	100.00	-0.59	0.02

En la Tabla 4-11 se puede observar que la distribución de la probabilidad de abandono histórica de los alumnos provenientes de una preparatoria dada en la universidad sigue una distribución normal con cola hacia la derecha, donde el grueso de los datos esta alrededor de 0.2 de probabilidad de abandono, la mediana es menor a la media. Es importante resaltar que se puede apreciar que tenemos datos atípicos que sobrepasan nuestro valor máximo que está por debajo de 0.6 de probabilidad.

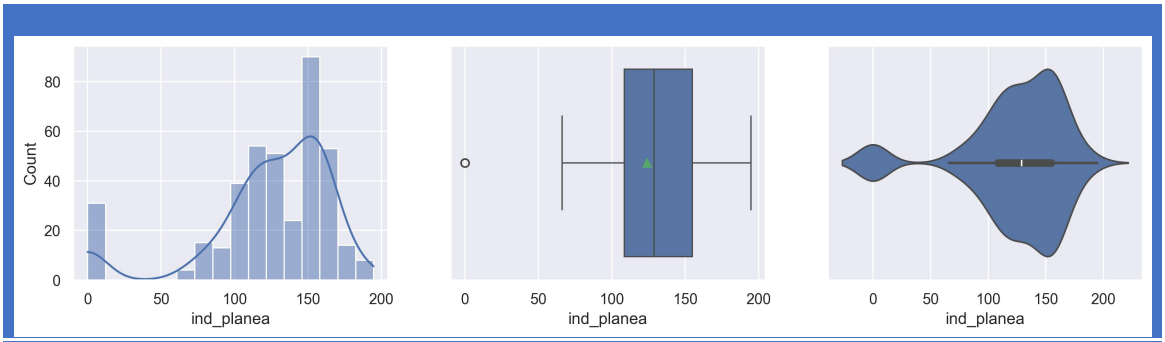
Tabla 4-11 Estadística descriptiva de Prepa_Encode_Prob_abandon



Columna	conteo	media	Desviación	min	25%	50%	75%	Max	curtosis	simetría
Prepa_Encode_Prob_abandon	396.00	0.248	0.154	0.000	0.167	0.194	0.297	1.000	5.769	2.065

En la Tabla 4-12 se muestra la distribución del índice de calidad que es una variable generada con "información de las preparatorias provista por autoridades educativas de gobierno, así como datos de alumnos en registros de la universidad" [1] de los resultados de PLANEA 2015. En la caja de cajas se puede observar que en el extremo inferior tenemos datos atípicos que se encuentran muy alejados del grueso de nuestros datos. Tanto en la gráfica de violín como en la distribución de kernel podemos observar una distribución binomial, lo cual muestra a los datos atípicos como un grupo separado.

Tabla 4-12 Estadística descriptiva de ind_planea



Columna	conteo	media	Desviación	min	25%	50%	75%	Max	curtosis	simetría
ind_planea	396.00	123.826	44.227	0.000	108.575	128.850	154.700	194.625	2.225	-1.548

4.2. Análisis Bivariado Descriptivo y Relacional

En este epígrafe se presentarán las distribuciones conjuntas entre pares de variables, ya sean ambas categóricas, ambas de intervalo/razón, o entre categóricas y de intervalo y razón. Además de ver cómo se distribuyen conjuntamente estos datos, se presenta un análisis de correlación entre los pares de variables utilizando diferentes coeficientes de correlación. Para los pares de variables categóricas se analizarán las tablas de frecuencia y la distribución normalizada. Para el análisis entre una variable categórica y una numérica, se analizará la correlación de punto biserial y la correlación de *Pearson*. Para los pares de variables numéricas se analizará la distribución mediante diagramas de dispersión y se calcularán los índices de correlación de *Pearson y Spearman*. Finalmente, se analizará la correlación de todas las variables mediante el cálculo del coeficiente de correlación phik [57] y la matriz de significancia.

4.2.1. Análisis entre dos Variables Categóricas (Nominales u Ordinales)

En esta sección se hizo la comparativa entre variables de tipo categoría tales como: CatPAA que indica los resultados obtenidos de la Prueba de Aptitud Académica. CatFin que identifica si el alumno cuenta con financiamiento. CatPrepa_Encode_Prob_abandon la cual muestra la probabilidad de abandono histórica de los alumnos provenientes de una preparatoria dada. EscolarizadaNacional que indica si una preparatoria es de tipo privada, pública, extranjera o abierta. Información adicional puede ser revisada en la Tabla 4-1 que muestra detalle de las columnas.

En la Figura 4-1 los alumnos que más egresan son los estudiantes provienen de preparatorias públicas o privadas representando el 66.4 % contra 1.5% que representa a los estudiantes que provienen preparatorias abiertas o extranjeras. De forma similar los estudiantes de preparatorias privadas o públicas egresan en su mayoría en el semestre Otoño más que los estudiantes de preparatorias abiertas o extranjeras; y en el semestre Primavera tienden a egresar casi en la misma cantidad, siendo por muy poco los extranjeros los que egresan más en este semestre. Se puede observar que referente al sexo, hay mayor cantidad de alumnos de sexo masculino que provienen de escuelas públicas o privadas lo que representa un 82.3 % y en el caso de preparatorias abiertas o extranjeras todos los estudiantes son del sexo femenino. En el caso de T12 una variable de salida que identifica cuando los alumnos egresan o son censurados, se puede apreciar que el mayor número de alumnos egresados y censurados pertenecen a preparatorias públicas o privadas que a extranjeras o abiertas. Con respecto a la categoría de probabilidad de abandono se puede observar que todos los estudiantes provenientes de preparatorias abiertas o extranjeras caen en la categoría 3 los cuales son los que mayor probabilidad de abandono tienen de los 4 grupos.

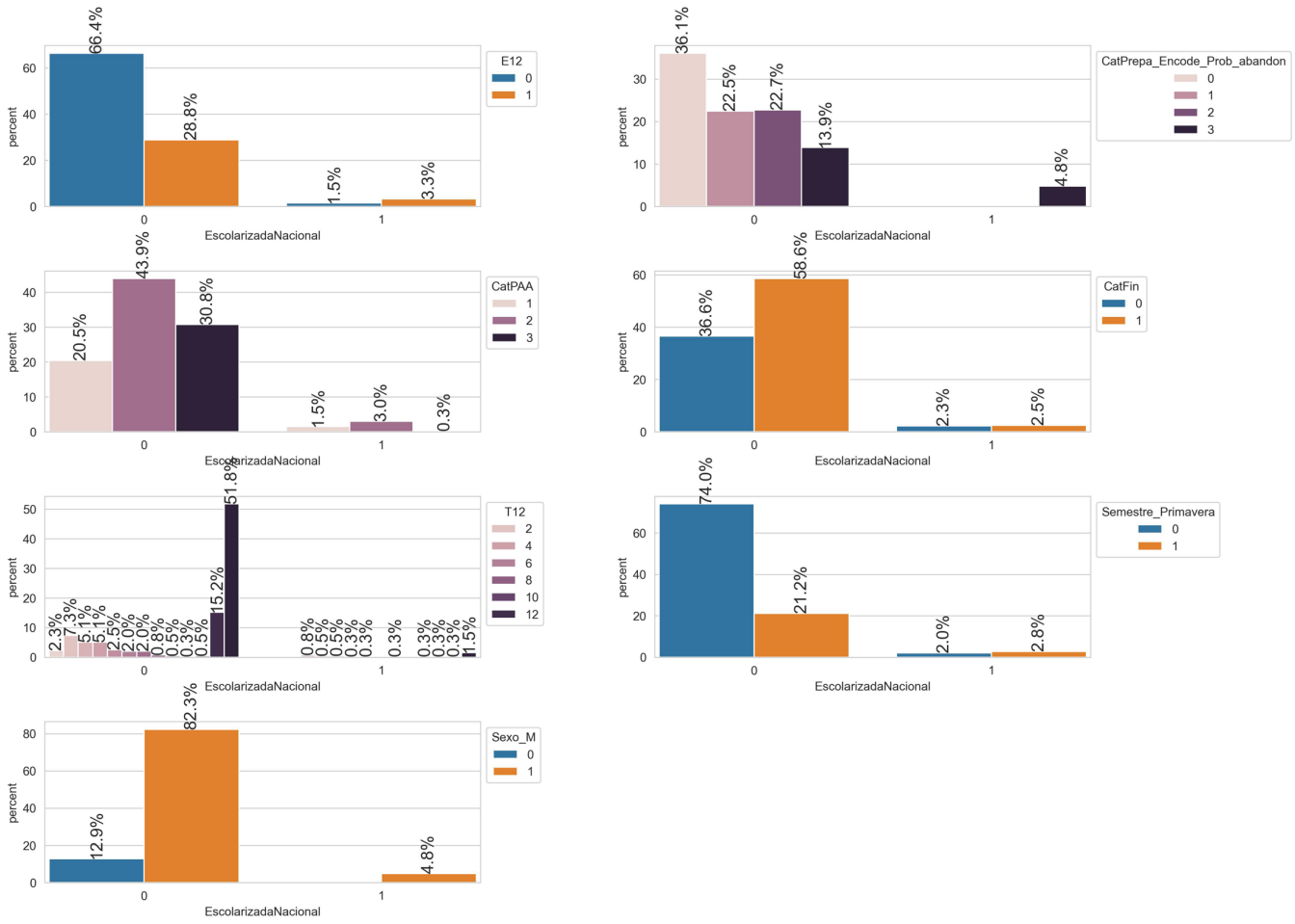


Figura 4-1 Distribuciones de frecuencia de EscolarizadaNacional contra el resto de las variables categóricas

En la figura Figura 4-2 se observa que los alumnos de sexo masculino egresan en mayor numero que los de sexo femenino, los alumnos que cambian de carrera o dejan la universidad también son un mayor número los alumnos de sexo masculino. En primavera egresan alrededor de 4 veces más alumnos que provienen de preparatorias públicas o privadas que alumnos que provienen del extranjero o de preparatorias abiertas; en contraste en otoño la cantidad de alumnos que cambian de carrera o abandonan la universidad es aproximadamente del doble para alumnos de preparatorias públicas o privadas que los extranjeros o de preparatorias abiertas. En el caso del total de semestres transcurridos hasta que el alumno egresó o hizo cambio de carrera o se dio de baja se puede observar que la censura se observa solo para los alumnos que egresaron debido a que sobrepasaron el tiempo de seguimiento. En el caso del financiamiento se puede observar que el porcentaje de alumnos que cambia de carrera es casi el mismo en ambos grupos de preparatorias, y para el caso de egreso los que tienen financiamiento tienden a egresar en 45.7% que un poco más del doble de los que no cuentan con financiamiento.

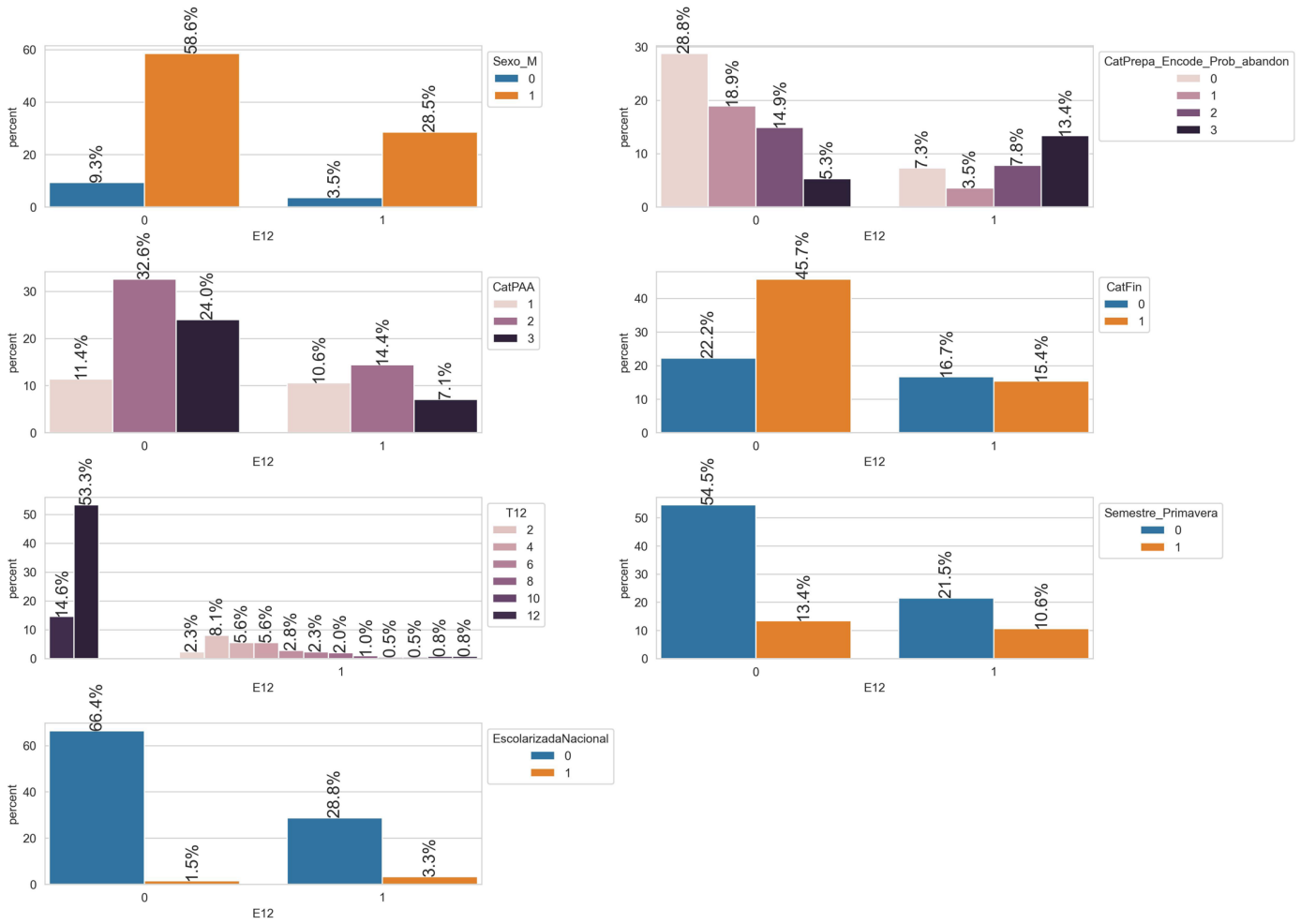


Figura 4-2 Distribuciones de frecuencia de E12 contra el resto de las variables categóricas

En la Figura 4-3 se puede observar que en otoño se encuentran la mayoría de los alumnos censurados es decir que no se les pudo dar seguimiento en el estudio que es alrededor de 4 veces más que en primavera. Los alumnos provenientes de preparatorias públicas o privadas egresan en un numero significativamente mayor en los semestres de otoño y primavera que los de preparatorias abiertas o extranjeras. Se puede apreciar que los alumnos del sexo femenino tienen una preferencia más marcada por ingresar en el semestre de otoño que por el de primavera. En el caso de los alumnos que no tienen pase directo en la prueba de aptitud en ambas categorías son alrededor de 3 veces más alumnos los que ingresan en otoño que los que ingresan en primavera y en el caso de los alumnos que tienen pase directo son alrededor de 5 veces más alumnos los que ingresan en otoño que los que ingresan en primavera.

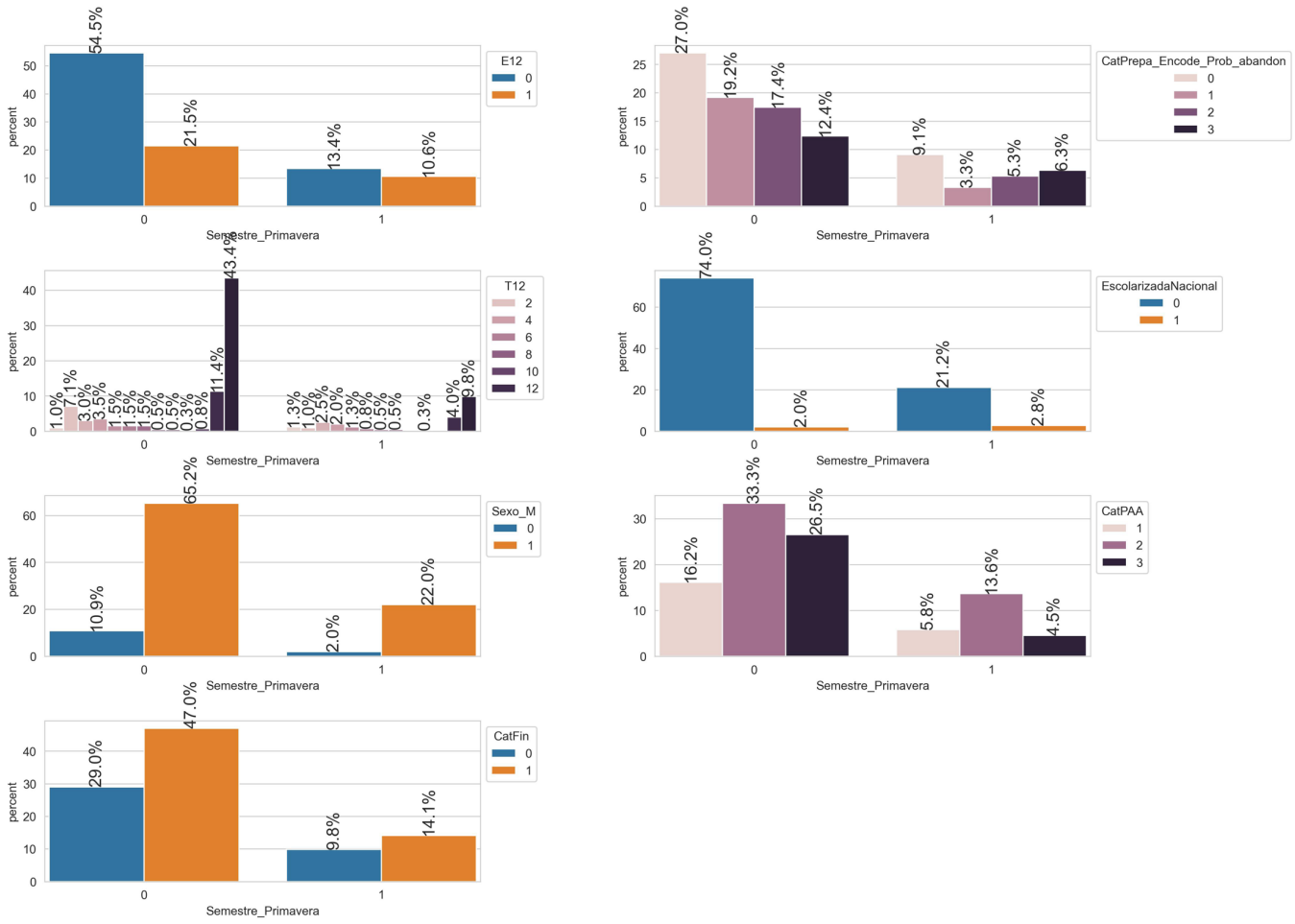


Figura 4-3 Distribuciones de frecuencia de Semestre_Primavera contra el resto de las variables categóricas

En la Figura 4-4 se observa la comparativa respecto al sexo del estudiante. El 58.6 % de los estudiantes que egresan son estudiantes del sexo masculino y solo el 9.3% del sexo femenino. Con respecto a la variable Escolarizada Nacional se puede observar que solo el 4.8% de todo el grueso de estudiantes provienen de preparatorias extranjeras o abiertas y corresponden al sexo masculino. Con respecto al financiamiento en proporción tienen mayor financiamiento los alumnos del sexo femenino que los alumnos del sexo masculino. Con respecto a la Prueba de Aptitud Académica en el caso de los alumnos del sexo masculino el porcentaje de alumnos con una puntuación menor a 1224 puntos es casi igual a los alumnos con pase directo, en el caso de los alumnos del sexo femenino en proporción se tiene un porcentaje mucho menor de alumnos con una puntuación menor a 1224 y se tiene un porcentaje mayor de los alumnos con pase directo. En la probabilidad de abandono los alumnos del sexo femenino tienden a tener menor probabilidad de riesgo de abandono.

