Instituto Tecnológico y de Estudios Superiores de Occidente

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

Departamento de Electrónica, Sistemas e Informática

MAESTRÍA EN SISTEMAS COMPUTACIONALES

CLASIFICACIÓN DE ABANDONO DE ALUMNOS EN UNA INGENIERÍA DE UNIVERSIDAD PRIVADA MEXICANA

Trabajo recepcional para obtener el grado de

MAESTRO EN SISTEMAS COMPUTACIONALES

Presentan: Carlos Fortún Manzano

Asesor: Dra. Gisel Hernández Chávez

San Pedro Tlaquepaque, Jalisco. mayo de 2025

AGRADECIMIENTOS

El autor desea dar las gracias al Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) por los recursos proporcionados para el desarrollo de este trabajo de obtención de grado. Adicionalmente, a la Asociación Universitaria Iberoamericana de Postgrado (AUIP) por el apoyo económico recibido a través de la Beca para cursar programas de Postgrado en ITESO - México. También se agradece el apoyo brindado por todos los profesores del claustro de la Maestría en Sistemas Computacionales que de una forma u otra aportaron los conocimientos necesarios para llevar a cabo este trabajo, con mención especial para la Dra. Gisel Hernández Chávez por su excelente desempeño como asesora durante todo el transcurso del posgrado, y a los coordinadores de la maestría, el Dr. Iván Esteban Villalón Turrubiates y el Dr. Luis Julián Domínguez Pérez.

RESUMEN

Actualmente el ITESO 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 emplearon técnicas estadísticas como regresión semi paramétrica de Cox, modelos de Cox extendidos (Li et al., 2017; Tay et al., 2022) y modelos de aprendizaje de máquina como los árboles de supervivencia aleatorios (RSF). También se cuenta con modelos de clasificación de abandono para las carreras de ingeniería usando regresión logística.

En este trabajo se presentan modelos de clasificación de abandono en la Ingeniería en Sistemas Computacionales (ISC) del ITESO para predecir si en el momento del ingreso un alumno abandonará o no la carrera al 2do semestre. Se emplearon 27 modelos de aprendizaje automático (AA) utilizando la librería H2O de Python y la metodología de Minería de Datos CRISP-DM. Se utilizaron variables predictoras asociadas directamente con el alumno o con la preparatoria, tanto demográficas, académicas, socioeconómicas como de comportamiento. Finalmente se comparó la calidad de los modelos en cuanto a área bajo la curva (AUC), sensibilidad, especificidad, puntuación F1, error cuadrático medio y precisión, encontrando que dos modelos de Máquina de Potenciación del Gradiente (GBM) y uno de Conjunto de modelos (Stacked Ensemble) arrojaron mejores resultados que el modelo base de Regresión Logística. De estos tres el mejor resultó el modelo GBM_1_AutoML_1 en el que se obtuvo una sensibilidad del 64.3 %, una especificidad del 85.1 %, AUC del 76 % y una precisión del 86.4 %. Las variables predictoras más importantes para este modelo fueron: Probabilidad histórica promedio de abandono de alumnos de una misma preparatoria en la universidad estudiada (Prepa_Encode_Prob_abandon), Índice de calidad de preparatoria en la prueba PLANEA 2015 (ind_planea), Promedio de preparatoria (PromPre), Categoría de beca (CatBeca), Local o foráneo (Local) y Edad de ingreso (EdadIng).

ABSTRACT

ITESO currently has models for estimating dropout, major change, and graduation at the university level and for student groups (women, men, by major, by age group, etc.), as well as predictive models at the individual level. For the latter, statistical techniques such as semiparametric Cox regression, extended Cox models (Li et al., 2017; Tay et al., 2022), and machine learning models such as random survival trees (RST) were used. Dropout classification models for engineering programs using logistic regression are also available.

This paper presents dropout classification models for the Computer Systems Engineering (CSE) program at ITESO to predict whether a student will drop out of the program in the second semester upon admission. Twenty-seven machine learning (ML) models were used using the H2O Python library and the CRISP-DM data mining methodology. Predictor variables directly associated with the student or the high school were used, including demographic, academic, socioeconomic, and behavioral variables. Finally, the quality of the models was compared in terms of area under the curve (AUC), sensitivity, specificity, F1 score, mean square error, and accuracy. It was found that two Gradient Boosting Machine (GBM) models and one Stacked Ensemble model yielded better results than the base Logistic Regression model. Of these three, the best was the GBM_1_AutoML_1 model, which obtained a sensitivity of 64.3%, a specificity of 85.1%, an AUC of 76%, and an accuracy of 86.4%. The most important predictor variables for this model were: Average historical probability of students dropping out of the same high school at the university studied (Prepa_Encode_Prob_abandon), High school quality index in the PLANEA 2015 test (ind_planea), High school average (PromPre), Scholarship category (CatBeca), Local or foreign (Local) and Entry age (EdadIng).

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1.INTRODUCCIÓN

1.1 Antecedentes

Para una mayor comprensión y entendimiento y con el objetivo de conocer mejor el alcance del trabajo a realizar se hizo un estudio de estado del arte cuyos resultados se presentan en el capítulo 2. Se revisaron trabajos de predicción de abandono clasificatorios en universidades entre los años 2018 y 2024, además de los trabajos del ITESO de González-Jiménez en el 2023 [1], Ramos-Rocha del 2024 [2] y Fuentes-García [3] y Hernández-Chávez [4] en el 2025, donde se utilizaron técnicas clasificatorias y análisis de supervivencia[1]. Con este análisis se identificaron variables comunes para el estudio de predicción de abandono y las técnicas más utilizadas.

1.2 Justificación

La deserción o abandono escolar, se entiende como la salida temprana y definitiva del sistema escolar regular o diurno de niños y jóvenes [5]. Las tasas de deserción actuales superan el 20% en las ingenierías del ITESO, después de 12 semestres. Predecir a tiempo la deserción contribuye con la identificación de factores de riesgo y la detección de alumnos con más peligro de abandonar para acompañarlos y evitar la salida de la universidad. Este problema justifica el desarrollo de modelos lo más precisos posibles.

1.3 Problema

En la universidad bajo estudio se cuenta con modelos de Regresión Logística que predicen si un estudiante de ingeniería en general abandonará o no después de 12 semestres [1], pero este tipo de modelos no son lo suficientemente precisos (la máxima Especificidad alcanzada por uno de estos modelos fue del 72.94 % y la máxima Sensibilidad del 67.53 %) para la correcta identificación de estudiantes en riesgo de abandono. También se cuenta con un estimador de riesgos en competencia de Aalen-Johansen [2] y un modelo para predecir los riesgos en competencia mediante el método de Fine & Gray con regularización [3] para predecir el abandono, el cambio de carrera o egreso a nivel de ingeniería, pero estos métodos no son técnicas avanzadas de AA. Además, se cuenta con un modelo que utiliza el método de Bosques de Supervivencia Aleatorios [4], pero este utiliza datos de todos los estudiantes del ITESO en doce semestres. En resumen, no se cuenta con modelos clasificatorios a nivel de programa de estudio y que utilicen técnicas avanzadas de Aprendizaje

[1] El Análisis de Supervivencia es un conjunto de métodos que permiten analizar datos a través del tiempo, desde el origen de una observación hasta la ocurrencia de un evento de interés o el fin de la observación. Incluye métodos estadísticos y de aprendizaje de máquina.

Automático para la predicción de abandono después del 1er semestre, que es el momento en el que se ha identificado el mayor abandono (no se inscriben en el 2do semestre).

1.4 Objetivos

1.4.1 Objetivo General:

Encontrar los modelos clasificatorios más precisos para determinar, en el momento del ingreso a la universidad, los alumnos de ISC que abandonarán después del primer semestre.

1.4.2 Objetivos Específicos:

1. Seleccionar variables predictoras del modelo.
2. Desarrollar modelos de clasificación de abandono después del primer semestre con datos de ingreso.
3. Evaluar la calidad de los modelos en cuanto a precisión de las predicciones.
4. Comparar los resultados de los modelos de Regresión Logística con los de Aprendizaje Automático.

1.5 Metodología de Minería de Datos

CRISP-DM (*Cross Industry Standard Process for Data Mining*) propone un modelo de proceso integral para llevar a cabo proyectos de minería de datos. La metodología CRISP-DM se describe en términos de un modelo de procesos jerárquico que comprende cuatro niveles de abstracción: fases, tareas genéricas, tareas especializadas e instancias de procesos.[6]

1.6 Aportación

Los modelos predictivos con que cuenta actualmente el ITESO para la predicción de abandono a nivel de programa de estudio no utilizan técnicas avanzadas de Aprendizaje Automático y no son lo suficientemente precisos. El aporte de este trabajo es encontrar los modelos de AA con mejor desempeño para predecir, en el momento del ingreso, el abandono de la carrera de ISC después del 1er semestre (inicio del 2do) y la selección de las variables predictoras más importantes para determinar este evento

2. ESTADO DEL ARTE

El objetivo de este estado del arte fue describir estudios de predicción de abandono en distintas universidades utilizando variadas técnicas de clasificación y análisis de supervivencia que se hayan publicado en los últimos años, identificando las variables predictoras de mayor importancia y las técnicas utilizadas. Estos hallazgos se resumen en la Tabla 2-1.

2.1 Predicción de abandono utilizando redes neuronales de Alban y colegas.

En el artículo de Alban et al. (2019) [7] los autores indican que en países de Latinoamérica existen altos porcientos de abandono de carreras universitarias. Se han realizado diversos estudios ahondando en los factores que causan este suceso, aunque se han dejado fuera factores del contexto académico, social, económico e institucional. El objetivo de ese trabajo fue el de proponer nuevos factores e investigar si estos afectan el abandono de las carreras, y si es posible predecir el abandono al analizar estos nuevos factores.

Se seleccionaron 11 factores los cuales se consideraron que podían ayudar a mitigar las tasas de deserción: Conocimiento limitado sobre el uso de software especializado en la carrera universitaria, Embarazo planeado/no planeado, Compromiso del profesor con los estudiantes, Compromiso financiero del primogénito con su familia, Acoso, Sexismo, Adicciones adquiridas por los estudiantes, Número de hijos de los estudiantes, Adaptación del estudiante al aprendizaje universitario, Ranking de la universidad o institución y Perspectiva del estudiante de su integración en el mercado laboral. Estos se clasificaron en cinco dimensiones: personal, académica, social, institucional y económica. Seis de estos factores se determinaron basándose en el análisis de teorías educativas, y las otras cinco relacionadas con el razonamiento lógico.

El conjunto de datos utilizado se construyó a partir de una encuesta realizada a 2670 estudiantes de las carreras de Ciencias Administrativas y Humanas, de la Universidad Pública de Ecuador.

Se construyeron dos modelos: uno utilizando la técnica de Función de Base Radial, y el otro mediante un Perceptrón Multicapas. Estos fueron evaluados mediante métricas de precisión como especificidad y sensibilidad. Al seleccionar los factores se le asignó un peso a cada uno en dependencia de la importancia de cada factor, y estas variables se utilizaron como predictores en la capa de entrada de los modelos de redes neuronales.

1. Para el Perceptrón Multicapas el conjunto de datos se particionó en un 60 % para entrenamiento, un 30 % para pruebas y un 10 % de validación. Durante la etapa de

entrenamiento de la red neuronal se utilizó la función sigmoide logarítmico y la función de entropía cruzada como función de error. El resultado obtenido se corresponde con el uso de los 11 factores como variables independientes y la deserción como variable dependiente. Se utilizó la función de tangente hiperbólica como función de activación de la capa oculta y la función de activación de la capa de salida fue *softmax*, cuyo modelo base es una función con valores vectoriales, y su componente individual consiste en un exponencial evaluado en un elemento de un vector, el cual está normalizado por la sumatoria del exponencial de todos los elementos de ese vector.

2. Para la Función de Base Radial se utilizó un 70 % del conjunto de datos para entrenamiento y un 30 % para pruebas. Se utilizaron las mismas variables dependiente e independientes. La función de activación de la capa oculta fue *softmax*. Para la capa de salida se utilizó la función de activación Identidad y como función de error la función de suma de cuadrados.

Se utilizaron las curvas *Receiver Operating Characteristics* (ROC) para medir la capacidad predictiva de los modelos. Se graficaron la sensibilidad y la especificidad de los dos modelos, arrojando como resultado que ambos modelos tuvieron un rendimiento óptimo en términos de predicción, sin diferencias significativas en la precisión de ambos modelos, por lo que ambos pueden ser considerados como opciones para la predicción de la deserción en las universidades.

2.2 Predicción de riesgo de deserción utilizando redes neuronales probabilísticas de Mason y colegas.

El estudio de Mason et al. (2018) [8] se enfocó en desarrollar un modelo para predecir el riesgo de abandono de estudiantes del Colegio de Ingeniería de Estados Unidos utilizando diferentes técnicas de aprendizaje automático. El objetivo principal fue identificar los factores que influyen en la decisión de los estudiantes de abandonar sus estudios y comparar la efectividad de diferentes algoritmos en la predicción de este riesgo.

Para lograr esto, los investigadores recolectaron datos de 682 estudiantes de primer año de ingeniería y separaron el conjunto de datos en un 70 % para entrenamiento y un 30 % para pruebas. Utilizaron tres técnicas diferentes: una Red Neuronal Probabilística (PNN), una Red Neuronal de Retro propagación (BPNN) y Regresión Logística. Las variables recopiladas incluyeron datos demográficos, académicos y socioeconómicos de los estudiantes.

Primero, se realizó un análisis exploratorio de los datos para identificar las características más relevantes, quedando estas reducidas a 14 variables predictoras: Promedio de calificaciones (GPA) acumuladas al final del primer semestre, GPA de Matemática al final del primer semestre, Horas acumuladas percibidas al final del primer semestre, Puntuación final calculada, Nota de matemática acreditada en el examen *American College Testing* (ACT), Horas inscritas por estudiantes de primer año al final del primer semestre, Créditos

totales del estudiante al final del primer semestre, Curso de matemática más alto en primer año, Edad al momento del día 20, Costo neto estudiantil en el primer semestre, Monto de ayuda financiera concedida en el semestre, Oferta de ayuda financiera aceptada, Niveles de ingreso y División de la especialidad universitaria. Se utilizaron como medidores de desempeño de las pruebas la precisión, la sensibilidad y la especificidad. Los investigadores compararon la precisión de los tres modelos en términos de su capacidad para predecir el riesgo de abandono de los estudiantes.

Los resultados mostraron que la Red Neuronal Probabilística (PNN) tuvo la mayor precisión en la predicción del riesgo de abandono (sensibilidad) de estudiantes de ingeniería (76.9 %) en comparación con la Red Neuronal de Retro Propagación (71.2 %) y la Regresión Logística (65.4 %). Esto sugiere que la PNN es una técnica efectiva para predecir el riesgo de abandono y puede ayudar a las instituciones educativas a identificar a los estudiantes que podrían estar en mayor riesgo de abandonar sus estudios de ingeniería.

La diferencia del modelo PNN con los otros reside principalmente en que el formato de salida puede ser normalizado, lo que proporciona resultados en un formato probabilístico: la probabilidad de retener al estudiante y la probabilidad de que abandone. Ofrecer una respuesta de salida en este formato se aproxima a la solución óptima de Bayes, lo que pudiera traducirse como el factor de riesgo de abandono. Este tipo de modelos puede utilizarse para seleccionar los estudiantes a los que intervenir para evitar su abandono. Se crearon diagramas de dispersión para analizar los datos de salida de cada uno de los 3 modelos, con la esperanza de ver tres segmentos horizontales: el segmento superior tenía pocos o ningún estudiante no retenido, el inferior tenía pocos o ningún estudiante retenido y luego una banda intermedia mezclada con retenidos y no retenidos. Los límites superior e inferior de la banda media representarían los umbrales superior e inferior en los que establecer los niveles "en riesgo". Sin embargo, en estos diagramas no se define una sección de riesgo clara, por lo que debe considerarse el número de estudiantes que se representen en cada rango de probabilidad. Para ello se evaluaron los rangos 20–80, 25–75, y 30–70% para cada modelo, arrojando cada uno un número distinto de estudiantes dentro del rango “en riesgo”. Por tanto, se deben tomar consideraciones para seleccionar el rango a ser usado, ya que se requieren tiempo, esfuerzos y recursos para tomar acciones de intervención hacia los estudiantes considerados en el rango de riesgo, por lo que las instituciones u organizaciones responsables de estos costos deben decidir un rango que resulte apropiado.

En resumen, el estudio demuestra que el uso de técnicas de aprendizaje automático, como la Red Neuronal Probabilística (PNN), puede ayudar a predecir de manera precisa el riesgo de abandono de los estudiantes de ingeniería. Estos resultados pueden ser utilizados por las instituciones educativas para implementar intervenciones tempranas y estrategias de retención que eviten el abandono de los estudiantes y fomenten su éxito académico.

2.3 Predicción de deserción universitaria utilizando redes neuronales convolucionales de Mezzini y colegas.

En el estudio de Mezzini et al. (2019) [9] los autores indican que, en la Unión Europea en el 2017, en promedio un 10.6 % de estudiantes de Educación Superior abandonan la carrera. Desarrollaron una red neuronal mediante la creación de Redes Neuronales Convolucionales (RNC), las cuales se centran en apilar decenas o cientos de capas neuronales convolucionales, con la que se obtiene una estructura de red profunda, la cual produce modelos de alta eficiencia.

Se probaron 3 arquitecturas de RNC para probar la efectividad del modelo predictivo: ResNetV2 (RNV2), InceptionResNetV4 (IRNV4) y DFSV1. El estudio fue desarrollado con un conjunto de datos de estudiantes obtenido de la oficina administrativa de la Universidad Roma Tre. El conjunto de datos fue transformado para obtener una tabla Estudiante, formados por los distintos atributos, los cuales fueron transformados usando una función arbitraria biyectiva, para tomar un dominio entero no negativo. La tabla fue separada en 9 tablas, cada una conteniendo los datos de cada estudiante para cada uno de los 9 años que le puede tomar a un estudiante graduarse, aunque para probar los modelos solo se utilizaron las tablas hasta el 3er año, ya que después de ese año el porciento de estudiantes que abandona es insignificante. Los datos de los estudiantes fueron separados en 3 grupos mutuamente disjuntos: un conjunto de entrenamiento formado por 4532 alumnos, uno de validación con 450, y uno de prueba formado por 447 alumnos. Los modelos con las 3 arquitecturas fueron entrenados con 100 épocas, lo que quiere decir que el conjunto de datos de entrenamiento le fue suministrado entero 100 veces a la RNC.

Para cada época se obtuvo la matriz de confusión de los sets de validación y prueba. Con cada una de estas matrices se calculó la precisión, evidenciándose un aumento drástico de la precisión en los años 2 y 3 evaluados, esto debido a que en los primeros años más estudiantes abandonan la carrera, y que en años siguientes se cuenta con más información de cada estudiante, concluyendo que mientras más datos se tengan y estos sean más exactos, las predicciones del modelo serán más exactas y efectivas.

2.4 Factores que influyen en las tasas de deserción en la educación superior de Kocsis & Molnár.

En el estudio de Kocsis & Molnár (2024) [10] se explora sistemáticamente las influencias multifacéticas sobre el éxito y el abandono de los estudiantes en el contexto de la educación superior contemporánea. El contexto de la investigación revela tasas crecientes de deserción escolar, con un promedio de alrededor del 30% a nivel mundial, un tema de profunda preocupación para los responsables de las políticas educativas, las instituciones y los estudiantes por igual. Los autores sostienen que las crecientes demandas de matrícula, junto con diversos paisajes sociopolíticos, complican el panorama de la educación superior, por lo que es crucial identificar y comprender las variables que contribuyen a la retención y el rendimiento de los estudiantes. Aprovechando los marcos históricos, incluidos los modelos de deserción escolar de Tinto y Bean, el artículo sintetiza ideas contemporáneas que incorporan dinámicas psicológicas, socioeconómicas y académicas en una comprensión coherente del éxito y el desgaste académico.

A partir de una revisión exhaustiva de 95 estudios revisados por pares publicados después de 2012, se encontró que los autores emplean metodologías de investigación sólidas, predominantemente algoritmos de minería de datos educativos (EDM) y modelado de ecuaciones estructurales (SEM). En el caso de los EDM estos se utilizaron en el 40 por ciento del conjunto de estudios empíricos seleccionados y los Algoritmos de Redes Neuronales Artificiales fueron los más frecuentemente utilizados en este tipo de estudio, aunque contradictoriamente se encontró que en general los algoritmos de Árboles de Decisión fueron los más comúnmente utilizados. Sus hallazgos arrojan luz sobre variables predictivas críticas, en particular el promedio de calificaciones (GPA), los créditos del Sistema Europeo de Transferencia y Acumulación de Créditos (ECTS) y el género, que emergen consistentemente como predictores influyentes del rendimiento académico. En particular, el análisis indica que las métricas de GPA y ECTS están significativamente mediadas por factores intrínsecos de los estudiantes, como la motivación y la autoeficacia, así como por factores externos de rendimiento, incluidas consideraciones financieras y compromiso académico.

