UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Julio 2024

DESERCIÓN UNIVERSITARIA EN POBLACIONES VULNERABLES. ESTUDIO DE CASO SOBRE EL PROGRAMA "JÓVENES A LA U" EN BOGOTÁ

UNIVERSITY DROPOUT RATES AMONG VULNERABLE POPULATIONS: "JÓVENES A LA U" CASE OF STUDY IN BOGOTÁ

Jose Daniel Pachón Ariza.[a]
josepachona@usantotomas.edu.co

Ph. D. Wilmer Pineda Ríos.[b]
wpinedar@unal.edu.co

Resumen

Diversos análisis sobre la deserción de estudiantes en la educación superior han permitido identificar que factores como el sexo, las características socioeconómicas, sociodemográficas y el desempeño académico en el colegio resultan determinantes para incrementar o reducir el riesgo de deserción. En este documento se pretende encontrar los factores que influyen en la deserción universitaria del programa gubernamental "Jóvenes a la U" en Bogotá mediante la aplicación del modelamiento con técnicas del machine learning como mecanismos de análisis y predicción del riesgo.

La primera parte del documento describirá los análisis realizados sobre la deserción universitaria haciendo especial énfasis en la aplicación de técnicas estadísticas, posteriormente, se realiza una exposición conceptual de las técnicas de machine learning para modelos de clasificación en línea con las utilizadas por la literatura de deserción. Seguidamente, se exponen hallazgos sobre los datos obtenidos del programa "Jóvenes a la U" y el tratamiento dado a estos que incluye la imputación de datos faltantes y la aplicación de técnicas de balance de clases. A continuación, se describen las características de los beneficiarios de este programa y finalmente, se desarrollará el análisis de la aplicación de un modelo logístico, un árbol de decisión, un bosque aleatorio, un clasificador por impulso de gradiente y una red neuronal como técnicas para el análisis de los factores que influyen en la deserción y para el modelamiento que permita predecir y clasificar a la población desertora.

Palabras clave: árboles, bosque aleatorio, deserción, factores, gradiente, jóvenes, logístico, machine learning, redes neuronales, riesgo, universitaria, vulnerables

1. Introducción

La educación superior en Colombia comprende los programas técnico profesionales, tecnológicos y profesionales universitarios y de acuerdo con el Sistema para la Prevención de la Deserción en Educación Superior SPADIES, la tasa de deserción en este nivel de educación alcanzó el 10.08 % para 2021 (última cifra disponible) y, al discriminar por nivel de formación esta tasa se ubica en promedio en 8.89 % para programas universitarios, 15.32 % para tecnologías y 18.79 % para técnicos profesionales (MEN 2023). Por sexo, esta tasa es superior en los hombres frente a las mujeres; así, en el nivel universitario esta tasa se ubicó en 9.96 % y 8.01 % respectivamente, en el nivel tecnológico alcanzó el 17.76 % y 15.04 % y

[a] Estudiante
[b] Director

finalmente para los técnicos profesionales la tasa de deserción fue de 19.94 % y 17.55 %. Por su parte, en Bogotá la tasa de deserción alcanzó el 9.14 % para el nivel universitario, el 18.54 % para el tecnológico y el 20.34 % para el técnico profesional.

"Jóvenes a la U" es un programa implementado por la Alcaldía de Bogotá que tiene por objetivo promover el acceso y permanencia en la educación superior de los bachilleres de la ciudad brindando la financiación del valor de la matrícula y otorgando un apoyo económico semestral equivalente a un salario mínimo mensual legal vigente sin generar endeudamiento alguno. El programa es financiado con recursos públicos y busca tener un impacto social destacado, toda vez que, hasta la fecha se han beneficiado más de 36.000 jóvenes.

El programa a lo largo de sus convocatorias ha beneficiado especialmente a población vulnerable, teniendo en cuenta que, según los documentos de términos y condiciones para la participación de los jóvenes, se prioriza la asignación de cupos a mujeres y madres, personas con discapacidad, personas pertenecientes a grupos étnicos (indígenas, afrocolombianos, afrodescendientes, raizales, palenqueros y rrom), con nivel SISBEN (Sistema de información de potenciales beneficiarios de programas sociales) en pobreza extrema (nivel A), moderada (nivel B) y vulnerabilidad económica (nivel C), así como, también prioriza a víctimas del conflicto, personas en proceso de reintegración y reincorporación, bachilleres provenientes de colegios oficiales rurales, entre otros grupos poblacionales vulnerables, de tal manera que, en condiciones tradicionales estas personas tendrían dificultades no solo para acceder y financiar este nivel de formación por sus propios medios, sino también para permanecer en este, situación en la cual cobra relevancia los factores económicos, sociales y académicos que se ha identificado en la literatura especializada (y que se describe posteriormente) por lo que, el éxito de "Jóvenes a la U" se sustentará principalmente en lograr la graduación de sus beneficiarios.

Con el desarrollo de este trabajo se espera identificar los factores que inciden en el riesgo de deserción de los beneficiarios de este programa modelando el uso de técnicas de clasificación contempladas en el machine learning como los modelos logísticos, árboles de decisión, bosques aleatorios, clasificador de gradiente y redes neuronales.