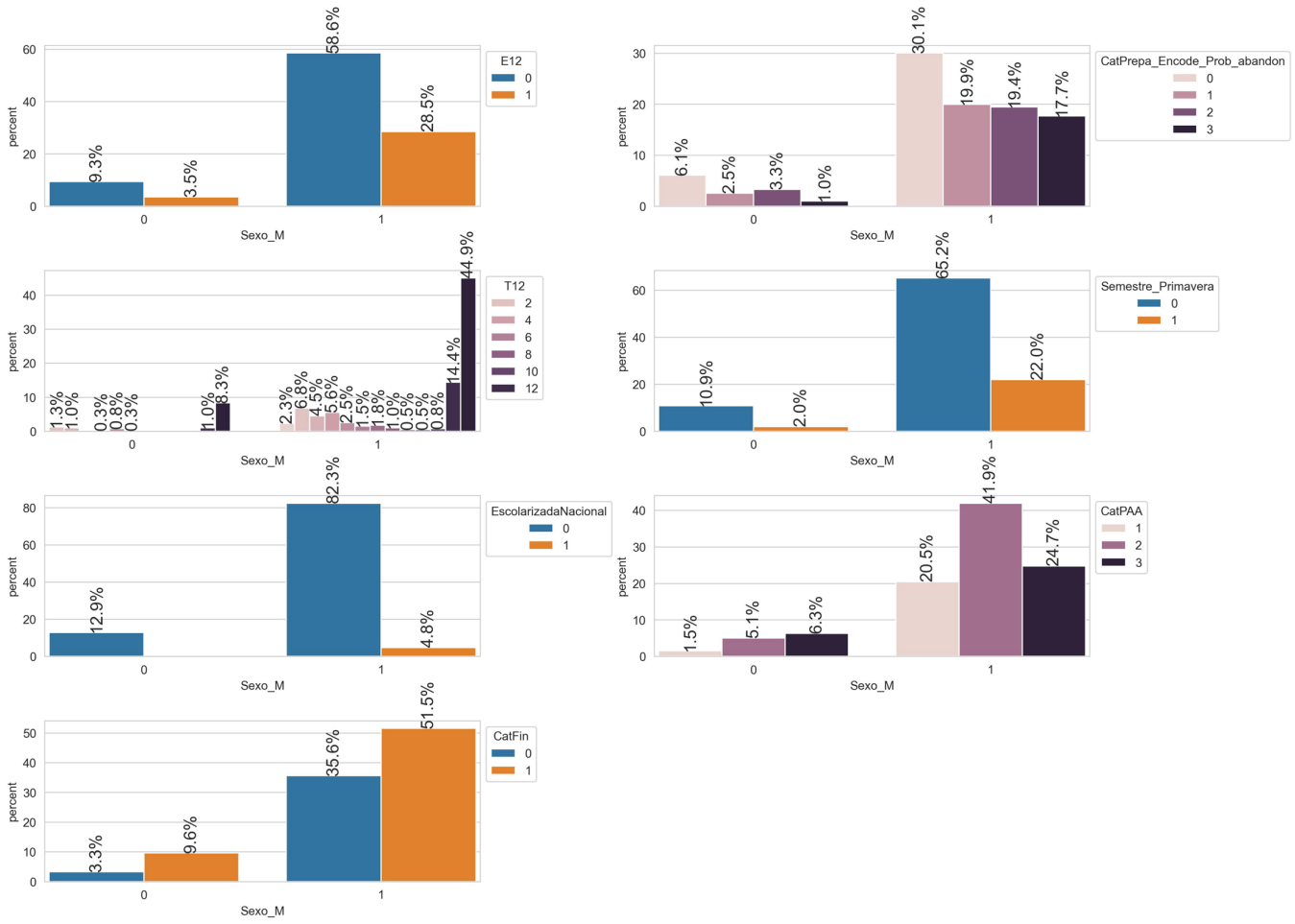


Figura 4-4 Distribuciones de frecuencia de Sexo_M contra el resto de las variables categóricas

En la Figura 4-5 se puede observar la comparación de la categoría de financiamiento. Los alumnos que egresan que tienen financiamiento son el doble de los que aprueban y no tienen financiamiento. Para los alumnos que cambian de carrera o abandonan la universidad tienen porcentajes muy similares, los que tienen financiamiento son el 15.4% y 16.7 % los que no tienen financiamiento. La probabilidad de abandono tiende a ser del doble para los que no tienen financiamiento, solo en aquellos que el riesgo es algo la diferencia es solo alrededor de un 2 % de los que no tienen financiamiento. Para alumnos que provienen de preparatorias extranjeras o abiertas se tiene financiamiento casi en la misma proporción que para los que no y para los alumnos de escuelas públicas y privadas el 58.6 % de los alumnos cuentan con financiamiento y el 36.6 % no cuentan con financiamiento. La mayor parte de los financiamientos se dan en otoño conformando el 47.0 % respecto a el semestre primavera que solo representa el 14.1% de los financiamientos. En proporción los alumnos del sexo femenino tienden a tener más financiamiento que los alumnos del sexo masculino.

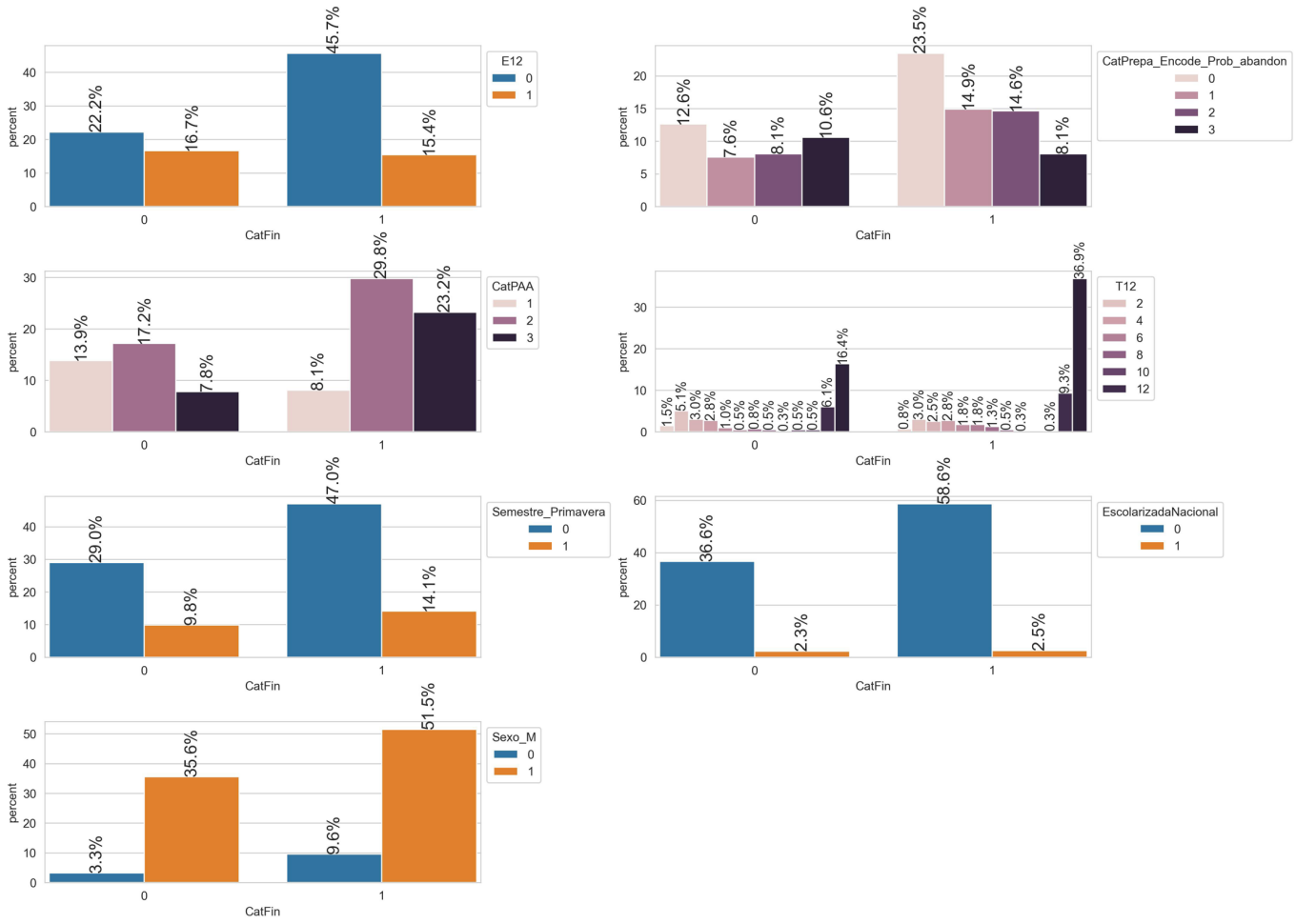


Figura 4-5 Distribuciones de frecuencia de CatFin contra el resto de las variables categóricas

En la Figura 4-6 se muestran las comparativas de la Prueba de Aptitud Académica categórica. Para los alumnos que un puntaje bajo en la Prueba de Aptitud Académica se muestran porcentajes similares de los alumnos que egresan 11.4% y los que no egresan o cambian de carrera 10.6 %. En el caso de los alumnos que tienen una puntuación alta se puede apreciar que la diferencia es de aproximadamente el doble de los que egresan 32.6 % y los que no egresan o cambian de carrera 14.4%. Para los alumnos que tienen pase directo se muestra una diferencia de alrededor de tres veces más los alumnos que egresan 24% y los que abandonan o cambian de carrera 7.1 %. Para los alumnos que tienen una alta probabilidad de abandono representado por la categoría de CatPrepaEnconde_Prob_abandon como el ordinal 3 se puede observar que los alumnos que tienen pase directo tienen menor probabilidad de estar en riesgo de abandono. En el caso de la prueba separada por sexos, se tienen muy pocos alumnos del sexo femenino con puntuación baja. En el caso del financiamiento los alumnos con puntuación alta y con pase directo tienen mayor cantidad de alumnos con financiamiento que los que tienen puntuación baja.

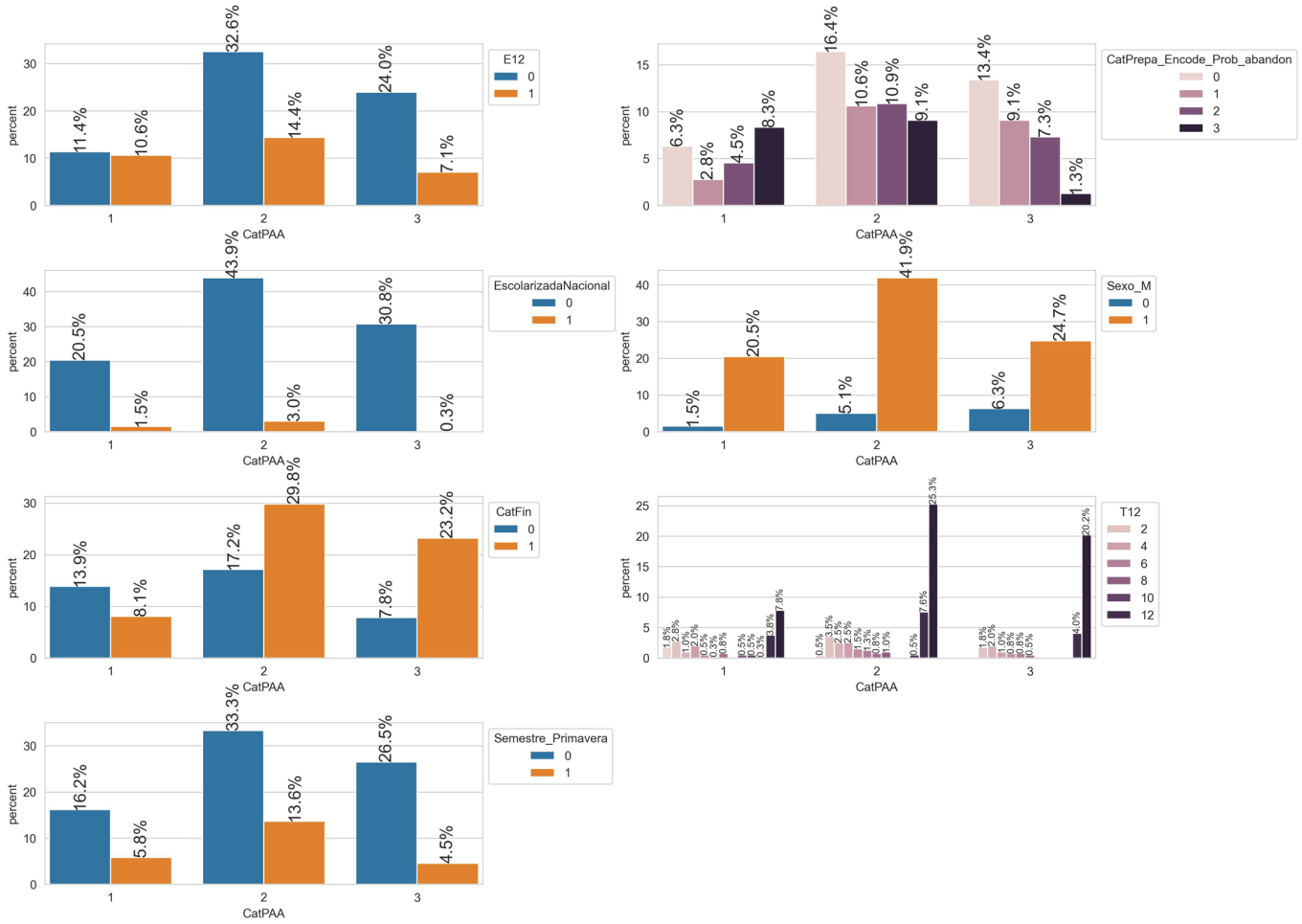


Figura 4-6 Distribuciones de frecuencia de CatPAA contra el resto de las variables categóricas

En la Figura 4-7 se muestran las comparaciones de la probabilidad de abandono histórica de los alumnos provenientes de una preparatoria dada en la Universidad. De forma ordinal de 0 a 3 siendo 3 el que mayor probabilidad de abandono representa. Para los alumnos que egresan de la carrera se puede observar que la cantidad de alumnos que egresa tiene un menor porcentaje conforme se incrementa la probabilidad de abandono lo que se puede ver representado en la gráfica de E12 y de forma correspondiente los alumnos que abandonan o cambian de carrera tienen un mayor porcentaje para la categoría 3. En el caso de la Escolarizada Nacional se puede ver que los alumnos con mayor riesgo son los extranjeros o provenientes de preparatorias abiertas. En el caso de los alumnos del sexo femenino tienen menor cantidad de alumnos en riesgo que los alumnos del sexo masculino. En el caso de la Prueba de Aptitud Académica de los alumnos con mayor riesgo los alumnos con pase directo representan un porcentaje bajo y los alumnos con una calificación alta tienen mayor probabilidad de abandono para esta categoría que los de una puntación baja. En el caso de los alumnos con mayor probabilidad de abandono tienen un 8.1 % de alumnos con financiamiento. De los alumnos que tienen mayor probabilidad de abandono se puede observar que aproximadamente el doble ingresa en otoño.