En conclusión, este estudio no sólo delinea los factores fundamentales que afectan el rendimiento académico, sino que también enfatiza la necesidad de que las instituciones adapten sus enfoques en respuesta al entorno educativo en evolución. Al articular la compleja interacción entre varios predictores y emplear marcos analíticos sólidos, los autores aportan conocimientos valiosos al diálogo en torno al éxito académico y las estrategias de mitigación de la deserción escolar. Esta investigación sirve como un recurso esencial para continuar los esfuerzos por mejorar la retención de estudiantes y fomentar mejores resultados educativos, respondiendo así a las demandas apremiantes del mercado laboral actual.

2.5 Revisión sistemática de Cardona y colegas.

En la revisión sistemática de Cardona et al. (2023) [11] se examina críticamente la literatura sobre la predicción de la retención de estudiantes en la educación superior a través de algoritmos de aprendizaje automático, centrándose específicamente en factores como el riesgo de deserción, el riesgo de abandono y el riesgo de finalización. La revisión emplea una metodología sólida que abarca un protocolo detallado para la selección de estudios y clasificación de artículos. Al abordar preguntas fundamentales de investigación sobre las técnicas empleadas para predecir las tasas de retención, el desempeño de estos métodos en diversos contextos, los factores que influyen y los desafíos que se enfrentan en este ámbito, el estudio contribuye significativamente al diálogo sobre la mejora de las tasas de graduación.

Los hallazgos de esta revisión revelan que las técnicas de aprendizaje automático pueden aprovechar los datos de manera efectiva para mejorar las estrategias de retención de estudiantes. Las más frecuentemente utilizadas fueron Redes Neuronales mostrando rangos de precisión en la clasificación entre 71.59% y 94%, un rango de 65.38%–81.36% para Árboles de Decisión, 50.18%–83.07% para Regresión Logística y un rango entre 57.69% y 86.4% de precisión para Máquinas de Vectores de Soporte. Además, estos estudios proporcionaron resultados más consistentes de Conjuntos Apilados de algoritmos, con un estrecho rango de

precisión en la clasificación entre 79.36% y 81.67%, lo que indica que estos conjuntos pudieran ser métodos más eficientes para predecir el abandono de estudiantes.

Estas técnicas no sólo proporcionan una comprensión matizada de los estudiantes en riesgo, sino que también iluminan los factores específicos que influyen en los resultados de retención. Aunque se encontraron disímiles factores en los diferentes trabajos estudiados, se identificó GPA como variable de gran importancia en 63% de los estudios, los factores medidos antes de iniciar la universidad (por ejemplo el GPA de escuela secundaria) fueron marcados como importantes en el 47% de los estudios, y la ayuda financiera fue encontrada como de importancia en el 37% de los estudios Es importante destacar que el estudio subraya la necesidad de utilizar metodologías integrales para perfeccionar las capacidades de predicción. Esto es particularmente relevante a la luz de los desafíos actuales que presentan las fluctuaciones económicas, que exigen una fuerza laboral equipada con títulos avanzados y habilidades técnicas, una necesidad que se satisface cada vez más a través de prácticas efectivas de educación superior.

Además, el estudio enfatiza el papel fundamental de los datos en los procesos de toma de decisiones dentro de las instituciones educativas. A medida que evoluciona el panorama de la educación superior, particularmente a raíz de la pandemia de COVID-19 y los impactos económicos resultantes, resulta vital que las instituciones aprovechen todo el potencial de los datos disponibles. Se subraya así la necesidad de mejorar los mecanismos de análisis de datos, lo que en última instancia sugiere que un enfoque colaborativo entre las partes interesadas en la educación superior es esencial para fomentar el éxito de los estudiantes y abordar los problemas multifacéticos que rodean las tasas de retención y finalización. En consecuencia, esta revisión no sólo presenta una base para futuras investigaciones, sino que también aboga por la implementación estratégica de la toma de decisiones basada en datos para mejorar los resultados educativos.

2.6 Estudio sobre la deserción estudiantil universitaria en México de Kuz y colegas.

El artículo de Kuz et al. (2023) [12] aborda un fenómeno crítico en el ámbito educativo contemporáneo: la deserción estudiantil en universidades privadas de México. A través del uso de técnicas de aprendizaje automático, en particular el algoritmo XGBoost, la investigación busca predecir la deserción escolar en el primer año de estudios universitarios. Los resultados preliminares indican que factores como el promedio estudiantil del primer período, el porcentaje de beca y la región del campus son determinantes significativos en las tasas de deserción.

La metodología del estudio, que incluye la selección cuidadosa de un conjunto de datos de estudiantes y la implementación de diversos modelos predictivos, demuestra un enfoque riguroso y sistemático. Al analizar un amplio conjunto de registros, el estudio no solo identifica patrones de comportamiento de los estudiantes, sino que también proporciona una base sólida para la formulación de políticas educativas más efectivas. De esta manera, las instituciones educativas pueden implementar estrategias de retención más informadas, basadas en datos concretos.

Un hallazgo clave de esta investigación es la alta precisión alcanzada en la predicción de deserción gestada por el algoritmo XGBoost, que supera el 99%. Esto posiciona a la ciencia de datos como una herramienta invaluable en la educación superior, contribuyendo a mitigar las preocupaciones sobre la deserción y promoviendo la sostenibilidad de las instituciones educativas. Este estudio no solo llena un vacío en la literatura existente, sino que también establece un modelo replicable que puede ser aplicado en otros contextos educativos, propiciando un impacto significativo en la retención de estudiantes y, en consecuencia, en la movilidad social.

2.7 Estudios anteriores del ITESO.

En el trabajo de González-Jiménez [1]se crearon dos modelos de Regresión Logística para predecir el abandono de estudiantes de Ingeniería del ITESO. Se emplearon los datos de todos los estudiantes que ingresaron en el semestre de otoño entre 2004 y 2011. Los datos se separaron en dos grupos: alumnos provenientes de preparatoria con convenio y alumnos de preparatorias sin convenio. Para el modelo de alumnos con convenio se seleccionaron las variables independientes edad, tipo de preparatoria de acuerdo con el volumen de alumnos que ingresan al ITESO y si la ingeniería inscrita es una de las siguientes: Electrónica, Alimentos, Financiera o Mecánica, y se alcanzó una sensibilidad del 65.85 %. El modelo para alumnos sin convenio utilizó las variables independientes tipo de preparatoria, existencia de algún tipo de apoyo económico, ingeniería inscrita, separada en tres grupos (Grupo 1: Industrial; Grupo 2: Civil, Sistemas, Ambiental, Química, Mecánica, Alimentos, Financiera y Redes; Grupo 3: Electrónica), puntaje del examen de admisión, promedio de la preparatoria y edad al ingresar, y se obtuvo una sensibilidad del 66.29 %.

Los trabajos de Ramos-Rocha [2] y Fuentes-García [3] utilizaron el mismo conjunto de datos empleado en el presente trabajo, 396 estudiantes de la Ingeniería en Sistemas Computacionales registrados entre primavera de 2006 y primavera de 2014.

En el primero de estos dos trabajos se utilizó el estimador de riesgos en competencia de Aalen-Johansen para estimar los eventos de deserción, cambio de carrera y egreso e identificar los grupos de alumnos con mayor o menor probabilidad de ocurrencia de esos eventos. Para el caso de mayor probabilidad de abandono o cambio de carrera se identificaron los grupos de alumnos sin financiamiento al ingresar, con un promedio menor a 80, mayores de 21 años al ingresar o con un puntaje menor a 1,224 en la prueba admisión. Los grupos identificados con mayor probabilidad de egresar fueron los alumnos con financiamiento al ingresar, provenientes de escuelas con convenio, con un promedio de preparatoria superior a 90 y aquellos que ingresan a la universidad antes de los 19 años. Las variables analizadas fueron sexo, financiamiento, tipo de preparatoria, promedio de preparatoria, desempeño en el examen PAA, pase directo y edad de ingreso.

En el segundo trabajo se predijeron los riesgos en competencia mediante el método de Fine & Gray con regularización. Se aplicaron las regularizaciones LASSO, Ridge y *Elastic Net* a cada uno de los eventos cambio de plan de estudio, baja académica y egreso. Se obtuvo que las variables más relevantes fueron edad de ingreso, promedio de preparatoria y probabilidad de abandono de preparatoria de procedencia, seleccionadas

mediante la regularización LASSO. Se crearon dos modelos para cada evento: uno con todas las variables y otro con las variables seleccionados por LASSO, concluyéndose que los modelos basados en LASSO reducían la complejidad de la solución sin comprometer la precisión de las predicciones.

El trabajo de Hernández-Chávez [4] utilizó el método de Bosques de Supervivencia Aleatorios (RSF por sus siglas en inglés). Se emplearon los datos de todos los estudiantes del ITESO y se crearon modelos para cada uno de los doce semestres analizados para predecir abandono de las carreras. Las variables independientes seleccionadas fueron probabilidad de abandono de preparatoria de procedencia, promedio de prepa, edad de ingreso y sexo y se obtuvo un valor de AUC para el segundo semestre del 71.6 %.

2.8 Resumen del estado del arte.

Tabla 2-1. Resumen del Estado del Arte

Trabajo	Variables predictoras de importancia	Algoritmo(s) más utilizado(s)
Neural Networks to Predict Dropout at the Universities	Conocimiento limitado sobre el uso de software especializado en la carrera universitaria, Embarazo planeado/no planeado, Compromiso del profesor con los estudiantes, Compromiso financiero del primogénito con su familia, Acoso, Sexismo, Adicciones adquiridas por los estudiantes, Número de hijos de los estudiantes, Adaptación del estudiante al aprendizaje universitario, Ranking de la universidad o institución y Perspectiva del estudiante de su integración en el mercado laboral	RBF, MLP
Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression	Promedio de calificaciones (GPA) acumuladas al final del primer semestre, GPA de Matemática al final del primer semestre, Horas acumuladas percibidas al final del primer semestre, Puntuación final calculada, Nota de matemática acreditada en el examen *American College Testing* (ACT), Horas inscritas por estudiantes de primer año al final del primer semestre, Créditos totales del estudiante al final del primer semestre, Curso de matemática más alto en primer año, Edad al momento del día 20, Costo neto estudiantil en el primer semestre, Monto de ayuda financiera concedida en el semestre, Oferta de ayuda financiera aceptada, Niveles de ingreso y División de la especialidad universitaria	PNN, BPNN, LR
Predicting university dropout by using Convolutional Neural Networks	Año de inicio de estudios, Año de nacimiento, Género, Año de nacimiento, País de nacimiento, Tipo de escuela secundaria, Puntaje de salida de la escuela secundaria, Puntaje máximo de salida de la escuela secundaria, Año de finalización de la escuela secundaria, Transferido de otra universidad, CFU (créditos) de otra universidad, Facultad, Año académico, Código del curso, Nombre del curso, Año del curso, Clase de ingreso familiar, Estatus laboral, Exención de impuestos, Tipo de exención de impuestos, Discapacidad, Estatus a tiempo parcial, CFU (créditos) de tiempo parcial, Tipo de renovación de matrícula	RNV2, IRNV4, DFSV1

Trabajo	Variables predictoras de importancia	Algoritmo(s) más utilizado(s)
Factors influencing academic performance and dropout rates in higher education	GPA, ECTS, Género	ANN, DT
Data Mining and Machine Learning Retention Models in Higher Education	GPA, Factores medidos antes de iniciar la universidad, Ayuda financiera	NN, DT, LR, SVM, Stacked Ensemble (SVM + DT + NN)
Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México	Promedio del primer período, Porcentaje de beca, Región del campus	XGBoost
Estudios anteriores del ITESO	Probabilidad de abandono de preparatoria de procedencia, Promedio de Prepa, Edad de Ingreso, Sexo, Desempeño en el examen PAA, Pase directo, Índice de calidad de preparatoria en la prueba PLANEA 2015, Financiamiento, Tipo de Preparatoria, Ingeniería inscrita, Local o Foráneo, Primavera u Otoño	LR, Fine & Gray, Aalen-Johansen, RSF

3. MARCO TEÓRICO

En el análisis de las revisiones de literatura como las de Kocsis & Molnár [10] y Cardona et al. [11] se identificaron variables importantes para la elaboración de modelos de AA para la predicción de abandono tales como Género y Promedio de preparatoria. Además, los modelos más utilizados y que arrojaron mejores resultados fueron algunos como las Redes Neuronales Artificiales, Árboles de Decisión, Perceptrón Multicapas, así como conjuntos formados por la combinación de varios tipos de modelos. En este capítulo se presentan las bases teóricas y conceptuales sobre los distintos tipos de modelos de aprendizaje automático que utiliza la librería H2O.

3.1 Modelos de Clasificación

La clasificación es un problema de asignar automáticamente una etiqueta a un ejemplo sin etiquetar. La detección de spam es un ejemplo famoso de clasificación.

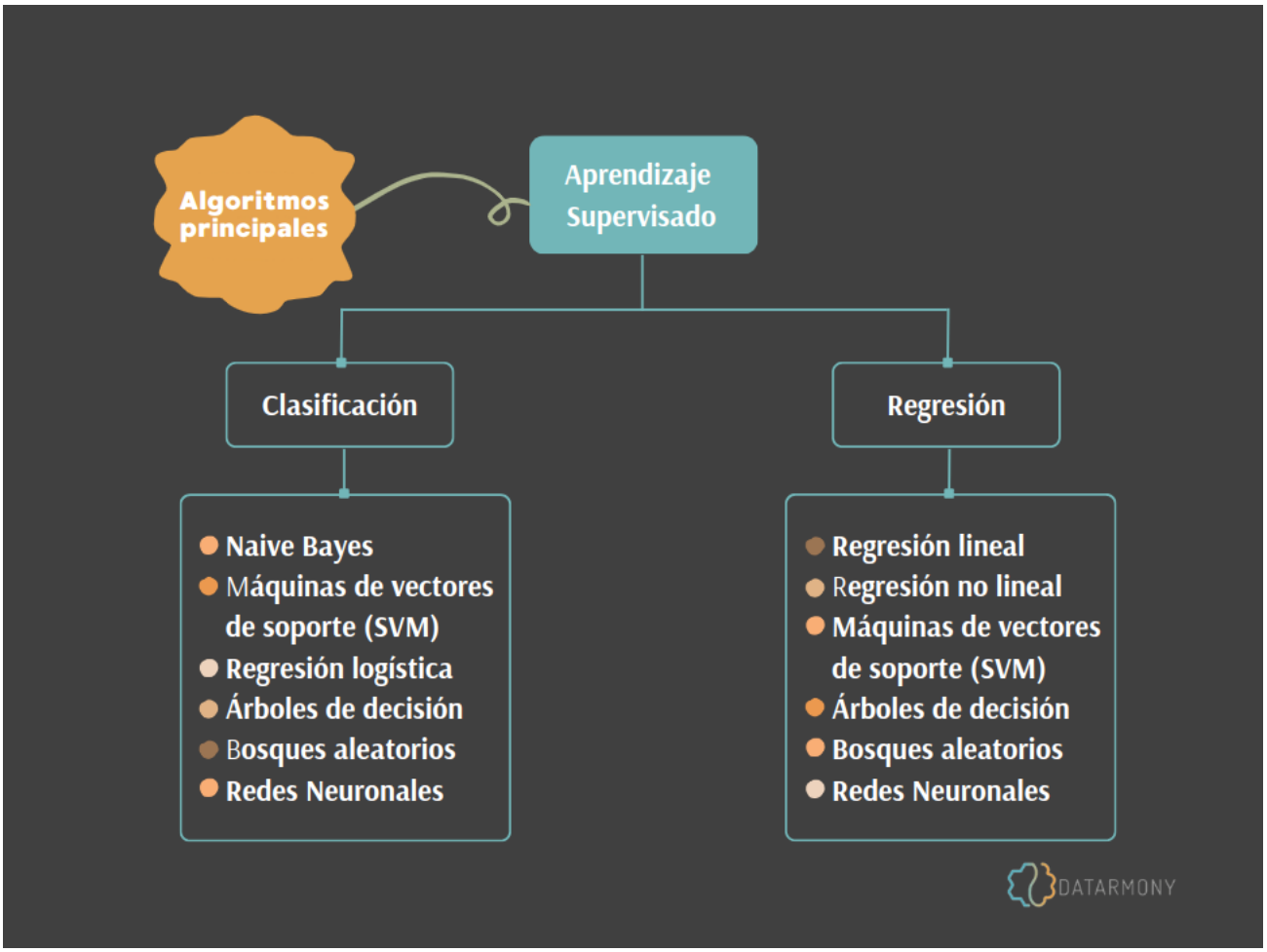
En el aprendizaje automático, el problema de clasificación se resuelve mediante un algoritmo de aprendizaje de clasificación que toma una colección de ejemplos etiquetados como entradas y produce un modelo que puede tomar un ejemplo sin etiquetar como entrada y generar directamente una etiqueta o generar un número que puede ser utilizado por el analista de datos para deducir la etiqueta fácilmente. Un ejemplo de tal número es una probabilidad.

En un problema de clasificación, una etiqueta es miembro de un conjunto finito de clases. Si el tamaño del conjunto de clases es dos ("enfermo" / "saludable", "spam" / "no spam"), hablamos de clasificación binaria (también llamada binomial en algunos libros).

La clasificación multiclase (también llamada multinomial) es un problema de clasificación con tres o más clases.

Mientras que algunos algoritmos de aprendizaje permiten naturalmente más de dos clases, otros son por naturaleza algoritmos de clasificación binaria.[13]

Figura 3-1. Taxonomía de algoritmos de aprendizaje automático supervisado



3.2 Modelo Lineal Generalizado y Modelo Logístico

El modelo lineal generalizado (GLM) es una generalización de la regresión lineal para el modelado de varias formas de dependencia entre el vector de características de entrada y la salida. El Modelo Logístico (o Modelo *Logit*), por ejemplo, es una forma de GLM. [13]

Los GLM estiman modelos de regresión para resultados que siguen distribuciones exponenciales. Además de la distribución gaussiana (es decir, normal), estas incluyen las distribuciones de Poisson, binomial y gamma. Cada uno tiene un propósito diferente y, dependiendo de la distribución y la función de enlace elegida, se puede utilizar para predicción o clasificación. [14]

El modelo logístico (de ahora en adelante regresión logística) según la define Sperandei (2014) [15] funciona de manera muy similar al modelo lineal, pero con una variable de respuesta binomial.

Una regresión logística modelará la probabilidad de que ocurra un resultado basado en características individuales. Dado que el azar es una proporción, lo que será realmente modelado es el logaritmo de la probabilidad dado por:

Ecuación 1. Regresión logística

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_m X_m$$

donde π indica la probabilidad de un evento y β_i son los coeficientes de regresión asociados a las variables explicativas x_i.

3.3 Bosque Aleatorio Distribuido

Según explica Andriy Burkov (2019) [13] hay dos paradigmas de aprendizaje en conjunto (*ensemble learning*): *bagging* y *boosting*. El *bagging* consiste en crear muchas "copias" de los datos de entrenamiento (cada copia es ligeramente diferente de otra) y luego aplicar el modelo débil a cada copia para obtener múltiples modelos débiles y luego combinarlos. El paradigma de *bagging* está detrás del algoritmo de aprendizaje bosque aleatorio distribuido (DRF).

El algoritmo de *bagging* "vainilla" funciona de la siguiente manera. Dado un conjunto de entrenamiento, creamos B muestras aleatorias S_b (para cada b = 1, . . ., B) del conjunto de entrenamiento y construimos un modelo de árbol de decisión f_b utilizando cada muestra S_b como conjunto de entrenamiento. Para la muestra S_b para alguna b, hacemos el muestreo con reemplazo. Esto significa que comenzamos con un conjunto vacío y luego elegimos al azar un ejemplo del conjunto de entrenamiento y se pone su copia exacta en S_b manteniendo el ejemplo original en el conjunto de entrenamiento original. Seguimos escogiendo ejemplos al azar hasta que $|S_b|$ = N. Después del entrenamiento, tenemos B árboles de decisión. La predicción para un nuevo ejemplo x se obtiene como el promedio de las predicciones B:

Ecuación 2. Función de promedio de predicciones de un modelo DRF

$$y \leftarrow f^{\wedge}(x) \stackrel{\text{def}}{=} \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

en el caso de regresión, o por mayoría de votos en el caso de clasificación.

