APRENDIZAJE SUPERVISADO EN EL DESARROLLO DE UN CLASIFICADOR DE INCUMPLIMIENTO CREDITICIO CON ENFOQUE EN LA AUTOMATIZACIÓN PARA EL CONJUNTO DE DATOS DE LOS CRÉDITOS EDUCATIVOS OTORGADOS A LOS ESTUDIANTES DE LA UNIVERSIDAD DE LA SABANA

Norberto A. Acero Benitez

Trabajo de grado profundización: Proyecto Aplicado.

Tutor: Felix Vivian Mohr

Facultad de Ingeniería

Maestría en Analítica Aplicada

Director Académico: Dr. Gonzalo Mejía Delgadillo

Resumen

La disponibilidad de datos y herramientas para el análisis exhaustivo ha abierto nuevas posibilidades en áreas donde el costo o la complejidad anteriormente limitaban el acceso. Un ejemplo de ello son los modelos de aprendizaje automático, los cuales brindan a las entidades no financieras herramientas competitivas para la predicción del riesgo crediticio en comparación con los modelos de puntuación tradicionales. Incluso, muchas entidades bancarias utilizan estos enfoques para optimizar sus sistemas de puntuación. En el caso de las entidades educativas que proporcionan su propio capital para respaldar a los estudiantes; estos modelos representan una excelente alternativa al azar frente al análisis de crédito. Este documento se desarrolló con el objetivo de conducir a la Universidad de La Sabana por la creación de una metodología a la medida de sus datos que, con ayuda de las técnicas de aprendizaje de maquina y la optimización automática de modelos determine si los datos de los estudiantes permiten la predicción de impago de cartera. Se realizó un análisis exploratorio de los datos sobre las 15 características mejor evaluadas de los estudiantes en los sistemas de la universidad, se diseñaron 2 atributos adicionales para la clasificación de estos y se incluyeron en análisis de correspondencias múltiples para los atributos categóricos. Se entrenaron con ayuda de una herramienta automatización de modelos de aprendizaje de máquina llamada NaiveAutoML y se utilizó la métrica AUC-ROC para la validación de los modelos. Con el ensamble de un algoritmo de aumento de gradiente por histograma se presentó una métrica AUC-ROC de 0.7648 sobre los datos entrenados y 0.5596 sobre los datos de prueba, lo cual indica que es pertinente contar con un modelo de clasificación de impago de cartera aun sin tener una gran cantidad de datos financieros de los estudiantes para tomar la decisión de crédito.

Palabras clave: aprendizaje de máquina – automatización – árboles de decisión – clasificación – puntaje de crédito.

Contenido

Introducción

La Universidad de La Sabana tiene como principal misión la búsqueda y preservación de la verdad a través de la investigación y la enseñanza [01]. Todos los pilares fundamentales de su proyecto educativo convergen en la realización de este objetivo, permitiendo que los estudiantes participen plenamente en su oferta académica. Una tarea de vital importancia es asegurar el acceso a esta oferta académica y garantizar la continuidad de los estudiantes en ella. Para lograrlo, ha implementado opciones de financiamiento que respaldan las necesidades económicas de las familias brindando el apoyo a aquellos estudiantes que enfrentan dificultades con el ingreso al mercado financiero con entidades de crédito tradicionales, ya sea debido a la complejidad de los requisitos de acceso, a las demoras en la aprobación del crédito o a la falta de confianza de las familias en estas entidades [02], siendo este último el diferencial para escoger a la Universidad sobre estas.

Esto ha llevado a la Universidad a exponerse a un riesgo crediticio que implica la posibilidad de que incurra en pérdidas y vea reducido el valor de sus activos [03] debido a los incumplimientos por parte de los deudores respecto a los términos acordados en las políticas de crédito de las líneas ofrecidas en el portafolio de la Universidad. Contra ello, la Universidad busca protegerse mediante el uso del pagaré como título valor y la figura del deudor solidario o codeudor, estas son garantías convencionales comúnmente empleadas en los productos de crédito en Colombia que ofrecen un respaldo legalmente eficaz [04] además de los documentos requisito que son estudiados por los analistas de crédito al momento de la aprobación. Sin embargo, el proceso de cobranza y jurídico necesario para ejecutar estos títulos valores resulta oneroso, costoso y en muchas ocasiones no se logra recuperar los valores adeudados.

A diferencia de las instituciones educativas, los bancos han implementado procesos de puntuación crediticia al momento de la solicitud que les permiten estimar la

probabilidad de impago en función de calificaciones internas de los atributos del cliente [05]. Estos métodos surgieron inicialmente en respuesta a las regulaciones bancarias establecidas en el marco del segundo acuerdo de Basilea por el Comité de Supervisión Bancaria [06] que tenía como propósito mejorar la gestión del riesgo crediticio y fortalecer la estabilidad financiera global en su momento. A pesar de que estas regulaciones surgieron como medidas locales en Europa, su adopción a nivel mundial fue impulsada por las normas internacionales de información financiera (IFRS). Con esta difusión, los bancos alrededor del globo comenzaron a apreciar las metodologías del análisis de datos para la construcción de estos modelos a tal punto que integraron procesos completos bajo su estructura organizacional con visión a largo plazo dentro de la propia estrategia corporativa [05] y un gran esfuerzo económico para la adquisición y mejoramiento de los datos de sus clientes.

Estos costos de infraestructura y compra de largos conjuntos de datos distancian a las instituciones de educación superior de los bancos ya que su enfoque principal no consiste en generar ingresos por intereses sino en llevar a cabo su labor académica. No obstante, surge la posibilidad de desarrollar una metodología que permita crear modelos basados en los datos existentes en la Universidad y que podrían proporcionar un valor adicional en el proceso de evaluación de las solicitudes de crédito por medio del aprendizaje automático supervisado [07]. En este contexto y considerando las tendencias de innovación digital enmarcadas en la transformación estratégica institucional de la Universidad De La Sabana, se plantea este estudio con el objetivo de explorar la viabilidad de construir dicha metodología que, mediante técnicas de aprendizaje automático, pueda categorizar a los estudiantes que solicitan créditos entre vencidos o corrientes, mejorando así el proceso de toma de decisión en la evaluación de crédito.

La sección 2 inicia con la definición del problema y los antecedentes de este, al igual que un marco teórico que acerca a las definiciones propias del análisis de datos. La sección 3, describe la metodología completa desde el análisis de los datos hasta la

construcción del modelo predictivo y un acercamiento a la automatización de la solución. En la sección 4 se presentan los resultados y la evaluación del modelo. Por último, se discuten las conclusiones y las futuras direcciones del proyecto aplicado en la sección 5.

2 Definición del problema

Una mirada al detalle del proceso de análisis de crédito, indica que la Universidad se apoya en los informes de centrales de riesgo para aprobar, o no, el crédito al estudiante quien queda en cabeza de la obligación apoyado por los padres como codeudores; haciendo de esta una deuda de carácter familiar. Estos informes recopilan información relacionada con las obligaciones financieras de las familias que muchas veces son insuficientes o no coinciden con la documentación presentada por el estudiante. Al igual, los formatos de estos reportes no permiten el ingreso de los datos de manera fácil o masiva a los sistemas de información. No obstante, se constituyen como la única herramienta de evaluación a disposición de los analistas de financiación al momento de otorgar el crédito. Ellos se enfocan principalmente en determinar si la capacidad de pago de las familias solicitantes es suficiente para cubrir la cuota de crédito calculada mensualmente.

El problema se centra en la necesidad de mejorar el proceso de evaluación de las solicitudes de crédito para los estudiantes por medio de una herramienta que permita clasificar a los deudores entre vencido y corriente a partir de la poca información que se conoce del estudiante. Esto vislumbra que se trata de un problema de clasificación binaria que espera predecir una clase entre los dos comportamientos para cada estudiante nuevo que solicita crédito con la Universidad con respecto de los datos históricos de los estudiantes que tienen cartera. Formalmente se puede representar el conjunto de datos D como una matriz de $n \; x \; d$, con n filas que representan cada una a un estudiante y d columnas que representan las características de estos estudiantes formando el espacio de

instancias de los atributos X (en las columnas) y también el espacio de las etiquetas Y de cada uno que clasifican las instancias (en las filas) binariamente. Donde:

X son las variables de los datos en función de la predicción.
Y es lo que queremos predecir, en nuestro caso, cualquiera de las dos etiquetas: Corriente y Vencido.

$$D = \begin{array}{|c|ccccc|c|} \hline & X_1 & X_2 & \dots & X_d & Y \\ \hline x_1 & X_{11} & X_{12} & \dots & X_{1d} & y_1 \\ x_2 & X_{21} & X_{22} & \dots & X_{2d} & y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_n & X_{n1} & X_{n2} & \dots & X_{nd} & y_n \\ \hline \end{array}$$

Dado el espacio de instancias X se espera estimar Y, algebraicamente se representa como la función:

$$f(h) = X \rightarrow Y.$$

Esta función $f(h)$ al igual que el conjunto de datos seleccionado D son los principales insumos para la herramienta, que se puede denominar algoritmo en términos de pensamiento computacional [08] [09], y que por medio de la metodología más adelante planteada busca el modelo con la mejor métrica sobre los datos para clasificar a los nuevos estudiantes. En el ámbito del aprendizaje automático supervisado la tarea principal es:

Dado el conjunto de datos entrenado, en nuestro caso $D_{entranamiento}$, donde Y se ha generado por una función desconocida de la forma: $y = f(x)$, se descubrirá una función $f(h)$ que se aproxima a la verdadera función $f(x)$ [10]. Aquí la función $f(h)$ es una hipótesis y el aprendizaje es una búsqueda a través del espacio de las posibles hipótesis para encontrar una función que funcione bien. Para medir la precisión de la hipótesis, le damos un conjunto de ejemplos de prueba, en nuestro caso D_{prueba}, distintos del conjunto de entrenamiento [10] y con esto se comparan los resultados

y se evalúa la hipótesis. Cuando el resultado Y es un valor finito entre un conjunto de valores, el problema se denomina clasificación y se llama binario si solo se tienen dos valores para escoger.

2.1 Antecedentes

Los estudios sobre los métodos de desarrollo de puntuación de crédito por medio del aprendizaje de maquina han sido trabajados durante décadas. De aquí que se tiene una revisión de literatura extensa [11] [12] con un enfoque no solo a la búsqueda de modelos que responden al problema de clasificación, sino que buscan expresar las puntuaciones existentes de los clientes en términos de probabilidades de impago con métodos regresivos e incluso aprendizaje automático no supervisado para encontrar los patrones necesarios y las probabilidades requeridas [13] [14]. De aquí que la revisión bibliográfica de este documento se centra en los clasificadores solamente.

Trabajos como el de Loterman [15] concluyen que los modelos no lineales como los árboles de decisión funcionan mejor para predecir indicadores financieros compuestos como la pérdida ajustada por defecto, que son métricas diseñadas a partir de los datos financieros del cliente. Estos modelos dieron métricas superiores sobre los modelos lineales. Al igual, el trabajo de Loterman enriquece estos modelos con procesos de ingeniería de características para cada instancia cuando no se tiene información completa.

También se resaltan trabajos como el de Pandey [16] que entrevé una metodología diferente, ya que compara conjuntos de datos financieros masivos de diferentes países y sobre ellos entrena diferentes modelos para comparar. En esta búsqueda, el mejor modelo entrenado fue el conocido como las máquinas de aprendizaje extremo que son redes neuronales de propagación hacia adelante [17] y que para

el estudio de Pandey se utilizó una sola capa de nodos ocultos dando una precisión por encima del 90% en promedio entre los datos alemanes y los datos australianos. A pesar de que numerosos métodos de puntuación de crédito han avanzado en el ámbito de la calificación crediticia, los datos empleados en las investigaciones actuales siguen representando solo una pequeña parte de este vasto conjunto de información basada en Big Data. Yang [18] concluye que muchas fuentes de datos, en especial los datos sociales dispersos en el ciberespacio no han sido aprovechados plenamente. Pero su trabajo enriquece este estudio trayendo al tema un tópico importante y es la automatización de las prácticas de aprendizaje de máquina para la búsqueda de la solución.

Partiendo de la revisión bibliográfica se presentan en la tabla [01] los modelos de aprendizaje automático más comunes hasta el 2023 para clasificar binariamente los deudores en el ámbito del análisis de riesgo de crédito:

Tabla 1. Modelos de aprendizaje automático en la revisión de literatura.