2. Marco conceptual

2.1. Deserción universitaria

El fenómeno de la deserción universitaria ha sido ampliamente investigado, por ejemplo, Tinto (1975), Páramo (1999), López (2004) y Díaz (2013) han encontrado que factores sociales, económicos, institucionales, personales, académicos, entre otros, explican este fenómeno. En Colombia, recientes investigaciones como las de Escarria (2010) y Castillo (2010) aducen esencialmente al puntaje obtenido en las pruebas de estado al terminar la educación media "Saber 11", al estrato socioeconómico de la vivienda, al nivel de ingresos y al programa de estudios, como los principales factores que determinan la permanencia y graduación de un estudiante de la educación superior.

Tinto considera la deserción desde una perspectiva individual en la que las habilidades académicas en matemáticas y lectura crítica son elementos fundamentales para el logro académico esperado y por consiguiente sobre el fenómeno del abandono (o deserción) si se tiene en cuenta que un joven con deficiencias conceptuales difícilmente se adaptará a la vida universitaria y a la propia exigencia de la formación impartida (Tinto 1975) .

Adicionalmente, Páramo y Correa (1999) destacan como factores de deserción, el nivel socioeconómico bajo, la edad (a mayor edad mayor riesgo), la inapetencia por el conocimiento, los ambientes educativos y los modelos pedagógicos implementados por los docentes. Otros análisis realizados como los de Rodríguez (2012), Arismendy (2018) y Fandiño (2022) consideran las notas iniciales del programa de formación y algunas variables sociodemográficas como el nivel educativo de los padres como determinantes de la deserción.

Castro et al (2021), realizaron un ejercicio de seguimiento a 679 estudiantes en una universidad privada de la ciudad de Medellín durante 14 períodos académicos y de quienes se obtuvo datos sociodemográficos, académicos y económicos, con el fin de encontrar causalidades entre estas variables y la deserción o graduación del estudiante. En este ejercicio se aplicó el modelo de riesgos competitivos propuesto por Kennedy (2005). En este tipo de modelos *"se siguen los sujetos a lo largo de los diferentes periodos hasta que experimentan uno de los dos resultados de interés. Lo anterior supone, que sólo puede ocurrir un resultado y que, una vez que éste se produzca, el sujeto ya no está en riesgo"* (Castro 2021).

Los resultados encontrados por los autores evidencian que, el 53 % de las mujeres lograron finalizar sus estudios respecto a un 40 % de los hombres. Por otro lado, con relación a la edad de los estudiantes, se observa que aquellos que iniciaron sus estudios siendo menores de 18 años, desertaron el 33 %, mientras que, frente al grupo de estudiantes que iniciaron sus estudios con 18 años o más, desertaron el 37 %. Finalmente, resaltan que *"la proporción de estudiantes graduados aumenta a medida que el estrato socioeconómico de sus viviendas y el ingreso familiar también lo hicieron"*.

Como se observa, si bien las perspectivas de análisis varían, las variables o factores que pueden incidir en la deserción universitaria llegan al mismo punto, factores sociales, económicos, individuales y académicos resultan determinantes en todas las investigaciones.

Ahora bien, las aplicaciones estadísticas sobre la deserción universitaria han contemplado principalmente la realizacion de modelos logísticos que permiten la interpretación de los coeficientes de las variables explicativas y, como mecanismos de predicción para la toma de decisiones se tiene la clasificación de estos modelos y más recientemente se han implementado los árboles de decisión, bosques aleatorios, vecinos más cercanos y máquinas de soporte vectorial.

Por ejemplo, Rodriguez (2012) a través de un modelo estocástico por cadenas de Markov analizó las variables de deserción para los estudiantes de los programas de ingenierías en la Escuela Colombiana de Ingeniería Julio Garavito, para el cual incorporó variables académicas (tipo de colegio de secundaria, puntaje Saber 11, notas asociadas a asignaturas del área de matemáticas y período de ingreso al programa académico), sociales (estudios de los progenitores, si el estudiante vive solo o no, tipo de vivienda, sexo, edad) y económicas (estrato de la vivienda, nivel de ingresos y valor del semestre). Ante la presencia de datos faltantes realizó la imputación por el método "MonteCarlo" y ajustó un modelo logístico semestral para identificar las variables significativas en términos de deserción.

Los resultados obtenidos indicaron que las mujeres tienen una mayor probabilidad de graduarse, no obstante, el tiempo de permanencia es mayor que el de los hombres. Por otro lado, los estudiantes con vivienda de estrato socioeconómico 1 demoran mucho menos tiempo en graduarse en comparación con los otros estratos y su probabilidad de graduarse es mayor que la de los demás estratos socioeconómicos, lo anterior, según la autora se debe a que las personas con vulnerabilidad económica propenden por no perder asignaturas, así como, no sobrepasar el tiempo normal de estudios para garantizar un acceso rápido al mercado laboral que le permita responder a las necesidades y expectativas económicas de sus hogares (Rodríguez Rios 2012).