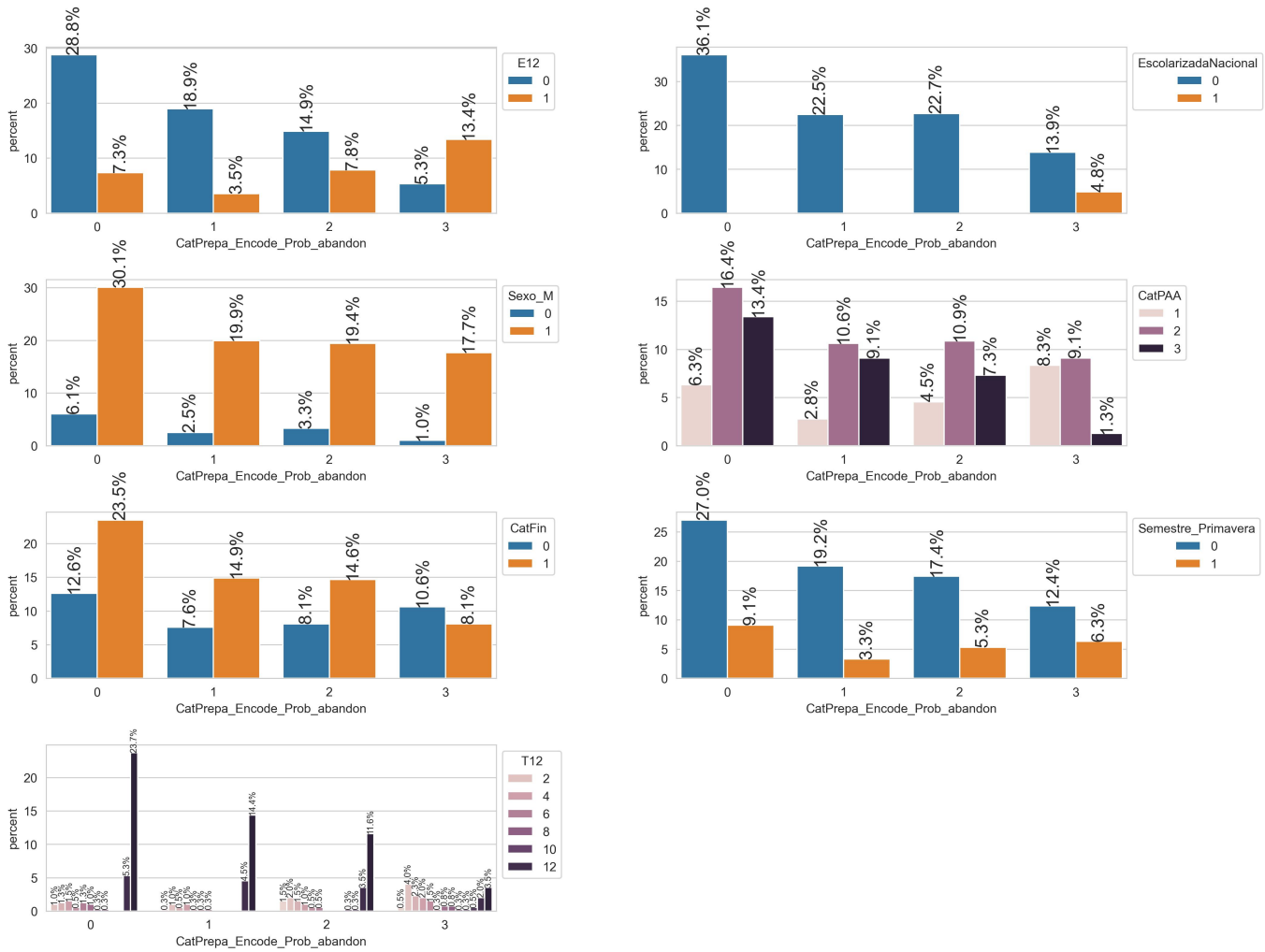


Figura 4-7 Distribuciones de frecuencia de CatPrepa_Encode_Prob_abandon contra el resto de las variables categóricas

4.2.2. Análisis entre dos Numéricas (Intervalo o Razón)

En la Figura 4-8 se hace una comparación haciendo uso de la gráfica pairplot (grafica por pares) de seaborn [56] en la diagonal central podemos observar una serie de graficas de densidad de kernel que se hacen contra la misma variable, en esta grafica también se pueden observar 2 tonalidades para los grupos de la variable E12 donde los egresados se muestran en una coloración azul y en naranja aquellos que cambiaron de carrera o abandonaron la universidad. Para la variable EdadIng que es la edad de ingreso se puede apreciar una distribución normal y que la mayoría de los que se cambian de carrera o abandonan la universidad están alrededor de los 20 años y algunos casos atípicos alrededor de los 35 años y de los que egresan se puede observar un gran número de alumnos por encima de los 20 años, lo que podría indicar que la edad también influye para el egreso de alumnos. En el caso de la variable PromPre que indica el promedio obtenido en la preparatoria también se puede apreciar una distribución normal para ambos grupos y para el caso de los alumnos que egresan se puede observar que aquellos que obtuvieron un promedio mayor a 80 tienen una mayor probabilidad de egresar de la carrera. Para la variable Prepa_Enconde_Prob_abandon que muestra la probabilidad de abandono histórica de los alumnos provenientes de una preparatoria dada en la universidad, que el grueso de los alumnos que egresan se encuentra alrededor de 0.248 y para los alumnos que cambian de carrera o abandonan la universidad se puede una distribución normal con cola hacia la derecha. Para la variable ind_planea[1] que muestra el índice de calidad de una preparatoria en la prueba planea 2015 se puede apreciar que tiene una distribución binomial para ambos grupos de estudiantes, los alumnos que egresan y los que cambian de carrera o abandonan la universidad. Para la variable T12 que indica el total de semestres transcurridos hasta que egresan o se dan de baja de la carrera o de la universidad se puede apreciar una distribución normal bimodal, esto debido a que el valor de 100 indica los alumnos que fueron censurados, es decir que aun continuaban en la carrera al fin del seguimiento del estudio por lo que no se pudo determinar si egresaron o no.

Figura 4-8 Comparación entre dos variables numéricas usando pairplot

En la Figura 4-9 podemos ver la covarianza de las columnas analizadas para ver su relación, esta grafica solo nos indica la dirección de la relación pero no que tan fuerte es la relación ni su significancia. La edad de Ingreso (EdadIng) tiene una relación negativa con el promedio de preparatoria (Prompre) lo que indica que a mayor edad menor promedio, una relación casi nula con Prepa_Encode_Prob_abandon y con E12. Una relación negativa con ind_planea. Prompre muestra una relación casi nula con Prepa_Encode_Prob_abandon, muestra una relación negativa con respecto a E12 que nos indica si el alumno egresó o no, es decir a menor promedio menor probabilidad de egreso. Y muestra una relación positiva con ind_planea. Prepa_Encode_Prob_abandon muestra una relación casi nula con EdadIng,

PromPre y E12, pero una relacion negativa con ind_planea. Ind_plane muesta una relación positiva con PromPre y relaciones negativas con EdadIng, PromPre, Prepa_Encod_Prob_abandon y E12.

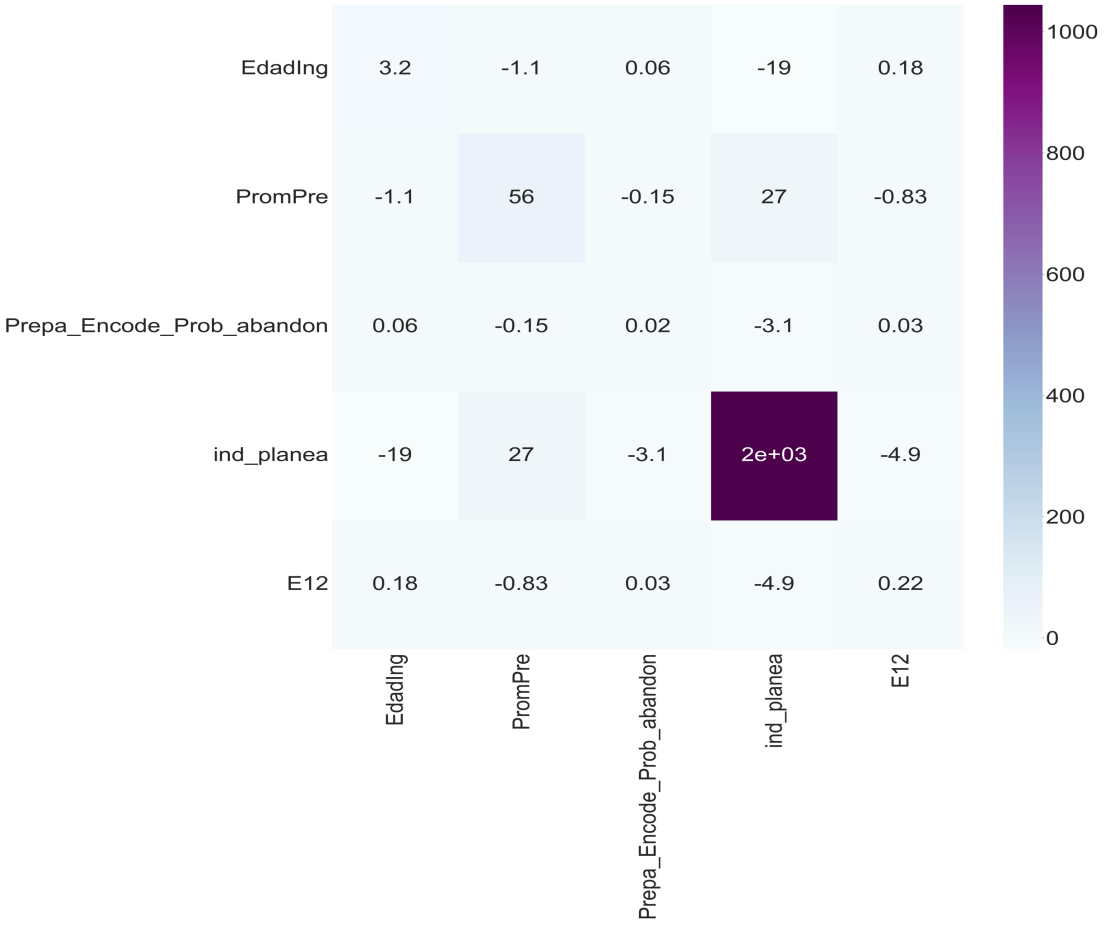


Figura 4-9 Grafica de la covarianza utilizando heatmap de seaborn

Correlación de Pearson

En esta sección se compararon variables dicotómicas contra variables de razón haciendo uso correlación de punto biserial , la cual mide la dirección y fortaleza de la relación entre una variable continua y una dicotómica para comprobar si son grupos diferentes, siendo un caso especial de correlación de Pearson cuando una variable es continua y otra dicotómica. En las comparaciones hechas en esta sección las tablas muestran con un "Sí" las correlaciones altas (mayor a 0.4) o bajas (menor a 0.4) significativas que son en las que se está interesado sí valor $p \leq 0.05$.

En la Figura 4-10 y Tabla 4-13 se puede observar la comparación de la columna E12 que muestra si un alumno cambio de carrera, abandonó la universidad o egresó. E12 tiene una correlación positiva de significancia baja con edad de ingreso (EdagIng), es decir tienen una mayor probabilidad de darse de baja los que tienen una mayor edad. El promedio de preparatoria (PromPre) tiene una correlación negativa de significancia baja, es decir que a menor promedio mayor su probabilidad de darse de baja. Con

Prepa_encode_Prob_abandon tiene una correlación positiva de significancia baja, es decir si la probabilidad de abandono incremente es más probable que un alumno abandone la universidad o cambie de carrera. Con ind_planea tiene una correlación negativa de significancia baja, es decir que si el índice de calidad aumenta es más probable que el alumno egrese.

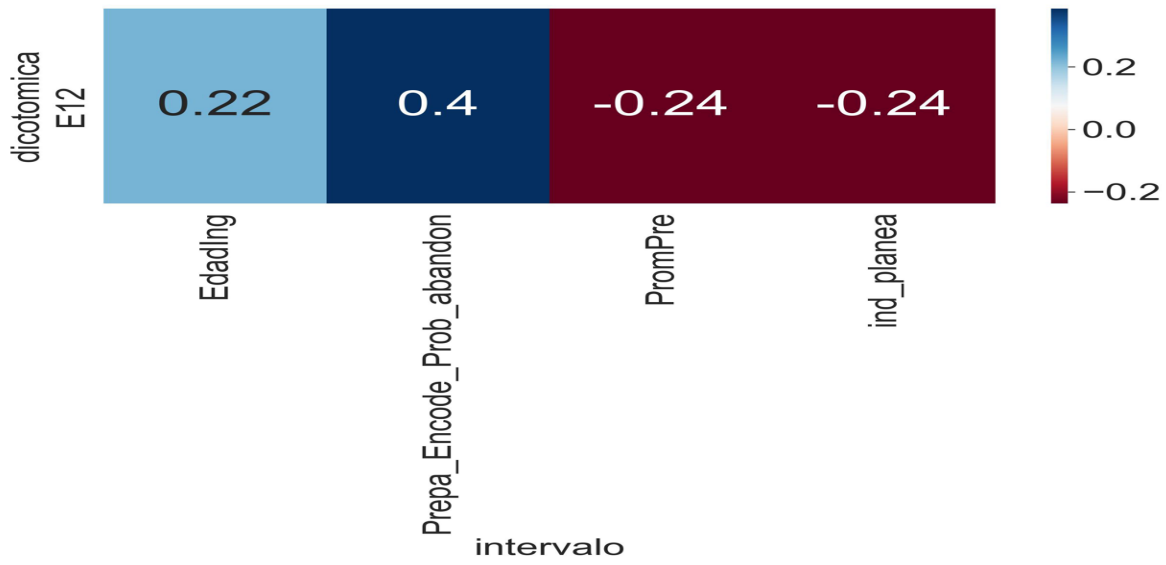


Figura 4-10 Comparación de variable dicotómica E12 contra variables de intervalo

Tabla 4-13 Comparación de variable dicotómica E12 contra variables de intervalo

Comparación de variable dicotómica E12 vs Intervalo			
Variable de intervalo	**corr**	**p-valor**	**Correlacion**
EdadIng	0.2192	0.0	Sí (baja)
PromPre	-0.2362	0.0	Sí (baja)
Prepa_Encode_Prob_abandon	0.3971	0.0	Sí (baja)
ind_planea	-0.2369	0.0	Sí (baja)

En la Figura 4-11 y Tabla 4-14 se puede observar la comparación de la columna EscolarizadaNacional que se utiliza para indicar si el alumno proviene de una preparatoria publica, privada o de una extranjera, abierta. La columna Edad de ingreso tiene una correlación positiva de significancia baja, es decir a mayor edad tiene mayor probabilidad de provenir de una preparatoria abierta o extranjera. La columna Promedio

de preparatoria tiene una correlación negativa de significancia baja, es decir a menor promedio tiene mayor probabilidad de provenir de una preparatoria abierta o extranjera. La columna Prepa_enconde_Prob_abandon tiene una correlación positiva de significancia alta, es decir que a mayor probabilidad de abandono es más probable que provenga de una preparatoria abierta o extranjera. La columna ind_planea tiene una correlación negativa de significancia baja, es decir que a mayor índice de calidad es más probable que el alumno provenga de una universidad privada o pública.

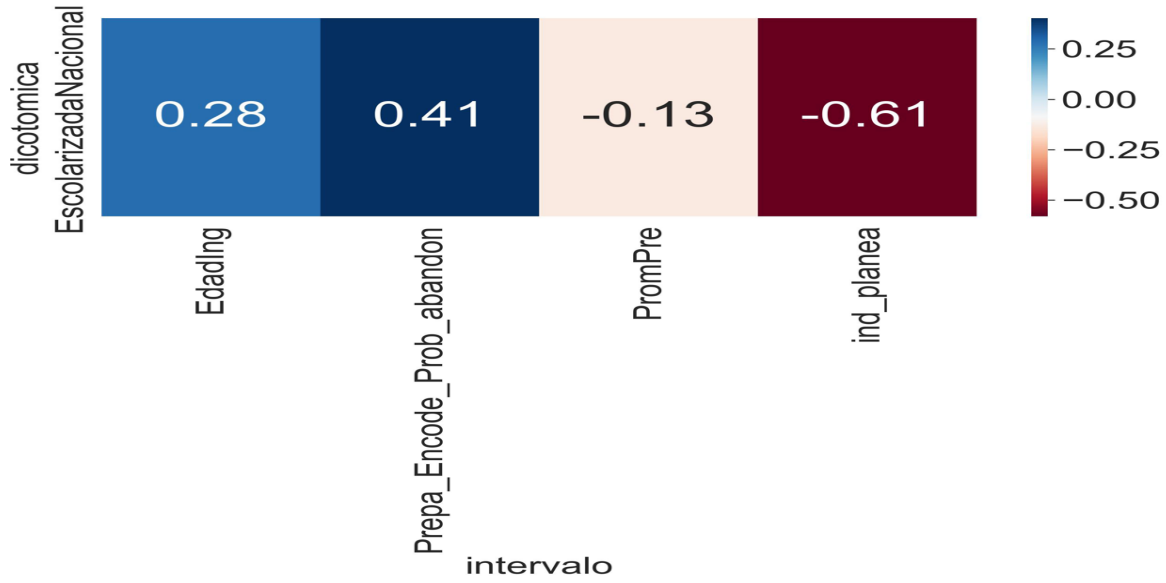


Figura 4-11 Comparación de variable dicotómica EscolarizadaNacional contra variables de intervalo

Tabla 4-14 Comparación de variable dicotómica EscolarizadaNacional contra variables de intervalo

Comparación de variable dicotómica EscolarizadaNacional vs **Intervalo**			
Variable de intervalo	**corr**	**p-valor**	**Correlacion**
EdadIng	0.2847	0.0000	Sí (baja)
PromPre	-0.1316	0.0088	Sí (baja)
Prepa_Encode_Prob_abandon	0.4100	0.0000	Sí (alta)
ind_planea	-0.6087	0.0000	Sí (baja)

En la Figura 4-12 y en la Tabla 4-15 se puede observar la comparación de la columna Semestre_Primavera. La columna probabilidad de abandono histórica (Prepa_Encode_Prob_abandon) no tiene correlación con el semestre primavera. La columna Edad de ingreso tiene una correlación positiva de significancia baja, es decir a mayor edad hay más probabilidad de que ingrese en el semestre primavera. La columna

Promedio de preparatoria tiene una correlación negativa de significancia baja, es decir que a menor promedio tiene una probabilidad mayor de ingresar en el semestre primavera. La columna ind_planea tiene una correlación negativa de significancia baja, es decir que a menor índice de calidad de la preparatoria es más probable que el alumno ingrese en el semestre primavera.

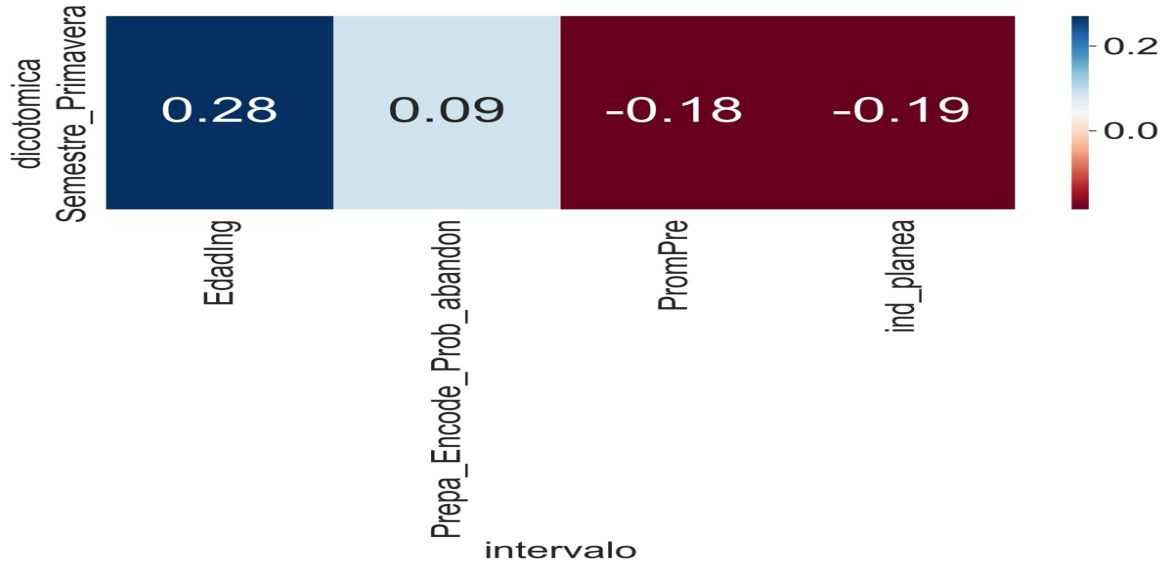


Figura 4-12 Comparación de variable dicotómica Semestre_Primavera contra variables de intervalo

Tabla 4-15 Comparación de variable dicotómica Semestre_Primavera contra variables de intervalo

Comparación de variable dicotómica Semestre_Primavera **vs Intervalo**			
Variable de intervalo	**corr**	**p-valor**	**Correlacion**
EdadIng	0.2820	0.0000	Sí (baja)
PromPre	-0.1850	0.0002	Sí (baja)
Prepa_Encode_Prob_abandon	0.0897	0.0745	No (baja)
ind_planea	-0.1859	0.0002	Sí (baja)

En la Figura 4-13 y Tabla 4-16 se puede observar la comparación de la variable Sexo_M que nos indica el sexo del alumno. Las columnas edad de ingreso (EdadIng), Probabilidad de abandono histórica (Prepa_Encode_Prob_abandon) e ind_planea no tienen correlación con el sexo del alumno. La columna Promedio de preparatoria (PromPre) tiene una correlación negativa de significancia baja, indicando que el

promedio de alumnos de sexo masculino y femenino son significativamente diferentes, es decir que si el promedio de la preparatoria es bajo es más probable que sea un alumno masculino.