El algoritmo de bosque aleatorio es diferente del *bagging* vainilla solo en un sentido. Se utiliza un algoritmo de aprendizaje de árbol modificado que inspecciona, en cada división del proceso de aprendizaje, un subconjunto aleatorio de las variables independientes. La razón para hacer esto es evitar la correlación de los árboles: si una o varias variables independientes son predictores muy sólidos para la variable dependiente, estas variables se seleccionan para dividir ejemplos en muchos árboles. Esto daría como resultado muchos árboles correlacionados en nuestro “bosque”. Los predictores correlacionados no pueden ayudar a mejorar la precisión de la predicción. La razón principal detrás de un mejor desempeño del ensamblaje de modelos es que los modelos que son buenos probablemente estarán de acuerdo en la misma predicción, mientras que los modelos malos probablemente no estén de acuerdo en diferentes predicciones. La correlación hará que los malos modelos tengan más probabilidades de estar de acuerdo, lo que obstaculizará la mayoría de voto o el promedio.

Los hiperparámetros más importantes a ajustar son el número de árboles, B, y el tamaño del Subconjunto aleatorio de las características a considerar en cada división.

El bosque aleatorio es uno de los algoritmos de aprendizaje por conjuntos más utilizados. ¿Por qué es tan efectivo? La razón es que, al utilizar múltiples muestras del conjunto de datos original, reducimos

la varianza del modelo final. Recuerde que una varianza baja significa un sobreajuste bajo. El sobreajuste ocurre cuando nuestro modelo intenta explicar pequeñas variaciones en el conjunto de datos porque nuestro conjunto de datos es sólo una pequeña muestra de la población de todos los ejemplos posibles del fenómeno que se intenta modelar. Si no tuviéramos suerte con la forma en que se muestreó nuestro conjunto de entrenamiento, entonces podría contener algunos artefactos indeseables (pero inevitables): ruido, valores atípicos y representación excesiva o insuficientes ejemplos. Al crear múltiples muestras aleatorias con reemplazo de nuestro conjunto de entrenamiento, podemos reducir el efecto de estos artefactos.

3.4 Árboles extremadamente aleatorios

Los árboles extremadamente aleatorios (XRT) son una variante del algoritmo bosques aleatorios. La base del algoritmo es la misma, pero presenta diferencias:

- En los bosques aleatorios, se utiliza un subconjunto aleatorio de variables independientes candidatas para determinar los umbrales más discriminativos que se eligen como regla de división.
- En los árboles extremadamente aleatorios, la aleatoriedad va un paso más allá en la forma en que se computan las divisiones. Al igual que en los bosques aleatorios, se utiliza un subconjunto aleatorio de variables independientes candidatas, pero en lugar de buscar los umbrales más discriminativos, los umbrales se extraen al azar para cada variable independiente candidata y el mejor de estos umbrales generados aleatoriamente se elige como regla de división. Esto usualmente permite reducir un poco más la varianza del modelo, a expensas de un aumento ligeramente mayor del *bias*. [14]
- La otra diferencia es que utiliza toda la muestra de aprendizaje (en lugar de una réplica de arranque) para hacer crecer los árboles.

Este algoritmo tiene dos parámetros: K, el número de atributos seleccionados aleatoriamente en cada nodo y n_{min}, el tamaño de muestra mínimo para dividir un nodo. Se utiliza varias veces con la muestra de aprendizaje original (completa) para generar un conjunto de modelos (denotamos por M el número de árboles de este conjunto). Las predicciones de los árboles se agregan para producir la predicción final, mediante voto mayoritario en problemas de clasificación y media aritmética en problemas de regresión.

El uso de la muestra de aprendizaje original completa en lugar de una réplica de arranque está motivado en función de minimizar el *bias*. [16]

3.5 Máquina de aumento de gradiente

Otro algoritmo eficaz de aprendizaje conjunto es la máquina de aumento de gradiente (GBM). En el caso de la regresión, para construir un fuerte regresor comenzamos con un modelo constante f = f_0:

Ecuación 3. Función de modelo de regresión

$$f = f_0(x) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{1=1}^{N} y_i$$

Luego se modifican las etiquetas de cada ejemplo i = 1, . . ., N en el conjunto de entrenamiento de la siguiente manera:

Ecuación 4. Función de Y residual

$$y_i^{\wedge} \leftarrow y_i - f(x_i)$$

donde $y_i^{\wedge}$, conocido como el residual, es la nueva etiqueta para el ejemplo x_i.

Luego se utiliza el conjunto de entrenamiento modificado, con residuos en lugar de etiquetas originales, para construir un nuevo modelo de árbol de decisión f_i. El modelo de *boosting* ahora se define como $f \stackrel{\text{def}}{=} f_0 + \alpha f_1$, donde α es la tasa de aprendizaje (un hiperparámetro).

Se vuelven a calcular los residuos y se reemplazan las etiquetas en los datos de entrenamiento una vez más, se entrena el nuevo modelo de árbol de decisión f_2, se redefine el modelo de *boosting* como $f \stackrel{\text{def}}{=} f_0 + \alpha f_1 + \alpha f_2$ y el proceso continúa hasta que se alcanza el máximo número M de árboles (otro hiperparámetro) que se combinan.

Al calcular los residuos, se descubre qué tan bien (o mal) la variable dependiente de cada ejemplo de entrenamiento es predicho por el modelo actual f. Entonces se entrena otro árbol para corregir los errores del modelo actual (es por eso que se usan residuos en lugar de las etiquetas reales) y se agrega este nuevo árbol al modelo existente con algo de peso α. Por lo tanto, cada árbol adicional agregado al modelo corrige parcialmente los errores cometidos por los árboles anteriores hasta que se combina el número máximo de árboles.

En el caso de aumento de gradiente, a diferencia de gradiente descendiente en donde se obtiene el gradiente directamente, usamos su representación en forma de residuos: muestra cómo se debe ajustar el modelo para que el error (el residual) se reduzca.

Los tres hiperparámetros principales para sintonizar el modelo de aumento de gradiente son la cantidad de árboles, la tasa de aprendizaje y la profundidad de los árboles: los tres afectan la precisión del modelo. La profundidad de los árboles también afecta la velocidad del entrenamiento y la predicción: cuanto más cortos, más rápido.

El algoritmo de aumento de gradiente para la clasificación es similar, pero los pasos son ligeramente diferentes. Asumiendo que se tienen M árboles de decisión de regresión, de manera similar a la regresión logística, la predicción del conjunto de árboles de decisión se modela utilizando la función sigmoide:

Ecuación 5. Función sigmoide

$$\Pr(y = 1|x, f) \stackrel{\text{def}}{=} \frac{1}{1+e^{-f(x)}},$$

donde $f(x) = \sum_{m=1}^{M} f_m(x)$ y f_m es un árbol de regresión.

Nuevamente, como en la regresión logística, se aplica el principio de máxima verosimilitud tratando de encontrar una f que maximice $L_f = \sum_{i=1}^{N} \ln(\Pr(y_i = 1 \mid x_i, f))$. Para evitar el desbordamiento numérico, se maximiza la suma de logaritmos de verosimilitudes en lugar del producto de las verosimilitudes.

El algoritmo comienza con el modelo constante inicial $f = f_0 = \frac{p}{1-p}$, donde $p = \frac{1}{N}\sum_{i=1}^{n} y_i$. Entonces en cada iteración m, se agrega un nuevo árbol f_m al modelo. Para encontrar el mejor f_m, primero se calcula la derivada parcial g_i del modelo actual para cada i = 1, . . ., N:

Ecuación 6. Función derivada parcial

$$g_i = \frac{dL_f}{df}$$

donde f es el conjunto de modelos clasificadores construidos en la iteración anterior m - 1. Para calcular gi se necesita encontrar las derivadas de ln (Pr ($y_i = 1|x_i, f$)) con respecto a f para todo i, donde la derivada parcial con respecto a f equivale a $\frac{1}{e^{f(x_i)}+1}$.

Luego se transforma el conjunto de entrenamiento reemplazando la etiqueta original y_i con la correspondiente derivada parcial g_i, y se construye un nuevo árbol f_m usando el conjunto de entrenamiento transformado. Entonces se encuentra el paso de actualización óptimo p_m como:

Ecuación 7. Función de actualización del paso óptimo

$$p_m = \arg_p \max L_{f+pfm.}$$

Al final de la iteración m, se actualiza el conjunto de modelos f agregando el nuevo árbol f_m :

Ecuación 8. Función de actualización del conjunto de modelos

$$f \leftarrow f + \alpha\, p_m\, f_m.$$

Se itera hasta que m = M, luego se devuelve el conjunto de modelos f.

El aumento de gradiente es uno de los algoritmos de aprendizaje automático más potentes. No sólo porque crea modelos muy precisos, pero también porque es capaz de manejar enormes conjuntos de datos con millones de ejemplos y características. Por lo general, supera al bosque aleatorio en precisión, pero, debido a su naturaleza secuencial, puede ser significativamente más lento en el entrenamiento. [13]

3.6 Redes Neuronales Artificiales

"Las Redes Neuronales Artificiales (*Artificial Neural Networks*, ANN) son unas estructuras computacionales que han sido inspiradas en sistemas neuronales biológicos. [...] Su objetivo principal es

transformar un conjunto de datos de entrada en un conjunto de datos de salida en donde entradas y salidas son datos numéricos."

Existen distintos tipos de Redes Neuronales: Perceptrón, Adaline, Perceptrón Multicapa, Máquina de Boltzmann, Máquina de Cauchy, LVQ '*Learning Vector Quantization*' (Red competitiva), Redes de Elman, Redes de Hopfield.

3.7 Aprendizaje Profundo

El aprendizaje profundo (*Deep Learning*) se refiere al entrenamiento de redes neuronales con más de dos capas que no sean de salida. En el pasado, se volvió más difícil entrenar tales redes a medida que crecía el número de capas. Los mayores desafíos se denominaron problemas de gradiente explosivo y gradiente evanescente al utilizarse gradiente descendente para entrenar los parámetros de la red.

Si bien el problema del gradiente explosivo era más fácil de abordar aplicando técnicas simples como el recorte de gradiente y la regularización L1 o L2, el problema del gradiente evanescente permaneció intratable durante décadas.

Para actualizar los valores de los parámetros en las redes neuronales se suele utilizar un algoritmo llamado retropropagación. Retropropagación es un algoritmo eficiente para calcular gradientes en redes neuronales utilizando la regla de la cadena. Esta regla es utilizada para calcular derivadas parciales de una función compleja. Durante gradiente descendente, los parámetros de la red neuronal reciben una actualización proporcional a la derivada parcial de la función de costos respecto del actual parámetro en cada iteración del entrenamiento. El problema es que, en algunos casos, el gradiente será extremadamente pequeño, evitando efectivamente que algunos parámetros cambien su valor. En el peor de los casos, esto puede impedir por completo que la red neuronal siga entrenándose.

Las funciones de activación tradicionales, como la función tangente hiperbólica, tienen gradientes en el rango (0, 1) y la retropropagación calcula los gradientes según la regla de la cadena. Esto tiene el efecto de multiplicar n de estos pequeños números para calcular los gradientes de las capas anteriores (más a la izquierda) en una red de n capas, lo que significa que el gradiente disminuye exponencialmente con n. Esto da como resultado el efecto de que las capas anteriores entrenan muy lentamente, en todo caso.

Sin embargo, las implementaciones modernas de algoritmos de aprendizaje de redes neuronales le permiten entrenar eficazmente redes neuronales muy profundas (hasta cientos de capas). La función de activación ReLU sufre mucho menos el problema del gradiente evanescente. Además, grandes redes de memoria a corto plazo (LSTM), así como técnicas como saltar las conexiones utilizadas en las redes neuronales residuales le permiten entrenar redes neuronales aún más profundas, con miles de capas.

Por lo tanto, hoy en día, dado que los problemas de gradiente evanescente y explosivo en su mayoría están resueltos (o su efecto disminuido) en gran medida, el término "aprendizaje profundo" se refiere al entrenamiento de redes neuronales que utilizan el moderno conjunto de herramientas algorítmicas y matemáticas independientemente de qué tan profunda es la red neuronal. En la práctica, muchos problemas de

negocios se pueden resolver con redes neuronales que tienen 2-3 capas entre las capas de entrada y salida. Las capas que no son ni la entrada ni la salida suelen denominarse capas ocultas.[13]

3.8 Conjuntos apilados

En la práctica, a veces se puede obtener una ganancia de rendimiento adicional combinando modelos potentes creados con diferentes algoritmos de aprendizaje. Una de las formas típicas de combinar modelos es la técnica de conjuntos apilados (*Stacked Ensemble*).

El promedio funciona para la regresión, así como para aquellos modelos de clasificación que devuelven puntuaciones de clasificación. Simplemente se aplican todos los modelos, llamémoslos modelos base, a la entrada x y luego se promedian las predicciones. Para ver si el modelo promediado funciona mejor que cada algoritmo individual, se prueba en el conjunto de validación utilizando la métrica que se prefiera.

El voto mayoritario funciona para los modelos de clasificación. Se aplican todos los modelos base a la entrada x y luego se devuelve la clase mayoritaria entre todas las predicciones. En caso de empate, se elige aleatoriamente una de las clases, o se devuelve un mensaje de error (si el hecho de clasificar erróneamente supondría un coste importante).

El apilamiento consiste en construir un metamodelo que toma el resultado de sus modelos base como entrada. Se desea combinar un clasificador f_1 y un clasificador f_2, ambos prediciendo el mismo conjunto de clases. Para crear un ejemplo de entrenamiento $(\hat{x_i}, \hat{y_i})$ para el modelo apilado, se configura $\hat{x_i} = [f_1(x), f_2(x)]$ y $\hat{y_i} = y_i$.

Si algunos de los modelos base devuelven no sólo una clase, sino también una puntuación para cada clase, también se pueden utilizar estos valores como variables independientes.

Para entrenar el modelo apilado, se recomienda utilizar ejemplos del conjunto de entrenamiento y ajustar los hiperparámetros del modelo apilado mediante validación cruzada. Obviamente, se debe asegurar que el modelo apilado funcione mejor, en el conjunto de validación, que cada uno de los modelos base que apiló.

La razón por la que la combinación de varios modelos puede generar un mejor rendimiento general es debido a la observación de que cuando varios modelos fuertes no correlacionados coinciden, es más probable que coincidan sobre el resultado correcto. La palabra clave aquí es "no correlacionados". Idealmente, diferentes modelos fuertes deben obtenerse usando diferentes variables independientes o usando algoritmos de diferente naturaleza, por ejemplo, Máquina de soporte de vectores (SVM) y Bosques Aleatorios. Combinando diferentes versiones del algoritmo de aprendizaje del árbol de decisiones, o varias SVMs con diferentes hiperparámetros pueden no dar como resultado un aumento significativo del rendimiento. [13]

3.9 Métricas de evaluación

Los modelos que se proponen como los mejores son los que arrojan mejores métricas de Área bajo la curva ROC (AUC), Error cuadrático medio (MSE) y Precisión (*Accuracy*) en el conjunto de datos de prueba.

3.9.1 Área Bajo la Curva ROC

La curva ROC (ROC significa "característica operativa del receptor", el término proviene de ingeniería

de radar) es un método comúnmente utilizado para evaluar el rendimiento de modelos de clasificación. Las curvas ROC utilizan una combinación de la tasa de verdaderos positivos (la proporción de ejemplos predichos correctamente, definidos exactamente como sensibilidad) y tasa de falsos positivos (la proporción de ejemplos negativos predichos incorrectamente) para construir una imagen resumida del desempeño de la clasificación.

La tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) se definen respectivamente como, $TPR = \frac{TP}{TP+FN}$ y $FPR = \frac{FP}{FP+TN}$.

Las curvas ROC sólo se pueden utilizar para evaluar clasificadores que arrojan alguna puntuación de confianza (o una probabilidad) de predicción. Por ejemplo, regresión logística, redes neuronales y árboles de decisión (y los conjuntos de modelos basados en árboles de decisión) se pueden evaluar utilizando curvas ROC.

Cuanto mayor sea el área bajo la curva ROC (AUC), mejor será el clasificador. Un clasificador con un AUC superior a 0,5 es mejor que un clasificador aleatorio. Si el AUC es inferior a 0,5, entonces algo anda mal con el modelo. Un clasificador perfecto tendría un AUC de 1. [13]

3.9.2 Error Cuadrático Medio

La métrica error cuadrático medio (MSE) mide el promedio de los cuadrados de los errores o desviaciones. MSE toma las distancias desde los puntos hasta la línea de regresión (estas distancias son los "errores") y las eleva al cuadrado para eliminar cualquier signo negativo. MSE incorpora tanto la varianza como el *bias* del predictor.

MSE también da más peso a diferencias más grandes. Cuanto mayor es el error, más se penaliza. Por ejemplo, si sus respuestas correctas son 2,3,4 y el algoritmo predice 1,4,3, entonces el error absoluto en cada una es exactamente 1, por lo que el error al cuadrado también es 1 y el MSE es 1. Pero si el algoritmo predice 2,3,6, entonces los errores son 0,0,2, los errores al cuadrado son 0,0,4 y el MSE es un valor mayor 1,333. Cuanto más pequeño sea el MSE, mejor será el rendimiento del modelo. [17]

Ecuación 9. MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

3.9.3 Precisión

La precisión (*Accuracy*) viene dada por el número de ejemplos clasificados correctamente dividido por el número total de ejemplos clasificados. En términos de la matriz de confusión, viene dada por:

Ecuación 10. Precisión

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

La precisión es una métrica útil cuando los errores al predecir todas las clases son igualmente importantes. En el caso de spam/no spam, es posible que este no sea el caso. Por ejemplo, se tolerarían falsos positivos menos que falsos negativos. Un falso positivo en la detección de spam es la situación en la que tu amigo te envía un correo electrónico, pero el modelo lo etiqueta como spam y no te lo muestra. Por otro lado, el falso negativo es un problema menor: si el modelo no detecta un pequeño porcentaje de mensajes spam, no es gran cosa [13].

3.9.4 Sensibilidad

La sensibilidad (*Recall|Sensitivity*) es la relación entre las predicciones positivas correctas y el número total de ejemplos positivos [13]. En términos de la matriz de confusión, viene dada por:

Ecuación 11. Sensibilidad

$$Sensibilidad = \frac{TP}{TP + FN}$$

3.9.5 Especificidad

La especificidad se calcula basada en cuantas personas no presentan una condición, rasgo, o evento. Puede ser calculada usando la ecuación: número de verdaderos negativos / (número de verdaderos negativos + número de falsos positivos). [18]

Ecuación 12. Especificidad

$$\text{Especificidad} = \frac{TN}{TN + FP}$$

3.9.6 Puntuación F1

La puntuación F1 mide la eficacia de un clasificador binario para clasificar casos positivos (dado un valor umbral). Se calcula a partir de la media armónica de la exactitud y la sensibilidad. Una puntuación F1 de 1 significa que tanto la exactitud como la sensibilidad son perfectas y que el modelo identificó correctamente todos los casos positivos y no marcó ningún caso negativo como positivo. Si la exactitud o la sensibilidad son muy bajas, se reflejará en una puntuación F1 cercana a 0. [17]

Ecuación 13. Puntuación F1

$$F1 = 2 * \frac{Exactitud * \text{Sensibilidad}}{Exactitud + \text{Sensibilidad}}$$

Donde:

La exactitud son las observaciones positivas (verdaderos positivos) que el modelo identificó correctamente de todas las observaciones que etiquetó como positivas (los verdaderos positivos + los falsos positivos).

4. ANÁLISIS DE DATOS EXPLORATORIO

Como resultado de una de las primeras fases de la minería de datos desarrollada en este trabajo, se presenta a continuación un análisis exploratorio de los datos, que se realizó con el objetivo de encontrar las características y patrones más importantes en estos datos, así como garantizar que estuvieran completos y correctos. También se buscó detectar datos atípicos y mostrar tanto en tablas como en gráficos análisis univariados y multivariados.

El conjunto de datos de estudio contiene la información de 396 estudiantes de una Licenciatura del ITESO. Las columnas para analizar del mismo, así como sus descripciones, tipos y rangos de valores se muestran en la Tabla 4-1. En la misma se describen once variables explicativas que caracterizan al alumno con información demográfica como el sexo, la edad, el lugar de procedencia de la preparatoria y el semestre de ingreso. También lo caracterizan académicamente con información financiera como si tienen o no beca, si ingresaron mediante pase directo y si cuentan con algún tipo de crédito. De estas variables seis son nominales (CatPrepa_Colapsed, CatPase, Sexo_M, Local, Semestre_Otonno y CatBeca) y cinco de razón (EdadIng, PromPre, Credito, Prepa_Encode_Prob_abandon e ind_planea), además de la variable nominal de salida E2 que permite clasificar si el alumno abandonó o no la carrera analizada.