El análisis de Rodríguez permitió concluir que las variables que más afectan la deserción (o probabilidad de graduarse) en la Escuela de Ingeniería son el género del estudiante (hombre o mujer), su desempeño en las ciencias básicas (coincidiendo con el estudio de Tinto (1975); el estrato socioeconómico, con quién vive actualmente el estudiante (la probabilidad de graduarse cuando el estudiante vive solo es apenas de 9.8 %) y su tipo de vivienda (si es arrendada o propia) en donde para este último escenario mejoraba la probabilidad de graduarse.

Amaya-Torrado et al (2014) propusieron un modelo predictivo de deserción para los estudiantes del programa de ingeniería de sistemas en la Universidad Simón Bolívar mediante árboles de decisión. Tuvieron en cuenta variables previamente determinadas de tipo demográficos como edad actual, ciudad de procedencia, estrato, sexo, ocupación, estado civil, nivel de estudios del padre y de la madre; variables económicas como valor de la matricula e ingresos, y de carácter académico como el semestre cursado, la jornada académica, materias cursadas, materias perdidas y promedio de notas semestrales. Si bien el volumen de desertores no es alto (5), el método utilizado clasificó adecuadamente a 4 de los 5 desertores

(Amaya Torrado et al. 2014).

Por otra parte, Arismendi et al (2018) analizaron la probabilidad de deserción temprana de los estudiantes que ingresaron a la Universidad de los Llanos entre las cohortes 2015-2 a 2018-1 mediante un modelo de regresión logística compuesto por trece variables iniciales identificadas de orden socioeconómico y académico que no difieren significativamente de las tomadas en los trabajos citados previamente; de estas variables siete resultan significativas (sexo, tipo de colegio, repitencia de años escolares, realización de estudios previos, la situación de los padres y las facultades del programa académico). El modelo propuesto clasificó correctamente al 54 % de los desertores (sensibilidad) y clasificó correctamente al 76 % de quienes permanecen vigentes (especificidad) para un total de predicciones correctas del 67 %. Los autores concluyen que un buen puntaje en la prueba Saber 11, el ser mujer, no haber reprobado años durante el bachillerato, el haber cursado estudios superiores previamente y si los padres conviven, disminuye la probabilidad de deserción temprana (primeros cuatro semestres) (Arismendy Fuentes et al. 2018).

Fandiño Benavides (2022), propuso comparar la regresión logística, máquinas de soporte vectorial, árboles de decisión, Bosques aleatorios y KNN (k vecinos más cercanos) para la predicción de deserción en los estudiantes de pregrado de la Fundación Universitaria Konrad Lorenz. Para ello utilizó datos académicos como histórico de notas, promedio semestral, promedio acumulado, asignaturas suspendidas, créditos del plan de estudios y graduación. La comparación de las técnicas utilizadas se basó en criterios como el nivel de accuracy, F1-score, AUC y Kappa. Así, los mejores modelos en orden de precisión (clasificación adecuada de los desertores y no desertores) fueron: Bosques aleatorios (+97 %), máquinas de soporte vectorial (+90 %), el modelo logístico (+82 %), y KNN (+80 %). Conlcuye el autor que el puntaje de la prueba de estado Saber 11, el promedio general de notas y la nota en las materias asociadas a las matemáticas resultaron ser las más significativas en la explicación de la deserción para todos los programas académicos (Fandiño Benavides 2022).

Frente a casos aplicados en otros países, Pérez y Rojas (2020) utilizaron máquinas de soporte vectorial con el fin de analizar los patrones de los estudiantes desertores en la Universidad Tecnológica del Perú entre 2017 y 2019, a partir de las notas obtenidas en las primeras evaluaciones de sus carreras junto con las variables sociodemográficas y económicas de estos. La técnica utilizada permitió clasificar correctamente al 90 % de los alumnos hayan sido desertores o no (Perez Bedia & Rojas Segovia 2020).

A partir de las investigaciones realizadas y como objetivo de este trabajo se aplicará un modelo logístico, para identificar e interpretar los factores que inciden en el riesgo de deserción de los beneficiarios de "Jóvenes a la U" y, como complemento a este análisis se aplicará un árbol de decisión, un bosque aleatorio, un clasificador por impulso de gradiente y una red neuronal. Se pretende comparar las métricas de desempeño de estos modelos, las variables que identifican como influyentes en el riesgo de deserción y determinar el más pertinente para la predicción del riesgo de ser o no desertor. Por ello, se describe el alcance de las técnicas de machine learning a utilizar:

2.2. Sobre el Machine learning

Traducido al español como "Aprendizaje de máquina" y derivado de la inteligencia artificial, se centra en el uso de los datos y de algoritmos para simular la forma de aprendizaje de los seres humanos. El desarrollo de estos algoritmos ha impactado sectores económicos, por ejemplo, el uso de estos para perfilamiento de clientes y ofrecimiento de productos hasta las recomendaciones de series y películas en plataformas digitales. Su uso se ha extendido a campos como la medicina, la biología, la geología entre otras ciencias y ha permitido innovar en procedimientos, diagnósticos y análisis de los millones de datos que se producen cada día en el mundo.

Dentro del Machine Learning se encuentran los modelos o algoritmos de clasificación cuyo aprendizaje se da a través del entrenamiento con datos, de manera que, su implementación represente una mejora sustancial en tiempos y precisión de procesos de la ciencia o el campo de conocimiento. Para efectos de este trabajo y teniendo en cuenta que la variable de interés se caracteriza por ser binaria, se hace mención

de los modelos de clasificación de respuesta binaria más representativos, considerados en la literatura de deserción y factibles de implementar.