Modelo	Referencia
Clasificador Bayesiano	[11] [16]
Clasificador Bayesiano Ingenuo	[11] [16]
Árboles de Decisión	[14] [16]
K Vecinos Cercanos	[16] [19]
Perceptor de multicapas	[16]
Máquinas de Soporte Vectorial	[11] [12] [15] [16] [18] [19]
Redes Neuronales Artificiales	[11] [14] [15] [16] [18] [19]
Bosques Aleatorios	[16] [18]

Sin embargo, son muchas las variaciones dentro de los mismos modelos. Por ejemplo, de los árboles de decisión se desprenden modelos como los bosques aleatorios, el refuerzo por gradiente y el refuerzo por gradiente basado en histograma. Esto se añade a los métodos empleados con el fin de adaptar los modelos a las situaciones particulares de cada uno de los conjuntos de datos.

De la misma forma que se presentan los modelos, se ha encontrado valiosos hallazgos en las metodologías de aprendizaje y análisis de datos que llevan a mejorar las predicciones. Es el caso de Paweł [19] quien plantea una metodología que se basa en el entrenamiento genético por capas mediante el uso de la validación cruzada estratificada ampliando significativamente los rangos de precisión sobre los mismos datos. Es de resaltar que en la mayoría de estas metodologías no escogen un solo modelo de entrada, sino que comparan varios modelos con los mismos datos para, por medio de la métrica, medirlos y tomar la decisión.

Dentro de las mejores técnicas a lo largo de estos años se encuentran las máquinas de soporte vectorial, las redes neuronales artificiales y los árboles de decisión. Sin embargo, los modelos basados en arboles de decisión responde mejor cuando los conjuntos de datos tienen un bajo porcentaje de datos numéricos, como el caso de la Universidad de La Sabana que, cuenta con un mínimo componente de datos financieros, dejando la mayoría de los atributos del estudiante como datos categóricos. Lo anterior da cuenta de que la calidad de los datos es primordial para sacar provecho del modelo y su predicción. Adicional a los modelos, en los antecedentes se pueden apreciar tópicos importantes para tener en cuenta dentro de la metodología como son: la reducción de dimensionalidad, técnicas de potencialización (boosting), el tratamiento de los datos y la valoración de las métricas de desempeño.

2.2 Marco teórico

El marco teórico nos acercará a la comprensión de los principales modelos utilizados en el aprendizaje automático supervisado y los conceptos de análisis de datos necesarios para el desarrollo de la metodología.

2.2.1 Aprendizaje de máquina

Mohri [20] define el aprendizaje de máquina como los métodos computacionales que permiten a un agente utilizar la experiencia para realizar un conjunto de tareas (entre ellas la clasificación) de manera más precisa. Aquí el concepto resalta la experiencia como la información obtenida previamente sobre dicha tarea para que el agente aprenda. Otros autores se acercan a esta definición, incluyendo que las técnicas y métodos basados en datos se encuentran inherentemente relacionados con las dinámicas conceptuales entre el análisis de datos y la estadística.

2.2.2 Aprendizaje automático supervisado

En el aprendizaje supervisado el agente observa dicha experiencia como entradas que son percepciones y salidas que son etiquetas dadas por un profesor que previamente ha emparejado entradas y salidas mapeando una función que le permite al agente aprender. Russell [21] incluye en esta definición que, si la etiqueta es un valor finito, el problema es llamado clasificación y si la etiqueta es un número (como podría ser el valor moratorio) el problema es llamado regresión.

2.2.3 Técnicas de clasificación

Diferentes técnicas para la evaluación de bases de datos de crédito son usadas en el análisis de riesgo crediticio. El **Clasificador Bayesiano**, por ejemplo, es un modelo que utiliza directamente el teorema de Bayes para predecir las clases para una nueva instancia [22]. Se puede representar por un gráfico acíclico directo donde cada nodo representa una variable aleatoria y los bordes representan una dependencia funcional entre las variables. En términos de probabilidad y verosimilitud a priori el teorema de Bayes establece que:

$$P(c_i/x) = \frac{P(x/c_i).P(c_i)}{P(x)}$$

Donde $P(x/c_i)$ es la probabilidad de observar x asumiendo que la verdadera clase es c_i, $P(c_i)$ es la probabilidad de la clase c_i y $P(x)$ es la probabilidad de observar x cuando se da alguna de las clases dadas. En palabras de Zaki [22], la predicción de la clase esencialmente depende de la probabilidad de que la clase tome la probabilidad a priori de x en cuenta.

El **Clasificador Bayesiano Ingenuo** es un clasificador basado en el anterior teorema de Bayes, se le conoce como ingenuo ya que asume que los atributos de una clase son independientes de los otros. Por esto no necesita grandes cantidades de datos para calcular los principales estadísticos necesarios para la clasificación [21].

Los **Árboles de Decisión** son modelos de predicción que alcanzan su decisión realizando una secuencia de pruebas por medio de nodos en ramas que mapean por medio de funciones los vectores de entrada hasta llegar a la clase. En esta técnica los nodos internos son clasificados con una etiqueta individual y una distribución de probabilidad sobre las clases que permite las particiones. Son algoritmos fáciles de interpretar, pero muy propensos al sobreajuste. Los **bosques aleatorios** son algoritmos formados por múltiples arboles de decisión individuales, se entrenan de manera aleatoria y se obtienen los pesos para cada predicción por las observaciones y las muestras de los datos.

K-Vecinos Cercanos se encuentra entre los modelos no paramétricos más usados para clasificación y regresión. Estos no intentan generalizar a partir de los datos de entrenamiento para producir una hipótesis que coincida con ellos, en su lugar los utiliza para determinar la clasificación basado en cada crédito [23]. Por lo general, se utiliza la distancia euclidiana para medir la distancia entre los vecinos $/x - x_i/: \sqrt{\sum}$

$(x, x_i)2$. Este puede ocupar mucha memoria computacionalmente siendo costoso para una implementación ideal. En clasificación, los K-vecinos cercanos pueden calcular la clase con mayor frecuencia para los K más similares vecinos del nuevo solicitante del crédito, según sus propios atributos.

Las **redes neuronales** actualmente juegan un papel importante en el control del riesgo de crédito a nivel mundial. Inspiradas en el funcionamiento de las redes neuronales biológicas, las redes neuronales artificiales contienen neuronas que transmiten información a la cual se aplica una función de activación para generar una salida. Los **perceptores multicapa** son ampliamente utilizados en el área de finanzas para el riesgo de crédito [16]. Son redes neuronales que por medio de una función de activación usa la propagación para activar los nodos de la red neuronal que se llaman perceptores. Pueden tener entre uno o varias capas escondidas para el procesamiento [10]. Las capas permiten a la red aprender sobre la relación entre las capas de entrada y las capas de salida.

Las **máquinas de soporte vectorial** son métodos de clasificación basados en los discriminantes lineales de máximo margen. El objetivo es optimizar el hiperplano que maximiza la brecha entre las clases [24] sobre un margen máximo que actúa como límite de decisión. El hiperplano se define como el conjunto de todos los puntos dados $w^T x = -b$, donde w es un vector dimensional de peso y b es un escalar llamado sesgo [22]. Estos modelos predicen la clase gracias a la función $h(x)$ y de acuerdo con la regla de decisión:

$$y = \begin{matrix} +1 \text{ si } h(x) > 0 \\ -1 \text{ si } h(x) < 0 \end{matrix}$$

2.2.4 Automatización del aprendizaje de máquina

Un ámbito importante en el desarrollo del tema es la automatización del aprendizaje de máquina que puntualmente busca el mejor rendimiento generalizado de los aprendices y los preprocesadores de los datos, generalmente por medio de tuberías (pipelines) que se encargan de la ejecución y la evaluación de estos modelos [25], a continuación, en la metodología se profundiza en esta tecnología.

3 Metodología

El ciclo de desarrollo para el modelaje en este documento se dividió en dos grandes actividades. La primera consiste en la minería de datos desde la descripción del modelo de extracción de los datos hasta el análisis de componentes principales. El segundo consiste en el ensamble del clasificador y la automatización del modelo.

3.1 Minería de los datos

3.1.1 Construcción del conjunto de datos

La Universidad de La Sabana cuenta con un sistema integrado de gestión académica y administrativa que almacena de manera tabular los datos en 3 módulos de información: Académica, Financiera y corporativa. Esta información fluye a un lago de datos de manera estructurada y relacional por medio de un proceso de extracción, transformación y cargue. Con lenguaje de consulta estructurada SQL se seleccionaron los atributos más completos históricamente del sistema. Con SQL se anidaron 7 tablas como dimensiones alrededor de una tabla de hecho que contiene la información de cartera del estudiante. Creando un modelo relacional que permite futuros accesos a la misma información. La tabla [02] presenta la descripción de las tablas del modelo relacional.

Tabla 2. Tablas en el modelo relacional del lago de datos de la Universidad de La Sabana.

Nombre	Tipo	Descripción
Cartera	Hechos	Atributos financieros del estudiante como valor de la cartera, número de días vencidos, línea de crédito, etc.
Admisiones	Dimensión	Registros de admisión como el tipo de admisión.
Asesorías	Dimensión	Asesorías académicas al estudiante.
Académica	Dimensión	Promedios académicos de notas del estudiante.
Beneficios	Dimensión	Beneficios económicos (diferentes a becas) otorgados.
Demográfica	Dimensión	Datos demográficos como género, estado civil, grupo étnico, estrato socioeconómico, etc.
Investigación	Dimensión	Participaciones en investigación o semilleros ofrecidos en la Universidad.
Movilidades	Dimensión	Movilidades internacionales o semestres cursados en el exterior.

De las relaciones entre las tablas y con ayuda de las cláusulas de Unión en SQL (Join) se crea una sola tabla en combinación de los atributos de los estudiantes necesarios para incluir en el modelo. Se tomó en consideración atributos de calidad para los datos como son completitud, validación, oportunidad y adecuación para el alcance de este modelo. Se exporta con formato CSV para el cargue en las herramientas de análisis. La tabla [03] muestra el total de los atributos seleccionados para los estudiantes.

Tabla 3. Atributos seleccionados para el conjunto de datos.

Atributo	Tipo	Descripción
Valor	Numérico	Monto adeudado en pesos del estudiante.
Días	Numérico	Días en mora de la cartera del estudiante.
Tipo de admisión	Categórico	Categoría interna de ingreso a la Universidad: estándar, por convenio, transferencia, etc.
Género	Categórico	Categorización del género del estudiante entre “hombre” o “mujer”.
Estado Civil	Categórico	Estado civil del estudiante: “casado”, “soltero”, “matrimonio”, etc.

Grupo Étnico	Categórico	Grupo étnico al que pertenece el estudiante: “Indígena”, “negritud”, etc.
Estrato	Categórico	Clasificación de estratificación socioeconómica del estudiante.
Fecha de nacimiento	Fecha	Fecha de nacimiento del estudiante.
Nivel educativo de los padres	Categórico	Nivel máximo de escolaridad alcanzado entre los dos padres.
Doble Programa	Booleano	Responde a la pregunta: ¿estudió doble programa?
Investigó	Booleano	Responde a la pregunta: ¿Ha realizado investigación?
Viajó Movilidad	Booleano	Responde a la pregunta: ¿Ha viajado en proceso de movilidad?
Beneficiario Pat	Booleano	Responde a la pregunta: ¿Ha trabajado bajo el programa PAT?
Nota Promedio	Numérico	Promedio académico ponderado del estudiante.

Se obtiene un conjunto inicial de datos con 2.944 instancias y 14 atributos. Es importante resaltar que la información obtenida no corresponde a información histórica de la cartera de los créditos, sino que se limita a reflejar la información financiera y académica de los estudiantes en un momento del tiempo. En el momento de este trabajo es a corte del 31 de diciembre de 2022.

3.1.2 Peculiaridades de los datos

Una primera revisión de la data presenta tres retos importantes a la luz de la revisión bibliográfica. El primero es la posibilidad del sobreajuste de los modelos. El segundo es la decisión sobre los datos ausentes, datos duplicados y los datos atípicos que se pueden encontrar en cada una de las columnas de la matriz de datos. Y el tercero es la construcción de la clase y demás características que den valor al modelo.