Tabla 4-1. Especificación de columnas del archivo AlumISC_afterKM_.csv utilizadas en la construcción de los modelos

Nombre del campo o columna	Descripción	Tipo csv	Tipo de dato según escala de medición estadística	Tipo sugerido en Python pandas	Valores válidos
CatPrepa_Colapsed	Preparatorias agrupadas en 2 grupos numéricos.	numérico	nominal	category	1: escolarizada publica o escolarizada privada 0: no escolarizada (abierta) o extranjera
E2	En análisis de supervivencia E es una variable de salida que indica la ocurrencia o no del evento bajo	numérico	nominal	int64	1: si cambió de carrera o no continuó en la universidad en el primer o segundo semestre. 0: lo contrario

Nombre del campo o columna	Descripción	Tipo csv	Tipo de dato según escala de medición estadística	Tipo sugerido en Python pandas	Valores válidos
	estudio y está asociada con la variable T que contiene el tiempo en el cual ocurre el evento de estudio o el tiempo de censura				
CatPase	Identifica si el alumno tiene pase directo	numérico	nominal	category	1: Con Pase directo 0: Sin pase directo
Sexo_M	Sexo del alumno	numérico	nominal	category	1: Masculino 0: Femenino
Local	Identifica si el alumno es local (de la Zona Metropolitana de Guadalajara) o foráneo (fuera de la zona metropolitana).	numérico	nominal	category	1: LOCAL 0: FORÁNEO
Semestre_Otonno	Identifica si el alumno ingresó en semestre de Otoño	numérico	nominal	category	1: Otoño 0: Primavera
CatBeca	Identifica si el alumno cuenta con beca o no.	numérico	nominal	category	1: Tiene beca 0: No tiene beca
EdadIng	Edad en el momento del ingreso	numérico	razón	float64	Valores entre 17.0 y 35.0
PromPre	Promedio obtenido por el alumno en la preparatoria	numérico	razón	float64	Valores entre 65.0 y 100.0
Credito	Porcentaje de crédito con el que entra a la universidad	numérico	razón	float64	Valores entre 0.0 y 45.0
Prepa_Encode_Prob_abandon	Probabilidad histórica promedio de abandono de alumnos de una	numérico	razón	float64	Valores entre 0 y 1

Nombre del campo o columna	Descripción	Tipo csv	Tipo de dato según escala de medición estadística	Tipo sugerido en Python pandas	Valores válidos
	misma preparatoria en la universidad estudiada				
ind_planea	Índice de calidad de preparatoria en la prueba PLANEA 2015	numérico	razón	float64	Valores entre 0.0 y 186.975

4.1 Análisis descriptivo univariado

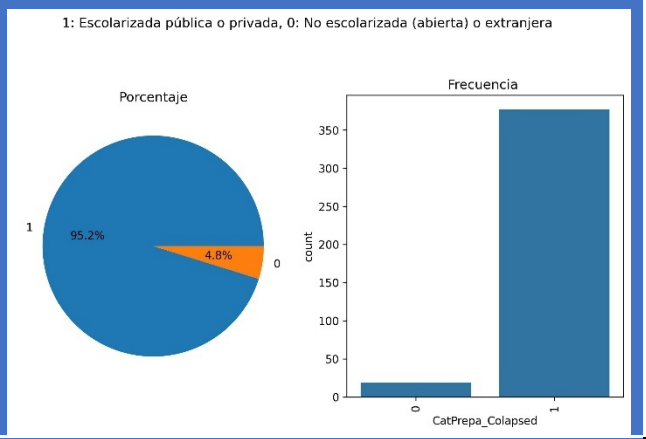
En esta sección exploramos cada una de las variables independientes y la variable dependiente E2 con el fin de garantizar la limpieza de estas, encontrar datos atípicos y corregir cualquier anomalía.

4.1.1 Análisis descriptivo univariado de datos nominales

Las preparatorias se clasificaron como escolarizadas públicas y privadas, no escolarizadas (abiertas) y extranjeras. En un trabajo anterior se determinó que no existían diferencias significativas en cuanto a abandono entre las escolarizadas ni diferencias entre las abiertas y extranjeras. Es por ello que se colapsaron solo en dos categorías: Escolarizadas o No Escolarizadas (abiertas y extranjeras). Esta variable se llamó CatPrepa_Colapsed y su análisis se muestra en la Figura 4-1.

Figura 4-1. Estadística descriptiva de CatPrepa_Colapsed

Valores	Frecuencia	%
1: escolarizada publica o escolarizada privada	377	95.2
0: no escolarizada (abierta) o extranjera	19	4.8



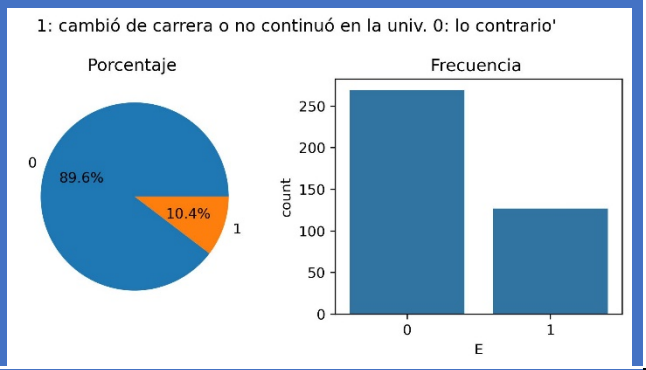
Columnas	conteo	valores únicos	moda
CatPrepa_Colapsed	396	2	1: escolarizada pública o escolarizada privada

Como se muestra en la Figura 4-1, la mayoría de los estudiantes provienen de escuelas privadas y públicas con 377 estudiantes, representando el 95.2 %. El resto de los estudiantes provienen de escuelas extranjeras y escuelas de tipo abiertas con solo 19 estudiantes, representando el 4.8 %.

En la Figura 4-2 se muestra información de la variable E2 que es la ocurrencia o no del evento de baja de carrera hasta el segundo semestre.

Figura 4-2. Estadística descriptiva del evento de abandono de carrera hasta el segundo semestre

Valores	Frecuencia	%
0: No cambio de carrera y se mantuvo en la universidad.	355	89.6
1: Cambio de carrera o no continuó en la universidad	41	10.4



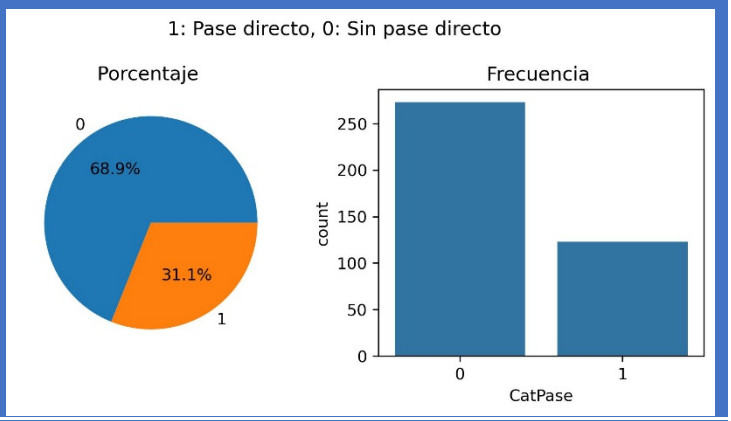
Columnas	conteo	valores únicos	moda
E2	396	2	0: No cambió de carrera y se mantuvo en la universidad.

Se observa que la mayoría de los estudiantes (el 89.65 %) continuó en la misma carrera y se mantuvo en la universidad. Dicho de otra manera, solo el 10.35 % dejó la universidad o se cambió de carrera.

La variable CatPase contiene información referente a si los estudiantes ingresaron a la ISC con pase directo. Esta indica si requirieron realizar pruebas de ingreso a la universidad o no. La información de esta variable se muestra en la Figura 4-3.

Figura 4-3. Estadística descriptiva de CatPase

Valores	Frecuencia	%
0: Sin pase directo	273	68.9
1: Con pase directo	123	31.1



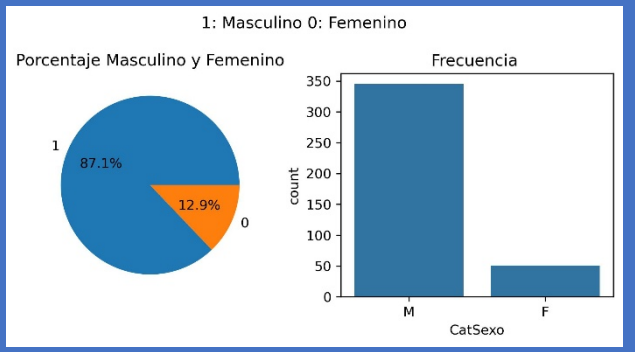
Columna	conteo	valores únicos	moda
CatPase	396	2	0 : Sin pase directo

Casi 2 terceras partes de los estudiantes (68.9%) no tienen pase directo (273 alumnos) y solo un 31.1 % tienen pase directo. Estos últimos son de preparatorias con convenio con el ITESO por su reconocida calidad, además de tener más de 5 de promedio.

En la Figura 4-4 se muestra información relevante sobre la variable Sexo_M, que indica si un estudiante es de sexo Masculino o no.

Figura 4-4. Estadística descriptiva de Sexo_M

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



Columna	conteo	valores únicos	moda
Sexo_M	396	2	1: Masculino

La mayoría de los estudiantes son hombres (87.1%) y solo el 12.9 % son mujeres.

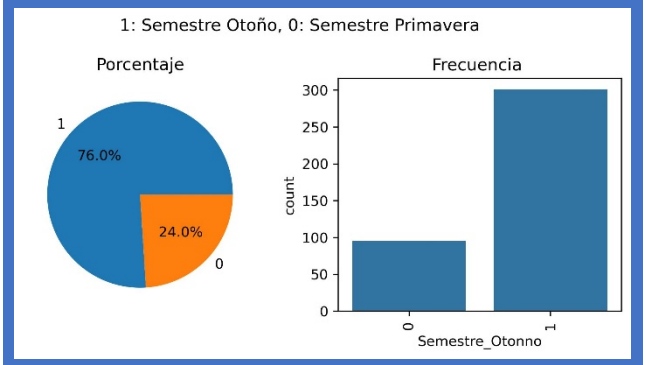
La

Figura 4-5 contiene información relativa a la variable Semestre_Otonno, la cual indica si un estudiante ingresó a la escuela en Semestre de Otoño o Semestre de Primavera.

Figura 4-5. Estadística descriptiva de Semestre_Otonno

Valores	Frecuencia	%
1: Semestre Otoño	301	76.0

0: Semestre Primavera	95	24.0



Columna	conteo	valores únicos	moda
Semestre_Otonno	396	2	1: Semestre Otoño

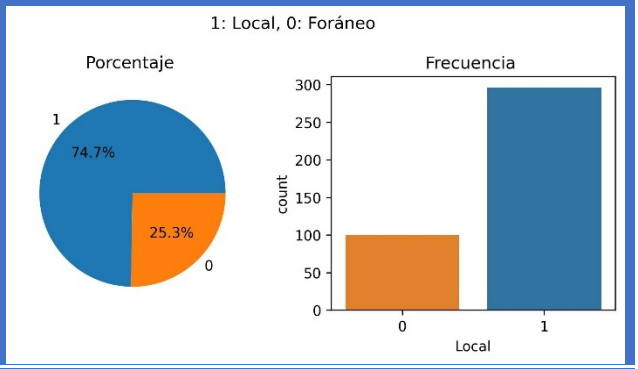
La mayoría de los estudiantes entran en otoño con un 76 % y solo el 24 % en primavera. Esto se podría explicar porque otoño es cuando los estudiantes salen de las preparatorias para unirse a la universidad.

La variable Local refiere si un estudiante es Local (que pertenece al Área Metropolitana de Guadalajara) o Foráneo (fuera del Área Metropolitana, incluyendo a estudiantes de otros estados y extranjeros). La

Figura 4-6 muestra las estadísticas descriptivas de la misma.

Figura 4-6. Estadística descriptiva para Local

Valores	Frecuencia	%
1: LOCAL	296	74.7
0: FORANEO	100	25.3



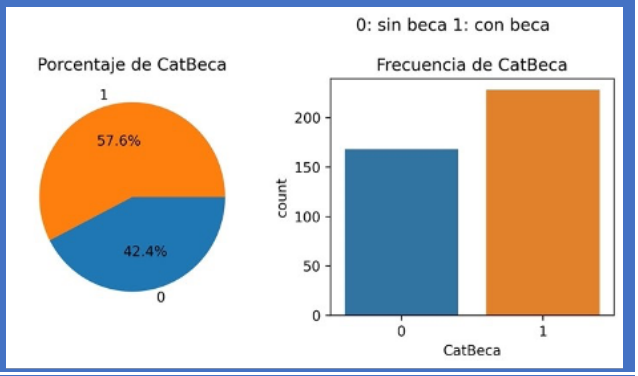
Columna	conteo	valores únicos	moda
Local	396	2	1: LOCAL

La mayoría de los estudiantes son locales con el 74.7 %, comparado con los foráneos que representan solo el 25.3 % de los datos.

CatBeca es la variable que contiene las categorías de Beca: 1 si al estudiante se le otorgó beca al ingreso y 0 en caso contrario. Su análisis se muestra en la Figura 4-7.

Figura 4-7. Estadística descriptiva de CatBeca

Valores	Frecuencia	%
1: Tiene beca	228	57.6
0: No tiene beca	168	42.4



Columna	conteo	valores únicos	moda
CatBeca	396	2	1: Tiene beca

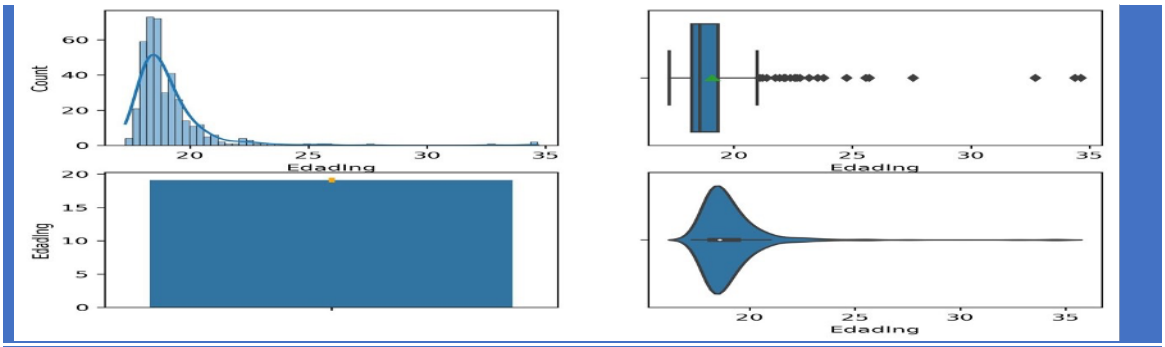
Puede ser una beca de excelencia académica, por promoción de nueva carrera, por situación socioeconómica difícil, por prestación a hijos de trabajadores del ITESO, etc.

No hay desbalance severo entre las muestras de CatBeca. Más de la mitad de los estudiantes cuentan con Beca.

4.1.2 Análisis descriptivo univariado de datos de razón

En el análisis descriptivo univariado de datos de razón se obtiene información relevante como la media, la mediana, la desviación estándar, los cuartiles, la curtosis y la simetría. Para la variable EdadIng, la cual representa la edad en el momento de ingreso a la ISC, este análisis se muestra en la Figura 4-8 que contiene en la parte superior izquierda el histograma con la función de densidad de Kernel estimada. En la parte superior derecha se presenta el diagrama de cajas y bigotes. En la parte inferior izquierda se presenta un diagrama de barras. En la parte inferior derecha se presenta un diagrama de violín (combinación de función de densidad de Kernel con cajas y bigotes).

Figura 4-8. Estadística descriptiva de EdadIng

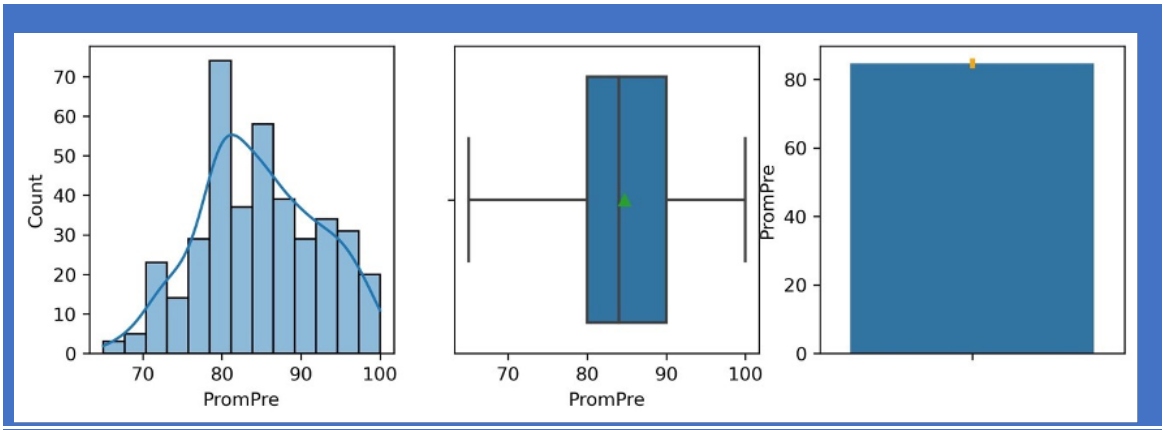


Colum-na	conteo	media	Desvia-ción	min	25%	50%	75%	Max	curtosis	simetría
Edad Ing	396.00	19.10	1.80	17.28	18.24	18.58	19.35	34.65	36.29	5.14

Se pueden observar valores atípicos arriba del cuarto cuartil por encima de los veintiún años. También se puede observar que la media es mayor a la mediana y que la distribución no es normal, está sesgada a la derecha, lo que se observa en la cola hacia la derecha.

El análisis de la variable PromPre, que contiene el promedio obtenido por el alumno en la preparatoria, se muestra a continuación en la Figura 4-9.

Figura 4-9. Estadística descriptiva de PromPre

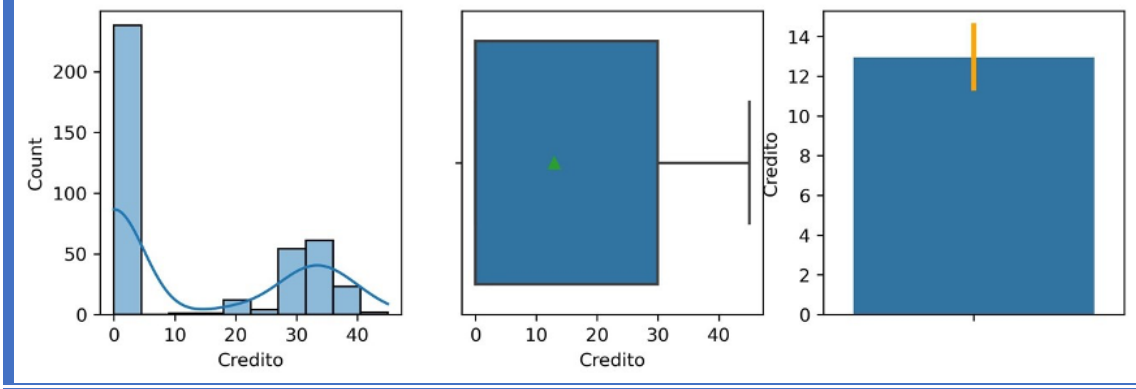


Colum-na	conteo	media	Desvia-ción	min	25%	50%	75%	Max	curtosis	simetría
Prom Pre	396.00	84.71	7.51	65.00	80.00	84.00	90.00	100.00	-0.59	0.02

El grueso de los registros está entre 80 y 90. La media y la mediana son cercanas. Hay un sesgo a la izquierda (negativo) ligero y la distribución parece cercana a la normal. No hay datos atípicos.

En la Figura 4-10 se muestra el análisis de la variable Crédito, que contiene el porcentaje de crédito con el que entra el alumno a la universidad.

Figura 4-10. Estadística descriptiva de Crédito

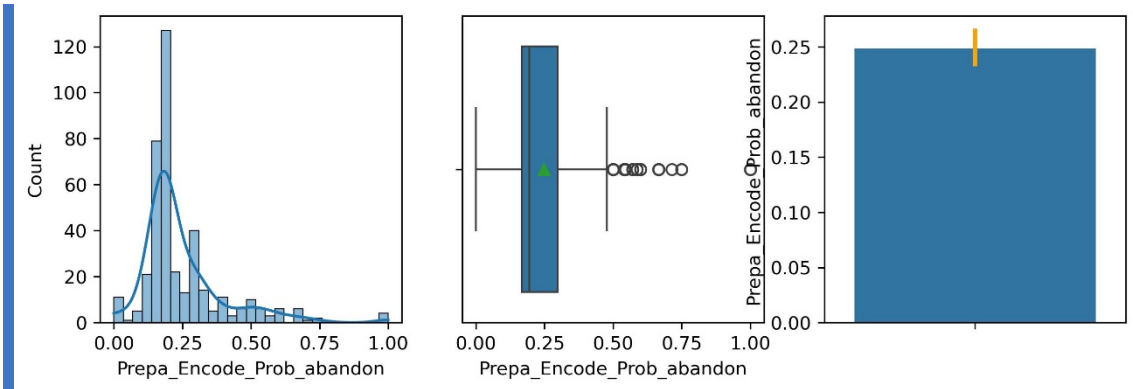


Colum-na	conteo	media	Desvia-ción	min	25%	50%	75%	Max	curtosis	simetría
Credito	396.00	12.95	16.33	0.00	0.00	0.00	30.00	45.00	-1.57	0.55

La media está muy por debajo de la mediana. Los datos tienen una distribución bimodal.