2.2.1. Modelos logísticos

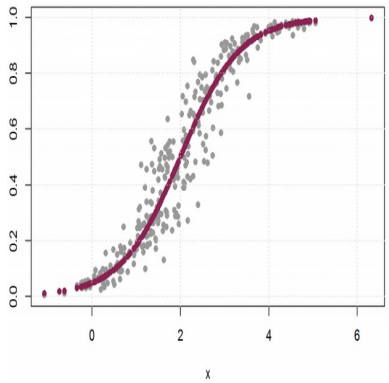
Son una extensión de los modelos lineales tradicionales. Esta técnica del machine learning permite analizar variables binarias a partir de variables explicativas que no necesariamente siguen una misma estructura (puede combinar variables continuas y categóricas) y, permite predecir la probabilidad de que la variable binaria tome un valor particular, es decir, las probabilidades de ocurrencia de un evento.

Esta técnica utiliza la función logística (o función sigmoide) para modelar la relación entre las variables independientes y la probabilidad de que la variable dependiente pertenezca a una clase. Así, el objetivo del entrenamiento de este modelo es encontrar los valores óptimos para los coeficientes que maximizan la verosimilitud de los datos observados, por lo que, una vez entrenada, la regresión puede predecir la probabilidad de pertenencia a las clases de interés. Si la probabilidad predicha es mayor que un umbral establecido, el modelo clasifica la observación como perteneciente a alguna de las clases. Algunas de las ventajas de la Regresión Logística son:

- **Interpretación**: Los coeficientes del modelo proporcionan información sobre la dirección y la fuerza de la relación entre las variables independientes y la probabilidad de pertenencia a una clase. Así mismo, el resultado del modelo es una probabilidad, lo que facilita su interpretación.
- **Robustez**: La regresión logística es robusta y eficaz, incluso cuando las suposiciones de normalidad y linealidad no se cumplen completamente.

La regresión logística se denomina función sigmoide porque refleja la evolución a través del tiempo de una curva de aprendizaje del modelo sobre los datos analizados y predichos, de tal forma que, como se observa en la figura la función tiene forma de "S" que muestra el tránsito temporal del aprendizaje con niveles bajos al inicio de esta, luego presenta un crecimiento acelerado para finalmente llegar a un punto "estacionario" de predicción.

Figura 1: Función logística



Fuente: www.statdeveloper.com (2022)

Ahora bien, el modelo logístico como se indicó previamente, ha sido ampliamente utilizado en los análisis de deserción universitaria. Esta técnica del machine learning permite analizar variables binarias a partir de variables explicativas que no necesariamente siguen una misma estructura. Este modelo permite predecir la probabilidad de que la variable binaria tome un valor particular, así, dado que la variable de

interés en este trabajo se caracteriza por ser dictotómica (estudiante desertor o no) es posible utilizar el modelo de regresión logística para ese análisis como se ha aplicado en los diferentes estudios de casos mencionados.

2.2.2. Árboles de decisión

De acuerdo con James (2013) los árboles de decisión pretenden clasificar adecuadamente una observación según reglas que se establecen o identifican (por el árbol) para los datos, así, estas reglas terminan guiando a través de "ramificaciones" hasta la ubicación de la observación en las "hojas" o segmento correspondiente. También se puede representar como un diagrama de flujo para la toma de decisiones. En todo caso, el árbol de decisión analiza las variables independientes suministradas en el conjunto de datos y su relación sobre la variable dependiente. (James et al. 2013)

Es importante mencionar que entre más grande sea el árbol de decisión más compleja resulta ser la clasificación o segmentación de los datos de interés, en ese sentido, el principio de parsimonia es aconsejable en esta técnica. El uso de los árboles conlleva la determinación de hipeparámetros que establecen la profundiad (extensión), cantidad de muestras con las cuales las ramificaciones o nodos tomarán la decisión de clasificación y el peso que cada muestra tendrá dentro del árbol para inclinar la decisión a la hoja o resultado correspondiente. Analicemos el alcance de cada uno:

- **Profundidad**: se refiere al número máximo de niveles que este puede tener, a mayor profundidad (o altura si se piensa en un árbol natural) más complejo será el árbol y más características de los datos podrá capturar, sin embargo, una profundidad excesiva puede llevar al sobreajuste porque el árbol se adaptaría o capturaría sesgadamente las características que determinan el comportamiento de la variable de interés en los datos de entrenamiento.

- **Cantidad mínima de muestras**: en cada nodo del árbol (en donde nacen y se dividen las ramas) se debe establecer un número mínimo de registros que se requieren para que el árbol pueda realizar una división. Un valor alto puede prevenir el sobreajuste del modelo al evitar divisiones con pocas muestras representativas que no permitan la identificación adecuada de las clases o resultados de interés.