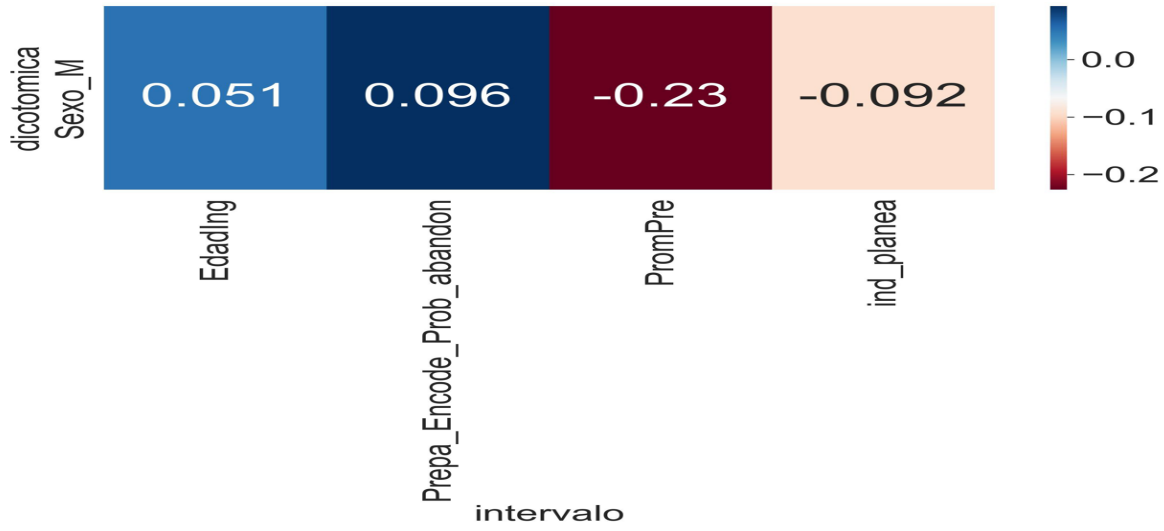


Figura 4-13 Comparación de variable dicotómica Sexo_M contra variables de intervalo

Tabla 4-16 Comparación de variable dicotómica Sexo_M contra variables de intervalo

Comparación de variable dicotómica Sexo_M **vs Intervalo**			
Variable de intervalo	**corr**	**p-valor**	**Correlacion**
EdadIng	0.0508	0.3135	No
PromPre	-0.2349	0.0000	Sí (baja)
Prepa_Encode_Prob_abandon	0.0955	0.0575	No
ind_planea	-0.0920	0.0676	No

En la Figura 4-14 y en la Tabla 4-17 se muestra la comparación de la variable CatFin que nos indica si el alumno tiene financiamiento o no. Las columnas edad de ingreso (EdadIng), e ind_planea no tienen correlación con el financiamiento del alumno. La columna Promedio de preparatoria (PromPre) tiene una correlación positiva de significancia alta, es decir que a mayor promedio obtenido en la preparatoria tiene

mayor probabilidad de obtener financiamiento. La columna de probabilidad de abandono histórico (Prepa_Encode_Prob_abandon) tiene una correlación negativa de significancia baja, es decir que a mayor probabilidad de abandono es menos probable que el alumno obtenga financiamiento.

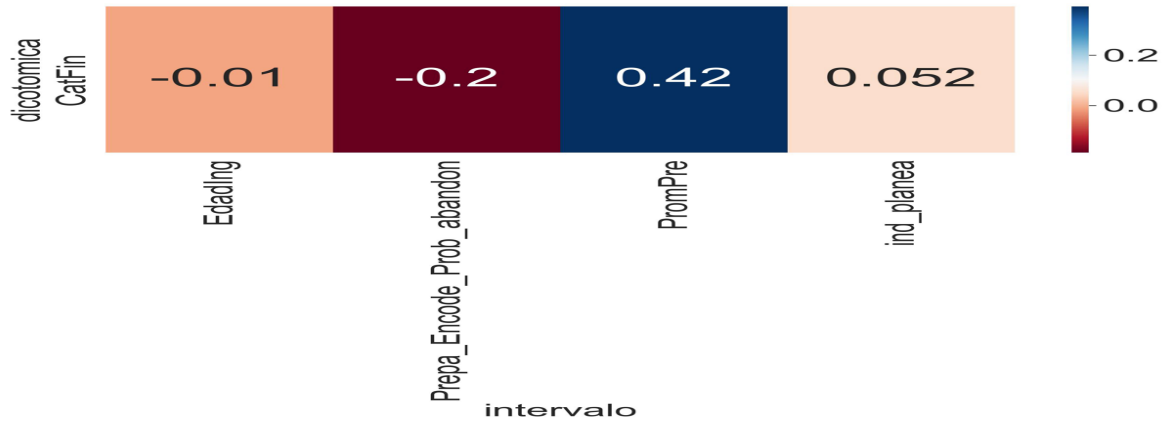


Figura 4-14 Comparación de variable dicotómica CatFin contra variables de intervalo

Tabla 4-17 Comparación de variable dicotómica CatFin contra variables de intervalo

Comparación de variable dicotómica CatFin **vs Intervalo**			
Variable de intervalo	**corr**	**p-valor**	**Correlacion**
EdadIng	-0.0105	0.835	No
PromPre	0.4155	0.000	Sí (alta)
Prepa_Encode_Prob_abandon	-0.1986	0.0001	Sí (baja)
ind_planea	0.0517	0.3049	No

4.3. Análisis Exploratorios entre todas las Variables

Para poder hacer un análisis entre todas las variables se hizo uso de una biblioteca de Python llamada phi_k [57] que permite hacer uso de una matriz de correlación de Pearson para comparar variables categorías, ordinales y de intervalo.

4.3.1. Coeficiente De Correlación Phik

Para poder comparar la Matriz Phik Figura 4-15 y la Matriz de significancia Figura 4-16, es necesario tomar los valores correspondientes a las coordenadas de las variables en ambas matrices y comparar los resultados. En este caso el valor mínimo tomado para la matriz phix es 0.4 y el valor mínimo de la matriz de significancia valido es valor absoluto de 1. Se puede encontrar una explicación más detallada en una respuesta dada por el autor de phik en un foro [58].

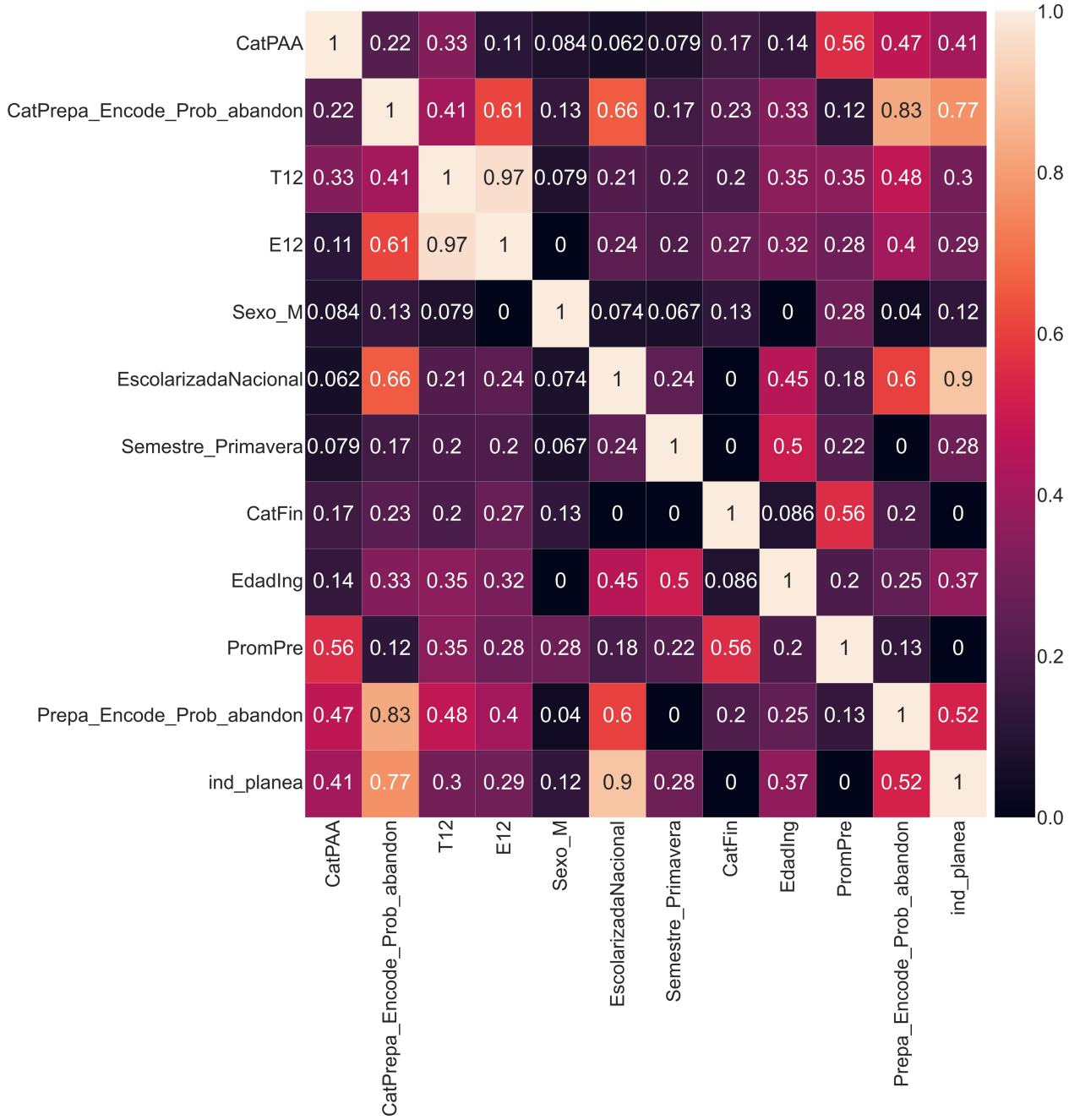


Figura 4-15 Matriz de phik usando una gráfica heatmap de seaborn

4.3.2. Matriz de Significancia

Al comparar la matriz de phik Figura 4-15 y la matriz de significancia Figura 4-16. Se obtuvieron las columnas con mayor significancia, las cuales se encuentran resumidas en la Tabla 4-18, Tabla 4-19, Tabla 4-20 y Tabla 4-21. En la tabla Tabla 4-18 se muestra que para E12, la columna T12 es significantes, lo que tiene sentido dado que E12 requiere de T12 para identificar si un alumno esta censurado. En la

Tabla 4-19 se muestra que para Edad de ingreso la columna EscolarizadaNacional es significante, esto se puede complementar con la información de la Figura 4-11 que muestra una correlación negativa de significancia baja con el Promedio de preparatoria.

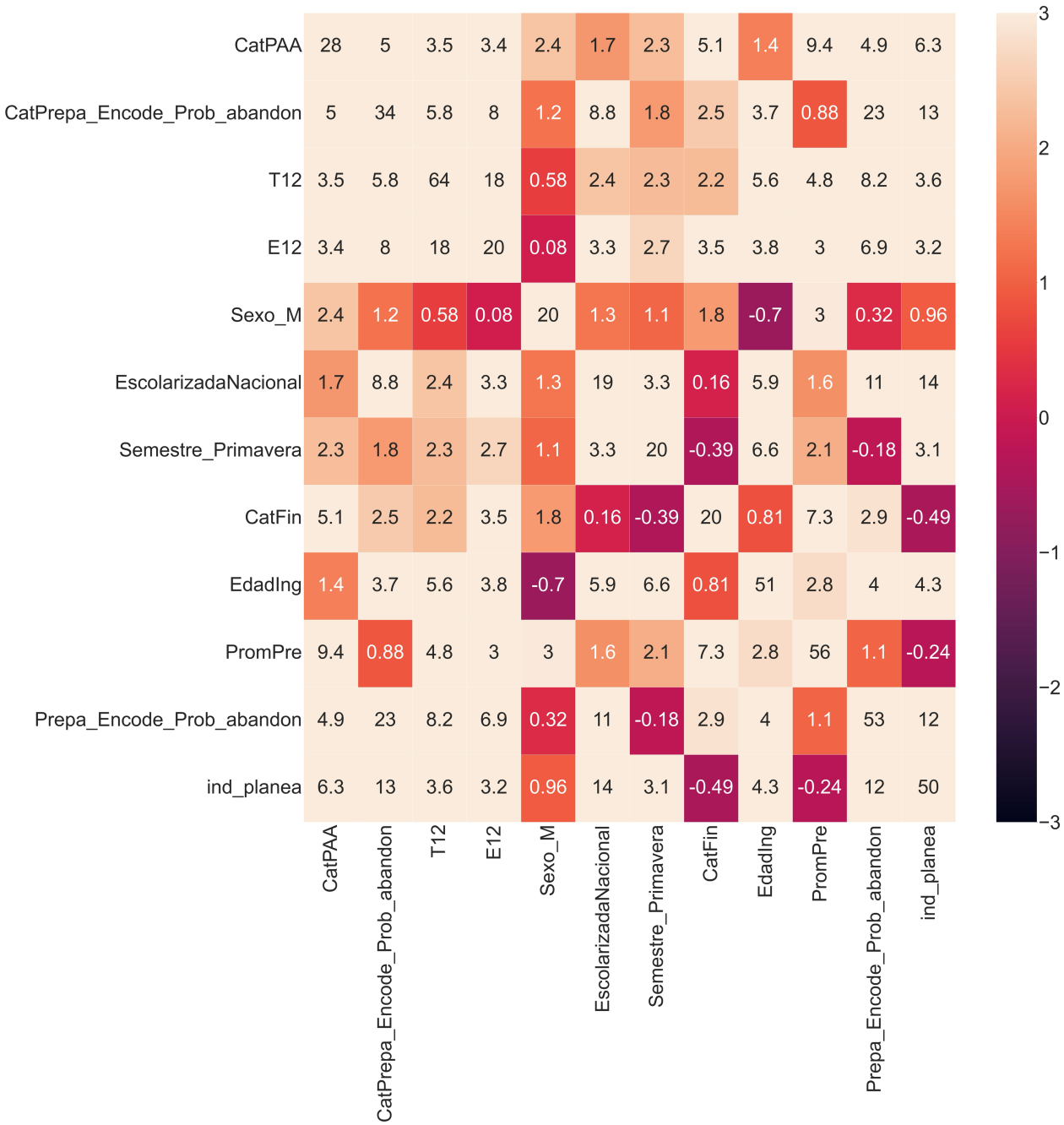


Figura 4-16 Matriz de significancia usando una gráfica heatmap de seaborn

Tabla 4-18 Variables que son significantes con E12 de acuerdo con comparativa de matriz de significancia y matriz phik

Comparación E12 vs las variables que son significantes		
	Matriz Phik	**Matriz de significancia**
CatPrepa_Encode_Prob_abandon	0.6114	7.96
T12	0.9668	18.26
Prepa_Encode_Prob_abandon	0.4026	6.88

Tabla 4-19 Variables que son significantes con EdadIng de acuerdo con comparativa de matriz de significancia y matriz phik

Comparación EdadIng **vs las variables que son significantes**		
	Matriz Phik	**Matriz de significancia**
EscolarizadaNacional	0.4529	5.83
Semestre_Primavera	0.4999	6.63

Tabla 4-20 Variables que son significantes con PromPre de acuerdo con comparativa de matriz de significancia y matriz phik

Comparación PromPre **vs las variables que son significantes**		
	Matriz Phik	**Matriz de significancia**
CatPAA	0.5566	9.39
CatFin	0.5546	7.26

Tabla 4-21 Variables que son significantes con ind_planea de acuerdo con comparativa de matriz de significancia y matriz phik

Comparación ind_planea **vs las variables que son significantes**		
	Matriz Phik	**Matriz de significancia**
CatPAA	0.4047	6.31

Comparación ind_planea **vs las variables que son significantes**		
	Matriz Phik	**Matriz de significancia**
CatPrepa_Encode_Prob_abandon	0.7710	13.16
EscolarizadaNacional	0.8984	13.55
Prepa_Encode_Prob_abandon	0.5242	11.81

5. RESULTADOS Y DISCUSIÓN

En este capítulo primeramente se presentan los resultados de aplicar 6 modelos MCM univariados (Con funciones Lognormal, Weibull, Generalized Gamma, Log-Logistic, Exponential y N Cubic Splines como distribuciones para las porciones que no se curan es decir los alumnos que no egresan), aplicando la ecuación y empleando la biblioteca de python lifelines.

Posteriormente se presentan 3 MCM multivariados (Lognormal, Weibull y Exponential), así como la comparación entre ellos atendiendo a las métricas AIC Y AUC de los mejores modelos. Haciendo uso de diversas bibliotecas de R [4] las cuales son: para modelos MCM, cuRe [45]; para el cálculo de la métrica BIC, flemix [59]*; para generar la combinatoria de variables a usar como predictoras, RcppAlgos [60]; para calcular la métrica AUC, Metrics* [61]*; para el cálculo del AIC se hizo uso de stats [4]*.