La variable Prepa_Encode_Prob_abandon se refiere a la probabilidad histórica promedio de abandono de alumnos de una misma preparatoria en la universidad estudiada. Las estadísticas de dicha variable se muestran en la Figura 4-11.

Figura 4-11. Estadística descriptiva de Prepa_Encode_Prob_abandon



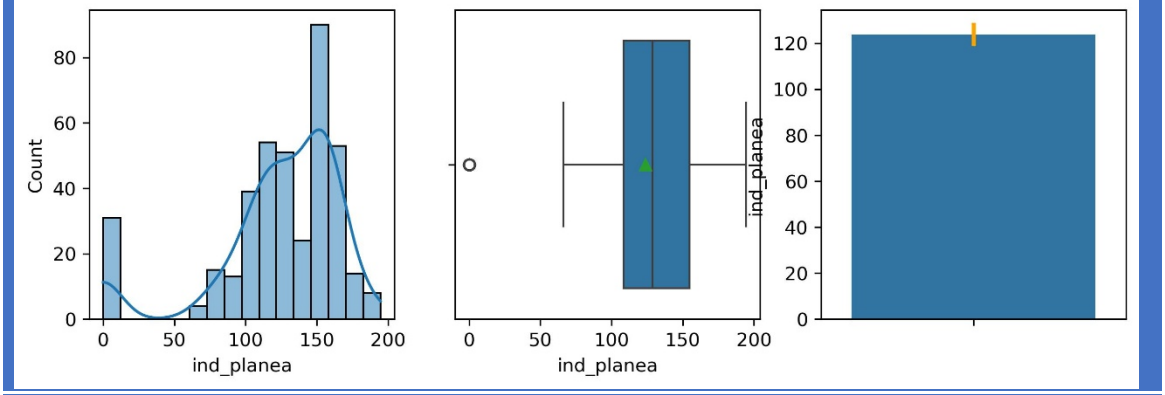
Colum-na	conteo	media	Desvia-ción	min	25%	50%	75%	Max	curtosis	simetría

Prepa_ Encode_ Prob_abandon	**396.00**	**0.2484**	**0.1539**	**0.00**	**0.1666**	**0.1944**	**0.2972**	**1.00**	**5.7685**	**2.0651**

La mediana es menor que la media. Después de 0.5 se observan valores atípicos. La distribución no es normal, está sesgada a la derecha, lo que se observa en la cola hacia la derecha.

En la Figura 4-12 se observan las estadísticas de la variable ind_planea, la cual se refiere al índice de calidad de preparatoria en la prueba PLANEA 2015. Esta variable en estudios anteriores se comprobó que es de mucha utilidad para predecir el abandono.

Figura 4-12. Estadística descriptiva de ind_planea



Colum-na	conteo	media	Desvia-ción	min	25%	50%	75%	Max	curtosis	simetría
ind_planea	**396.00**	**123.82**	**44.22**	**0.00**	**108.57**	**128.85**	**154.70**	**194.62**	**2.22**	**-1.54**

La media es menor que la mediana. En el valor cero se observan valores atípicos. Los datos tienen una distribución bimodal.

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/razón. Además de ver cómo se distribuyen conjuntamente estos datos, presentaremos 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

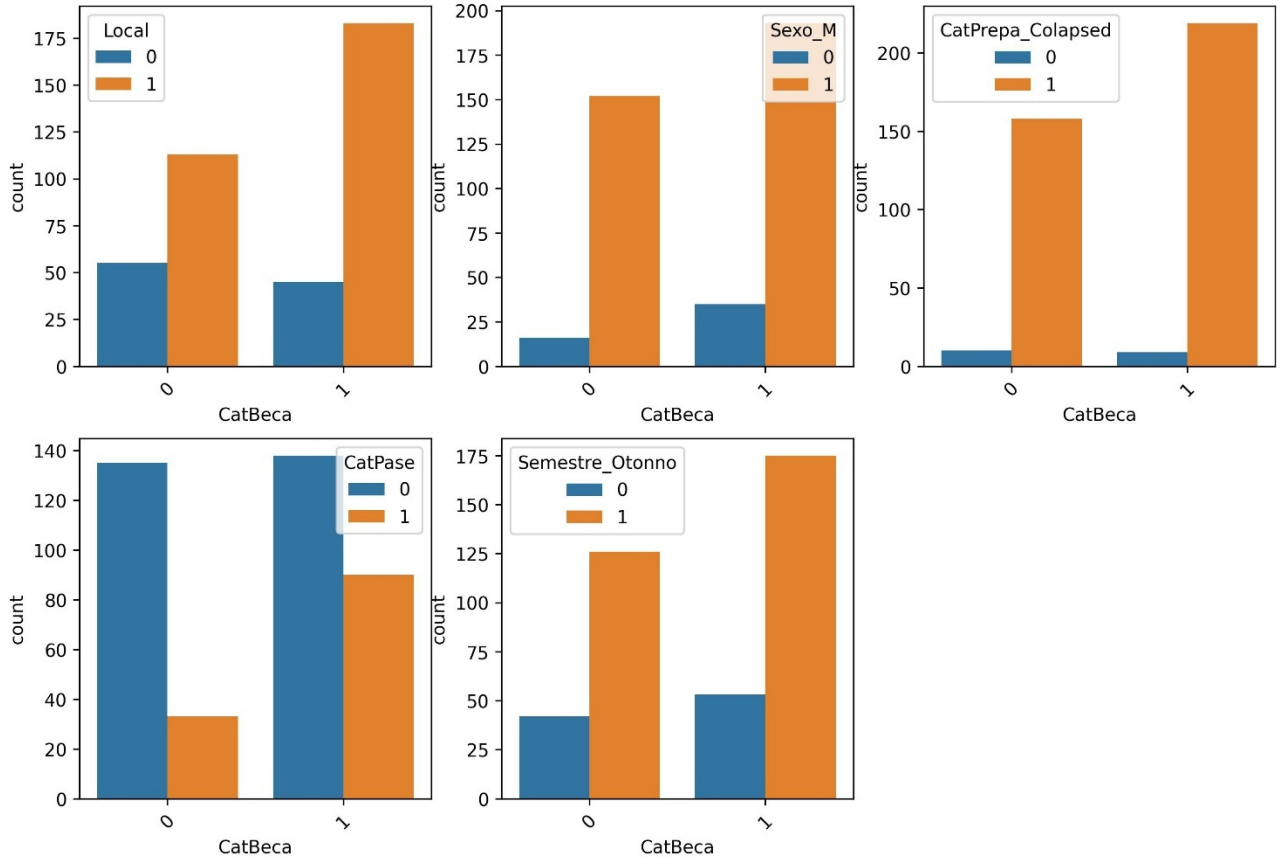
analizará la distribución mediante diagramas de dispersión y se calcularán los índices de correlación de Pearson y Spearman. Finalmente se analizará la correlación de todas las variables mediante el cálculo del coeficiente de correlación phi-k y la matriz de significancia.

4.2.1 Entre dos variables categóricas (nominales u ordinales)

Se utilizaron los métodos *countplot* de la librería seaborn, el cual nos muestra las distribuciones de frecuencia de una variable contra el resto de variables categóricas, y el método *crosstab* de la librería pandas el cual construye una tabla de frecuencias, en este caso entre cada par de variables mostradas en el resultado de *countplot*. Estas frecuencias fueron normalizadas y llevadas a porciento para un mejor entendimiento.

La Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable CatBeca contra las demás variables categóricas.

Figura 4-13. Distribuciones de frecuencia de CatBeca contra el resto de las variables categóricas



La Tabla 4-2 muestra el resultado del método *crosstab* para las categorías de beca y las categorías de pase directo.

Tabla 4-2. Porcentaje de alumnos por categoría de beca y categoría de pase directo

CatPase	0	1	Total
CatBeca			
0	34.09%	8.33%	42.42%
1	34.85%	22.73%	57.58%
Total	68.93%	31.06%	100.00%

Al 57.58 % de los alumnos les fue otorgada una beca, mientras que al 42.42 % restante no. De estos que no recibieron beca, la mayoría no ingresaron por pase directo, con un 34.09 % del total de estudiantes en este grupo.

En la Tabla 4-3 se muestra la distribución de alumnos por categoría de beca y semestre de otoño.

Tabla 4-3. Porcentaje de alumnos por categoría de beca y semestre de otoño

Semestre_Otonno	0	1	Total
CatBeca			
0	10.60%	31.82%	42.42%
1	13.38%	44.19%	57.57%
Total	23.99%	76.01%	100.00%

Poco más de tres cuartos de los estudiantes ingresaron en otoño. En primavera solo ingresaron el 23.99 % de los estudiantes, y de estos solo el 13.38 % del total recibieron beca. Se concluye que la mayoría de estudiantes que reciben beca ingresan en el semestre de otoño.

El porcentaje de alumnos que reciben beca o no según la categoría de preparatoria colapsada se muestra en la Tabla 4-4.

Tabla 4-4. Porcentaje de alumnos por categoría de beca y categoría de Prepa colapsada

CatPrepa_Colapsed	0	1	Total
CatBeca			
0	2.53%	39.90%	42.43%
1	2.27%	55.30%	57.57%
Total	4.80%	95.20%	100.00%

La mayoría de los estudiantes que reciben beca provienen de prepas Privadas y Públicas.

La distribución en porciento de estudiantes que reciben beca o no según el género se muestra en la Tabla 4-5.

Tabla 4-5. Porcentaje de alumnos por categoría de beca y sexo masculino

Sexo_M	0	1	Total
CatBeca			
0	4.04%	38.38%	42.42%
1	8.83%	48.74%	57.57%
Total	12.87%	87.12%	100.00%

Los estudiantes de género femenino representan solamente el 12.87 %, y de estos la mayoría reciben beca. En el caso de los estudiantes de género masculino poco más de la mitad reciben beca, representado por el 48.74 % del total.

La distribución de alumnos locales y foráneos que reciben beca se muestra en la Tabla 4-6.

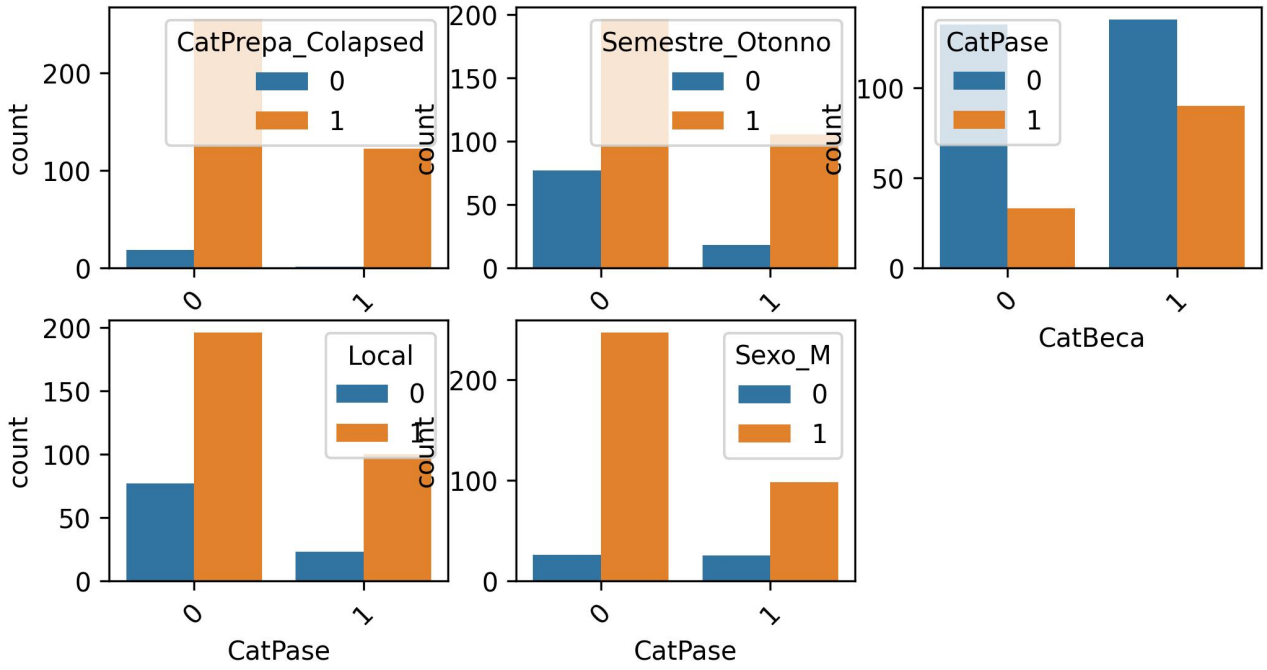
Tabla 4-6. Porcentaje de alumnos por categoría de beca y Local

Local	0	1	Total
CatBeca			
0	13.89%	28.54%	42.43%
1	11.36%	46.21%	57.57%
Total	25.25%	74.75%	100.00%

Los estudiantes locales representan casi tres cuartos de los estudiantes, por tanto, este grupo es el que más becas reciben. Solo el 11.36 % de los estudiantes son foráneos y reciben beca.

La Figura 4-14Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable CatPase contra las demás variables categóricas.

Figura 4-14. Distribuciones de frecuencia de CatPase contra el resto de las variables categóricas



Los estudiantes con pase directo o no, según la categoría de prepa, se muestran en la Tabla 4-7.

Tabla 4-7. Porcentaje de alumnos por categoría de pase directo y categoría de Prepa colapsada

CatPrepa_Colapsed	0	1	Total
CatPase			
0	4.55%	64.39%	68.94%
1	0.25%	30.81%	31.06%
Total	4.80%	95.20%	100.00%

El 68.94 % de los estudiantes no ingresó a la carrera mediante pase directo. Del 31.06 % que sí ingresó con pase directo, la mayoría provienen de preparatorias Públicas y Privadas.

La Tabla 4-8 muestra el porcentaje de alumnos según el semestre en que ingresaron y la categoría de pase directo.

Tabla 4-8. Porcentaje de alumnos por categoría de pase directo y semestre Otoño

Semestre_Otonno	0	1	Total
CatPase			
0	19.44%	49.49%	68.93%
1	4.55%	26.52%	31.07%
Total	23.99%	76.01%	100.00%

Se otorga mayor número de pases directos en otoño con el 26.52 % y solo el 4.55 % en primavera.

En la Tabla 4-9 se muestra la distribución de alumnos por sexo y pase directo.

Tabla 4-9. Porcentaje de alumnos por categoría de pase directo y sexo Masculino

Sexo_M	0	1	Total
CatPase			
0	6.57%	62.37%	68.94%
1	6.31%	24.75%	31.06%
Total	12.88%	87.12%	100.00%

Se otorgan más pases directos a hombres con 24.75 % que a mujeres con el 6.31 %. Aunque en la población de mujeres es casi el mismo porcentaje para mujeres con o sin pase.

El porcentaje de alumnos según la categoría de local o foráneo y si contaron con pase directo o no se muestra en la Tabla 4-10Tabla 4-10.

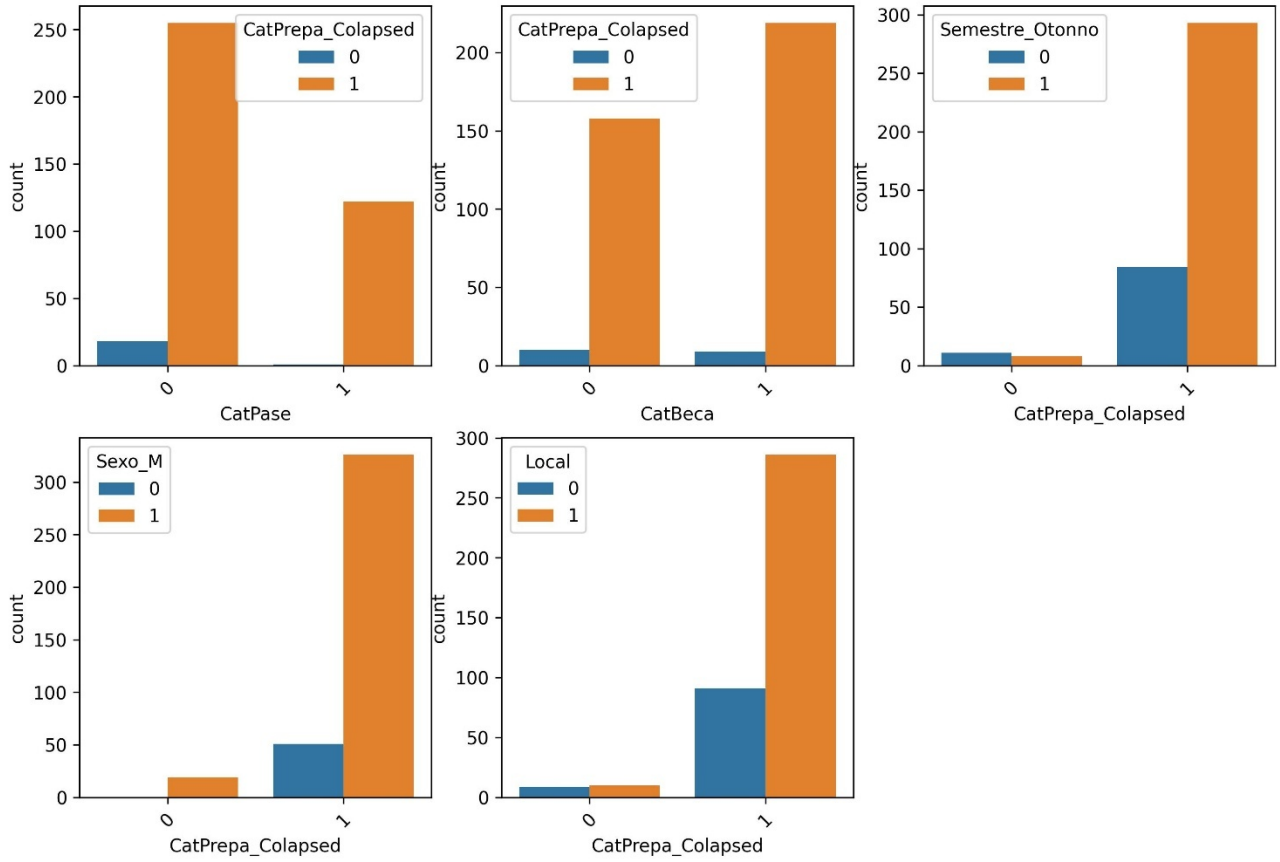
Tabla 4-10. Porcentaje de alumnos por categoría de pase directo y Local

Local	0	1	Total
CatPase			
0	19.44%	49.49%	68.93%
1	5.81%	25.26%	31.07%
Total	25.25%	74.75%	100.00%

Solo el 31.07 % de los estudiantes ingresaron mediante pase directo. De estos casi todos son estudiantes locales, representado por el 25.26 % del total.

La Figura 3-1Figura 4-15Figura 4-16Figura 4-18Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable CatPrepa_Colapsed contra las demás variables categóricas.

Figura 4-15. Distribuciones de frecuencia de CatPrepa_Colapsed contra el resto de las variables categóricas



La Tabla 4-11Tabla 4-11 contiene información relevante a la categoría de local o foráneo y la categoría de prepa colapsada.

Tabla 4-11. Porcentaje de alumnos por categoría de Prepa colapsada y Local

Local	0	1	Total
CatPrepa_Colapsed			
0	2.27%	2.53%	4.80%
1	22.98%	72.22%	95.20%
Total	25.25%	74.75%	100.0%

El 72.22 % de los estudiantes son locales y provienen de preparatorias Privadas y Públicas. El segundo grupo que más estudiantes aporta a la carrera es el de estudiantes foráneos de este mismo tipo de preparatorias con el 22.98 %.

En la Tabla 4-12 se muestra el porciento de estudiantes según la categoría de prepa colapsada y el semestre de ingreso.

Tabla 4-12. Porcentaje de alumnos por categoría de Prepa colapsada y semestre Otoño

Semestre_Otonno	0	1	Total
CatPrepa_Colapsed			
0	2.78%	2.02%	4.80%
1	21.21%	73.99%	95.20%
Total	23.99%	76.01%	100.00%

El 76.01 % de los estudiantes ingresa en otoño. De estos la gran mayoría provienen de prepas Públicas y Privadas. Solo el 2.02 % de los estudiantes de otoño son de prepas Abiertas y Extranjeras.

La distribución de alumnos por categoría de prepa colapsada y sexo se muestra en la Tabla 4-13.