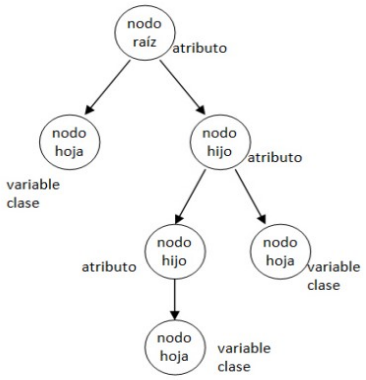
- **Peso mínimo de las muestras**: el árbol proporcionará un peso o importancia relativa las muestras muestras o registros, de manera que, si la suma de estos pesos es menor que el peso mínimo establecido entonces el nodo o inicio de ramificación donde se ubican las muestras no se dividirá más y se convertirá en un nodo hoja o de resultado.

Estos hiperparámetros trabajan en conjunto y no de manera secuencial. Así, los arboles de decisión han tenido amplias aplicaciones en diferentes áreas de conocimiento, por ejemplo, Barrientos et al (2009) los aplicaron como técnica de diagnóstico médico para el cáncer de seno, conjunto de datos que contempló 692 pacientes con con 12 variables correspondientes a sintomatologías diagnosticadas. El árbol logró clasificar adecuadamente al 94 % de los pacientes.

Cardona (2004) utilizó los árboles de decisión en modelos de riesgo crediticio para entidades financieras, así, tomando atributos como la cartera comercial, hipotecaria, de consumo, los bonos, acciones e inversiones de la entidad, inmuebles, el riesgo legal, operativo y de liquidez, con el uso del árbol de decisión logró determinar los clientes que presentaban riesgo de pago para la entidad.

Timarán, Caicedo e Hidalgo (2019) buscaron predecir el buen o mal desempeño en la prueba Saber 11 mediante factores académicos, individuales y socioeconómicos de estudiantes de bachillerato de colegios colombianos para los años 2015 y 2016. Se aplicaron árboles para un conjunto de datos con 1'300.000 registros y 49 variables de análisis. los árboles implementados clasificaron correctamente en promedio al 61 % de los estudiantes con buen o mal desempeño en la prueba Saber 11.

Figura 2: Ejemplo de un árbol de decisión

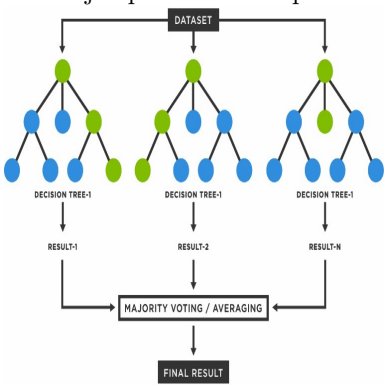


Fuente: Barrientos et al (2009)

2.2.3. Bosques aleatorios

Es una técnica que implica el uso de los árboles de decisión en masa. Es un algoritmo que genera y combina varios árboles de decisión encontrando características comunes para la clasificación de las observaciones de una manera más robusta y eficiente. Así, se espera que varios árboles tengan mejores resultados (en el bosque aleatorio) que un árbol de decisión individual tal y como se evidencia en la figura a continuación:

Figura 3: Ejemplo de un bosque aleatorio



Fuente: www.tibco.com (2022)

Los bosques aleatorios se desarrollan en tres pasos esenciales:

1. Se crea una muestra aleatoria (con reemplazo) del conjunto de datos de entrenamiento para un conjunto de árboles individuales que conformarían el bosque. Cada árbol se construye de manera independiente utilizando el proceso convencional.

2. Una vez construidos todos los árboles, cada uno puede realizar predicciones individuales basadas en las características del conjunto de datos (variable dependiente e independientes).

3. Se realiza una predicción combinada utilizando los árboles individuales para determinar la clasificación final de la variable u observación de interés. Así, la predicción combinada ayuda a mejorar la precisión y la capacidad del modelo al reducir la influencia de variaciones aleatorias presentes en los datos de entrenamiento.

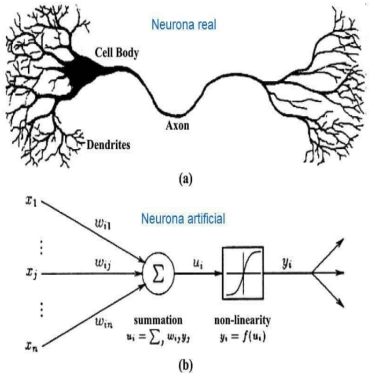
Céspedes (2022) implementó seis bosques aleatorios para predecir la anemia en niños del Perú; para ello tomó información demográfica y de salud familiar de la encuesta realizada por el instituto nacional de estadísticas del Perú entre los años 2015 y 2019. Los bosques se construyeron con 300 árboles de decisión (sin una justificación clara del valor). Los resultados de sensibilidad y especificidad indican un acierto promedio del 65 % de las observaciones sometidas a testeo.

Torres et al (2023) modelaron mediante bosque aleatorio el diagnóstico de Covid-19 a través de biomarcadores genéticos de 158 pacientes sospechosos de covid-19. Las métricas de desempeño AUC, F1-SCORE y precisión (91 %, 92 % y 89 %) permitieron concluir que la técnica era adecuada para este tipo de análisis.