Finalmente se presentan las comparativas de los 10 MCM seleccionados contra modelos Cox PH uso de la biblioteca survival [43] incluida como dependencia en la biblioteca cuRe [45];

5.1. Resultados

Se seleccionaron 64 modelos MCM paramétricos multivariados de un total de 1501 modelos probados para predicción de abandono escolar, los cuales se explican por número de variables. En el Apéndice A de la página 92 se muestra una gráfica de dispersión de los 64 modelos que muestra el comportamiento de las funciones de distribución Log Normal, Weibull y Exponencial, en esta grafica se puede observar que los mejores modelos de acuerdo con la métrica de AIC son los Log Normal, lo que se termina de corroborar en la selección de los 10 mejores modelos.

5.1.1. Comparación de MCM Paramétricos

Se compararon 2 tipos de MCM paramétricos. Univariado que se utilizó para encontrar la bondad de ajuste de este tipo de modelos y al cual se refiere como base line o modelo base. Multivariado para seleccionar los modelos que finalmente se utilizaron para la predicción de abandono escolar.

5.1.1.1. Comparación de MCM Paramétrico Univariado

Como resultado de ajustar los modelos Weibull, Exponential, Log Normal, Log Logistic, Generalized Gamma y N Cubic Splines se obtienen las gráficas de la Figura 5-1 que muestra una gráfica de riesgos estimados de los modelos de curación paramétricos y Nelson Aalen un modelo no paramétrico. Nelson Aalen nos dio la función de riesgo estimada sin diferenciar los curados de los no curados. Pero como se puede apreciar en la gráfica subestima el riesgo de abandono en comparación con los modelos mejor calificados que dan una mejor estimación de riesgo.

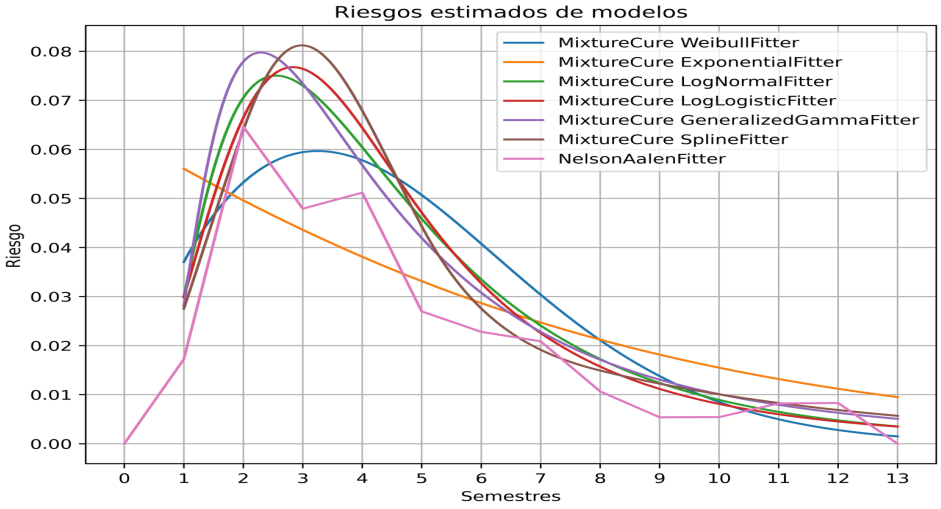


Figura 5-1 Gráfica de riesgos estimados por los modelos MCM univariado

En la Tabla 5-1se muestran los resultados de las comparativas haciendo uso de las métricas AIC y LL. Los 2 modelos que mejor fueron calificados por la métrica AIC son Log Normal y Generalized gamma puesto que tienen una diferencia significativa con respecto a los demás modelos comparados, de aproximadamente 2, 4, 17 y 49 unidades de AIC con respecto a Log Logistic, Spline, Weibull y Exponencial respectivamente. Así mismo los 2 modelos mejor calificados por LL son Generalized gamma y Log Normal con la diferencia en el orden con respecto a AIC, pero como ya se observó en la métrica de AIC esta diferencia es mínima y no significativa. Con los modelos seleccionados se pueden generar modelos MCM multivariados de los cuales se esperan den buenos resultados.

Tabla 5-1 Comparación de MCM (Mixture Cure Model) paramétrico univariado

Modelo de curación MixtureCure	Criterio de Información de Akaike (AIC)	Logaritmo de la Verosimilitud (LL)	Orden por AIC	Orden por LL
LogNormalFitter	1050.89	-522.45	1	2
GeneralizedGammaFitter	1051.35	-521.68	2	1
LogLogisticFitter	1053.60	-523.80	3	3
SplineFitter	1055.85	-523.93	4	4
WeibullFitter	1068.33	-531.16	5	5
ExponentialFitter	1100.72	-548.36	6	6

5.1.1.2. Comparación entre modelos MCM Paramétricos Multivariado

Los modelos MCM que se generaron fueron log normal (ecuación 13), Weibull (ecuación 15) y exponencial (ecuación 19) para la proporción de los que abandonan. Se utilizaron las funciones de enlace logit y loglog complementario, con las funciones antes mencionadas. De igual manera se probaron combinatorias de 1 hasta 8 variables como predictoras para obtener los resultados de distintas métricas. La descripción de las variables puede ser encontrada en la Tabla 4-1.El MCM de forma automática provee la información de LL pero otras métricas tales como AIC, BIC y AUC se tuvieron que calcular con otras bibliotecas de R. Para la parte de selección de modelo las métricas que se tomaron en cuenta principalmente son AIC y AUC. En el caso de los resultados obtenidos con la métrica AUC se hizo uso de los resultados de predicción arrojados por el modelo utilizando la opción "*curerate*" y haciendo uso de la librería de R llamada cuRe [45] la cual esta descrita por Wang y colegas [29]. Para el cálculo de la métrica AUC se hizo uso de la función "auc" de librería Metrics [61] sobre los datos de prueba.

Un total de 1501 modelos MCM fueron los que se probaron y se filtraron los mejores de cada modelo por número de variables dando como resultado 64 modelos los cuales 39.1 % fueron de función exponencial, 34.4 % Log normal y 26.6 % Weibull información que se puede encontrar en la Figura 5-2 y en la Figura 5-3 se pueden encontrar esta misma información pero por número de variables predictoras. De estos modelos finalmente se seleccionaron solo los 10 mejores que utilizan la función log normal de los cuales 4 usan 1 y 2 variables y solo 2 modelos usan 3 variables información que se puede observar en la Figura 5-4. La información detallada se encuentra en las diferentes secciones de este capítulo de resultados.

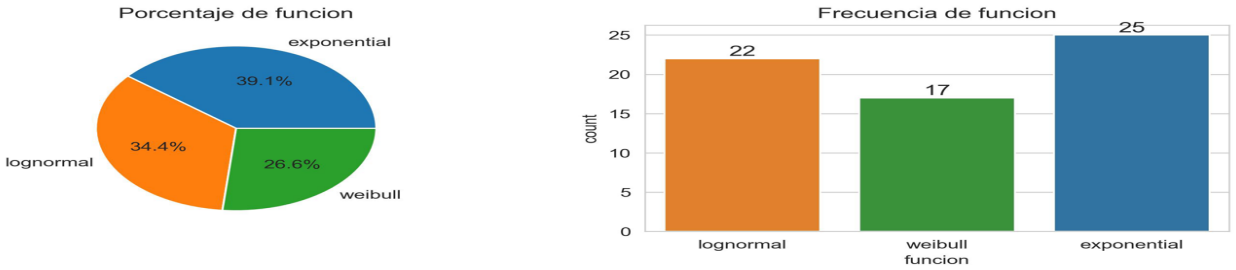


Figura 5-2 Frecuencia y conteo de todos los modelos seleccionados por funciones de distribución

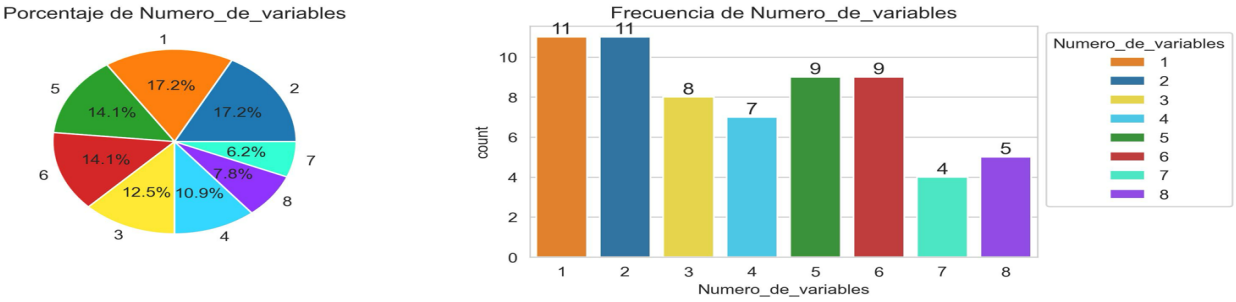


Figura 5-3 Frecuencia y conteo de todos los modelos seleccionados por número de variables

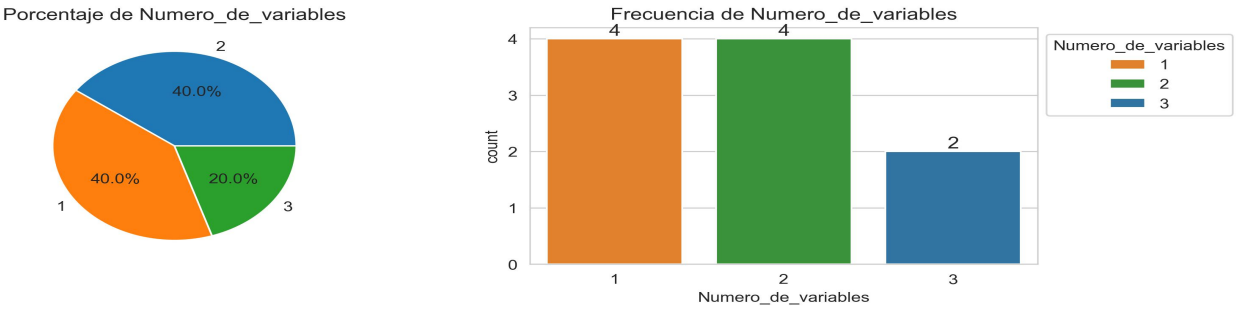


Figura 5-4 Frecuencia y conteo de los modelos finales por número de variables

Resultados de MCM de una variable

En la Tabla 5-2 Modelos seccionados de una variable usando logit y Tabla 5-3 Modelos seccionados de una variable usando loglog se muestran los resultados de modelos con una variable que tienen un AUC

mayor o igual a 0.65 ordenados por mejor AIC de mejor a peor. Lo que deja ver que ambos casos los primeros 2 mejores modelos son los que utilizan la función de distribución log normal para ambas funciones de enlace. También se puede observar que los modelos seleccionados solo hacen uso de las variables PromPre y CatPAA. Esta misma información se puede observar en las gráficas de la Figura 5-5 y Figura 5-6.

Tabla 5-2 Modelos seccionados de una variable usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ PromPre**	**lognormal**	**0.652**	**714.474**	**724.970**	**-351.237**
2	**logit**	**Surv(T12,E12)~ CatPAA**	**lognormal**	**0.651**	**721.601**	**732.097**	**-354.801**
3	**logit**	**Surv(T12,E12)~ CatPAA**	**weibull**	**0.651**	**730.173**	**740.669**	**-359.086**
4	**logit**	**Surv(T12,E12)~ PromPre**	**exponential**	**0.652**	**743.650**	**751.522**	**-367.325**
5	**logit**	**Surv(T12,E12)~ CatPAA**	**exponential**	**0.651**	**750.886**	**758.758**	**-370.943**

Tabla 5-3 Modelos seccionados de una variable usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ PromPre**	**lognormal**	**0.652**	**714.562**	**725.058**	**-351.281**
2	**loglog**	**Surv(T12,E12)~ CatPAA**	**lognormal**	**0.651**	**721.368**	**731.864**	**-354.684**
3	**loglog**	**Surv(T12,E12)~ PromPre**	**weibull**	**0.652**	**723.174**	**733.670**	**-355.587**
4	**loglog**	**Surv(T12,E12)~ CatPAA**	**weibull**	**0.651**	**729.949**	**740.445**	**-358.974**
5	**loglog**	**Surv(T12,E12)~ PromPre**	**exponential**	**0.652**	**743.733**	**751.605**	**-367.367**
6	**loglog**	**Surv(T12,E12)~ CatPAA**	**exponential**	**0.651**	**750.624**	**758.496**	**-370.812**

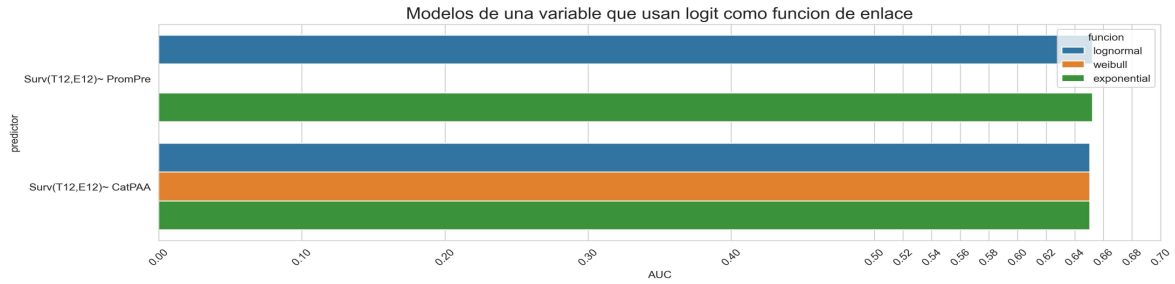


Figura 5-5 Modelos seleccionados con una variable usando logit

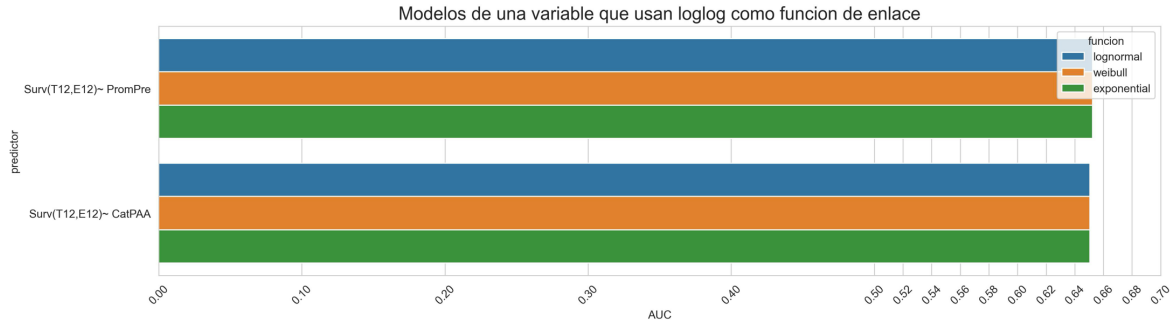


Figura 5-6 Modelos seleccionados con una variable usando loglog

Resultados de MCM de dos variables

En la tabla Tabla 5-4 y Tabla 5-5 se muestran los resultados de modelos de 2 variables que tienen un AUC mayor o igual a 0.69 ordenados por AIC de mejor a peor. Lo que deja ver que al igual que el modelo de una variable en ambos casos los 2 primeros modelos son los que utilizan la función de distribución log normal para ambas funciones de enlace. También se puede observar que todos los modelos seleccionados hacen uso de la variable ind_planea más la variable PromPre o la variable CatPAA. En la Figura 5-7 y Figura 5-8 se puede apreciar visualmente que CatPAA con ind_planea tienen un desempeño menor.

Tabla 5-4 Modelos seccionados de dos variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ PromPre+ind_planea**	**lognormal**	**0.725**	**713.874**	**721.995**	**-346.937**
2	**logit**	**Surv(T12,E12)~ CatPAA+ind_planea**	**lognormal**	**0.695**	**721.930**	**730.051**	**-350.965**
3	**logit**	**Surv(T12,E12)~ CatPAA+ind_planea**	**weibull**	**0.695**	**730.517**	**738.637**	**-355.259**
4	**logit**	**Surv(T12,E12)~ PromPre+ind_planea**	**exponential**	**0.724**	**742.045**	**748.541**	**-363.023**
5	**logit**	**Surv(T12,E12)~ CatPAA+ind_planea**	**exponential**	**0.693**	**750.168**	**756.664**	**-367.084**

Tabla 5-5 Modelos seccionados de dos variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ PromPre+ind_planea**	**lognormal**	**0.725**	**714.102**	**722.223**	**-347.051**
2	**loglog**	**Surv(T12,E12)~ CatPAA+ind_planea**	**lognormal**	**0.699**	**721.894**	**730.014**	**-350.947**
3	**loglog**	**Surv(T12,E12)~ PromPre+ind_planea**	**weibull**	**0.725**	**722.741**	**730.861**	**-351.370**
4	**loglog**	**Surv(T12,E12)~ CatPAA+ind_planea**	**weibull**	**0.699**	**730.483**	**738.604**	**-355.242**
5	**loglog**	**Surv(T12,E12)~ PromPre+ind_planea**	**exponential**	**0.726**	**742.228**	**748.724**	**-363.114**
6	**loglog**	**Surv(T12,E12)~ CatPAA+ind_planea**	**exponential**	**0.696**	**750.120**	**756.616**	**-367.060**

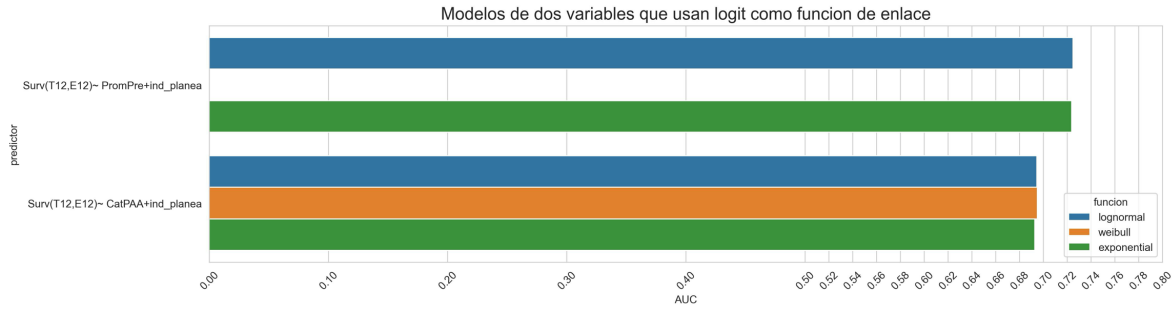


Figura 5-7 Modelos seleccionados con dos variables usando logit

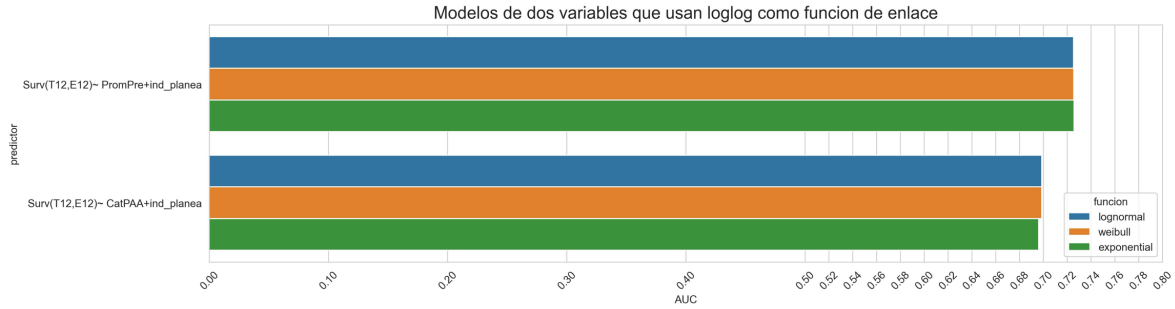


Figura 5-8 Modelos seleccionados con dos variables usando loglog

Resultados de MCM de tres variables

En la Tabla 5-6 y Tabla 5-7 se muestran los resultados de modelos de 3 variables que tienen un AUC mayor o igual a 0.73 ordenados por AIC de mejor a peor. Lo que deja ver que al igual que los modelos anteriores los 2 mejores modelos son los que utilizan la función de distribución log normal para ambas funciones de enlace. También se puede observar que todos los modelos incluyen PrompPre e ind_planea más la variable EdadIng o CatPAA. En la Figura 5-9 y Figura 5-10 se pueden observar desempeños similares para todos los modelos.