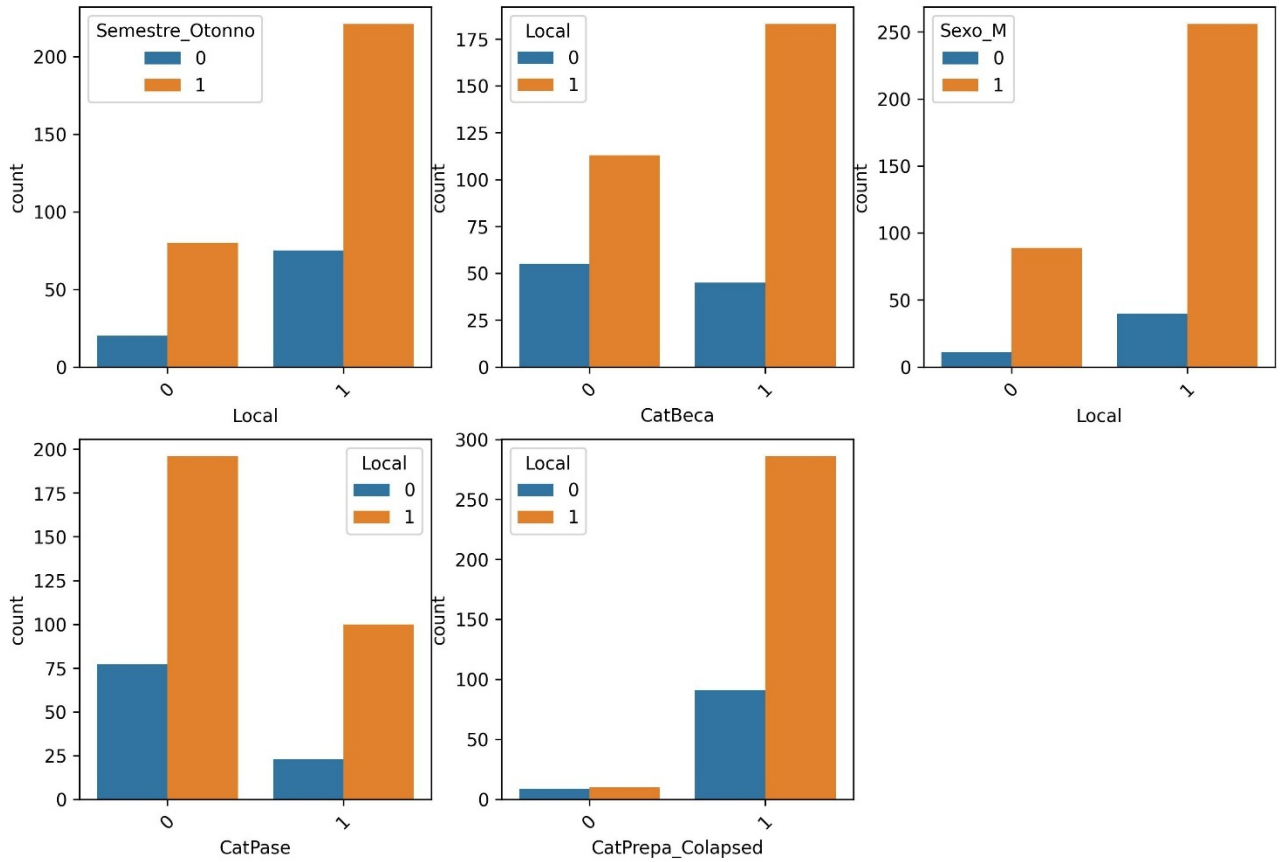
Tabla 4-13. Porcentaje de alumnos por categoría de Prepa colapsada y sexo masculino

Sexo_M	0	1	Total
CatPrepa_Colapsed			
0	0.00%	4.80%	4.80%
1	12.88%	82.32%	95.20%
Total	12.88%	87.12%	100.00%

La gran mayoría de estudiantes son hombres y provienen de prepas Públicas y Privadas, representando el 82.32 % del total. Solo el 12.88 % son mujeres provenientes de este mismo tipo de prepas, y no hay mujeres que provengan de prepas Abiertas y Extranjeras.

La Figura 4-16Figura 4-18Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable Local contra las demás variables categóricas.

Figura 4-16. Distribuciones de frecuencia de Local contra el resto de las variables categóricas



El porcentaje de alumnos por semestre de ingreso y categoría local o foráneo se muestra en la Tabla 4-14.

Tabla 4-14. Porcentaje de alumnos por categoría de Local y semestre Otoño

Semestre_Otonno	0	1	Total
Local			
0	5.05%	20.20%	25.25%
1	18.94%	55.81%	74.75%
Total	23.99%	76.01%	100.00%

La mayoría de los alumnos locales ingresan en otoño, representado por el 74.75 %. De los foráneos el 20.20 % del total también ingresa en otoño.

En la Tabla 4-15 se presenta el porcentaje de estudiantes por sexo y categoría de local o foráneo.

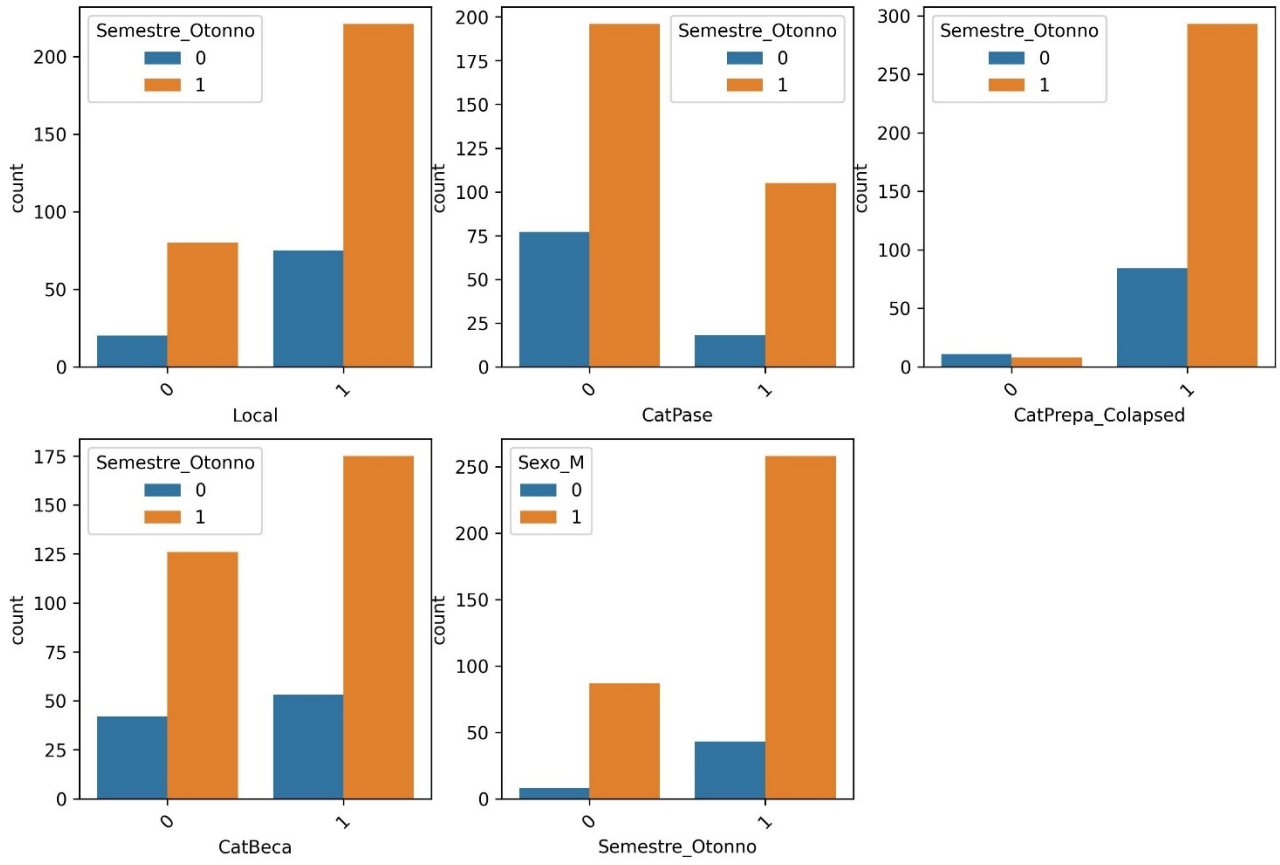
Tabla 4-15. Porcentaje de alumnos por categoría de LocalForaneo Local y sexo Masculino

Sexo_M	0	1	Total
Local			
0	2.78%	22.47%	25.25%
1	10.10%	64.65%	74.75%
Total	12.88%	87.12%	100.00%

La mayoría de los alumnos locales son del género masculino. Solo el 10.10 % del total son mujeres locales.

La Figura 4-17Figura 4-18Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable Semestre_Otonno contra las demás variables categóricas.

Figura 4-17. Distribuciones de frecuencia de Semestre_Otonno contra el resto de las variables categóricas



En la Tabla 4-16 se representa el porcentaje de alumnos que ingresaron a la ISC por sexo y semestre.

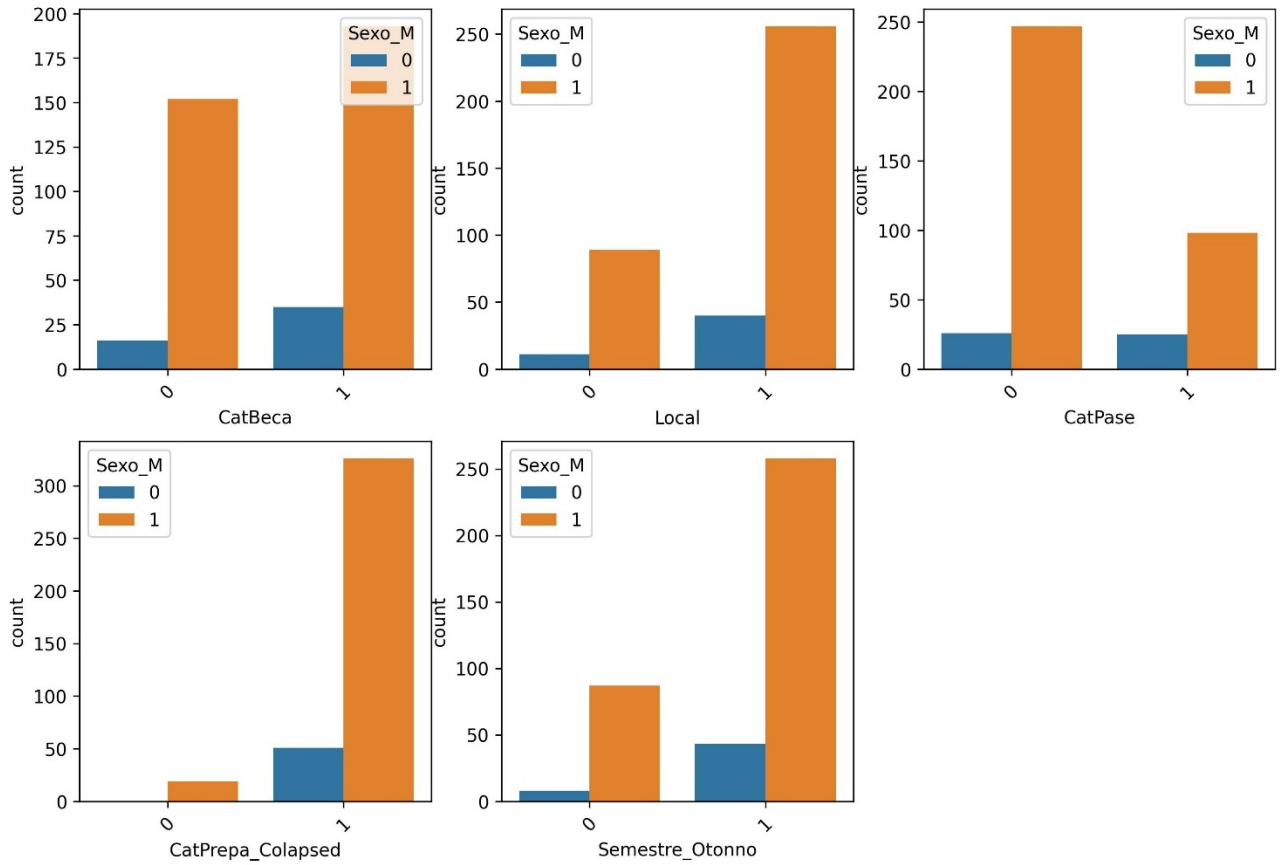
Tabla 4-16. Porcentaje de alumnos por semestre Otoño y sexo Masculino

Sexo_M	0	1	Total
Semestre_Otonno			
0	2.02%	21.97%	23.99%
1	10.86%	65.15%	76.01%
Total	12.88%	87.12%	100.00%

Solo el 2.02 % de los estudiantes son del sexo femenino e ingresan en primavera. La mayoría son hombres e ingresan en otoño.

La Figura 4-18Figura 4-13 nos muestra el resultado de comparar mediante el método *countplot* la variable Sexo_M contra las demás variables categóricas.

Figura 4-18. Distribuciones de frecuencia de Sexo_M contra el resto de variables categóricas



Las distribuciones de alumnos según el sexo en comparación con el resto de las variables categóricas ya fueron analizadas en tablas anteriores.

4.2.2 Entre una categórica (nominal u ordinal) y una numérica (de intervalo o razón)

Para analizar las correlaciones entre variables de este tipo se utilizó la correlación de punto biserial[2]. Los resultados de las pruebas de significancia de estas correlaciones se muestran en el apéndice 7.1. En la Tabla 4-17Tabla 4-17 se muestra un resumen de este análisis en el que se compararon las variables categóricas contra las numéricas de dos en dos y se analizó si los grupos de una variable tenían iguales medias en comparación con los grupos de la otra.

Tabla 4-17. Resumen de correlaciones de punto biserial entre variables categóricas y numéricas

Variables numéricas	EdadIng	PromPre	Credito	Prepa_Encode_Prob_abandon	ind_planea
Variables categóricas					
CatBeca	No	Si (0.43)	Si (0.53)	No	No
CatPase	No	Si (0.44)	No	No	No
Sexo_M	No	No	No	No	No
Semestre_Otonno	No	No	No	No	No
CatPrepa_Colapsed	No	No	No	No	Si (0.61)
Local	No	No	No	No	No

Se estableció como umbral de decisión el p-valor de 0.05. Si el p-valor es menor a 0.05 los grupos de variables no tienen igual media y por tanto se estableció que si existe una correlación significativa. Además, se observó si los grupos de variables tenían correlación significativa alta entre sí, considerándose alta la correlación superior a 0.4.

4.2.3 Entre dos numéricas (de intervalo o razón)

Para analizar las relaciones entre variables numéricas se utilizó el método *pairplot* de la librería seaborn el cual genera un diagrama de dispersión y mediante el método *corr* de la librería pandas el cual calcula la correlación de Pearson entre las columnas de un conjunto de datos.

En la Figura 4-19 se muestra el diagrama de dispersión de todas las variables numéricas.

[2] Esta correlación mide la fortaleza y dirección de la relación entre una variable continua y una dicotómica. Es un caso especial de correlación de Pearson donde una variable es continua y la otra dicotómica.

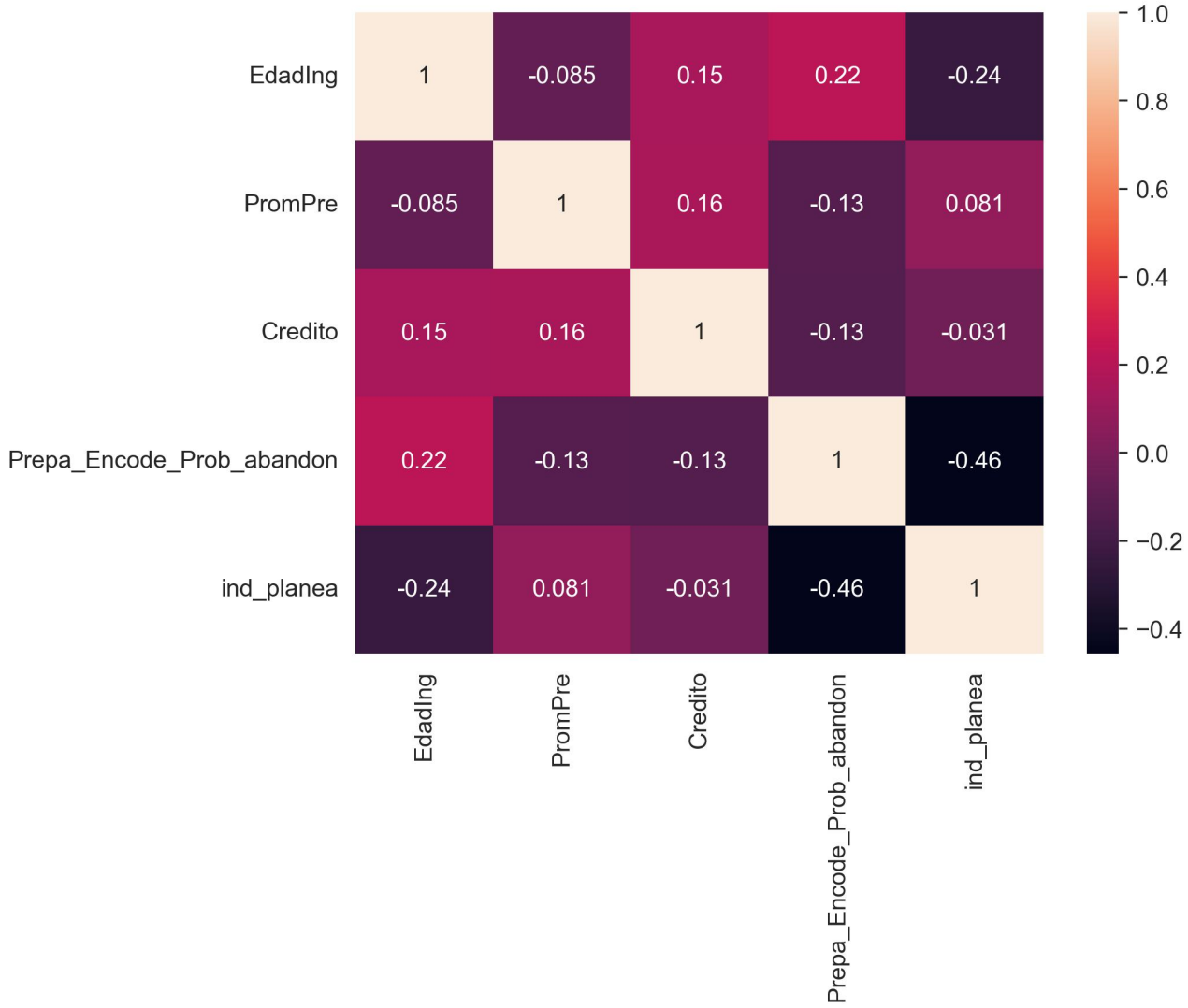
Figura 4-19. Diagrama de dispersión entre variables numéricas

Todas las variables representadas son de razón. Crédito, Prepa_Encode_Prob_abandon e ind_planea son discretas, el resto son continuas. No se observan relaciones lineales entre los pares de variables. En la diagonal principal tenemos la estimación de densidad de kernel que refleja las distribuciones de densidad de las variables de razón. En el caso de Crédito, Prepa_Encode_Prob_abandon e ind_planea parecen ser bimodales.

4.2.3.1 Correlación de Pearson

La correlación de Pearson indica la existencia de correlaciones lineales entre las variables. Los valores de correlaciones entre las variables de razón se muestran en la Figura 4-20.

Figura 4-20. Correlación de Pearson entre variables de razón



Valores cercanos a uno indican una correlación lineal fuerte, valores cercanos a 0.45 indican que existe una correlación lineal, pero esta es débil, mientras que el símbolo negativo indica una correlación lineal inversa, o sea que mientras una variable crece la otra disminuye. Para este grupo de variables solo se observa una correlación negativa débil entre ind_planea y Prepa_Encode_Prob_abandon.