3.1.2.1 Sobreajuste

Tradicionalmente en la evaluación de modelos de aprendizaje automático se encuentra un problema de sobreajuste a los modelos cuando se selecciona la hipótesis h que intenta predecir $Y: h(x) \in Y$ midiendo dicha predicción con una función de error tal que, $e(y, \ddot{y}) = [y \neq \ddot{y}]$ siendo $\ddot{y}$ la clase predicha diferente de y la clase correcta, pero utilizando todo el conjunto de datos D [27]. Cuando dicha hipótesis tiene un error dentro del conjunto E_{in} menor que el error fuera E_{out}, indica que E_{in} ya no es una buena guía para el aprendizaje.

Esto se aborda por medio de la partición del conjunto de datos entre un subconjunto de entrenamiento $D_{entrenamiento}$ y un subconjunto para prueba D_{prueba}. Esto permitirá evaluar el rendimiento del modelo y sortear dicha sobreestimación. Por ello, se dividió el conjunto de datos en:

$$D_{entrenamiento} = D \times 0.7$$
$$D_{prueba} = D \times 0.3$$

Dicha división se realizó sin ningún criterio en particular a excepción de la capacidad de incluir la mayor cantidad de datos en modelos que requieren para el aprendizaje el mayor volumen posible. Al igual, la división se realiza de manera aleatoria con un muestreo estratificado sobre la variable Clase, esto garantiza la misma distribución de la Clase entre ambos conjuntos de datos.

3.1.2.2 Valores ausentes, duplicados y atípicos

Aunque en el proceso de extracción de datos de la sección anterior se utilizaron funciones de agrupación para las instancias por medio de SQL garantizando la no duplicidad de los datos, sí se encuentra datos faltantes en la columna de la nota promedio, para algunos estudiantes. Esto debido a que al momento de la toma de

los datos no se contaba con el cargue de estos en el sistema. Siendo este un faltante real, no por error en la extracción, y constituyendo un porcentaje menor al 10% de los datos, se toma la decisión de prescindir de las instancias.

3.1.2.3 Selección de características

Una rama de la ingeniería de características en el análisis de los datos permite seleccionar, construir o transformar las características en la matriz de datos por medio del conocimiento empresarial de los datos sin procesar [26]. Partiendo de la definición anterior, se construye la clase necesaria para cada instancia donde:

$$y = \begin{matrix} Si\ \text{“días de vencimiento”} = 0\ Entonces\ \text{“Corriente”} \\ Si\ \text{“días de vencimiento”} > 0\ Entonces\ \text{“Vencido”} \end{matrix}$$

Al igual, se toma la decisión de construir el atributo “Edad” a partir de la fecha de nacimiento que permite una caracterización más apropiada de los estudiantes y facilita el manejo de los datos ya que la fecha no es un tipo de datos adecuado para el análisis.

Estos cambios se realizan nuevamente desde el lago de datos con el mismo código de consulta SQL y se exporta a un repositorio para la consulta desde las herramientas de análisis de datos más adelante.

3.1.3 Análisis exploratorio de los datos

3.1.3.1 Análisis univariado

En adelante se buscará con el análisis interpretar gráficamente si existe una separación evidente de las clases por cada una de las variables por medio de su distribución para el conjunto de entrenamiento $D_{entrenamiento}$.

3.1.3.1.1 La clase

La dimensionalidad del nuevo conjunto de datos para entrenamiento es de: los mismos 15 atributos por 2060 instancias. Cada uno clasificado entre corriente y vencido dada la construcción de la clase en la sección de selección de características. De este total, el 75.73% de los estudiantes se clasifican como CORRIENTES y el 24.27% restante como VENCIDOS se evidencia en la Fig. 1.

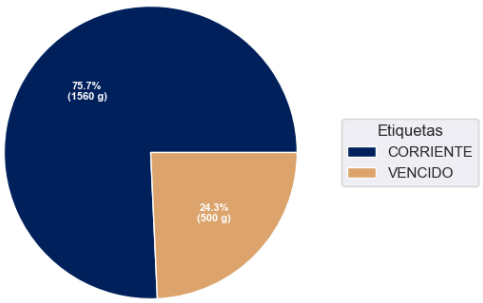

Fig. 1. Distribución de la clase.

Esto evidencia un desbalanceo en la clase binaria para tener en cuenta en el ensamble del modelo. Aunque existen técnicas de construcción de datos a partir la muestra no se tendrán en cuenta para este documento.

3.1.3.1.2 Atributos numéricos

El valor adeudado, la edad y la nota promedio son los únicos atributos numéricos del conjunto de datos. La tabla [04] muestra los estadísticos descriptivos para estas características.

Tabla 4. Estadística descriptiva para los atributos numéricos del conjunto de datos.

Tipo	Medida	Valor	Edad	Nota
Tendencia central	Media	13.319.062	24	3,74
	Mediana	7.653.650	22	3,81
Distribuciones de frecuencia	Valor mínimo	132.000	16	0,52
	Primer cuartil	4.998.052	20	3,48
	Segundo cuartil	7.653.500	22	3,81
	Tercer cuartil	14.411.230	25	4,13
	Valor máximo	202.739.266	66	4,85
Dispersión	Desviación estándar	17.302.877,59	6,55	0,56

El valor promedio adeudado por los estudiantes es de 13,3 millones de pesos colombianos. Esto concuerda con el valor promedio de matrícula en general para los programas de pregrado en 2022 de la Universidad. Los valores fuera de la distribución como el valor máximo de 202,7 millones de pesos corresponden a los créditos mediano plazo, donde el estudiante paga una parte en época de estudio y el restante se acumula para pagar al final de la carrera. Sin ellos, se refleja en la Fig. 2 una distribución normal sesgada a la derecha.

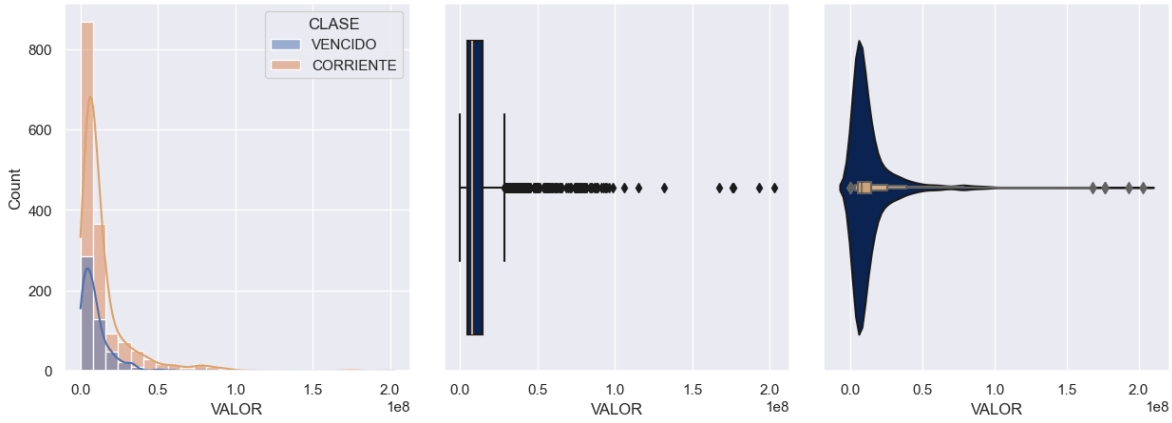

Fig. 2. Distribución del valor por clase.

El promedio de edad para los estudiantes es de 24 años en general. Aunque la edad promedio de los estudiantes en la clase vencido es 27 años y en corriente es 22, significando que son mayores en edad los estudiantes que se vencen en cartera. Ambas clases se distribuyen de manera asimétrica positiva ver Fig. 3.

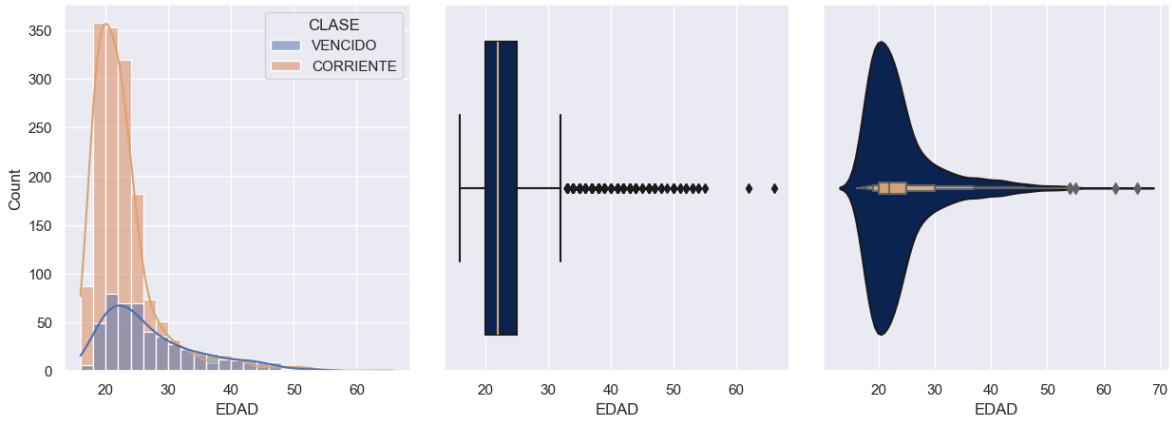

Fig.3. Distribución de la edad del estudiante por clase.

La nota media promedio por estudiante 3.74 en general. Para la clase corriente 3.77 y para la clase vencida 3.65 la diferencia entre ambas no presenta una diferencia amplia y se infiere que el nivel académico es independiente del estado económico del estudiante, ya que no refleja un riesgo de deserción por mal rendimiento académico y que pueda sumar a la pérdida de cartera. La Fig. 4 nos muestra la distribución gráfica.

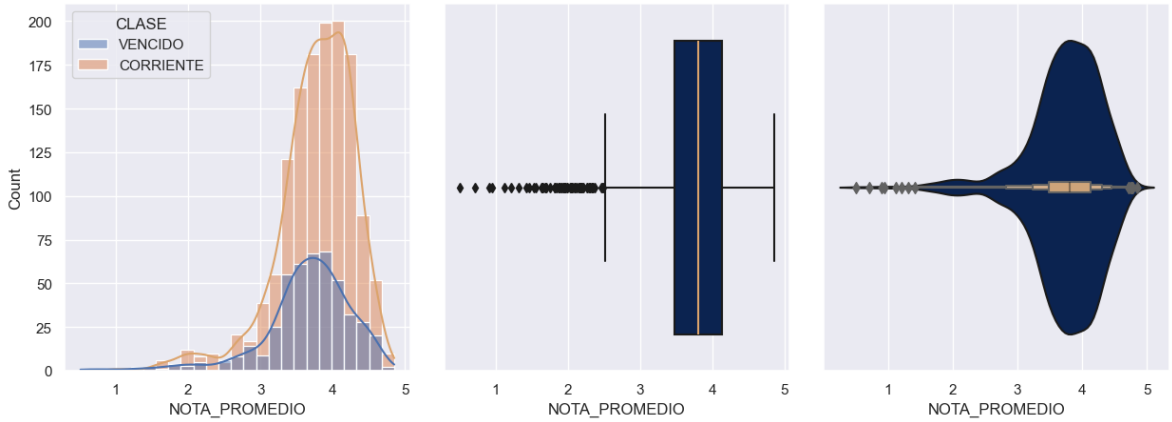

Fig. 4. Distribución de la nota promedio de los estudiantes por clase.

3.1.3.1.3 Atributos categóricos

La mayoría de los atributos en el conjunto de datos son categóricos de tipo nominal, esto se refiere a que no cuentan con un orden especifico y por lo tanto solo tienen sentido las comparaciones de igualdad [28] a excepción del estrato socioeconómico.

El programa de inscripción, admisión, matrícula e inducción (PIAMI) es general para todos los estudiantes de pregrado. Sin embargo, la Universidad inicia el ingreso por tres diferentes líneas de admisión, estándar, programas con colegios en convenio y por medio de programas del gobierno (nacional, departamental o municipal). Los estudiantes con crédito en su mayoría ingresan por el programa estándar. Ver Fig. 5.