2.2.4. Redes neuronales

Son una técnica estadística de aprendizaje automático la cual busca emular el comportamiento de las neuronas del cerebro humano en la comprensión y análisis de información. Así, las redes neuronales pretenden encontrar patrones y asociaciones en un conjunto de datos, aprender de este, y generar un resultado coherente para su interpretación. El aprendizaje de estas redes puede ser supervisado o profundo (deep learning), el primero implica (al igual que los árboles, modelo logístico y bosque aleatorio) suministrar la variable de interés y las variables independientes, de manera que, se indica el camino a seguir o el tipo de predicción a realizar (se supervisa el ejercicio), por su parte, el deep learning pretende que las redes neuronales identifiquen por su cuenta la correlación de un conjunto de variables.

Figura 4: Ejemplo de una red neuronal



Fuente: www.platzi.com (2023)

Las redes neuronales se organizan en "capas" o niveles que desarrollan el aprendizaje; así, el modelo más sencillo contempla tres capas: la primera (de entrada) contiene los datos o variables del conjunto de datos. La segunda capa (la oculta) le asigna a cada variable de entrada un peso aleatorio e interactúa con los datos de manera iterativa hasta lograr un punto de convergencia en el aprendizaje o interacción de estos, es decir, en esta capa se genera el aprendizaje de las "neuronas". Posteriormente se ejecuta una "función de activación" que sobre los pesos ponderados permitirá obtener el resultado de la capa final o de salida.

Las redes neuronales se pueden clasificar según la característica de interés:

- Número de capas (monocapas y multicapas) la primera es la estructura más sencilla de una red neuronal como se ha explicado previamente, por su parte, la multicapa contempla varias capas ocultas las cuales pueden o no interactuar entre sí.
- Conexiones: se tienen conexiones recurrentes en las cuales la red neuronal se "alimenta" nuevamente con la información generada por sus capas y/o la alimentación se da entre capas, de manera que,

estas conexiones generan un aprendizaje más robusto al producirse un efecto de "memoria" sobre los datos; de hecho este tipo de conexión es la más utilizada en la implementación de una red. Por otra parte, las conexiones no recurrentes (usadas en el nacimiento de las redes neuronales) trabajan en un solo sentido, carecen de efecto memoria y aunque sus resultados pueden ser adecuados, no logra el mismo desempeño que una conexión recurrente.

La arquitectura de la red neuronal define en gran medida su funcionalidad. Actualmente se cuenta con redes "transformers" que usadas por inteligencias artificiales como ChatGPT que, a través de este tipo de redes pueden hacer la interpretación de lenguajes naturales y su contexto. Por su función, estas redes consumen una gran cantidad de energía y recursos físicos para su adecuado desarrollo. Por otro lado, se tienen las redes convuncionales, utilizadas para el reconocimiento facial y en tareas como la vigilancia por medio de cámaras inteligentes que han sido implementadas en las grandes ciudades del mundo para la lucha contra el crimen. También estas redes se han implementado para la conducción autónoma de automóviles. Otro tipo de redes comúnmente utilizadas son las "siamesas" que permiten hacer comparaciones y encontrar similitudes, su uso se ha dado en el reconocimiento facial (para desbloquear telefónos) o para verificar la autenticidad de documentos al reconocer el texto, sellos o características únicas que permitan su validación.

El proceso de entrenamiento de cada tipo de red neuronal dependerá del modelo que más se ajuste a su propósito, a la estructura de los datos y el análisis deseado, es así como, el uso de algortimos para implementar dichas redes parte de una variada gama de estos, entre los más comunes se encuentran:

- Descenso por gradiente: consiste en un algoritmo de optimización que pretende que la red neuronal alcance el punto mínimo de error en su proceso de aprendizaje. La minimización del error se da entonces en los resultados generados. Este algoritmo se ha aplicado también en regresión lineal, logística y máquinas de soporte vectorial.
- Sigmoide: al igual que la regresión logística (y en forma de S) este algoritmo contempla la evolución del aprendizaje a través del tiempo hasta encontrar un punto estacionario en donde se minimiza el error y se mejora la predicción del modelo implementado.
- Propagación hacia atrás: este algortimo implica el uso de una red multicapa al pretender minimizar la tasa de error, remitiendo la información del error en la capa de salida nuevamente al inicio de la red o a la capa oculta, logrando así una re ponderación de los pesos que calcula la red y por tanto una mejora en el resultado obtenido.

Las redes neuronales también contemplan funciones de pérdida cuyo uso depende del tipo de problema (regresión, clasificación binaria, clasificación multiclase, etc.). Así, el objetivo del entrenamiento de la red es minimizar la función de pérdida para que las predicciones se acerquen lo más posible a los valores reales. En modelos de clasificación binaria se usan comúnmente las siguientes funciones de pérdida:

- **Entropía Cruzada Binaria (Binary Cross-Entropy)**: mide la discrepancia entre la distribución de probabilidad predicha por el modelo y la distribución de probabilidad real de los datos.
- **Entropía Cruzada Categórica (Categorical Cross-Entropy)**: se utiliza en problemas de clasificación multiclase, donde hay más de dos clases posibles.
- **Entropía Cruzada Categórica Esparsa (Sparse Categorical Cross-Entropy)**: se utiliza cuando las etiquetas están codificadas como enteros (por ejemplo píxeles en imagenes clasificados de 0 a 9). Esta función de pérdida acepta etiquetas enteras y las convierte internamente en vectores one-hot (1 y 0) antes de calcular la pérdida.