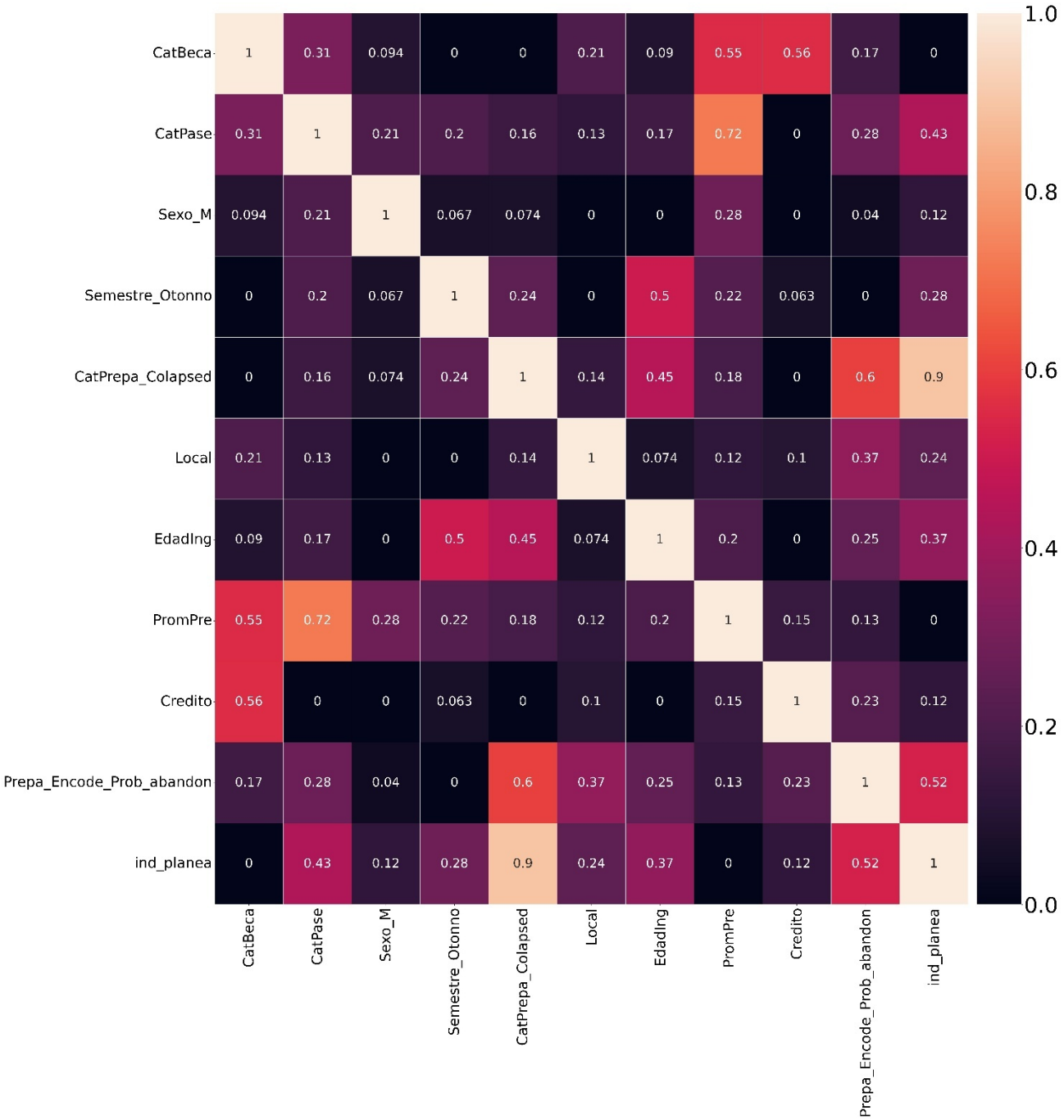
4.2.4 Análisis exploratorios entre todas las variables

En esta sección se analizará la correlación entre todas las variables mediante el cálculo del coeficiente de correlación phi-k y luego se analizará el nivel de significancia de estas correlaciones mediante la matriz de significancia. Finalmente se hará una comparación resumen entre estos valores.

4.2.4.1 Coeficiente de correlación phi-k

En la Figura 4-21 se muestra el mapa de calor generado con los valores de las correlaciones de phi-k mediante el método *heatmap* de la librería seaborn.

Figura 4-21. Coeficiente de correlación phi-k entre todas las variables



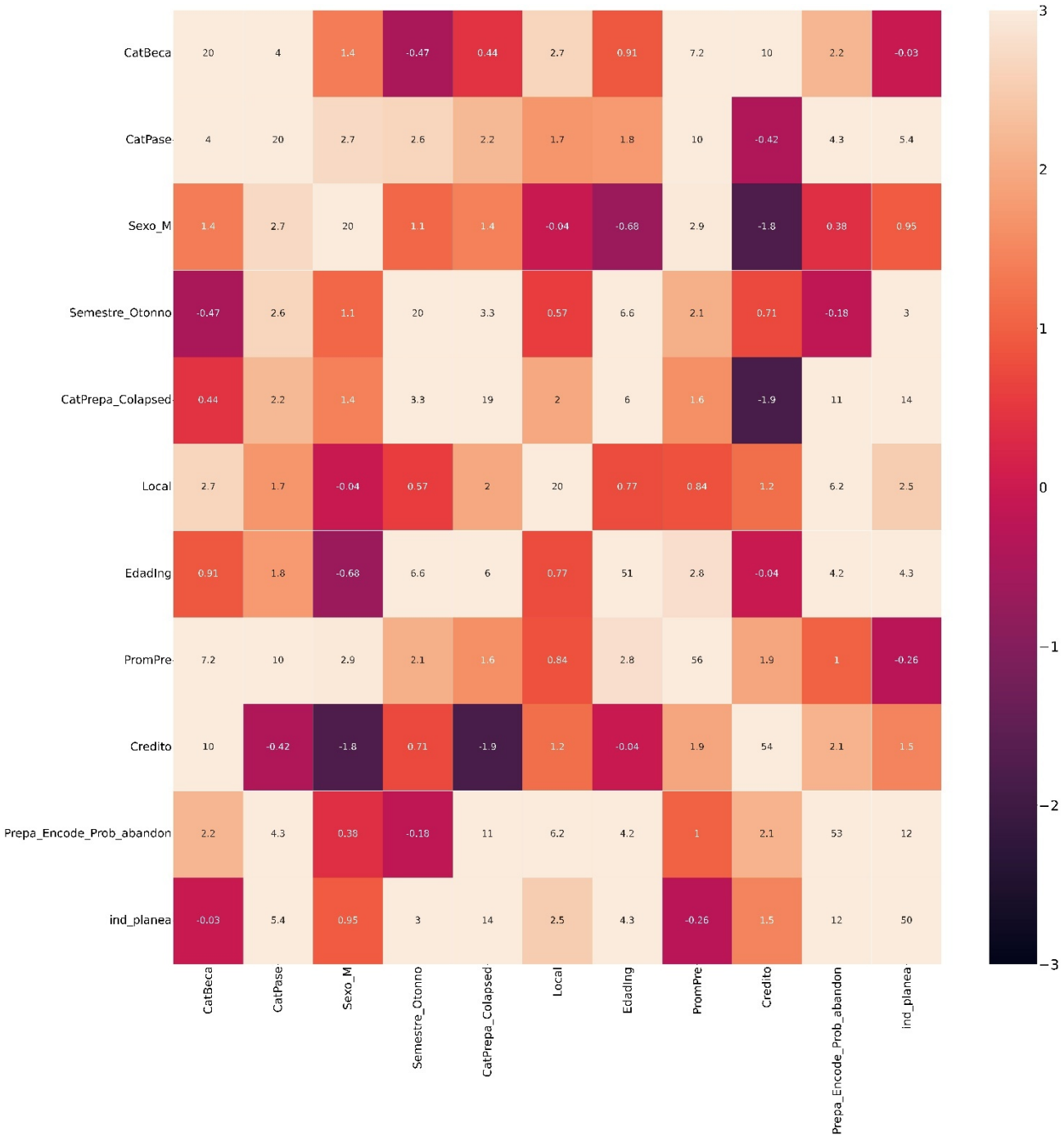
Existe una correlación fuerte entre la variable CatBeca y las variables PromPre y Crédito que indica que los estudiantes con mayor promedio de preparatoria tienen más posibilidades de contar con algún tipo de financiamiento. CatPase tiene una correlación muy fuerte con PromPre ya que los estudiantes con muy buen promedio de preparatoria tienen más posibilidades de ingresar a la carrera mediante pase directo. Hay una

correlación muy fuerte de CatPrepa_Colapsed contra la variable ind_planea que indica que los estudiantes de preparatorias privadas y públicas tienen un mejor índice de calidad de preparatoria en la prueba PLANEA 2015.

4.2.4.2 Matriz de significancia

La matriz de significancia que se muestra en la Figura 4-22 nos muestra el mapa de calor generado con los valores de significancia que arroja el método *significance_matrix* de pandas.

Figura 4-22. Matriz de significancia de las correlaciones de todas las variables



Se define la importancia de la correlación mediante el mapa de calor, y el nivel de significancia se satura al rango [-3, +3] desviaciones estándar. Los valores en este rango indican una mayor significancia de la correlación.

4.2.4.3 Comparación de los valores de la matriz de phi-k contra la matriz de significancia

En el apéndice 7.2 se muestran las comparaciones de los valores de la matriz de phi-k contra la matriz de significancia de cada variable contra el resto, que indica si es significante esta correlación.
En la Tabla 4-18 se indica el resumen de este análisis.

Var1: CatBeca

Var2: CatPase

Var3: Sexo_M

Var4: Semestre_Otonno

Var5: CatPrepa_Colapsed

Var6: Local

Var7: EdadIng

Var8: PromPre

Var9: Credito

Var10: Prepa_Encode_Prob_abandon

Var11: ind_planea

Tabla 4-18. Resumen de significancia de la correlación entre variables categóricas y numéricas

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11
Var1	No	No	No	No	No	No	No	Si	Si	No	No
Var2	No	No	No	No	No	No	No	Si	No	No	Si
Var3	No	No	No	No	No	No	No	No	No	No	No
Var4	No	No	No	No	No	No	Si	No	No	No	No
Var5	No	No	No	No	No	No	Si	No	No	Si	Si
Var6	No	No	No	No	No	No	No	No	No	No	No
Var7	No	No	No	Si	Si	No	No	No	No	No	No
Var8	Si	Si	No	No	No	No	No	No	No	No	No
Var9	Si	No	No	No	No	No	No	No	No	No	No
Var10	No	No	No	No	Si	No	No	No	No	No	Si
Var11	No	Si	No	No	Si	No	No	No	No	Si	No

5. RESULTADOS Y DISCUSIÓN

En este capítulo se presentan los resultados obtenidos del desarrollo de este trabajo y una discusión sobre los puntos a mejorar para encontrar los modelos clasificatorios más precisos para predecir el abandono de estudiantes.

5.1 Selección de variables

Una vez realizada la limpieza y procesamiento de una muestra de 396 alumnos y determinada la variable dependiente E2 - Evento de abandono en el segundo semestre (1: abandona, 0 no abandona), se procedió a seleccionar las variables independientes. Se utilizó el método de filtrado de modelos para seleccionar las variables finales mediante la función Ganancia de Información[3]. De las variables resultantes de esta función organizadas de mayor a menor según el coeficiente, se hizo un filtrado de variables con el apoyo de los valores de correlación previamente calculados. En la

Tabla 5-1 se presenta el resumen de estos valores de Ganancia de Información.

Tabla 5-1. Ganancia de Información de las variables independientes seleccionadas

Variables independientes	Ganancia de Información
Prepa_Encode_Prob_abandon	0.0410
Credito	0.0383
PromPre	0.0195
CatBeca	0.0180
Local	0.0179
CatPase	0.0169
ind_planea	0.0165
CatPrepa_Colapsed	0.0061
EdadIng	0.0037
Sexo_M	0.0031
Semestre_Otonno	0.0014

Se descartaron variables que presentaban una alta correlación, como la que se presentó entre variables

[3] La ganancia de información mide la reducción de la entropía lograda al dividir el conjunto de datos en función de una característica particular.

continuas y sus derivadas categóricas. Tales fueron los casos de los pares de variables EdadIng y CatEdad, Credito y CatCredito, y Beca y CatBeca. De estos pares las variables descartadas fueron CatEdad, CatCredito y Beca.

5.2 Selección de parámetros del objeto AutoML

Luego de separar los datos en un 80 % para entrenamiento y un 20 % para pruebas se creó un objeto AutoML de la clase H2O encargado de la creación y entrenamiento de los modelos. Los parámetros escogidos durante la creación del objeto AutoML se muestran en la Tabla 5-2.

Tabla 5-2. Parámetros de entrada del objeto AutoML

Parámetro	Valor
max_models	25
balance_classes	True
class_sampling_factors	[1., 4.]
nfolds	5
exclude_algos	["XGBoost"]
sort_metric	AUC
seed	42

El parámetro *max_models* que define la cantidad de modelos a crear y entrenar es de 25, aunque se crearon 27 modelos (ver Apéndice 7.3), dos de ellos de tipo *Stacked Ensemble*, ya que los de este tipo no se incluyen en el conteo. Se decidió que se balancearan las clases mediante el parámetro *balance_classes* ya que en la muestra se cuentan con 355 estudiantes con valor cero en la columna que representa la variable dependiente E2 y solo 42 estudiantes con valor uno, lo que representa un desbalance significativo. El parámetro *class_sampling_factors* define la razón de balanceo por clase. La clase 0 se mantiene igual y para la clase 1 se aumentará 4 veces el número de muestras, resultando en 168 muestras con valor 1 en la clase dependiente. El parámetro *nfolds* define si se aplicará validación cruzada o no. El valor escogido define el número de pliegues, en este caso cinco. El algoritmo XGBoost no funciona en el sistema operativo *Windows,* por lo que se excluyó directamente con el parámetro *exclude_algos*. Finalmente, el parámetro *seed* define la semilla aleatoria para la creación de los modelos.

5.3 Selección de modelos

Luego del proceso de entrenamiento, el cual incluyó validación cruzada, se seleccionaron los mejores modelos de la etapa de pruebas. Para la comparación se utilizaron las métricas de discriminación Sensibilidad, Especificidad, AUC y Puntuación F1. Los valores de estas métricas se compararon contra los que arrojó un modelo de Regresión Logística. El resumen de esta comparación se muestra en la

Tabla 5-3, donde se observa que los modelos seleccionados arrojan mejores valores en todas las métricas

en comparación con el modelo de Regresión Logística.

Tabla 5-3. Métricas de discriminación de los modelos seleccionados y su comparación contra un modelo de Regresión Logística

Algoritmo	Modelo	Sensibilidad	Especificidad	AUC	Puntuación F1
GBM	GBM_1_AutoML_1	0.643	0.851	0.760	0.546
GBM	GBM_grid_1_AutoML_1_model_6	0.5	0.910	0.754	0.518
Stacked Ensemble	StackedEnsemble_AllModels_1 _AutoML_1	0.643	0.791	0.761	0.486
LR	GLM_1_AutoML_1	0.5	0.761	0.596	0.378

Se realizó una comparación adicional tomando en cuenta la Precisión (Accuracy) y el Error Cuadrático Medio (MSE por sus siglas en inglés) de los modelos en el conjunto de prueba, lo que se muestra en la Tabla 5-4.

Tabla 5-4. Métricas de Precisión y Error Cuadrático Medio y su comparación contra un modelo de Regresión Logística

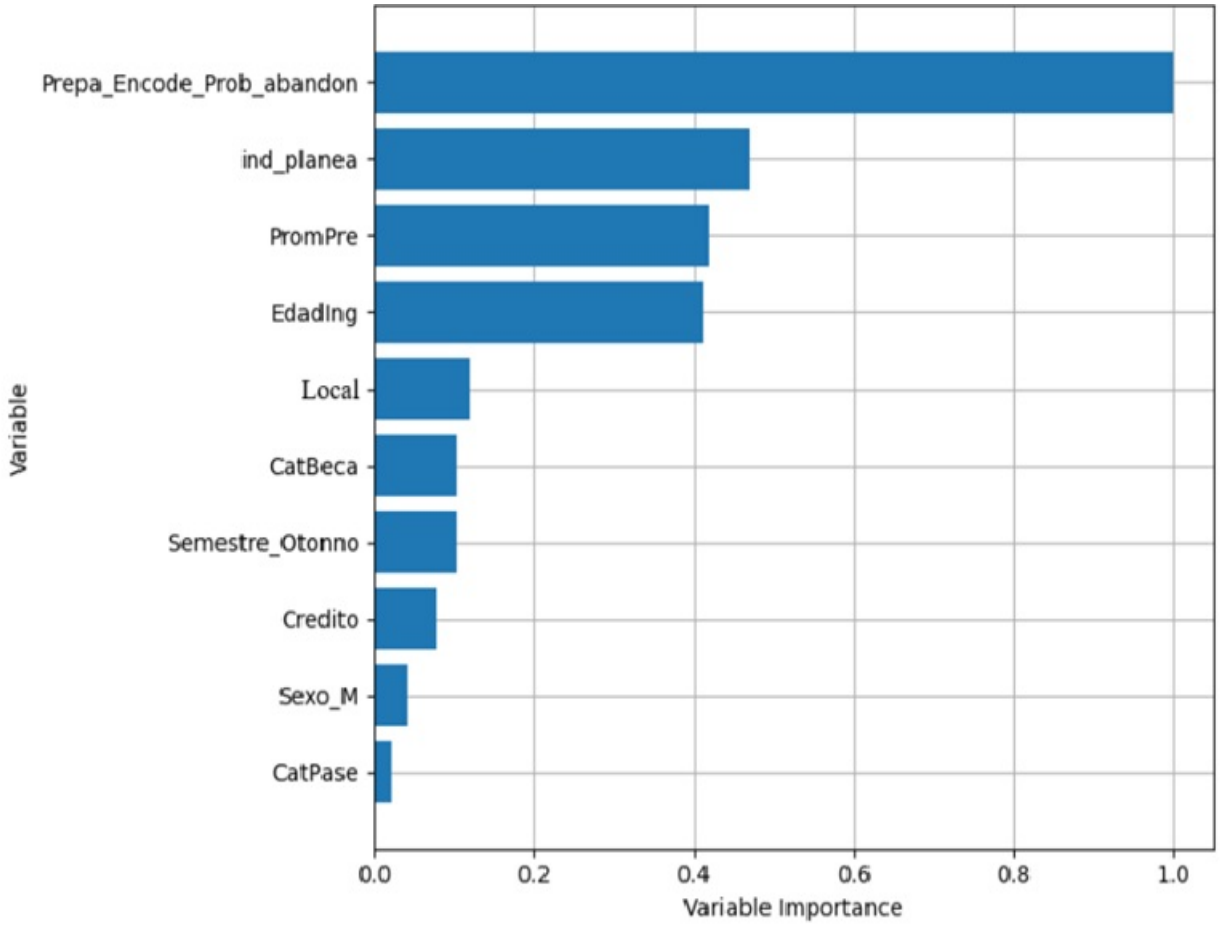
Algoritmo	Modelo	Precisión	Error Cuadrático Medio
GBM	GBM_1_AutoML_1	0.864	0.137
GBM	GBM_grid_1_AutoML_1_model_6	0.851	0.137
Stacked Ensemble	StackedEnsemble_AllModels_1 _AutoML_1	0.827	0.147
LR	GLM_1_AutoML_1	0.827	0.149

Los dos modelos de Máquina de Potenciación del Gradiente (GBM) y el de Conjunto de modelos (*Stacked Ensemble*) seleccionados tienen menor MSE y son más precisos con respecto al modelo de Regresión Logística, con lo que se logra el objetivo general de este trabajo.

5.4 Interpretación del modelo GBM_1_AutoML_1

De los tres modelos seleccionados el mejor fue el GBM_1_AutoML_1. Las variables más importantes[4] para este modelo fueron Probabilidad histórica promedio de abandono de alumnos de una misma preparatoria en la universidad estudiada (Prepa_Encode_Prob_abandon), Índice de calidad de preparatoria en la prueba PLANEA 2015 (ind_planea), Promedio de preparatoria (PromPre), Categoría de beca (CatBeca), Local o foráneo (Local) y Edad de ingreso (EdadIng), lo que se muestra en la Figura 5-1. El resto de variables seleccionadas también arrojaron un nivel de relevancia que, aunque no muy alto, igual es de importancia para los modelos, a excepción de la variable CatPrepa_Colapsed con un valor de importancia de cero.

Figura 5-1. Importancia de variables según modelo GBM_1_AutoML_1



5.5 Discusión

[4] La importancia de la variable se determina calculando la influencia relativa de cada variable: si esa variable fue seleccionada para dividirse durante el proceso de construcción del árbol y cuánto mejoró (disminuyó) el error cuadrático (en todos los árboles) como resultado. El error cuadrático de cada nodo individual es la reducción de la varianza del valor de respuesta dentro de ese nodo. Luego de obtenidos estos valores para cada variable H2O los escala entre 0 y 1.

El modelo seleccionado presentó la mayor precisión con un valor de 86.4 %, lo que representa su capacidad de clasificar tanto los alumnos que van a abandonar la carrera como los que no lo harán. Aunque este valor representa solo una ligera mejoría en comparación con el modelo de Regresión Logística, el cual arrojó una precisión del 82.7 %, se analizaron otras métricas que miden la capacidad del modelo de identificar correctamente la tasa de verdaderos positivos y la tasa de verdaderos negativos. Estas métricas son la Sensibilidad y Especificidad, las cuales son calculadas en base a los valores de Verdaderos Negativos (TN por sus siglas en inglés), Falsos Positivos, Falsos Negativos y Verdaderos Positivos (TP por sus siglas en inglés) los cuales se muestran en la Tabla 5-5.