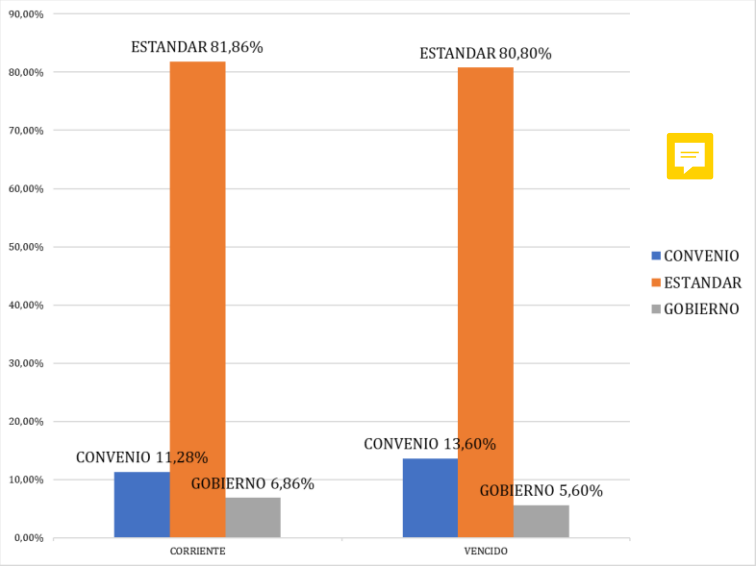

Fig. 5. Tipo de admisión a la Universidad por clase.

La similitud en las columnas no permite una separabilidad de las clases por tipo de admisión para los estudiantes. En ambas, más del 80% se encuentra en admisión estándar.

Respecto del género, más del 51% de los casos son las mujeres quienes encabezan los datos y se distribuyen similarmente entre ambas clases ver Fig. 6.

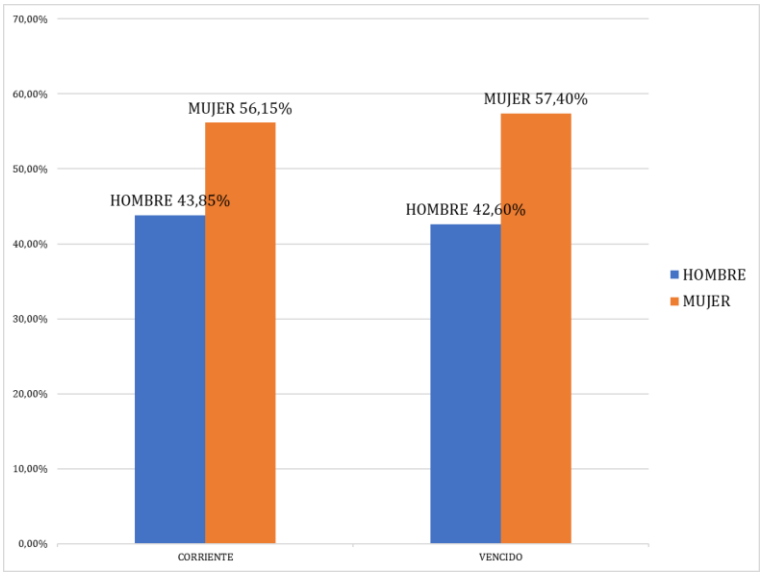

Fig. 6. Clase por género.

Más del 90% de los estudiantes son solteros siendo la categoría más alta entre toda la población y en coherencia con la edad de ellos. Se registran casados como segunda categoría más alta seguidos por unión libre, separado, divorciado, viudo y religioso. Comparando ambas clases en la Fig. 7 no se ve una distribución diferente entre corriente y vencido.

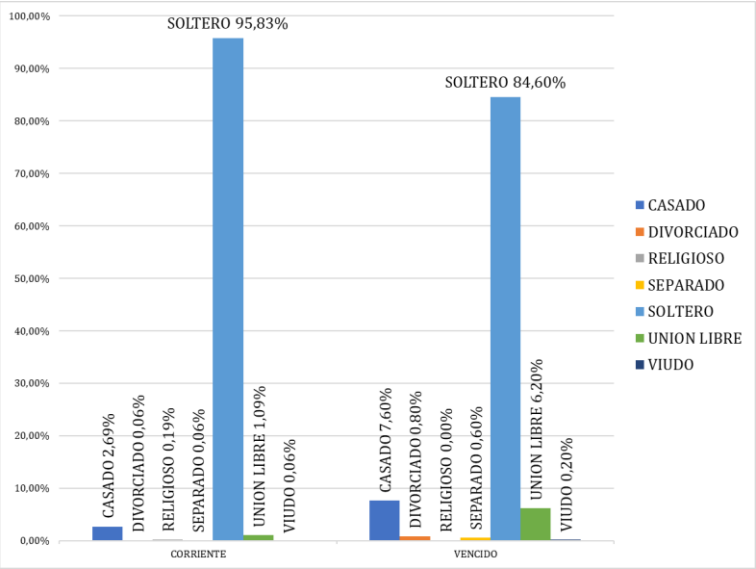

Fig.7. Estado civil de los estudiantes por clase.

Solamente un 1% de los estudiantes con crédito de la Universidad pertenece a algún grupo étnico representativo o minoría. Sin embargo, se ve participación de la comunidad negra, del pueblo RROM (Gitano) y de pueblos indígenas. Las clases comparten la misma distribución Fig. 8.

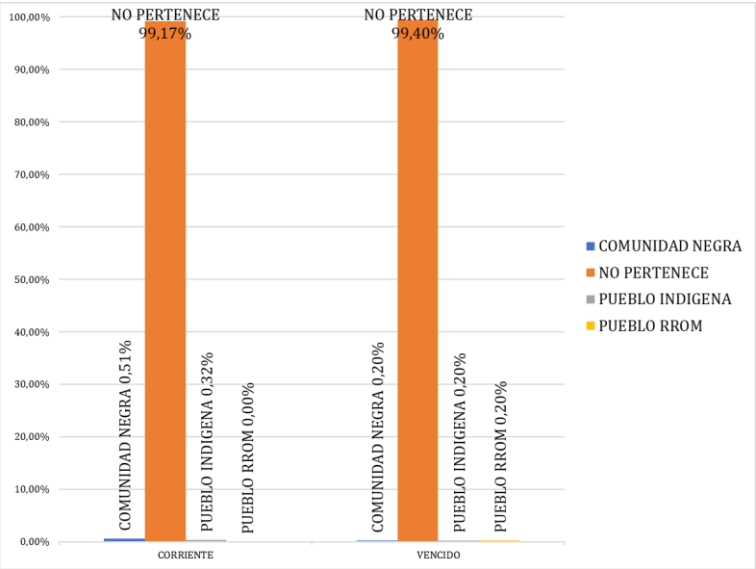

Fig. 8. Grupo étnico de los estudiantes por clase.

El estrato socioeconómico de los estudiantes se solicita al ingreso del estudiante como la estratificación de la vivienda familiar o domicilio de este se marca como un atributo categórico ordinal, ya que entre mayor el estrato se percibe mayor la capacidad económica de la familia, y se evidencia la construcción de la característica “No Indicado” para los estudiantes que no completan el campo, al no ser obligatorio. Las frecuencias más altas son estratos cuatro, tres y dos. Aunque se puede apreciar en la Fig. 9 que el estrato 4 es el de mayor peso en la cartera sin vencer con el 37% y el no indicado en la cartera vencida con el 38%, esto no permite una separabilidad entre las clases por medio de este atributo.

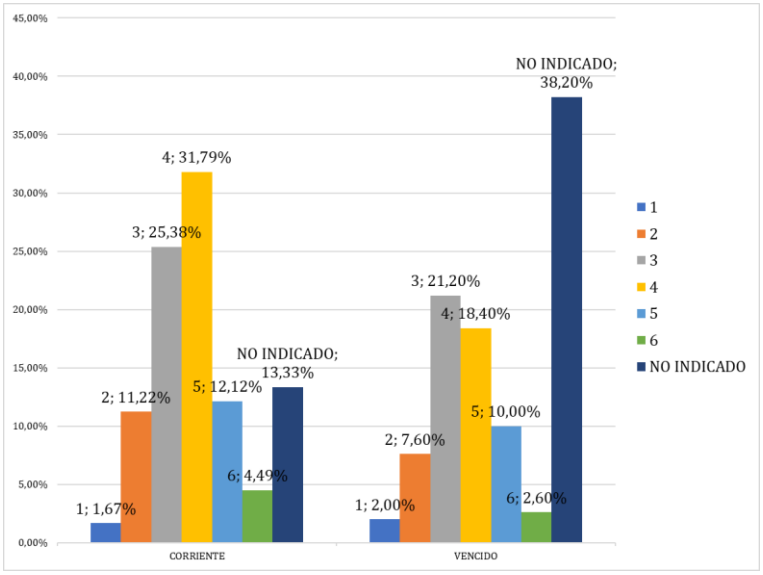

Fig. 9. Estrato socioeconómico.

La característica más representativa para el nivel académico alcanzado por los padres es la ausencia del dato en la mayoría de los estudiantes. Este se imputa con una nueva categoría “No indicado” desde el DWH. Sin embargo, el nivel académico más alto de alguno de los dos padres (o de los dos) para los estudiantes es el pregrado, seguido del bachillerato y el técnico. Ambas distribuciones entre las clases son similares ver Fig. 10.

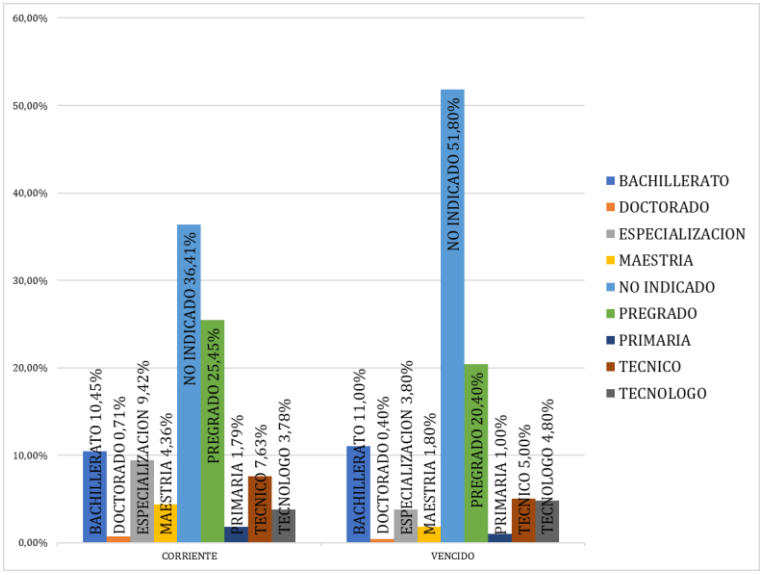

Fig. 10. Nivel académico alcanzado por los padres.

Solo el 4% de los estudiantes que solicitan crédito se encuentran cursando doble programa. Esto se ve coherente con la carga económica que implica un segundo programa para la familia del estudiante. Ver Fig. 11.

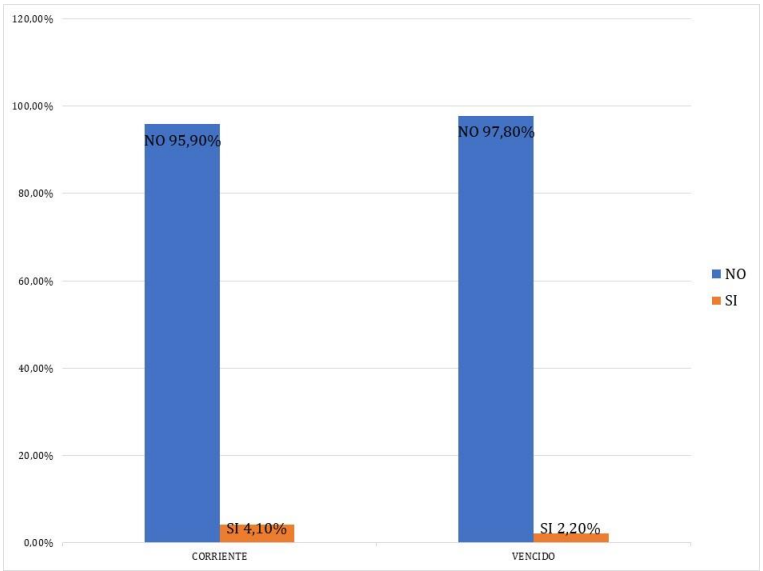

Fig. 11. Estudiante de doble programa.