Gonzalez et al (2014) realizaron la aplicación de un algoritmo de redes neuronales retro propagadas con el fin de analizar su aporte en la clasificación de una población de potenciales deportistas encuestados. El

modelo pretendía establecer si el individuo era hombre, mujer o sin respuesta a partir del tiempo dedicado al deporte, el tiempo usado para responder la encuesta y variables sociodemográficas recolectadas. Al algoritmo implementado se le incluyó la técnica "cross validation" en 10 segmentos. Los resultados arrojaron una tasa de aciertos (sensibilidad) de la red neuronal de 80 % para quienes habían respondido ser hombres o mujeres, para el caso de quienes no indicaron su sexo el modelo no fue efectivo. Los autores atribuyen este último resultado a la baja cantidad de personas que no indicaron su sexo (3 % de los encuestados). (Gonzalez.et.all 2014)

Una de las conclusiones más interesantes de este análisis aplicado es que los autores consideran que las redes neuronales reponden adecuadamente ante datos con cierto sesgo en la información (dado que se trataba de una encuesta en la que no necesariamente había certeza de la respuesta adecuada, por ejemplo, en la edad se suele presentar un sesgo por parte de las mujeres), así mismo, encuentran que el uso de datos complementarios (tiempo usado en el diligenciamiento de la encuesta) puede resultar en información útil para el desempeño del modelo propuesto. Este tipo de información adicional no siempre es captada eficazmente por otros algoritmos de aprendizaje automático según los autores y por consiguiente las redes neuronales representan la oportunidad en el uso de información variable.

Siguiendo los resultados previos Gould (1993) utilizó las redes retro propagadas para predecir mapas de incidencia del SIDA en Estados Unidos (predecir el número de casos de SIDA). Después de entrenar la red con información de los años 1986 a 1989, comprobó su eficacia con 39 situaciones reales como punto de control del aprendizaje de la red y por tanto su generalización. Destacó Gould que los mejores resultados se obtuvieron cuando se utilizaban variables de entrada "dinámicas" (efectos epidemiológicos del SIDA) en comparación con redes construidas con variables "estáticas" (demográficas), es decir, las redes neuronales logran captar mejor información de datos complejos, de manera que, su aprovechamiento debe contemplar este tipo de información por sobre asociaciones que se consideren usualmente más lógicas y que con modelos más sencillos pueden tener resultados aceptables.(Peter Gould n.d.)

Por otro lado, McGinnis (1994) aplicó redes neuronales para la predicción de nevadas en la cuenca del Río Colorado en Estados Unidos, utilizando 47 variables como la presión atmosférica, la temperatura, número de precipitaciones, velocidad del viento, entre otras, concluyendo que estas logran predecir las nevadas con una precisión del 70 % comparado con métodos de regresión lineal en donde el porcentaje de aciertos se reduce aproximadamente a la mitad.

Las redes neuronales podrán tener tantas capas profundas como se considere necesario; recientes desarrollos como los algoritmos de reconocimiento de voz, datos no estructurados (redes sociales), perfilamientos publicitarios en internet, algoritmos musicales, entre otros, han sido desarrollados gracias a las redes neuronales dadas las complejidades de esos tipos de información y que los modelos convencionales de machine learning tendrían dificultades de procesamiento. Particularmente cuando se trata de datos no estructurados o lo que es lo mismo, que no siguen una tendencia de comportamiento fácilmente predecible y que solamente son comprensibles desde la óptica humana.

2.2.5. Gradient Bossting Classifier o clasificador de refuerzo por gradiente

Propuesto por Jerome Friedman (2000) aunque inicialmente explorado por Shapire en 1990 y Freund en 1995, es una técnica que puede combinar las virtudes de los árboles de decisión y el descenso gradiente de las redes neuronales que hace predicciones en secuencia y combina múltiples modelos "débiles" (como los denominó Friedman) trabajando con sus errores residuales para generar un proceso de aprendizaje acumulado de dichos errores y, por tanto, generando un resultado más fiable si se habla en términos de modelos de clasificación. En esencia funciona de la siguiente manera:

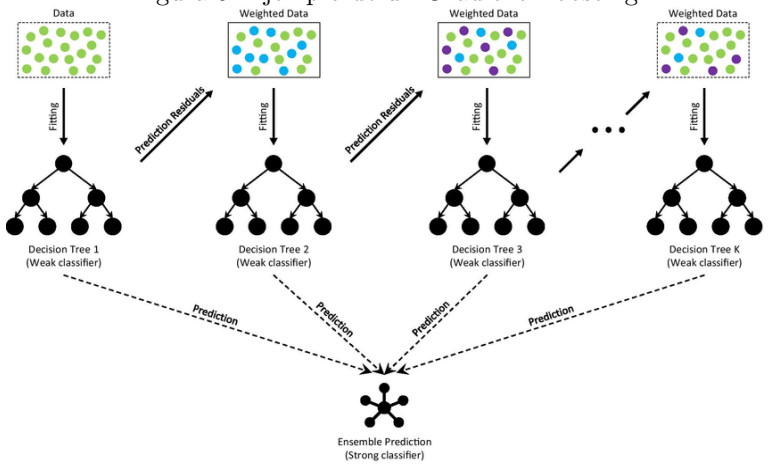
1. Inicia con un modelo simple (como un árbol de decisión usado por defecto, aunque puede utilizar otros modelos), lo entrena y genera unos errores residuales.