Tabla 5-6 Modelos seccionados de tres variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ EdadIng+PromPre+ind_planea**	**lognormal**	**0.734**	**716.243**	**719.987**	**-343.121**
2	**logit**	**Surv(T12,E12)~ CatPAA+PromPre+ind_planea**	**lognormal**	**0.730**	**723.599**	**727.343**	**-346.799**
3	**logit**	**Surv(T12,E12)~ EdadIng+PromPre+ind_planea**	**weibull**	**0.731**	**726.026**	**729.770**	**-348.013**
4	**logit**	**Surv(T12,E12)~ CatPAA+PromPre+ind_planea**	**exponential**	**0.730**	**750.788**	**753.908**	**-362.894**

Tabla 5-7 Modelos seccionados de tres variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ EdadIng+PromPre+ind_planea**	**lognormal**	**0.733**	**716.671**	**720.415**	**-343.336**
2	**loglog**	**Surv(T12,E12)~ CatPAA+PromPre+ind_planea**	**lognormal**	**0.733**	**723.511**	**727.255**	**-346.756**
3	**loglog**	**Surv(T12,E12)~ CatPAA+PromPre+ind_planea**	**weibull**	**0.733**	**732.112**	**735.856**	**-351.056**
4	**loglog**	**Surv(T12,E12)~ CatPAA+PromPre+ind_planea**	**exponential**	**0.732**	**750.733**	**753.853**	**-362.866**

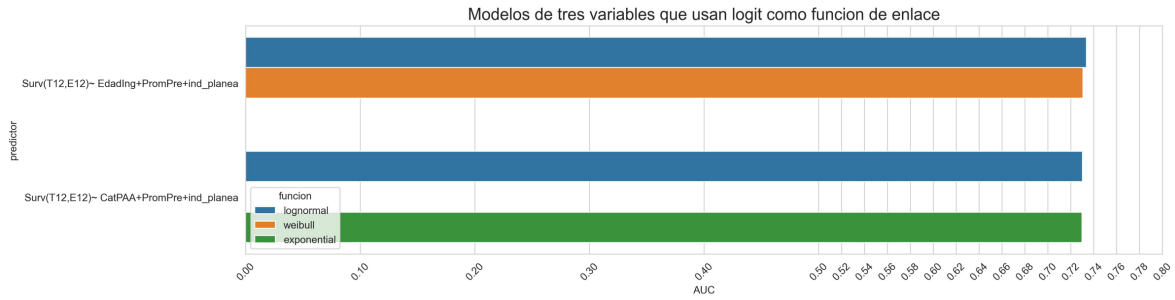


Figura 5-9 Modelos seleccionados con tres variables usando logit

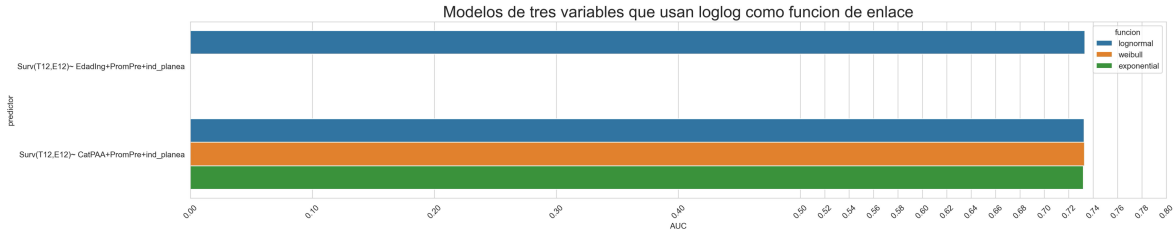


Figura 5-10 Modelos seleccionados con tres variables usando loglog

Resultados de MCM de cuatro variables

En la Tabla 5-8 y Tabla 5-9 se muestran los resultados de modelos de 4 variables que tienen un AUC mayor o igual a 0.734 ordenados por AIC de mejor a peor. Lo que deja ver que en este caso los 2 mejores modelos para la función de enlace logit son los que usan la función de distribución Weibull y para la función de enlace loglog el primer lugar es log normal seguido de un exponencial en segundo y tercer lugar. Además se pueden observar que para todos los modelos se Incluyen las variables PromPre e ind_planea y otras 2 variables como Edading, Sexo_M, EscolarizadaNacional, Semestre_Primavera y CatPAA siendo esta ultima la que más se repite. En la Figura 5-11 y Figura 5-12 se observan desempeños similares para todos los modelos y la selección de 2 funciones de distribución distintas para cada función de enlace.

Tabla 5-8 Modelos seccionados de cuatro variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ EdadIng+CatPAA+PromPre+ind_planea**	**weibull**	**0.736**	**736.062**	**733.430**	**-347.031**
2	**logit**	**Surv(T12,E12)~ Sexo_M+Semestre_Primavera+PromPre+ind_planea**	**weibull**	**0.737**	**748.651**	**746.019**	**-353.325**
3	**logit**	**Surv(T12,E12)~ EdadIng+CatPAA+PromPre+ind_planea**	**exponential**	**0.734**	**754.201**	**751.945**	**-359.100**
4	**logit**	**Surv(T12,E12)~ Semestre_Primavera+CatPAA+PromPre+ind_planea**	**exponential**	**0.737**	**760.118**	**757.862**	**-362.059**

Tabla 5-9 Modelos seccionados de cuatro variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ EscolarizadaNacional+EdadIng+PromPre+ind_planea**	**lognormal**	**0.734**	**728.632**	**726.000**	**-343.316**
2	**loglog**	**Surv(T12,E12)~ Semestre_Primavera+CatPAA+PromPre+ind_planea**	**exponential**	**0.738**	**759.906**	**757.650**	**-361.953**
3	**loglog**	**Surv(T12,E12)~ Sexo_M+CatPAA+PromPre+ind_planea**	**exponential**	**0.735**	**762.412**	**760.157**	**-363.206**

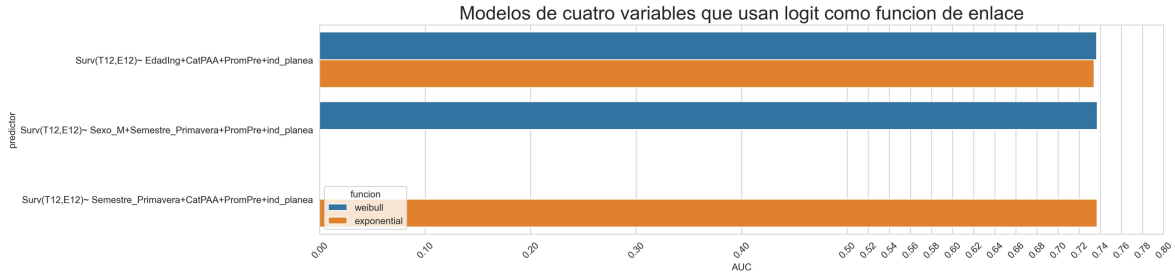


Figura 5-11 Modelos seleccionados con cuatro variables usando logit

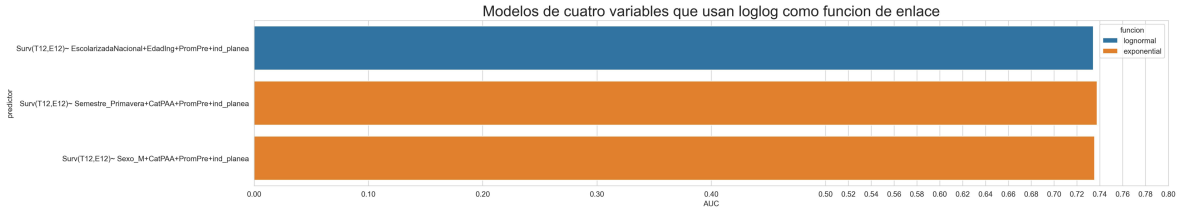


Figura 5-12 Modelos seleccionados con cuatro variables usando loglog

Resultados de MCM de cinco variables

En la Tabla 5-10 y Tabla 5-11 se muestran los resultados de modelos de 5 variables que tienen un AUC mayor o igual a 0.734 ordenados por AIC de mejor a peor. Lo que deja ver que en este caso los 2 mejores modelos para la función de enlace logit son los que usan la función de distribución log normal y en el caso de la función de enlace loglog el primer lugar es Log normal seguido de un Weibull . Además se puede observar que las variables PromPre e ind_planea se observan en todos los modelos seleccionados para ambas funciones de enlace , sumando algunas de las siguientes variables CatPAA, EdadIng, Semestre_Primavera, EscolarizadaNacionaly Sexo_M. En la Figura 5-13 y Figura 5-14 se puede apreciar que la selección de modelos fue muy diferente para modelos con función de enlace logit que para loglog, mostrando una mayor selección para logit.

Tabla 5-10 Modelos seccionados de cinco variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng +PromPre+ind_planea**	**lognormal**	**0.734**	**741.902**	**730.894**	**-342.951**
2	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+EdadIng+PromPre+i nd_planea**	**lognormal**	**0.737**	**742.456**	**731.448**	**-343.228**
3	**logit**	**Surv(T12,E12)~ EscolarizadaNacional+EdadIng+CatPAA+PromPre+i nd_planea**	**weibull**	**0.735**	**750.078**	**739.071**	**-347.039**
4	**logit**	**Surv(T12,E12)~ Semestre_Primavera+EdadIng+CatPAA+PromPre+in d_planea**	**exponential**	**0.738**	**767.150**	**757.518**	**-359.075**
5	**logit**	**Surv(T12,E12)~ EscolarizadaNacional+EdadIng+CatPAA+PromPre+i nd_planea**	**exponential**	**0.735**	**767.155**	**757.523**	**-359.078**

Tabla 5-11 Modelos seccionados de cinco variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**lognormal**	**0.737**	**741.079**	**730.071**	**-342.539**
2	**loglog**	**Surv(T12,E12)~ Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**weibull**	**0.735**	**749.325**	**738.317**	**-346.662**
3	**loglog**	**Surv(T12,E12)~ Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**exponential**	**0.735**	**766.503**	**756.872**	**-358.752**
4	**loglog**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+EdadIng+PromPre+ind_planea**	**exponential**	**0.734**	**767.802**	**758.170**	**-359.401**

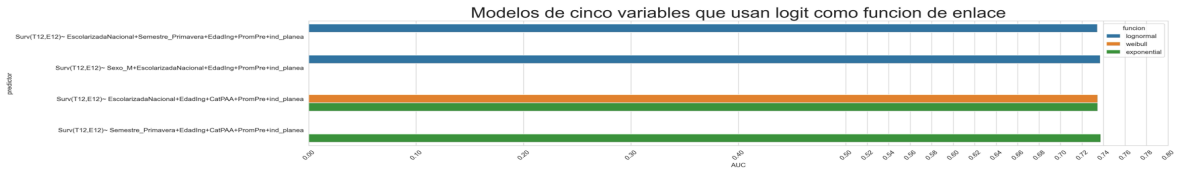


Figura 5-13 Modelos seleccionados con cinco variables usando logit

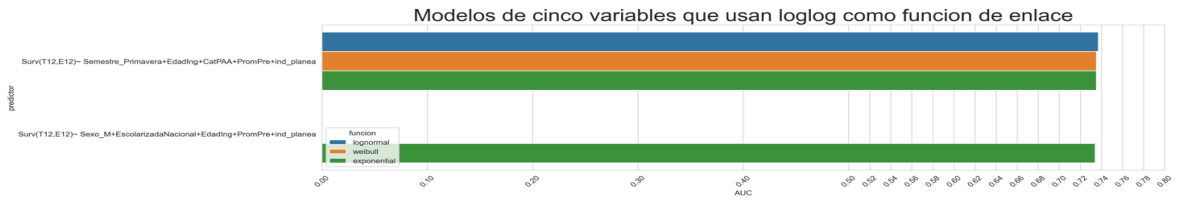


Figura 5-14 Modelos seleccionados con cinco variables usando loglog

Resultados de MCM de seis variables

En la Tabla 5-12 y Tabla 5-13 se muestran los resultados de modelos de 6 variables que tienen un AUC mayor o igual a 0.734 ordenados por AIC de mejor a peor. Se puede observar que este caso los 2 mejores modelos para ambas funciones de enlace, son los que utilizan la función de distribución lognormal. También se observa que para todos los modelos las variables Semestre_Primavera , PromPre e ind_planea son usadas en todos los modelos, combinando las variables Sexo_M, EdadIng, CatPAA y EscolarizadaNacional; esta última variable aparece en casi todos los modelos exceptuando el segundo mejor modelo para enlace con logit. En la Figura 5-15 y Figura 5-16 se puede apreciar que hay una mayor selección de modelos para función de enlace loglog.

Tabla 5-12 Modelos seleccionados con seis variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	logit	Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	lognormal	0.739	757.335	735.951	-342.668
2	logit	Surv(T12,E12)~ Sexo_M+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	lognormal	0.744	757.557	736.173	-342.778
3	logit	Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	weibull	0.735	768.425	747.041	-348.212
4	logit	Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	exponential	0.739	782.659	763.651	-359.329

Tabla 5-13 Modelos seleccionados con seis variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	loglog	Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	lognormal	0.736	757.178	735.794	-342.589
2	loglog	Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+CatPAA+PromPre+ind_planea	lognormal	0.740	763.805	742.421	-345.903
3	loglog	Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+PromPre+ind_planea	weibull	0.740	767.420	746.036	-347.710
4	loglog	Surv(T12,E12)~ EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea	exponential	0.741	781.570	762.562	-358.785
5	loglog	Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+CatPAA+PromPre+ind_planea	exponential	0.746	788.189	769.181	-362.095

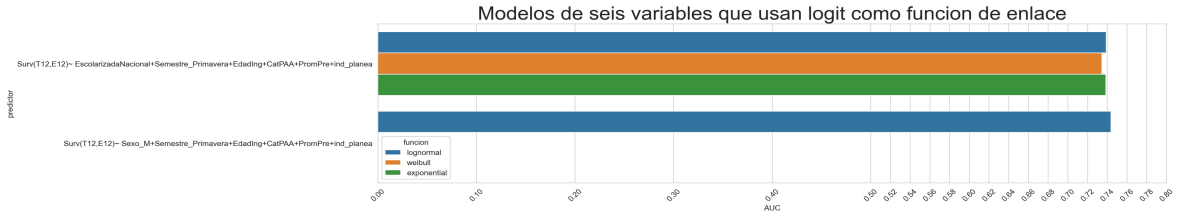


Figura 5-15 Modelos seleccionados con seis variables usando logit

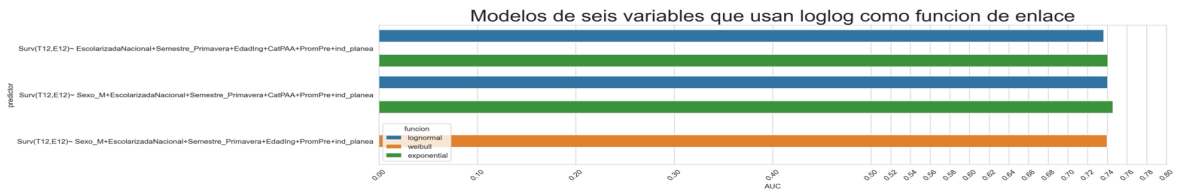


Figura 5-16 Modelos seleccionados con seis variables usando loglog

Resultados de MCM de siete variables

En la Tabla 5-14 y Tabla 5-15 se muestran los resultados de modelos de 7 variables que tienen un AUC mayor o igual a 0.734 ordenados por AIC de mejor a peor. Se puede observar que para este caso los resultados son más variables dado que para el primer lugar usando función de enlace logit tenemos una función de distribución log normal, seguido de un Weibull y un exponencial respectivamente; pero en la función de enlace loglog tenemos un exponencial en primer y único lugar. En cuestión de AUC hay una mejora con respecto a los modelos de 6 variables pero en cantidad de modelos dando un AUC mayor a 0.734 tenemos menos modelos. En la Figura 5-17 y Figura 5-18 se aprecia que tenemos más número de modelos para la función de enlace logit en comparación a loglog donde solo tenemos uno.