Menos del 2% de los estudiantes del conjunto de datos se han presentado a semilleros de investigación, o ayudado en investigaciones formales dentro de sus facultades. Esto se da por igual en ambas clases. Ver Fig. 12.

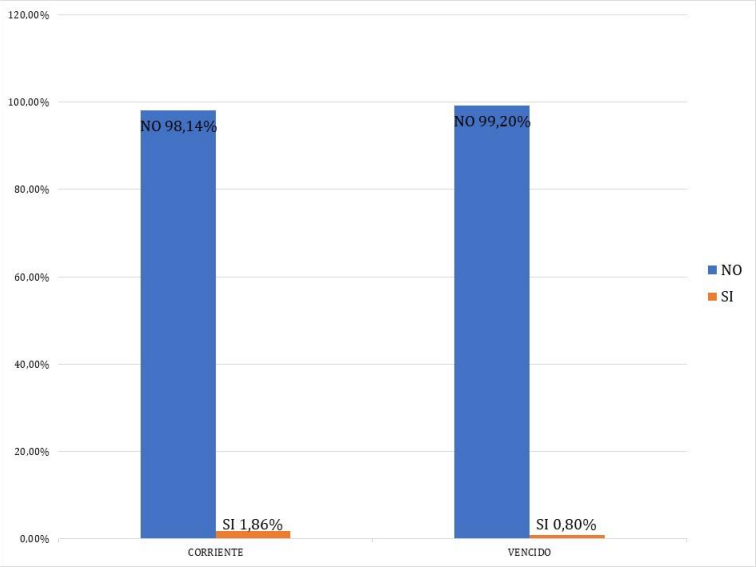

Fig. 12. Estudiantes investigadores.

El 10% de los estudiantes con crédito accedió a uno o más programas de movilidad ofrecidos por la Universidad a nivel internacional para el intercambio académico se evidencia en la Fig. 13. Esta igualdad en ambas clases no permite una separabilidad entre ellas.

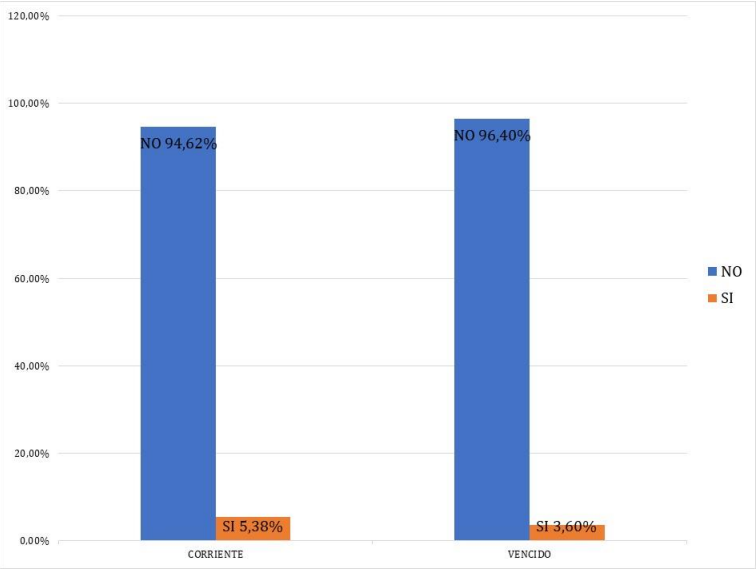

Fig. 13. Estudiantes que han realizado movilidad.

El programa Aprendiendo a Trabajar (PAT) promueve la inclusión temprana de los estudiantes en la vida laboral y a su vez permite a los estudiantes de la Universidad contar con recursos económicos para el sustento de su vida universitaria. Menos del 20% en ambas clases se ayudó con este programa para apalancar los gastos de la matrícula esto se ve en Fig. 14. No se evidencia la posibilidad de separar las clases por medio de esta variable.

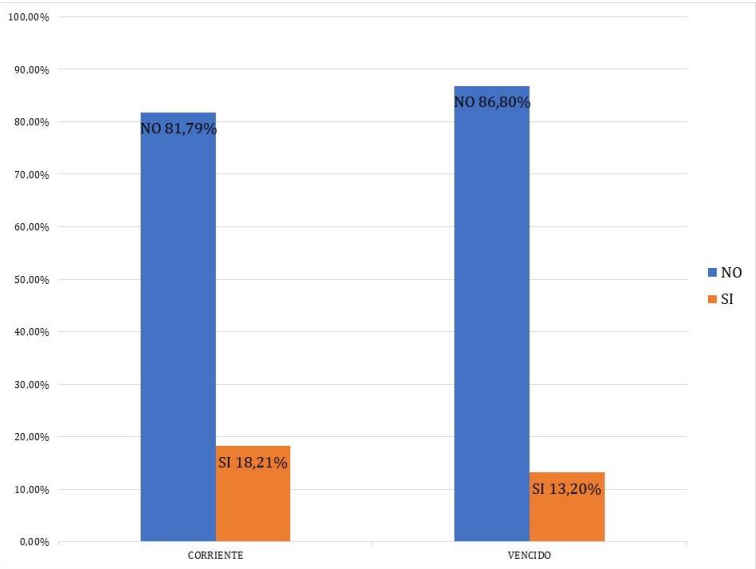

Fig. 14. Estudiantes en el programa PAT.

A partir del análisis univariado de las variables se puede concluir que no existe un atributo por el cual se pueda establecer una relación clara con la variable objetivo. En otras palabras, no es factible la separabilidad de las clases en ninguna individualmente.

3.1.3.2 Análisis multivariado

En el análisis multivariado se comprobará si existe una relación entre la variable objetivo y más de un atributo. Se asume la independencia de las variables numéricas y con ellas se espera entender las relaciones lineales entre ellas.

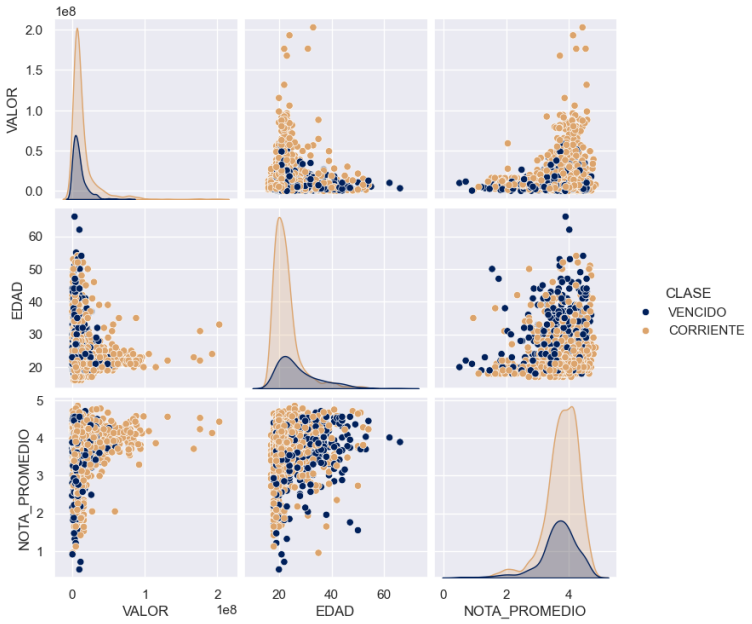

Fig. 15. Dispersión correlacionada de las variables numéricas.

La relación entre las variables numéricas en disposición de la clase no refleja gráficamente (ver Fig. 15) una separación entre estas concluyendo que se dificulta un posible modelo de regresión sobre el valor, la edad y la nota promedio del estudiante.

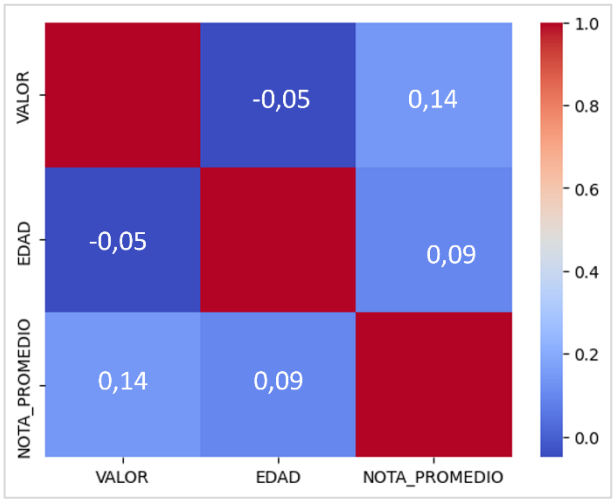

Fig. 16. Correlación de las variables numéricas.

Con el mapa de calor en la Fig. 16 se evidencia que no existe una correlación lineal importante entre las variables numéricas que permita adicionar información relevante a este trabajo.

3.1.4 Análisis de componentes principales

Dado que el objetivo del análisis de componentes principales es el de encontrar las direcciones principales en las cuales los datos tienen la mayor varianza para reducir la alta dimensionalidad de los datos, se puede aplicar un método de correspondencias múltiples dado el alto número de atributos categóricos que se ven en el conjunto de datos [29]. En general, busca analizar la asociación entre las variables cualitativas a través de cada categoría.

Este método codifica las 11 variables categóricas convirtiéndolas en columnas binarias para cada categoría. Las n observaciones de las variables binarizadas se recogen en una tabla de contingencia donde las modalidades de X se recogen en $x1, x2, ...xn$ y las de Y son $y1, y2, ...yn$, e.g., en dos columnas *clase.corriente* y *clase.vencido* donde es 0 cero en *clase.corriente* pero 1 uno en *clase.vencido* para

el estudiante vencido. En la tabla de contingencia ubica las ocurrencias de las categorías en diferentes variables y calcula la inercia que es la variabilidad o dispersión de la tabla de contingencia.

El grafico de nube de variables (ver Fig. 17) muestra las variables cualitativas activas dada el cuadrado de la tasa de correlación entre cada variable y el eje; representando las dos primeras dimensiones en que explican 12.31% de la inercia total. El estrato y el nivel educativo de los padres se asocia a la dimensión 2 y las variables académicas como presentó saber pro o beneficiario PAT se asocian con la variable 1. Al igual, gráficamente la nube de individuos (ver Fig. 18) refleja los que están cercanos es porque tienen atributos similares y la mayoría se relacionan con las dos dimensiones de manera parecida, es decir, cuando los mismos individuos toman valores en una categoría lo hacen también en la otra.

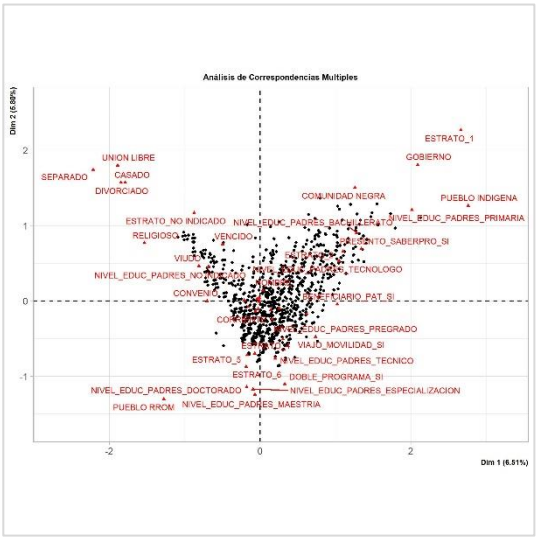

Fig. 18. Gráfico de nube de individuos

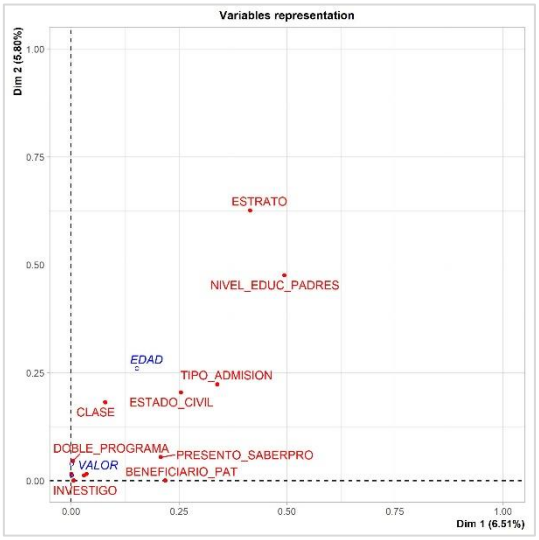

Fig. 17. Gráfico de nube de variables

Las dos primeras dimensiones de análisis expresan el 11,49% de la inercia total del conjunto de datos, lo que significa que el plano explica el 11,49% de la variabilidad total de la nube de individuos (o variables). Se trata de un porcentaje muy pequeño

y el primer plano representa una pequeña parte de la variabilidad de los datos. Este valor es superior al valor de referencia que es igual al 7,44%, la variabilidad explicada por este plano es pues significativa (el valor de referencia es el 0,95-cuantil de la distribución de los porcentajes de inercia obtenidos simulando 2244 tablas de datos de tamaño equivalente sobre la base de una distribución uniforme).