Tabla 5-5. Valores de TN, FP, FN y TP de los modelos analizados

Modelo	TN	FP	FN	TP
GBM_1_AutoML_1	57	10	5	9
GLM_1_AutoML_1	51	16	7	7

El modelo GBM_1_AutoML_1 presentó una Especificidad del 85.1 %, mayor a la del modelo de LR que fue de 76.1 %. Con esta métrica se mide la capacidad del modelo de identificar correctamente los alumnos que no van a abandonar la carrera (los Verdaderos Negativos), con lo que la escuela se ahorraría recursos y esfuerzos que se estarían desperdiciando si se emplearan en la atención de estudiantes que representen Falsos Positivos, o sea que sean clasificados como posibles casos de abandono cuando en realidad no lo son. En el caso de la Sensibilidad el modelo seleccionado presentó un valor de 64.3 % mientras que el modelo de LR obtuvo solo el 50 %. Esta métrica se centra en los Falsos Negativos, que representan a los estudiantes que sí abandonaron la carrera, pero los modelos predijeron que no lo harían, lo que representaría uno de los peores escenarios posibles para el objetivo del uso de este tipo de modelos de Aprendizaje Automático, el cual es disminuir el abandono de estudiantes de la carrera mediante la correcta identificación de alumnos en riesgo de deserción. Aunque el modelo seleccionado mejora considerablemente al de LR, esto plantea una posible área de mejora para el desarrollo de modelos en trabajos futuros, que no solo se centren en aumentar la precisión sino también en la mejora de la Sensibilidad mediante la disminución de la tasa de Falsos Negativos.

En general este modelo mejora las métricas de AUC, Error Cuadrático Medio, Precisión, Sensibilidad, Especificidad y Puntuación F1, en comparación con el modelo de Regresión Logística. La Sensibilidad es necesario mejorarla, lo que también se resalta en el trabajo de González-Jiménez, lo que denota la importancia de esta métrica para este tipo de modelos. Las variables predictoras EdadIng, PromPre, Prepa_Encode_Prob_abandon, ind_planea, Local, CatPrepa_Colapsed y Sexo_M se presentan como variables comunes en este tipo de estudios, ya que se identificó su uso en los trabajos del ITESO de González-Jiménez [1] y Fuentes-García [3], aunque su nivel de significancia varió según el tipo de estudio.

6. CONCLUSIONES

En este capítulo se presentan las conclusiones y trabajo futuro en relación a encontrar los modelos clasificatorios más precisos para determinar los alumnos de ISC que abandonarán después del primer semestre.

6.1 Conclusiones

Mediante el uso de la librería H2O de Python se crearon 27 modelos de Aprendizaje Automático para la predicción de abandono de alumnos del programa ISC después del primer semestre. Se seleccionaron dos de Máquina de Potenciación del Gradiente y uno de *Stacked Ensemble* como los mejores modelos basándose en las métricas de discriminación Sensibilidad, Especificidad, AUC y Puntuación F1, además del Error Cuadrático Medio y la Precisión, y el modelo GBM_1_AutoML_1 destacó sobre los otros dos.

Se encontraron las variables EdadIng, PromPre, Prepa_Encode_Prob_abandon, ind_planea, Local, CatPrepa_Colapsed y Sexo_M como variables en común con trabajos anteriores del ITESO para la predicción de abandono. Estas variables junto con Credito, CatBeca, CatPase y Semestre_Otonno fueron las variables independientes más importantes para este modelo de tipo GBM.

Este modelo presentó una Sensibilidad de 64.3 % y una Especificidad de 85.1 %, además de una Precisión del 86.4 %, lo que supera los valores que estas métricas arrojaron en el modelo de Regresión Logística, que fueron del 50 %, 76.1 % y 82.7 % respectivamente, con lo que se cumple el objetivo general de este trabajo: encontrar los modelos clasificatorios más precisos para determinar, en el momento del ingreso a la universidad, los alumnos de ISC que abandonarán después del primer semestre.

Este trabajo es de relevancia para el ITESO ya que dicha institución no contaba con modelos avanzados de AA para la predicción de abandono a nivel de programa de estudio después del primer semestre, que es el momento con mayor tasa de abandono para el programa de ISC. No obstante, el mismo presenta las limitantes que la librería H2O no cuenta con las herramientas para el cálculo de métricas de bondad de ajuste para todos los tipos de modelos que puede generar, además de que tampoco permite el cálculo de métricas de calibración de los modelos de forma directa, por lo que el cálculo de estas métricas se recomienda para trabajos futuros.

6.2 Trabajo Futuro

Para profundizar en la búsqueda de los modelos de AA más eficientes para la detección de abandono de la carrera de ISC se ofrecen las siguientes recomendaciones:

1. Ampliar el conjunto de datos con la información de nuevas generaciones de estudiantes de la

carrera ISC, ya que más datos contribuyen a un mejor proceso de entrenamiento de los modelos.

2. Evaluar la calibración de los modelos ya que H2O no posee funcionalidades para obtener estas métricas directamente.
3. Valorar la incorporación de datos adicionales de cada estudiante. La recopilación de información de estudiantes relacionada a los LMS (*Learning Management Systems*), tal como el tiempo de uso de los portales de la biblioteca y canvas, ya que la importancia de estos datos como variables para predecir *dropout* se resalta en varios estudios escolares.
4. Elaborar un modelo *Stacked Ensemble* personalizado y probar distintas combinaciones de modelos bases fuertes que incluya al menos un modelo GBM, para evaluar si mejoran las métricas del modelo en comparación con el que se crea mediante la función H2OAutoML, la cual no permite personalizar los modelos base que utiliza.

7.APÉNDICES

7.1 Correlaciones de punto biserial

Figura 7-1. Correlaciones de variable CatBeca contra numéricas

```
Correlación de punto biserial entre variables CatBeca - EdadIng y correlación de Pearson
SignificanceResult(statistic=-0.015082993455564127, pvalue=0.7647752519275469)
'Los grupos de CatBeca tienen iguales medias que EdadIng'
'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatBeca - PromPre y correlación de Pearson
SignificanceResult(statistic=0.4268467810476121, pvalue=5.76134961123419e-19)
'Los grupos de CatBeca NO tienen iguales medias que PromPre.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatBeca - Credito y correlación de Pearson
SignificanceResult(statistic=0.5267250380735039, pvalue=1.1961063531047947e-29)
'Los grupos de CatBeca NO tienen iguales medias que Credito.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatBeca - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=-0.18507109046524703, pvalue=0.00021284151021109643)
'Los grupos de CatBeca NO tienen iguales medias que Prepa_Encode_Prob_abandon.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatBeca - ind_planea y correlación de Pearson
SignificanceResult(statistic=0.06213451672503041, pvalue=0.21729794238695233)
'Los grupos de CatBeca tienen iguales medias que ind_planea'
'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
```

Figura 7-2. Correlaciones de variable CatPase contra numéricas

```
----------
Correlación de punto biserial entre variables CatPase - EdadIng y correlación de Pearson
SignificanceResult(statistic=-0.07426534840608766, pvalue=0.1401513033232695)
'Los grupos de CatPase tienen iguales medias que EdadIng'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPase - PromPre y correlación de Pearson
SignificanceResult(statistic=0.43899184377979633, pvalue=4.379250216749746e-20)
'Los grupos de CatPase NO tienen iguales medias que PromPre.'

'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPase - Credito y correlación de Pearson
SignificanceResult(statistic=0.03399110980555205, pvalue=0.5000117792130483)
'Los grupos de CatPase tienen iguales medias que Credito'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPase - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=-0.2154626483215632, pvalue=1.5253005099646975e-05)
'Los grupos de CatPase NO tienen iguales medias que Prepa_Encode_Prob_abandon.'

'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPase - ind_planea y correlación de Pearson
SignificanceResult(statistic=0.25763734121496107, pvalue=2.006533943854485e-07)
'Los grupos de CatPase NO tienen iguales medias que ind_planea.'

'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
```

Figura 7-3. Correlaciones de variable Sexo_M contra numéricas

```
----------
Correlación de punto biserial entre variables Sexo_M - EdadIng y correlación de Pearson
SignificanceResult(statistic=0.05077708994849102, pvalue=0.3134989022078915)
'Los grupos de Sexo_M tienen iguales medias que EdadIng'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Sexo_M - PromPre y correlación de Pearson
SignificanceResult(statistic=-0.23485984457499987, pvalue=2.299674126924728e-06)
'Los grupos de Sexo_M NO tienen iguales medias que PromPre.'

'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Sexo_M - Credito y correlación de Pearson
SignificanceResult(statistic=-0.006618165364690946, pvalue=0.895549903687011)
'Los grupos de Sexo_M tienen iguales medias que Credito'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Sexo_M - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=0.09554868438140444, pvalue=0.05746821384746236)
'Los grupos de Sexo_M tienen iguales medias que Prepa_Encode_Prob_abandon'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Sexo_M - ind_planea y correlación de Pearson
SignificanceResult(statistic=-0.09195598495811472, pvalue=0.0675508862712998)
'Los grupos de Sexo_M tienen iguales medias que ind_planea'

'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
```

Figura 7-4. Correlaciones de variable Semestre_Otonno contra numéricas

```
----------
Correlación de punto biserial entre variables Semestre_Otonno - EdadIng y correlación de Pearson
SignificanceResult(statistic=-0.2819593133499204, pvalue=1.133531517207987e-08)
'Los grupos de Semestre_Otonno NO tienen iguales medias que EdadIng.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Semestre_Otonno - PromPre y correlación de Pearson
SignificanceResult(statistic=0.18499363839909921, pvalue=0.0002141691449832458)
'Los grupos de Semestre_Otonno NO tienen iguales medias que PromPre.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Semestre_Otonno - Credito y correlación de Pearson
SignificanceResult(statistic=-0.13209311218037956, pvalue=0.008492079673015119)
'Los grupos de Semestre_Otonno NO tienen iguales medias que Credito.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Semestre_Otonno - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=-0.08973824600795924, pvalue=0.07446856656062464)
'Los grupos de Semestre_Otonno tienen iguales medias que Prepa_Encode_Prob_abandon'
'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables Semestre_Otonno - ind_planea y correlación de Pearson
SignificanceResult(statistic=0.1859027585995267, pvalue=0.00019906260793464844)
'Los grupos de Semestre_Otonno NO tienen iguales medias que ind_planea.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
```

Figura 7-5. Correlaciones de variable CatPrepa_Colapsed contra numéricas

```
----------
Correlación de punto biserial entre variables CatPrepa_Colapsed - EdadIng y correlación de Pearson
SignificanceResult(statistic=-0.284665251579882, pvalue=8.088875147121537e-09)
'Los grupos de CatPrepa_Colapsed NO tienen iguales medias que EdadIng.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPrepa_Colapsed - PromPre y correlación de Pearson
SignificanceResult(statistic=0.13157551029572373, pvalue=0.00875569618117305)
'Los grupos de CatPrepa_Colapsed NO tienen iguales medias que PromPre.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPrepa_Colapsed - Credito y correlación de Pearson
SignificanceResult(statistic=0.022556140554972938, pvalue=0.6545144154475038)
'Los grupos de CatPrepa_Colapsed tienen iguales medias que Credito'
'Aceptamos la H0 y decimos que NO existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPrepa_Colapsed - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=-0.40996622594441096, pvalue=1.748874011239042e-17)
'Los grupos de CatPrepa_Colapsed NO tienen iguales medias que Prepa_Encode_Prob_abandon.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
----------
Correlación de punto biserial entre variables CatPrepa_Colapsed - ind_planea y correlación de Pearson
SignificanceResult(statistic=0.6086755762755862, pvalue=1.6663059431966332e-41)
'Los grupos de CatPrepa_Colapsed NO tienen iguales medias que ind_planea.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa (H1)'
```

Figura 7-6. Correlaciones de variable Local contra numéricas

```
Correlación de punto biserial entre variables Local - EdadIng y correlación de Pearson
SignificanceResult(statistic=np.float64(0.00046291080142070104), pvalue=np.float64(0.9926733770904836))
'Los grupos de Local tienen iguales medias de EdadIng'
'Aceptamos la H0 y decimos que NO existe una correlación significativa baja (H1)'
----------
Correlación de punto biserial entre variables Local - PromPre y correlación de Pearson
SignificanceResult(statistic=np.float64(-0.011458783348183774), pvalue=np.float64(0.8201799426426605))
'Los grupos de Local tienen iguales medias de PromPre'
'Aceptamos la H0 y decimos que NO existe una correlación significativa baja (H1)'
----------
Correlación de punto biserial entre variables Local - Credito y correlación de Pearson
SignificanceResult(statistic=np.float64(0.1284544127528037), pvalue=np.float64(0.010505270673395456))
'Los grupos de Local NO tienen iguales medias de Credito.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa baja (H1)'
----------
Correlación de punto biserial entre variables Local - Prepa_Encode_Prob_abandon y correlación de Pearson
SignificanceResult(statistic=np.float64(-0.27811544316757153), pvalue=np.float64(1.8193737065391046e-08))
'Los grupos de Local NO tienen iguales medias de Prepa_Encode_Prob_abandon.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa baja (H1)'
----------
Correlación de punto biserial entre variables Local - ind_planea y correlación de Pearson
SignificanceResult(statistic=np.float64(0.15351782068878347), pvalue=np.float64(0.002187728414575554))
'Los grupos de Local NO tienen iguales medias de ind_planea.'
'Rechazamos la H0 y decimos que Sí existe una correlación significativa baja (H1)'
----------
```

7.2 Comparación de los valores de la matriz de phi-k contra la matriz de significancia

Tabla 7-1. Comparación de valores de phi-k y significancia de la variable CatBeca

	CatBeca	significancia	significante
PromPre	0.558414	7.353858	Yes
Credito	0.555945	10.185921	Yes

Tabla 7-2. Comparación de valores de phi-k y significancia de la variable CatPase

	CatPase	significancia	significante
PromPre	0.721215	10.072211	Yes
ind_planea	0.432739	5.407172	Yes

Tabla 7-3. Comparación de valores de phi-k y significancia de la variable Semestre_Otonno

	Semestre_Otonno	significancia	significante
EdadIng	0.499904	6.615845	Yes

Tabla 7-4. Comparación de valores de phi-k y significancia de la variable CatPrepa_Colapsed

	CatPrepa_Colapsed	significancia	significante
EdadIng	0.452938	5.90892	Yes
Prepa_Encode_Prob_abandon	0.558452	10.121209	Yes
ind_planea	0.898401	13.558659	Yes

Tabla 7-5. Comparación de valores de phi-k y significancia de la variable EdadIng

	EdadIng	significancia	significante
Semestre_Otonno	0.499904	6.630129	Yes
CatPrepa_Colapsed	0.452938	5.90892	Yes

Tabla 7-6. Comparación de valores de phi-k y significancia de la variable PromPre

	PromPre	significancia	significante
CatBeca	0.558414	7.353858	Yes
CatPase	0.721215	10.072211	Yes

Tabla 7-7. Comparación de valores de phi-k y significancia de la variable Credito

	Credito	significancia	significante
CatBeca	0.555945	10.185921	Yes

Tabla 7-8. Comparación de valores de phi-k y significancia de la variable Prepa_Encode_Prob_abandon

	Prepa_Encode_Prob_abandon	significancia	significante
CatPrepa_Colapsed	0.599007	10.969665	Yes
ind_planea	0.524239	11.783771	Yes

Tabla 7-9. Comparación de valores de phi-k y significancia de la variable ind_planea

	ind_planea	significancia	significante
CatPase	0.432739	5.381197	Yes
CatPrepa_Colapsed	0.898401	13.558659	Yes
Prepa_Encode_Prob_abandon	0.516175	11.480869	Yes

7.3 Listado de modelos generados por H2O

Tabla 7-10. Distribución de modelos creados por tipo de Algoritmo

Algoritmo	Modelo
Stacked Ensemble	StackedEnsemble_AllModels_1_AutoML_1 StackedEnsemble_BestOfFamily_1_AutoML_1
Deep Learning	DeepLearning_grid_2_AutoML_1_model_1 DeepLearning_grid_3_AutoML_1_model_3 DeepLearning_1_AutoML_1 DeepLearning_grid_1_AutoML_1_model_2 DeepLearning_grid_2_AutoML_1_model_2 DeepLearning_grid_2_AutoML_1_model_3 DeepLearning_grid_1_AutoML_1_model_3 DeepLearning_grid_3_AutoML_1_model_1 DeepLearning_grid_1_AutoML_1_model_1 DeepLearning_grid_3_AutoML_1_model_2
XRT	XRT_1_AutoML_1 DRF
GBM	GBM_1_AutoML_1 GBM GBM_grid_1_AutoML_1_model_6 GBM_4_AutoML_1 GBM GBM_grid_1_AutoML_1_model_1 GBM GBM_grid_1_AutoML_1_model_4 GBM GBM_grid_1_AutoML_1_model_7 GBM GBM_grid_1_AutoML_1_model_5 GBM GBM_grid_1_AutoML_1_model_2 GBM GBM_5_AutoML_1 GBM GBM_2_AutoML_1 GBM_3_AutoML_1 GBM_grid_1_AutoML_1_model_3
DRF	DRF_1_AutoML_1
GLM	GLM_1_AutoML_1

8.BIBLIOGRAFÍA

[1] J. N. González-Jiménez, “Predicción de abandono y egreso tardío en alumnos de ingeniería de la universidad ITESO,” Trabajo de obtención de grado, Maestría en Informática Aplicada, ITESO, Tlaquepaque, Jalisco, 2023.

[2] E. R. Ramos-Rocha, “Estimaciones de deserción, cambio de carrera y egreso para la Ingeniería en Sistemas Computacionales del ITESO,” Trabajo de obtención de grado, Maestría en Sistemas Computacionales, ITESO, Tlaquepaque, Jalisco, 2024.

[3] S. Fuentes-García, “MODELO PREDICTIVO DE RIESGOS EN COMPETENCIA PARA EVENTOS DE HISTORIA DE VIDA DE ALUMNOS DE UNA INGENIERÍA EN EL ITESO,” Trabajo de obtención de grado, Maestría en Informática Aplicada, ITESO, Tlaquepaque, Jalisco, 2025. Accessed: May 01, 2025. [Online]. Available: https://hdl.handle.net/11117/11454

[4] G. Hernández-Chávez, “Aplicación de Bosques de Supervivencia Aleatorios a la Predicción de Abandono Universitario,” Tesis Doctoral, Universidad de Guadalajara, Guadalajara, México, 2025.

[5] M. Fuentes Peralta and C. Chávez Preisler, “Abandono escolar: circunstancias que caracterizan el momento del abandono escolar temprano,” *Revista INTEREDU*, vol. 1, no. 1, pp. 74–93, Jan. 2021, doi: 10.32735/s2735-65232019000181.

[6] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATION OF KNOWLEDGE DISCOVERY AND DATA MINING*, 2000, pp. 29–39. Accessed: Nov. 24, 2024. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133

[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] C. Mason, J. Twomey, D. Wright, and L. Whitman, “Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression,” *Res High Educ*, vol. 59, no. 3, pp. 382–400, May 2018, doi: 10.1007/s11162-017-9473-z.

[9] M. Mezzini, G. Bonavolontà, and F. Agrusti, “PREDICTING UNIVERSITY DROPOUT BY USING CONVOLUTIONAL NEURAL NETWORKS,” Mar. 2019, pp. 9155–9163. doi: 10.21125/inted.2019.2274.

[10] Á. Kocsis and G. Molnár, “Factors influencing academic performance and dropout rates in higher education,” *Oxf Rev Educ*, 2024, doi: 10.1080/03054985.2024.2316616.

[11] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, “Data Mining and Machine Learning Retention Models in Higher Education,” *J Coll Stud Ret*, vol. 25, no. 1, pp. 51–75, May 2023, doi: 10.1177/1521025120964920.

[12] A. Kuz and R. Morales, “Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México,” *Education in the Knowledge Society (EKS)*, vol. 24, p. e30080, Jun. 2023, doi: 10.14201/eks.30080.

[13] B. Andriy Burkov, “The Hundred-Page Machine Learning.”

[14] “H2O AutoML: Automatic Machine Learning.” Accessed: Oct. 23, 2024. [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

[15] S. Sperandei, “Understanding logistic regression analysis,” *Biochem Med (Zagreb)*, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.
[16] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach Learn*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.
[17] “H2O AutoML: Performance and Prediction.” Accessed: Nov. 24, 2024. [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/performance-and-prediction.html#mse-mean-squared-error
[18] A. Swift, R. Heale, and A. Twycross, “What are sensitivity and specificity?,” *Evid Based Nurs*, vol. 23, no. 1, Jan. 2020, doi: 10.1136/ebnurs-2019-103225.