Tabla 5-14 Modelos seccionados de siete variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**lognormal**	**0.741**	**775.634**	**741.874**	**-342.817**
2	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**weibull**	**0.740**	**784.659**	**750.899**	**-347.329**
3	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**exponential**	**0.742**	**799.280**	**768.896**	**-359.140**

Tabla 5-15 Modelos seccionados de siete variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+PromPre+ind_planea**	**exponential**	**0.747**	**799.84**	**769.456**	**-359.42**

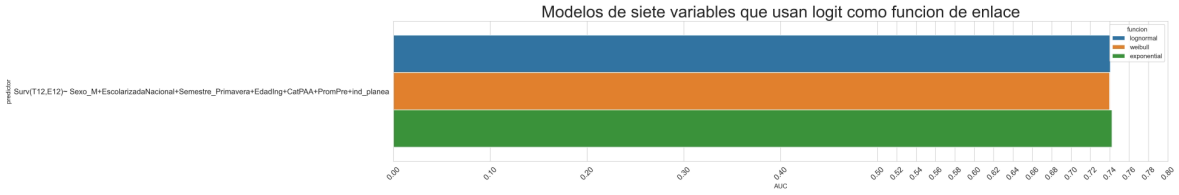


Figura 5-17 Modelos seleccionados con siete variables usando logit

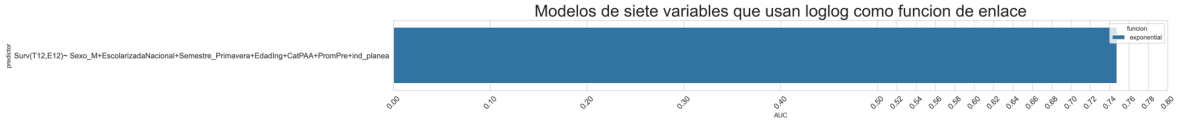


Figura 5-18 Modelos seleccionados con siete variables usando loglog

Resultados de MCM de ocho variables

En la Tabla 5-16, Tabla 5-17 y en la Figura 5-19 se muestran los resultados de los modelos de 8 variables ordenados por AIC de mejor a peor. Se puede observar que tenemos en primer lugar a una distribución Weibull, seguido de exponencial para la función de enlace logit; en el caso de la función de enlace loglog tenemos en primer lugar una función de distribución log normal, seguido de Weibull y exponencial respectivamente. Algo que es importante mencionar y que no se había mencionado es que algunos de los modelos que se generaron no pudieron converger y dar resultados por lo que no aparecen en la tabla de resultados como es el caso de la distribución log normal para la función de enlace logit.

Tabla 5-16 Modelos seccionados de ocho variables usando logit

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+CatFin+PromPre+ind_planea**	**weibull**	**0.678**	**803.573**	**755.437**	**-346.786**
2	**logit**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+CatFin+PromPre+ind_planea**	**exponential**	**0.705**	**812.738**	**768.978**	**-356.369**

Tabla 5-17 Modelos seccionados de ocho variables usando loglog

Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
1	**loglog**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+CatFin+PromPre+ind_planea**	**lognormal**	**0.704**	**792.931**	**744.795**	**-341.465**
2	**loglog**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+CatFin+PromPre+ind_planea**	**weibull**	**0.750**	**805.993**	**757.857**	**-347.997**
3	**loglog**	**Surv(T12,E12)~ Sexo_M+EscolarizadaNacional+Semestre_Primavera+EdadIng+CatPAA+CatFin+PromPre+ind_planea**	**exponential**	**0.698**	**816.553**	**772.794**	**-358.277**

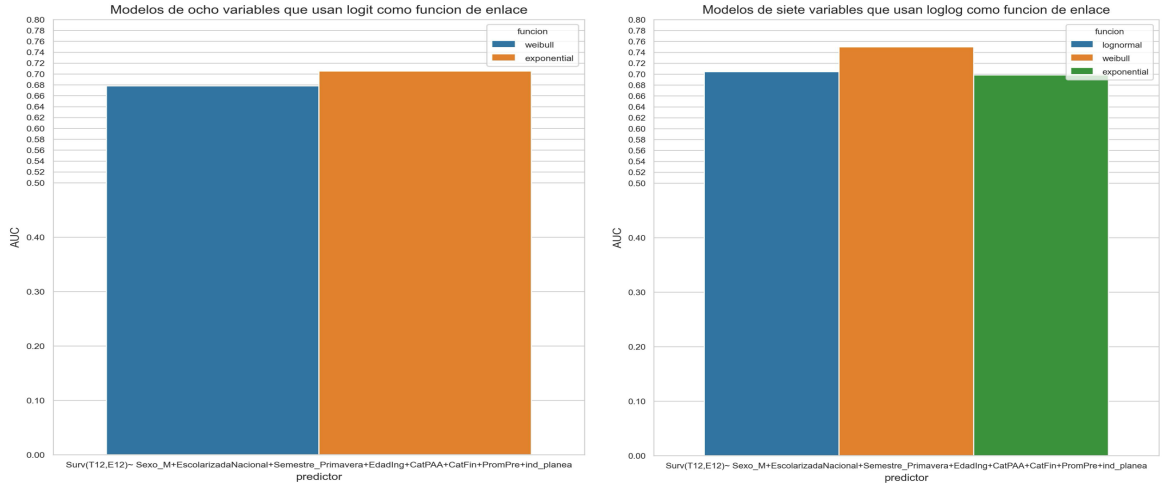


Figura 5-19 Todos los resultados de los modelos probados con ocho variables usando logit y loglog

5.2. Discusión

Para la selección de los 10 mejores modelos se creó una tabla con los resultados de los mejores modelos por variables y se ordenó por AIC y AUC de mejor a peor. Como se puede observar en la Tabla 5-18. los 10 mejores modelos usan la función distribución log normal, de estos modelos los mejores 2 marcados en verde tienen el mejor AIC con un buen valor de AUC, con una diferencia no significativa entre ellos con el uso de 2 variables; pero que contrasta con el siguiente par de modelos que tienen el mejor AUC que están marcados en color azul, que son el tercer y cuarto lugar con una separación de 2 unidades de AIC lo que los hacen significativamente peor que los primeros 2 modelos. Además se incluyen una gráfica de riesgo y otra de supervivencia para el modelo base que muestra la bondad de ajuste en la Figura 5-20 y para todos los modelos seleccionados la predicción de riesgo y supervivencia en el orden correspondiente de selección en la Tabla 5-18, en primer lugar Figura 5-21 que muestra un riesgo máximo de 0.23 en el tercer semestre y una supervivencia mínima de alrededor de 0.25; en segundo lugar la Figura 5-22 que muestra resultados muy similares al primer modelo; en tercer lugar la Figura 5-23 que muestra un riesgo de 0.20 que es 0.3 menor que los 2 primeros modelos y una supervivencia mínima para un alumno de alrededor de 0.26 y en cuarto lugar Figura 5-24 que muestra un riesgo máximo de 0.17 que es mucho menor que los primeros 3 modelos y con una supervivencia mínima de alrededor de 0.39.

Tabla 5-18 Los 10 mejores modelos seccionados

Num. variables	Orden	enlace	predictor	funcion	AUC	AIC	BIC	LL
2	1	logit	Surv(T12,E12)~ PromPre+ind_planea	lognormal	0.725	713.874	721.995	-346.937
2	2	loglog	Surv(T12,E12)~ PromPre+ind_planea	lognormal	0.725	714.102	722.223	-347.051
3	3	logit	Surv(T12,E12)~ EdadIng+PromPre+ind_planea	lognormal	0.734	716.243	719.987	-343.121
3	4	loglog	Surv(T12,E12)~ EdadIng+PromPre+ind_planea	lognormal	0.733	716.671	720.415	-343.336
1	5	logit	Surv(T12,E12)~ PromPre	lognormal	0.652	714.474	724.970	-351.237
1	6	loglog	Surv(T12,E12)~ PromPre	lognormal	0.652	714.562	725.058	-351.281
1	7	loglog	Surv(T12,E12)~ CatPAA	lognormal	0.651	721.368	731.864	-354.684
1	8	logit	Surv(T12,E12)~ CatPAA	lognormal	0.651	721.601	732.097	-354.801
2	9	loglog	Surv(T12,E12)~ CatPAA+ind_planea	lognormal	0.699	721.894	730.014	-350.947
2	10	logit	Surv(T12,E12)~ CatPAA+ind_planea	lognormal	0.695	721.930	730.051	-350.965

En la Tabla 5-19 se puede observar una comparativa de los 10 MCM seleccionados con los modelos Cox PH en la cual se compara haciendo uso de AUC y se muestra como referencia el índice de concordancia el cual no se pudo obtener para MCM por las limitantes de la implementación utilizada. Como se puede observar para los 4 primeros modelos seleccionados se muestra un mejor desempeño de los modelos MCM, para los demás modelos se muestran desempeños similares, los modelos MCM siguen teniendo mejor desempeño en la mayoría de los casos a excepción de los modelos 7 y 8 marcados en amarillo donde Cox PH muestra mejor desempeño que los MCM.

Tabla 5-19 Comparación de modelos MCM vs CoxPH

Orden	predictor	AUC MCM	AUC Cox PH	C Index Cox PH
1	Surv(T12,E12)~ PromPre+ind_planea	0.725	0.525	0.692
2	Surv(T12,E12)~ PromPre+ind_planea	0.725	0.525	0.692
3	Surv(T12,E12)~ EdadIng+PromPre+ind_planea	0.734	0.515	0.702
4	Surv(T12,E12)~ EdadIng+PromPre+ind_planea	0.733	0.515	0.702
5	Surv(T12,E12)~ PromPre	0.652	0.632	0.623
6	Surv(T12,E12)~ PromPre	0.652	0.632	0.623
7	Surv(T12,E12)~ CatPAA	0.651	0.662	0.607
8	Surv(T12,E12)~ CatPAA	0.651	0.662	0.607
9	Surv(T12,E12)~ CatPAA+ind_planea	0.699	0.602	0.663
10	Surv(T12,E12)~ CatPAA+ind_planea	0.695	0.602	0.663

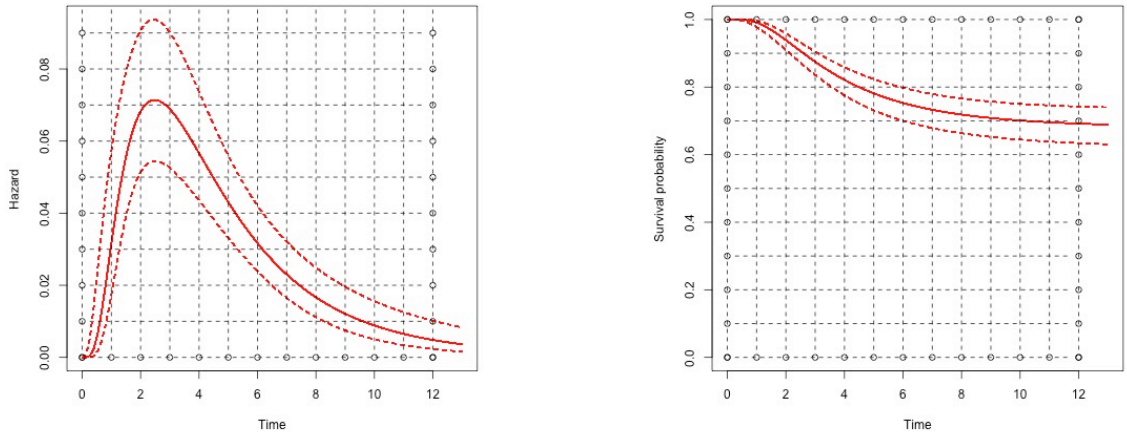


Figura 5-20 Riesgo y supervivencia para modelo base

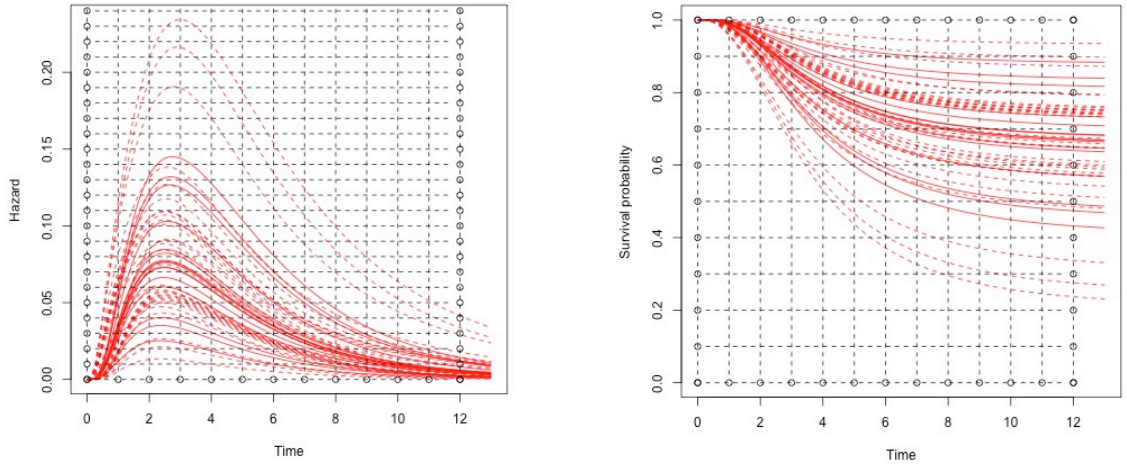


Figura 5-21 Riesgo y supervivencia para primer modelo seleccionado Surv(T12,E12)~ PromPre+ind_planea con logit

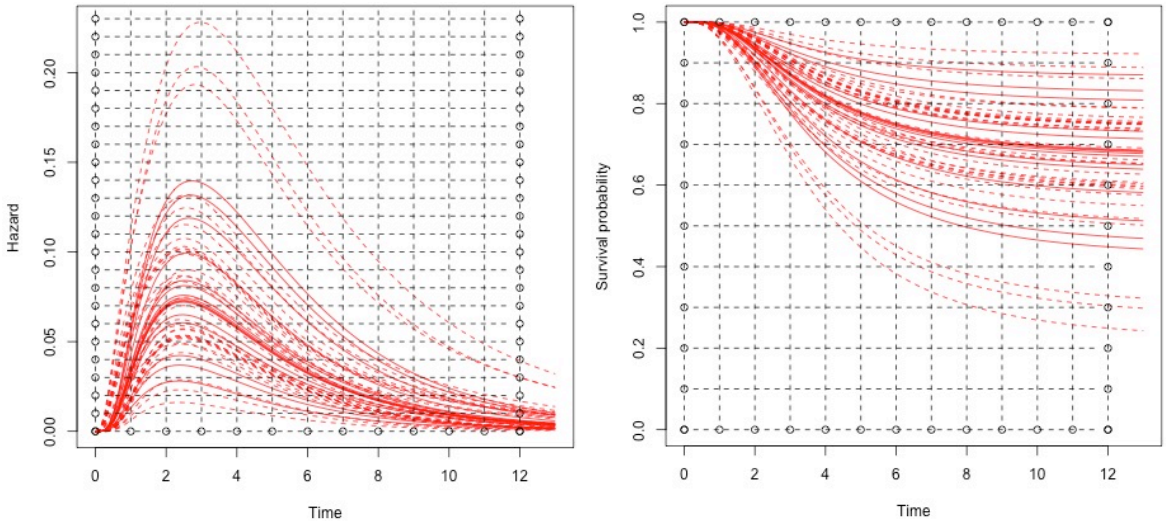


Figura 5-22 Riesgo y supervivencia para segundo modelo seleccionado Surv(T12,E12)~ PromPre+ind_planea con loglog

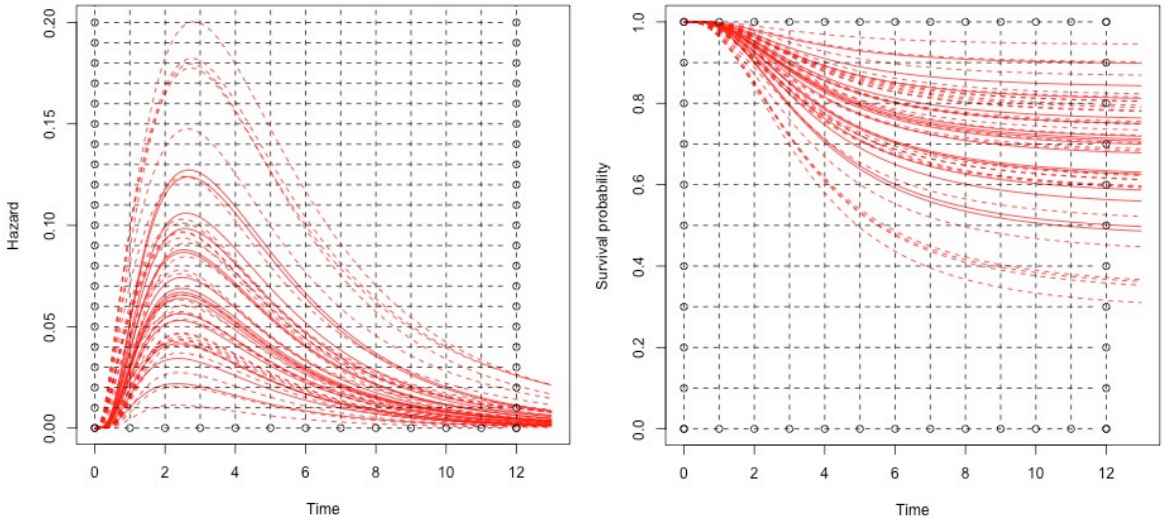


Figura 5-23 Riesgo y supervivencia para tercer modelo seleccionado Surv(T12,E12)~ EdadIng+PromPre+ind_planea con logit

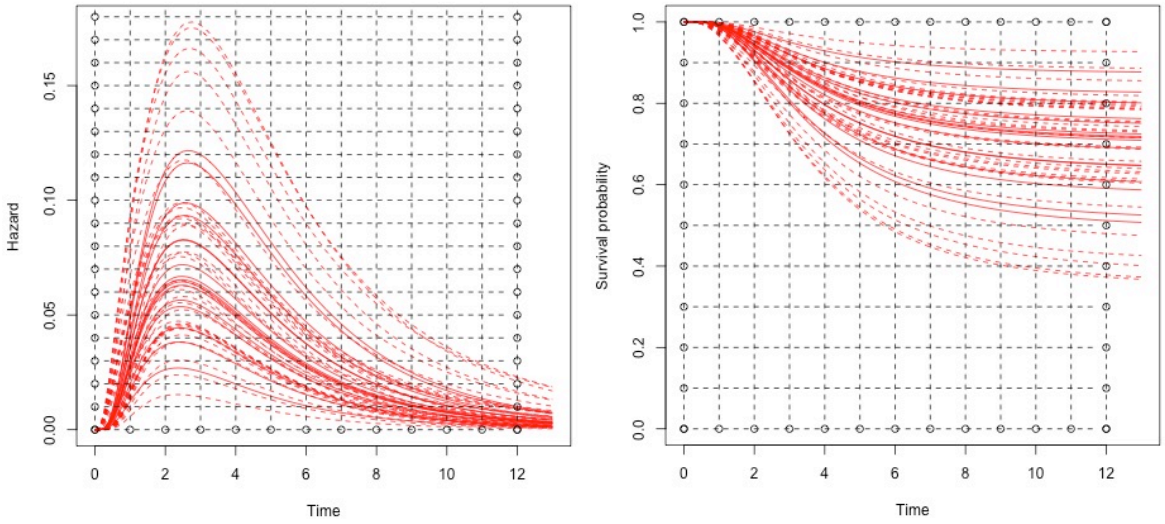


Figura 5-24 Riesgo y supervivencia para cuarto modelo seleccionado Surv(T12,E12)~ EdadIng+PromPre+ind_planea con loglog

6. CONCLUSIONES

En este capítulo se presentan las conclusiones y trabajo futuro en relación con los modelos de curación mixtos multivariados para la predicción de abandono escolar y trabajo futuro que se desprende de las conclusiones de este trabajo.