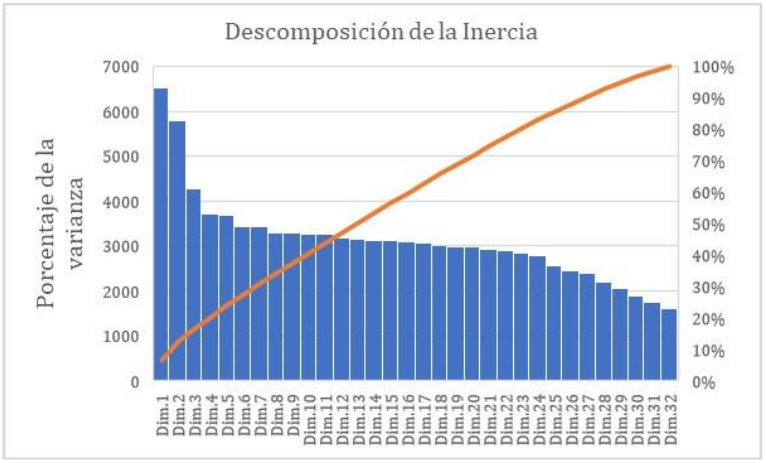

Fig. 19. Inercia de las correspondencias por la distribución de la varianza.

Una estimación del número adecuado de ejes a interpretar sugiere restringir el análisis a la descripción de los 5 primeros ejes se ve en la Fig. 19. Estos ejes presentan una cantidad de inercia superior a la obtenida por el cuantil 0,95 de las distribuciones aleatorias (23,03% frente a 18,11%). Esta observación sugiere que sólo estos ejes son portadores de una información real. Por consiguiente, la descripción se centrará en estos.

La minería de los datos nos concluye dos aspectos relevantes. El primero es que la Universidad, aunque contiene pocos datos para el ejercicio de este documento, si son de alta calidad. En el segundo se concluye que, con el análisis de los datos las categorías académicas tienen una correspondencia o asociación entre sus categorías y con dos dimensiones se puede obtener hasta el 12% de la varianza entre los datos.

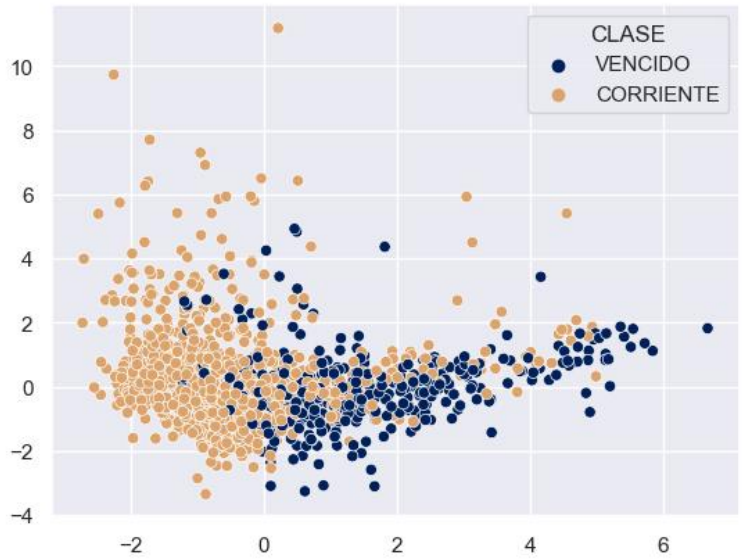

Fig. 20. Dispersión de las clases por componentes principales.

Finalmente, a partir del análisis de los datos y en la Fig. 20 se pone de manifiesto la viabilidad de establecer una separabilidad de los estudiantes en sus respectivas clases mediante los componentes principales en los datos, lo que conlleva a validar la pertinencia de un modelo para el análisis previo antes de la decisión de crédito.

3.2 Ensamble del modelo

Para ensamblar un modelo se toma la decisión desde un conjunto de clasificadores base que se entrenan individualmente. En algunas investigaciones se realizan por enfoques de agregación de modelos de forma que el rendimiento del clasificador combinado sea mayor que el clasificador individual [30]. Sin embargo, se toma el ensamble básico sin combinación para medir cada modelo por separado en este trabajo, pero se utiliza un algoritmo de automatización de aprendizaje de máquina que permite entrenar los modelos respecto de los datos.

3.2.1 Naive AutoML

El optimizador ingenuo de automatización de aprendizaje de máquina (Naive AutoML) consiste en 3 fases: la selección del algoritmo, el ajuste de los hiperparámetros y la definición y entrenamiento de la tubería final [25].

3.2.1.1 Selección del algoritmo

En esta fase, para cada ranura y componente del modelo, el algoritmo construye una tubería que contiene solamente un hiperparámetro y una técnica de preprocesamiento y una métrica arbitraria y de esta manera computa la puntuación con la función de validación. Se encuentra la mejor puntuación y se almacena en memoria.

3.2.1.2 Ajuste de los hiperparámetros

Luego el algoritmo se ejecuta en ciclo comparando cada hiperparámetro por cada componente (por separado). Cada iteración busca el mejor rendimiento del componente sobre el valor de los hiperparámetros de manera aleatoria. Esto se repite hasta que el tiempo de espera se agota.

3.2.1.3 Entrenamiento de la tubería final

Escogida la tubería con la decisión local se entrena con los datos y devuelve el modelo. Este método de automatización asume que cada tubería se puede optimizar localmente y que el algoritmo mejor ajustado para una ranura es también el algoritmo que mejor funciona si se utiliza con los parámetros por defecto.

3.2.1.4 Tubería para los datos de los estudiantes

Los datos crudos de entrenamiento luego de la división se corrieron con un tiempo de ejecución de 600 segundos y de forma predeterminada se utiliza la medida AUC-ROC para clasificación binaria. Con una mejor puntuación en general de 0.7648 el modelo escogido es un árbol de clasificación de aumento de gradiente basado en histograma, dentro de una tubería que se muestra en la Fig. 21.

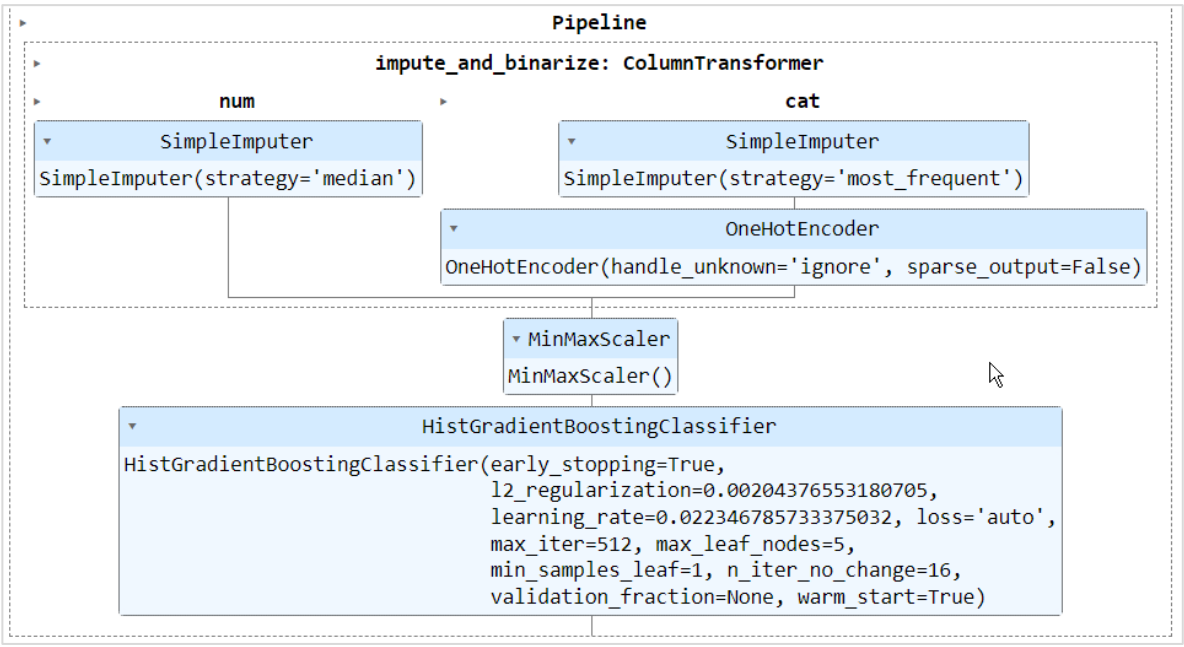

Fig. 21. Modelo propuesto por la iteración.

Para el tratamiento de los datos faltantes en las variables numéricas la optimización escoge un imputador simple por medio de la media del atributo rellena los datos faltantes. Para las variables categóricas escoge el mismo imputando con cada atributo con el dato de mayor frecuencia. En estas variables categóricas, realiza una codificación por medio de la clase OneHotEncoder de la librería SciKitLearn la cual asigna un valor numérico como código para cambiar las variables categóricas e incluirlas en el modelo. De la misma manera, incluye una normalización para todas las variables donde, se escala y traduce cada característica por separado de tal

manera que se encuentre en el rango entre cero y uno, esto le permite al modelo realizar comparaciones entre atributos por medio de la varianza de los datos. Un número máximo de nodos en las hojas de 5 para controlar la complejidad del modelo.

Por último, para los hiperparámetros del clasificador HistGraientBoostingClassifier propuso agregar una regularización rígida (l2 regularización) a la función de pérdida, que, para los problemas de clasificación binaria se utiliza comúnmente la pérdida logarítmica también conocida como desviación binomial, para ayudar a prevenir el sobreajuste del modelo. Un control de la tasa de aprendizaje del algoritmo para cada árbol de 0.0223, entre más bajos los valores, más robusto es el modelo. Un número máximo de árboles en el ensamblaje de 512, entre más arboles mejor el rendimiento, pero aumenta el tiempo de entrenamiento. Y por último propuso un número de 16 iteraciones sin mejorar la métrica de validación antes de detener el entrenamiento temprano.

El puntaje de la recomendación por componentes se determina así: 0.7637 para el clasificador, 0.7648 para el preprocesador de los datos y 0.7648 para el preprocesador de las características. La tabla [05] nos muestra el historial de la ejecución del algoritmo durante los 600 segundos.

Tabla 5. Historial de ejecución por métrica de área bajo la curva.

Clasificador	Preprocesador de Datos	Preprocesador de Características	roc_auc
HistGradientBoostingClassifier	MinMaxScaler		0,7648
GradientBoostingClassifier	QuantileTransformer		0,7647
HistGradientBoostingClassifier			0,7645
GradientBoostingClassifier	PowerTransformer		0,7644
GradientBoostingClassifier	StandardScaler		0,7643
GradientBoostingClassifier	RobustScaler		0,7641
GradientBoostingClassifier	VarianceThreshold		0,7637

GradientBoostingClassifier	MinMaxScaler	PolynomialFeatures	0,7569
RandomForestClassifier			0,729
LinearDiscriminantAnalysis			0,7233
GradientBoostingClassifier	MinMaxScaler	PCA	0,7035
GradientBoostingClassifier	MinMaxScaler	SelectPercentile	0,6887
GradientBoostingClassifier	MinMaxScaler	GenericUnivariateSelect	0,6885
ExtraTreesClassifier			0,6742
BernoulliNB			0,6698
HistGradientBoostingClassifier			0,6667
GradientBoostingClassifier	MinMaxScaler	Nystroem	0,6618
MultinomialNB			0,6566
KNeighborsClassifier			0,656
DecisionTreeClassifier			0,6037
GradientBoostingClassifier	Normalizer		0,5854
SVC			0,5825
MLPClassifier			0,5021
GaussianNB			0,4785

Con esto se puede observar que, al igual que en la revisión bibliográfica, los modelos basados en árboles de decisión permiten una mejor clasificación de los cuentahabientes de créditos. Y con un pequeño presupuesto de recursos computacionales y de tiempo se puede obtener una buena predicción incluyendo algoritmos de optimización como el NaiveAutoML dentro de las metodologías del análisis de los datos. A continuación, se profundiza en el modelo recomendado por la herramienta.

3.2.2 Árbol de clasificación de aumento de gradiente basado en histograma

El aumento de gradiente es un algoritmo de aprendizaje de máquina que agrega modelos de árboles de decisión a un conjunto de forma secuencial donde cada modelo de árbol agregado al conjunto intenta corregir los errores de predicción cometidos por los modelos ya presentes en el conjunto [31] [32]. En Fig. 22 se presenta gráficamente.

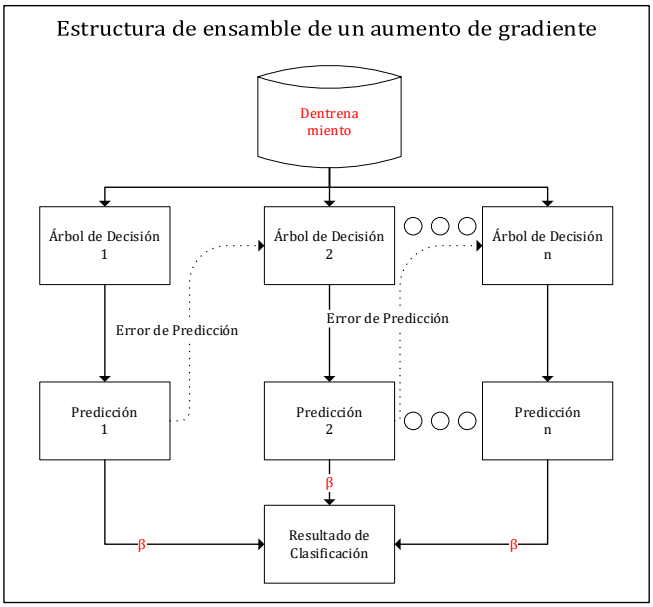

Fig. 22. Estructura de ensamble de un aumento de gradiente.

Por este mecanismo la función de pérdida que indica la tasa de error a la hora de la clasificación de todo el ensamble se disminuye gradualmente durante la fase de entrenamiento del modelo, mejorando así la predicción para los datos nuevos.

Internamente el árbol de decisión consiste en nodos que representan decisiones correspondientes a los hiperplanos y de nodos de hoja que representan regiones en el espacio de los datos. Primero selecciona el mejor atributo que separa las muestras de los datos, calculando una medida de impureza como la entropía o el índice Gini para cada atributo. Luego divide el conjunto de datos en subconjuntos más pequeños en función de la medida. Esto se repite de manera recursiva creando más subárboles hasta que se cumpla una profundidad máxima.

Para realizar la predicción el árbol comienza con el nodo raíz y desciende a través de las ramas siguiendo las decisiones basado en las características entrenadas.

4 Resultados

En esta sección se presentan los resultados del modelo entrenado sobre los datos de prueba. Se presenta la medición AUC-ROC y otras métricas utilizadas en los problemas de clasificación. Se revisa la pertinencia de incluir más datos con la curva de aprendizaje del modelo. Y por último se revisan los atributos relevantes de los árboles de decisión entrenados.

4.1 AUC-ROC

La curva ROC (Receptor Operative Characteristic por sus siglas en inglés) se constituye como un método estadístico para determinar la exactitud diagnostica [33]. Esta cuantifica el rendimiento global del modelo incluyendo las probabilidades de la sensibilidad de cada estudiante en la clase dándole importancia sobre métricas como el F1 score para este estudio.

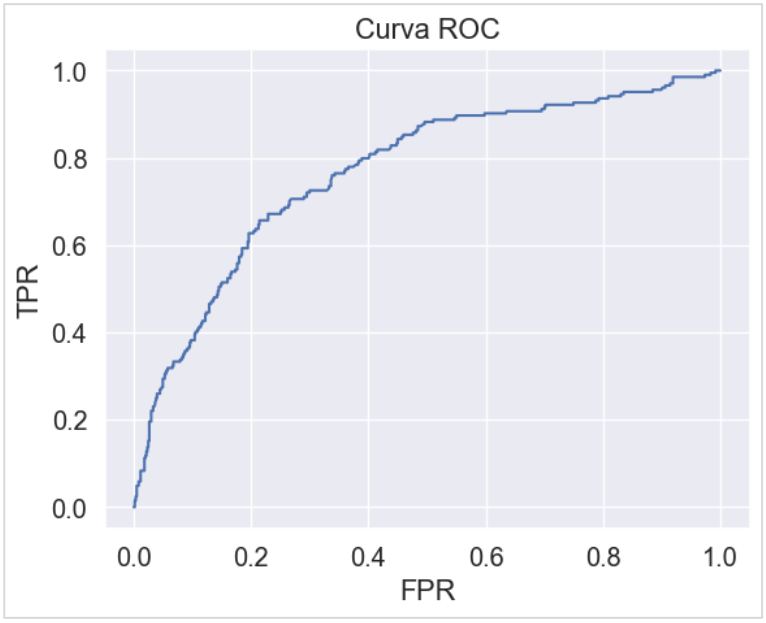

Fig. 23. Curva ROC para los datos de entrenamiento.

Entre mayor sea el AUC-ROC, mayor es el poder del modelo de discriminar entre estudiantes corrientes o vencidos. Esto indica que el modelo con un AUC-ROC de 0.7648 (76,48%) alcanza un valor importante para los datos de entrenamiento (ver Fig. 23 gráficamente). Sin embargo, este valor no indica que necesariamente es un rendimiento perfecto, ya que todavía existe cierta superposición entre las distribuciones de probabilidad de las dos clases.

Una vez entrenado el modelo, se adaptó el conjunto de datos de prueba D_{prueba} al modelo, este contiene 844 instancias con los mismos atributos de la partición inicial de los datos. Para X_{prueba} se predice la probabilidad de que, Y_{prueba} sea 1 en el caso de la binarización. Con la métrica AUC-ROC-SCORE arroja un resultado de 0.5596 (55.96%) para la predicción. Aunque es menor que los datos entrenados, indica que el modelo con los datos de prueba tiene iguales probabilidades de predecir un estudiante tanto vencido como corriente. Sin embargo, se presentan a continuación, otras métricas importantes para el entendimiento de AUC-ROC.

4.2 matriz de confusión

La curva ROC traza la tasa de verdaderos positivos (TPR) en el eje Y, frente a la tasa de faltos positivos (FPR) en el eje X. Estas se pueden derivar de una matriz de confusión que resume todas las correctas y falsas predicciones, para el conjunto de datos de prueba se genera [34].

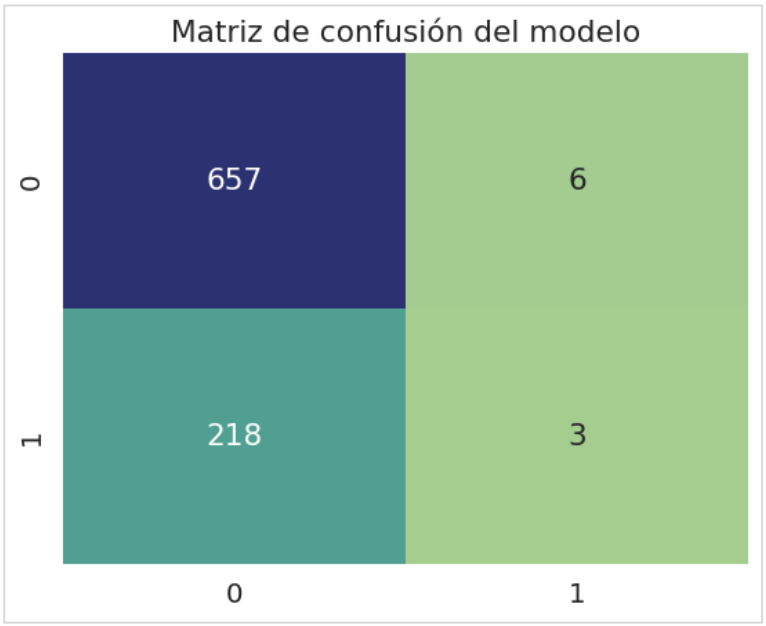

Fig. 24. Matriz de confusión del modelo con datos de prueba.

El resultado de la matriz en Fig. 24 es 657 son los verdaderos positivos (que predijo como “corriente” siendo “corriente”). 218 son los falsos positivos, el modelo los predijo corriente siendo vencidos, 6 estudiantes los predijo vencidos siendo corrientes, estos son los falsos negativos y 3 estudiantes los predijo vencidos siendo vencidos, estos son los verdaderos negativos. A continuación, se analizan las métricas más populares en la revisión bibliográfica, aunque estas no fueron sujeto de optimización.

La exactitud (accuracy) mide el rendimiento del modelo en relación de todas las instancias correctas sobre todas las instancias de los datos de prueba, así:

$$\frac{\text{Verdaderos Positivos + Verdaderos Negativos}}{\text{Total Instancias}} = \frac{657 + 3}{884} = 0.7466$$

Esto indica que el modelo es un 74% exacto sobre las predicciones de los datos de prueba.

La precisión (precision) nos mide cuan precisas son las predicciones positivas del modelo:

$$\frac{\text{Verdaderos Positivos}}{\text{Verdaderos positivos + Falsos Positivos}} = \frac{657}{657+218} = 0.7509$$

La métrica de precisión nos muestra que el modelo es 75% preciso en las predicciones positivas que realizó.**La sensibilidad (recall)** mide la eficacia en identificar las instancias positivas del conjunto de datos:

$$\frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos + Falsos Negativos}} = \frac{657}{657+6} = 0.9910$$

Esto indica que el modelo es muy bueno prediciendo los estudiantes corrientes.
En contraste, **la especificidad (specifity)** mide la eficacia en identifica las instancias negativas del conjunto de datos:

$$\frac{\text{Verdaderos Negativos}}{\text{Verdaderos Negativos + Falsos Positivos}} = \frac{3}{218+3} = 0.0136$$

Esto nos indica que el modelo no es bueno prediciendo la clase vencidos.

El F1-score evalúa el rendimiento general del modelo de clasificación armonizando la precisión y la sensibilidad anteriores:

$$\frac{2 \text{ . Precisión . Sensibilidad}}{\text{Precisión + Sensibilidad}} = \frac{2 \text{ x } 0.7509 \text{ x } 0.9910}{0.7509 + 0.9910} = 0.8544$$

Esta última métrica refleja un buen rendimiento del modelo para la clase corrientes en coherencia con el desbalanceo de la clase vista en la sección 3.

4.2 Curva de aprendizaje

La curva de aprendizaje emplea una técnica de validaciones cruzadas entre el entrenamiento del modelo y el rendimiento de la prueba. En este proceso se utilizan subconjuntos de los datos de entrenamiento con diferentes tamaños para entrenar el clasificador y calcular una medida de puntuación que se promedian y presentan para analizar el rendimiento del modelo si se incluyen más datos sobre este [35].

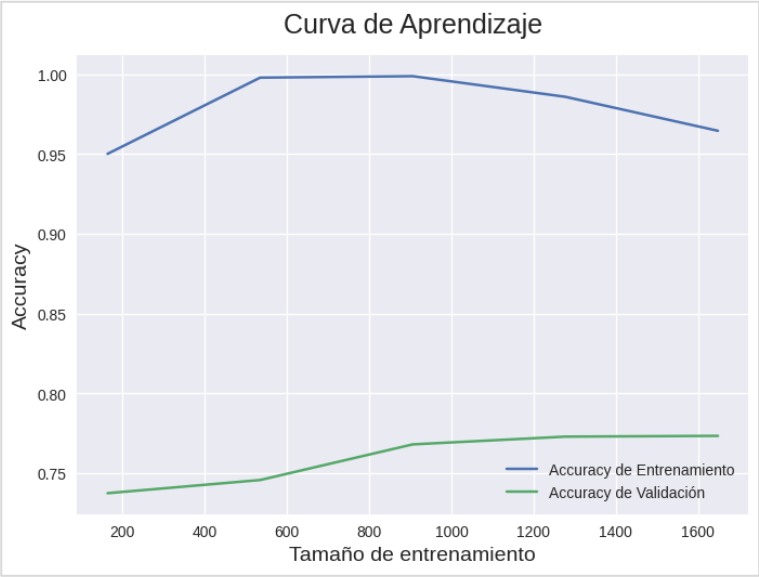

Fig. 25. Curva de aprendizaje para el modelo.