6.1. Conclusiones

Al probar bibliotecas de Python [3] se encontraron dificultades al no encontrar una alternativa de uso de MCM Paramétrico multivariado dado que la biblioteca lifelines [44] que para este momento es la única que soporta modelos de curación mixtos, solo tiene la opción para utilizar modelos univariados; es decir que solo se puede probar la bondad de ajuste pero no generar modelos que utilicen otras variables además de las de tiempo y eventos; debido a esto se optó por utilizar R [4] haciendo uso de una librería llamada cuRe [45] que soporta los modelos MCM Paramétricos multivariados; además cuenta con documentación aceptable, aunque le falta incluir más ejemplos de uso.

El objetivo de generar modelos MCM paramétricos multivariados para la predicción de abandono escolar se ha cumplido. Un total de 1501 modelos MCM generados fueron comparados con funciones de distribución log normal, Weibull y exponencial; así como con 2 funciones de enlace las cuales son logit y loglog. Estos modelos se compararon por número de variables; en este caso en específico se probaron modelos MCM que hacen uso de 1 variable hasta un máximo de 8 variables, con lo cual resulto un total de 64 modelos seleccionados y de los cuales al final se seleccionaron 10 modelos que muestran buen desempeño pero de los cuales la sugerencia es quedarse con el primer y segundo modelo de la lista en la Tabla 5-18, dado que proveen el AIC más bajo que indica el mejor modelo con un valor de 713.874 y 714.102 respectivamente; y el AUC más alto que indica el mejor rendimiento con un valor de 0.725 para ambos modelos. Los 2 modelos sugeridos tienen un AIC menor a 2 unidades de diferencia entre ambos lo que indica que entre estos 2 modelos no hay diferencia significativa entre cual es mejor; pero si hay diferencia con los demás modelos; adicional a esto la función de distribución que mostró mejor desempeño para los modelos MCM es la función log normal sin importar la función de enlace.

Los 10 modelos MCM seleccionados se compararon contra modelos de Cox PH en los cuales los modelos MCM mostraron mejor desempeño que los de Cox PH al compararlos con la métrica de AUC; de estos 10 modelos solo 2 modelos de Cox PH mostraron un mejor desempeño que los MCM los cuales son los modelos 7 y 8 los cuales no están dentro de los modelos recomendados para utilizar.

Durante las pruebas se pudo encontrar que las variables utilizadas por los modelos con mejores resultados son: 1) el índice de calidad de una preparatoria en la prueba planea 2015 (ind_planea) variable generada en el trabajo de Hernández-Chávez (2025)[1]; 2) el promedio obtenido por el alumno en preparatoria (PromPre) ; 3) categoría para los alumnos dependiendo la puntación en la Prueba de Aptitud Académica (CatPAA) y 4) la edad en el momento del ingreso (EdadIng). Que corrobora la información encontrada durante el análisis exploratorio de las variables, que sin duda es una parte importante para la generación de modelos.

Teniendo todo lo anterior en cuenta los modelos MCM muestran potencial para la predicción de abandono escolar comparados contra los modelos de Cox PH que actualmente están siendo utilizados.

6.2. Trabajo Futuro

Mis principales recomendaciones para trabajo futuro son:

- Recomiendo generar modelos NMCM y MCM semi-paramétricos y compararlos con los resultados de este trabajo.
- Realizar comparación de grupos (por cada variable categórica) en el mejor modelo haciendo uso de la métrica RMST (Restricted Mean Survival Time) que es una alternativa para medida para resumir el perfil de supervivencia [62] .
- Construir estos modelos no solo en el ingreso, sino después del primer semestre que es el momento de mayor abandono, pudiendo incorporar variables como el promedio del primer semestre, la cantidad créditos matriculados, etc.
- Construir modelos para otros programas de estudio y un modelo de curación general para la universidad.

BIBLIOGRAFÍA

[1] G. Hernández Chávez, “Aplicación de Bosques de Supervivencia Aleatorios a la Predicción de Abandono Universitario,” UNIVERSIDAD DE GUADALAJARA, Zapopan, Jalisco, 2025.

[2] M. Amico, I. Van Keilegoms, M. Maïlis Amico, and I. Van Keilegom, “Cure models in survival analysis,” 2017.

[3] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[4] R Core Team, “R: A Language and Environment for Statistical Computing,” 2024, *Vienna, Austria*. [Online]. Available: https://www.R-project.org/

[5] A. G. Hernandez Gonzalez, R. A. Melendez Armenta, L. A. Morales Rosales, A. Garcia Barrientos, J. L. Tecpanecatl Xihuitl, and I. Algredo, “Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico,” *IEEE Latin America Transactions*, vol. 14, no. 11, pp. 4573–4578, 2016, doi: 10.1109/TLA.2016.7795831.

[6] B. R. Cuji Chacha, W. L. Gavilanes López, V. X. Vicente Guerrero, and W. G. Villacis Villacis, “Student Dropout Model Based on Logistic Regression,” in *Applied Technologies*, M. and T.-C. P. and M. L. S. and P. V. G. and D. B. Botto-Tobar Miguel and Zambrano Vizuete, Ed., Cham: Springer International Publishing, 2020, pp. 321–333.

[7] M. Alban and D. Mauricio, “Neural networks to predict dropout at the universities,” *Int J Mach Learn Comput*, vol. 9, no. 2, pp. 149–153, Apr. 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[8] M. Escobar-Bach and I. Van Keilegom, “Nonparametric estimation of conditional cure models for heavy-tailed distributions and under insufficient follow-up,” Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.08436

[9] P. Mthisi, “MSc Dissertation An Application of Survival Analysis in Credit Risk Management,” University of the Witwatersrand, Johannesburg, 2022. Accessed: Nov. 24, 2023. [Online]. Available: https://wiredspace.wits.ac.za/items/3ecf191e-1f45-497c-bdf0-6488675ef749

[10] Z. Al-Khayat, T. Vincken, A. Verhoek, B. Heeg, and M. Ouwens, “POSC322 The Impact of Treatment Coefficient Parameterization for Mixture Cure Models on Cost-Effectiveness Models,” *Value in Health*, vol. 25, no. 1, p. S210, Jan. 2022, doi: 10.1016/j.jval.2021.11.1026.

[11] K. Thompson, X. Koolman, and F. Portrait, “Height and marital outcomes in the Netherlands, birth years 1841-1900,” *Econ Hum Biol*, vol. 41, p. 100970, May 2021, doi: 10.1016/j.ehb.2020.100970.

[12] Y. Peng and J. M. G. Taylor, “Cure models,” in *Handbook of survival analysis*, vol. 34, J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, and T. H. Scheike, Eds., CRC Press, Boca Raton, FL, 2014, pp. 113–134.

[13] C. Raúl and F. Regalado, “Bayes’s theorem and its use in diagnostic test lectures in clinical laboratory,” *Revista Cubana de Investigaciones Biomédicas*, vol. 28, no. 3, pp. 158–165, 2009, [Online]. Available: http://scielo.sld.cu158

[14] M. Amico, “Cure models in survival analysis: from modelling to prediction assessment of the cure fraction,” PhD thesis, Louvain Catholic University and Leuven Catholic University, 2018.

[15] V. Patilea and I. Van Keilegom, “A general approach for cure models in survival analysis,” *The Annals of Statistics*, vol. 48, no. 4, pp. 2323–2346, Aug. 2020, doi: 10.1214/19-AOS1889.

[16] S. Glen, “Heavy Tailed Distribution & Light Tailed Distribution: Definition & Examples,” StatisticsHowTo.com: Elementary Statistics for the rest of us! Accessed: Nov. 24, 2023. [Online]. Available: https://www.statisticshowto.com/heavy-tailed-distribution/

[17] N. Bonginkosi Duncan, “Modelling Time To Graduation of Durban University of Technology Students Using Event History Analysis(Doctoral dissertation),” Durban University, Kwa-Zulu Natal, 2015.

[18] S. Glen, “Multinomial Distribution: Definition, Examples,” StatisticsHowTo.com: Elementary Statistics for the rest of us! Accessed: Nov. 25, 2023. [Online]. Available: https://www.statisticshowto.com/multinomial-distribution/

[19] N. S. L. Mathews de, J. B. Fachini Gomes, M. Holanda, C. C. Koike, and M. T. Leao Costa, “Study on Computer Science Undergraduate Students Dropout at the University of Brasilia,” in *2023 IEEE Frontiers in Education Conference (FIE)*, IEEE, Oct. 2023, pp. 1–7. doi: 10.1109/FIE58773.2023.10343503.

[20] H. Shi and Y. Zhou, “Stay or Leave? Exploring Student Factors Associated with Dropout Patterns in Massive Open Online Courses,” in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, IEEE, Jul. 2023, pp. 26–30. doi: 10.1109/ICALT58122.2023.00013.

[21] M. KLITZKE and F. CARVALHAES, “STUDENT DROPOUT IN A BRAZILIAN PUBLIC UNIVERSITY: A SURVIVAL ANALYSIS,” *Educação em Revista*, vol. 39, 2023, doi: 10.1590/0102-469837576t.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases) (© AAAI),” 1996. doi: https://doi.org/10.1609/aimag.v17i3.1230.

[23] W. A. Shewhart *et al.*, “WILEY SERIES IN PROBABILITY AND STATISTICS,” 2002. [Online]. Available: www.copyright.com.

[24] D. G. Kleinbaum and Mitchel. Klein, *Survival analysis : A self-learning text*, Second. Springer, 2005.

[25] J. P. Klein, H. C. Van Houwelingen, J. G. Ibrahim, and T. H. Scheike, *Handbook of Survival Analysis*. Chapman and Hall/CRC, 2014. [Online]. Available: https://www.taylorfrancis.com/books/9781466555679

[26] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br J Cancer*, vol. 89, pp. 232–238, 2003, doi: 10.1038/sj.bjc.6601118.

[27] O. O. Aalen, Ø. Borgan, and H. K. Gjessing, “Nonparametric analysis of survival and event history data,” in Statistics for Biology and Health. , New York, NY: Springer New York, 2008, pp. 69–130. doi: 10.1007/978-0-387-68560-1_3.

[28] Ø. Borgan, “Three contributions to the encyclopedia of biostatistics: the Nelson-Aalen, Kaplan-Meier, and Aalen-Johansen,” *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*, 1997.

[29] P. Wang, Y. Li, and C. K. Reddy, “Machine Learning for Survival Analysis: A Survey,” Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.04649

[30] J. M. Bland and D. G. Altman, “The logrank test,” *BMJ*, vol. 328, no. 7447, p. 1073, May 2004, doi: 10.1136/bmj.328.7447.1073.

[31] G. Shmueli, “To Explain or to Predict?,” *Statistical Science*, vol. 25, no. 3, pp. 289–310, Aug. 2010, doi: 10.1214/10-STS330.

[32] E. O. Ogundimu, D. G. Altman, and G. S. Collins, “Adequate sample size for developing prediction models is not simply related to events per variable,” *J Clin Epidemiol*, vol. 76, pp. 175–182, Aug. 2016, doi: 10.1016/j.jclinepi.2016.02.031.

[33] G. S. Collins, E. O. Ogundimu, and D. G. Altman, “Sample size considerations for the external validation of a multivariable prognostic model: a resampling study,” *Stat Med*, vol. 35, no. 2, pp. 214–226, Jan. 2016, doi: 10.1002/sim.6787.

[34] F. B. GONÇALVES and P. FRANKLIN, “On the definition of likelihood function,” *arXiv preprint arXiv:1906.10733*, 2019, doi: https://doi.org/10.48550/arXiv.1906.10733.

[35] ¿Qué es log-verosimilitud?, “¿Qué es log-verosimilitud? - Minitab.” Accessed: Nov. 23, 2024. [Online]. Available: https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/regression/supporting-topics/regression-models/what-is-log-likelihood/

[36] J. Aldrich, “R.A. Fisher and the making of maximum likelihood 1912-1922,” *Statistical Science*, vol. 12, no. 3, Sep. 1997, doi: 10.1214/ss/1030037906.

[37] S. Portet, “A primer on model selection using the Akaike Information Criterion,” *Infect Dis Model*, vol. 5, pp. 111–128, Jan. 2020, doi: 10.1016/j.idm.2019.12.010.

[38] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans Automat Contr*, vol. 19, no. 6, pp. 716–723, Dec. 1974, doi: 10.1109/TAC.1974.1100705.

[39] Bevans Rebecca, “Akaike Information Criterion | When & How to Use It (Example).” Accessed: Nov. 23, 2024. [Online]. Available: https://www.scribbr.com/statistics/akaike-information-criterion/

[40] G. Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics*, vol. 6, no. 2, Mar. 1978, doi: 10.1214/aos/1176344136.

[41] E. A. Mohammed, C. Naugler, and B. H. Far, “Chapter 32 - Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics,” in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*, Q. N. Tran and H. Arabnia, Eds., in Emerging Trends in Computer Science and Applied Computing. , Boston: Morgan Kaufmann, 2015, pp. 577–602. doi: https://doi.org/10.1016/B978-0-12-802508-6.00032-6.

[42] Martí Pol, “Entendiendo la curva ROC y el AUC: Dos medidas del rendimiento de un clasificador binario que van de la mano. – Pol Martí Sanahuja.” Accessed: Nov. 23, 2024. [Online]. Available: https://polmartisanahuja.com/entendiendo-la-curva-roc-y-el-auc-dos-medidas-del-rendimiento-de-un-clasificador-binario-que-van-de-la-mano/

[43] T. M. Therneau, “A Package for Survival Analysis in R,” 2024. [Online]. Available: https://CRAN.R-project.org/package=survival

[44] Cam Davidson-Pilon, “Estimating univariate models — lifelines 0.30.0 documentation.” Accessed: Mar. 01, 2025. [Online]. Available: https://lifelines.readthedocs.io/en/latest/Survival%20analysis%20with%20lifelines.html

[45] Lasse Hjort Jakobsen, “cuRe: Parametric Cure Model Estimation,” 2023. [Online]. Available: https://CRAN.R-project.org/package=cuRe

[46] A. Chakrabarty, S. Mannan, and T. Cagin, “Chapter 7 - Equipment Failure,” in *Multiscale Modeling for Process Safety Applications*, A. Chakrabarty, S. Mannan, and T. Cagin, Eds., Boston: Butterworth-Heinemann, 2016, pp. 309–338. doi: https://doi.org/10.1016/B978-0-12-396975-0.00007-3.

[47] M. Verma, “Wind Farm Repowering Using WAsP Software – An Approach for Reducing CO2 Emissions in the Environment,” in *Encyclopedia of Renewable and Sustainable Materials*, S. Hashmi and I. A. Choudhury, Eds., Oxford: Elsevier, 2020, pp. 844–859. doi: https://doi.org/10.1016/B978-0-12-803581-8.11001-X.

[48] O. Gomes, C. Combes, and A. Dussauchoy, “Parameter estimation of the generalized gamma distribution,” *Math Comput Simul*, vol. 79, no. 4, pp. 955–963, 2008, doi: https://doi.org/10.1016/j.matcom.2008.02.006.

[49] D. D. Hanagal, “Chapter 9 - Frailty Models in Public Health,” in *Disease Modelling and Public Health, Part B*, vol. 37, A. S. R. Srinivasa Rao, S. Pyne, and C. R. Rao, Eds., in Handbook of Statistics, vol. 37. , Elsevier, 2017, pp. 209–247. doi: https://doi.org/10.1016/bs.host.2017.07.004.

[50] C. S. Young, “Chapter 2 - The fundamentals of security risk measurements,” in *Metrics and Methods for Security Risk Management*, C. S. Young, Ed., Boston: Syngress, 2010, pp. 19–43. doi: https://doi.org/10.1016/B978-1-85617-978-2.00008-3.

[51] P. Royston and M. K. B. Parmar, “Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects,” *Stat Med*, vol. 21, no. 15, pp. 2175–2197, Aug. 2002, doi: 10.1002/sim.1203.

[52] R. K. Jensen, M. Clements, L. K. Gjærde, and L. H. Jakobsen, “Fitting parametric cure models in R using the packages cuRe and rstpm2,” *Comput Methods Programs Biomed*, vol. 226, p. 107125, Nov. 2022, doi: 10.1016/j.cmpb.2022.107125.

[53] N. R. Latimer and M. J. Rutherford, “Mixture and Non-mixture Cure Models for Health Technology Assessment: What You Need to Know,” *Pharmacoeconomics*, vol. 42, no. 10, pp. 1073–1090, Oct. 2024, doi: 10.1007/s40273-024-01406-7.

[54] T. pandas development team, “pandas-dev/pandas: Pandas,” Sep. 2024, *Zenodo*. doi: 10.5281/zenodo.3509134.

[55] J. Hunter, D. Dale, E. Firing, and M. Droettboom, “Pyplot tutorial — Matplotlib 3.10.1 documentation.” Accessed: Mar. 07, 2025. [Online]. Available: https://matplotlib.org/stable/tutorials/pyplot.html

[56] M. Waskom, “seaborn: statistical data visualization,” *J Open Source Softw*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/JOSS.03021.

[57] M. Baak, R. Koopman, H. Snoek, and S. Klous, “A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics,” *Comput Stat Data Anal*, vol. 152, p. 107043, 2020.

[58] Making sense of the ϕ_k correlation coefficient - Cross Validated, “Making sense of the ϕ_k correlation coefficient - Cross Validated.” Accessed: Nov. 23, 2024. [Online]. Available: https://stats.stackexchange.com/questions/497742/making-sense-of-the-phi-k-correlation-coefficient

[59] B. Gruen and F. Leisch, “flexmix: Flexible Mixture Modeling,” 2025. [Online]. Available: https://CRAN.R-project.org/package=flexmix

[60] J. Wood, “RcppAlgos: High Performance Tools for Combinatorics and Computational Mathematics,” 2024. [Online]. Available: https://CRAN.R-project.org/package=RcppAlgos

[61] B. Hamner and M. Frasco, “Metrics: Evaluation Metrics for Machine Learning,” 2018. [Online]. Available: https://CRAN.R-project.org/package=Metrics

[62] C. Guo and Y. Liang, “Analyzing Restricted Mean Survival Time Using SAS/STAT,” in *SAS global forum proceedings Paper SAS3013-2019, SAS Institute Inc*, 2019.

APÉNDICE A. Graficas de dispersión de 64 modelos seleccionados separados por función de distribución.