La curva de entrenamiento en la Fig. 25 mejora cuando se incluyen más muestras en el modelo al igual que la puntuación de validación. Sin embargo, la convergencia entre las dos curvas indica un posible sobre ajuste del modelo al incluir más muestras. Esto nos concluye que se pueden incluir más muestras, pero no tantas que sobre ajusten el modelo.

4.3 Atributos relevantes del clasificador

En el ensamble del modelo una salida importante para el estudio es reconocer los atributos que el árbol de decisión esta puntuando como más importantes para el modelo. Esto se analiza con la ayuda de la figura que contiene los nodos desde la raíz hasta la decisión.

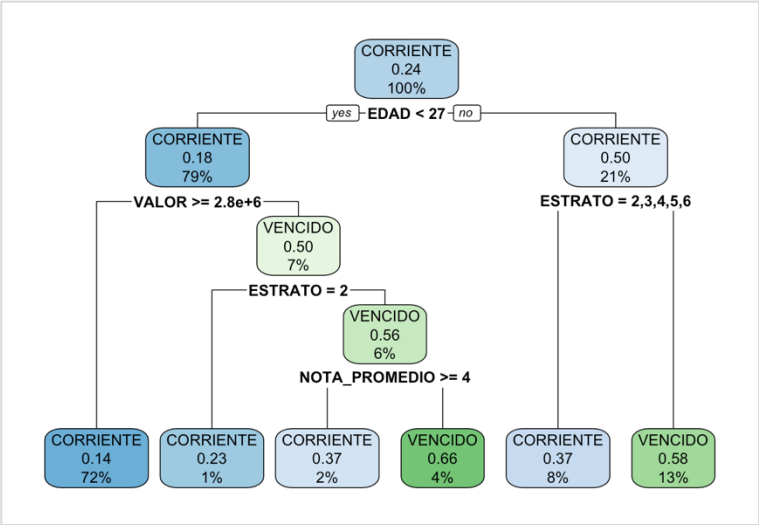

Fig. 26. Árbol de decisión principal de los datos de prueba.

La Fig. 26 muestra que, para una muestra aleatoria de los datos D_{prueba}, la probabilidad global de ser un estudiante buena paga. El 79% de los estudiantes de la muestra son menores de 27 años y estos tienen una probabilidad del 18% de ser clasificados como corrientes. Si el valor del préstamo es mayor o igual a 2.8 millones de pesos tiene una probabilidad de 14% de ser corriente. Si la respuesta es negativa, tiene una probabilidad de 50% de ser vencido. De estos, si la nota promedio es mayor a 4 tiene una probabilidad de 37% de ser corriente, en caso contrario se clasifica vencido con una probabilidad de 66%.

Esto nos ayuda a concluir que las características que afectan la probabilidad de ser un cuentahabiente “corriente” son la edad, el valor del préstamo, el estrato y la nota promedio, sobre los demás atributos.

5 Conclusiones y trabajos futuros

5.1 Conclusiones

En este documento se ha abordado la viabilidad de entrenar un modelo de clasificación para predecir el impago de los créditos educativos ofrecidos por la Universidad de La Sabana utilizando datos limitados de los estudiantes y sin contar con información financiera completa. Con el despliegue de la metodología planteada, se evidencia que se logró desarrollar un componente basado en el procesamiento de los datos y las técnicas de aprendizaje automático que devuelve una probabilidad de impago al analista de crédito de cada estudiante que permitirá añadir valor a su decisión.

Es importante incluir dentro en el ensamble del modelo una metodología de automatización de tuberías como el NaiveAutoML que permita explorar eficazmente incluso los modelos altamente sofisticados para los datos.

Se identificó que las características más relevantes dentro del análisis de los datos y que afectan la probabilidad de ser un buen deudor (corriente) son la edad, el valor del préstamo, el estrato y la nota promedio, sobre los demás atributos. Aunque no existe un atributo por el cual se pueda establecer una relación clara con la clase y que no existe una correlación lineal importante entre las variables numéricas, se pone de manifiesto la viabilidad de establecer una separabilidad de los estudiantes en sus respectivas clases mediante los componentes principales, lo que conlleva a validar la pertinencia del modelo para el análisis previo antes de la decisión de crédito.

Los algoritmos basados en árboles de decisión permiten clasificar los estudiantes con un mejor rendimiento global sobre otros modelos como redes neuronales o discriminantes lineales dada su mayoría de datos categóricos. Con una puntuación AUC-ROC de 76.48%, el modelo de aumento de gradiente basado en histograma alcanza un valor importante para los datos de entrenamiento. Sin embargo, este valor no indica que necesariamente es un rendimiento perfecto, ya que todavía existe cierta superposición entre las distribuciones de probabilidad de las dos clases y esto se evidencia en los datos de prueba. También se concluye que es pertinente incluir más instancias para mejorar la curva de validación del modelo, pero luego de 1000 muestras adicionadas el modelo se sobre ajusta.

Al igual, es importante tener en cuenta las probabilidades generadas por el modelo y no solamente la predicción ya que esta se encuentra limitada por los datos entrenados y no corresponde totalmente a la realidad del solicitante. A demás, el modelo debe ofrecer una recomendación a la persona del analista de crédito, más que tomar la decisión por sí mismo.

5.2 Trabajos Futuros

Un despliegue en producción del modelo permite llevar a la realidad lo planteado en este documento para generar el sistema de recomendación en una plataforma o interfaz de programación de aplicaciones que permita la fácil utilización del modelo y su recomendación. A futuro la Universidad de La Sabana puede incluir más datos de los estudiantes en relación con aspectos tan diversos como las redes sociales, los desempeños deportivos, la creación de contenidos, los criterios y valoraciones en música, comida o televisión que, sin ser financieros, amplían significativamente los datos para proponer más modelos de clasificación. Y más apremiadamente, desarrollar un módulo de información financiera que guarde esta información de manera automática.

6 Referencias

[01] Universidad De La Sabana, documentos institucionales. Proyecto Educativo Institucional PEI. https://www.unisabana.edu.co/fileadmin/Archivos_de_usuario/Documentos/Documentos_la_Universidad/Docs_Institucionales/2._Proyecto_Educativo_Institucional_-PEI.pdf

[02] Vélez, C. (15 de abril de 2010). Boletín Informativo Educación Superior, Financiar la educación, un compromiso de todos. https://www.mineducacion.gov.co/1621/articles-92779_archivo_pdf_Boletin15.pdf

[03] Superintendencia Financiera de Colombia. (01 de noviembre de 1995) Circular Básica Contable y Financiera (circular externa 100 del 95). https://fasecolda.com/cms/wp-content/uploads/2019/08/ce100-1995-cap-ii.pdf

[04] Suárez Ortiz, G. (2020). Del Pagaré al Pagaré de Consumo. Un nuevo panorama para el derecho del consumo colombiano. Revista De La Facultad De Derecho De México, 70(278-2), 863–888. https://doi.org/10.22201/fder.24488933e.2020.278-2.77495

[05] Siddiqi, N. (2017). Intelligent Credit Scoring (2° ed.). John Wiley & Sons, Inc.

[06] Castillo, M. Pérez, F. (2008). Gestión del riesgo crediticio: un análisis comparativo entre Basilea II y el sistema de administración del Riesgo Crediticio Colombiano, SARC. https://revistas.javeriana.edu.co/index.php/cuacont/article/view/3249/2471

[07] Cano, J. (2021). Aprendizaje supervisado en la construcción de un modelo de Credit Scoring para cooperativas de ahorro y crédito en Colombia. https://repositorio.unal.edu.co/bitstream/handle/unal/81003/1035424538.2021.pdf?sequence=1&isAllowed=y

[08] Hill, R.K. What an Algorithm Is. Philos. Technol. 29, 35–59 (2016). https://doi.org/10.1007/s13347-014-0184-5

[09] Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. Big Data & Society, 10(1). https://doi.org/10.1177/20539517221145372

[10] Russell, S. Norving, P. (2010). Artificial Intelligence A Modern Approach (3° ed.). Pearson Education, Inc.

[11] Lessman, S. Baesens, B. Seow, H. Thomas, I. (2015) Benchmarking state of the art classification algorithm for credit scoring: An update of research. European Journal of Opertional Research. Vol. 247. DOI: 10.1016/j.ejor.2015.05.030

[12] Goh, Y. Lee, S. (2019). Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. Advances in Operations Research. Hindawi Limited. DOI: https://doi.org/10.1155/2019/1974794

[13] Kumar, A. Ramesh, S. Rahul, S. (2017). A technology on credit score system assessing public perception in Bengaluru city. International conference on intelligent sustainable systems Palladam. DOI: 10.1109/ISS1.2017.8389442

[14] Abdou, H. Pointon, J. (2011). Credit Scoring Statistical techniques and evaluation criteria: a review of the literature. Institute System in Accounting, Finance and Management DOI: 10.1002/isaf.325

[15] Loterman, G. Brown, I. Martens, D. Mues, C. Baesens, B. (2012), “Benchmarking regression algorithms for loss given default modeling”, International Journal of Forecasting, no. 28, pp. 161–170.

[16] Pandey, T. Jagadev, A. Mohapatra, S. Dehuri, S. (2018). Credit Risk Analysis Using Machine Learning Classifiers. http://doi.org/10.1109/ICECDS.2017.8389769

[17] Chen, M. Huang, S. Credit scoring and rejected instances reassigning through evolutionary computation techniques, Expert Systems with Applications, Vol. 24(4), pp. 433–441, 2003.

[18] Yang, F., Qiao, Y., Qi, Y. et al. BACS: blockchain and AutoML-based technology for efficient credit scoring classification. Ann Oper Res (2022). https://doi.org/10.1007/s10479-022-04531-8

[19] Paweł, P. Abdar, M. Pławiak, J. Makarenkovc, V. Acharya, R. (2020). DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. https://doi.org/10.1016/j.ins.2019.12.045

[20] Mohri, M. Rostamizadeh, A. Talwalkar, A. (2018). Foundations of Machine Learning. (2° ed.). The MIT Press.

[21] Russell, S. Norving, P. (2010). Artificial Intelligence A Modern Approach (pp. 694-695. 3° ed.). Pearson Education, Inc.

[22] Zaki, M. Meira, W. (2020). Data Mining and Machine Learning Fundamental Concepts and Algorithms. (2° ed.). Cambridge University Press.

[23] Collin, B. (2004). Artificial Intelligence Illuminated. Jones and Bartlett Publishers, Inc.

[24] Danenas, P., Garsva, G., Gudas, S. (2011). Credit risk evaluation using SVM classifier, International Conferences On Computational Science. (pp.1699-1709).

[25] Mohr, F. Wever, M. Naive automated machine learning. (2022). https://doi.org/10.1007/s10994-022-06200-0

[26] Ustundag, A., Sivri, M. S., & Menguc, K. (2022). Feature Engineering. In Springer Series in Advanced Manufacturing (pp. 153–169). Springer Nature. https://doi.org/10.1007/978-3-030-93823-9_6

[27] Abu-Mostafa, Magdon-Ismail, M., & Lin, H.-T. (2012). Learning from data: a short course. AMLbook.

[28] Zaki, & Meira Jr, W. (2017). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

[29] L. G. Díaz Monroy y M. A. Morales Rivera. “Análisis estadístico de datos categóricos”. Universidad Nacional de Colombia. (2009).

[30] Li, F.C., Wang, P.K., Wang, G.E., Comparison of primitive classifier with ELM for credit scoring, Procceding of IEEE IEEM, pp. 685-688, 2009.

[31] Brownlee, J. (27 de abril de 2021). Histogram-Based Gradient Boosting Ensembles in Python. https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/

[32] Hoang, N. Tran, V. (2023). Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification. https://doi.org/10.1016/j.autcon.2023.104767

[33] Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16(1). https://doi.org/10.1186/s13040-023-00322-4

[34] Fahmy Amin, M. (2022). Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial. Journal of Engineering Research, 6(5), 0–0. https://doi.org/10.21608/erjeng.2022.274526

[35] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.learning_curve.html