2. Construye un segundo modelo (segundo árbol) que predice los errores residuales del primer modelo.

3. Se combinan ambos modelos y se repite el proceso de modelamiento (con nuevos árboles), prediciendo los errores de manera acumulativa hasta alcanzar un número determinado de modelos o por un parámetro dado de optimización.

El resultado de este proceso es un aprendizaje de cada árbol sobre los errores cometidos por su predecesor y, cada árbol mejorado aportará de manera ponderada a la predicción final de la variable de interés, logrando así reducir el error conjunto de los modelos (árboles) desarrollados y alcanzando un mejor grado de precisión de la predicción en comparación con los árboles de decisión individuales.

Figura 5: Ejemplo de un Gradient Boosting



Fuente: www.researchgate.com (2023)

El método del gradiente es considerado oportuno para el manejo de datos heterogéneos, inclusive contribuye para la reducción de los efectos derivados de la multicolinealidad si el conjunto de datos lo presenta.

Nicolás Rugeles comparó el desempeño de clasificación del método del gradiente con respecto a un bosque aleatorio con hasta 100 árboles. El gradiente tuvo un desempeño ligeramente superior al bosque aleatorio con un modelo que contempló 5 y 10 árboles con una profundidad máxima de 3 niveles para cada uno. Es decir, con un modelo más sencillo que aprende de los errores residuales se logró un mejor desempeño de clasificación. (Rugeles 2022)

Pablo Flores y Giorgina Congacha aplicaron la regresión logística y los árboles de decisión para predecir la incidencia de desnutrición crónica infantil en Ecuador, sin embargo, las variables utilizadas si bien teóricamente eran pertinentes para el análisis (grado de escolaridad, grupo étnico, tipo de parto, mes de nacimiento, vacunación de la madre, entre otras) los modelos empleados arrojaron un desempeño "muy baj" (52 %) al momento de los resultados de la clasificación. Los autores intuyeron que los modelos no eran suficientemente consistentes con la captación de la información en datos heterogéneos y por consiguiente propusieron el uso del algoritmo del gradiente como solución. Así, este modelo logró clasificar adecuadamente al 62 % de los niños en condición de desnutrición crónica.(Flores-Congacha 2021)

Vinicio Andrade y Pablo Flores aplicaron árboles de decisión, bosques aleatorios y el método de gradiente para predecir la satisfacción laboral de un conjunto de empresas en Ecuador. Las variables analizadas consideraron la satisfacción laboral, el sexo, la edad, estado civil, años de trabajo, duración del último contrato, tamaño del establecimiento y horas dedicadas al trabajo al día. Si bien los modelos aplicados obtuvieron tasas similares de aciertos (30 %) los autores destacan que el gradiente fue el modelo con menor tiempo de desarrollo y menor consumo de recursos computacionales. Mientras el bosque aleatorio demoró 52 segundos y el árbol 34, el gradiente tomó apenas 12 segundos. En el mismo orden se presentó el consumo de memoria ram del equipo utilizado. (Andrade-Flores 2018)

Mariano Romero, Pedro Carmona y José Pozuelo aplicaron el método del gradiente extremo para predecir el fracaso empresarial de las cooperativas españolas. Este modelo es una versión mejorada del gradiente común pero su costo computacional es mucho mayor al contemplar imputación de datos, regularización y desbalance de clases durante el desarrollo del modelo. Los resultados fueron muy favorables con una tasa de 100 % de aciertos en datos de entrenamiento y del 86 % en datos no vistos por el algoritmo.(Romero-Carmona-Pozuelo 2021)

2.3. Sobre el tratamiento de bases de datos

El análisis de datos debe contemplar una serie de procesos que confluyan para la obtención de cantidad, calidad y coherencia de la información para una adecuada aplicación de la estadística y por tanto de los resultados de dichos análisis. Así, particularmente el tratamiento de una base de datos implica conocer la estructura de la información, el contexto de su origen, los desafíos a priori identificados con su uso original y la aplicación de las diferentes técnicas que la estadística provee para superar umbrales deseables de la cantidad, calidad y coherencia de la información.

Como se notará en el desarrollo de este documento, se hace necesario el uso de técnicas estadísticas para el tratamiento de datos faltantes, así como, el uso de técnicas para el balance de clases en la cantidad de datos de las variables de interés cuyo desequilibrio afectaría no solo la confiabilidad de los modelos propuestos de análisis de deserción sino naturalmente la insterpretación de sus resultados.

2.3.1. Técnicas para la imputación de datos

La ausencia de información en un conjunto de datos representa un desafío técnico y operativo para su análisis. Por el lado técnico, se debe identificar el origen y el contexto de la ausencia de la información y en consecuencia se debe determinar si el elemento técnico es subsanable o no; en este último escenario es donde surge la concepción de técnicas de imputación (completitud de los datos) con diferentes grados de complejidad y de acuerdo con la naturaleza de los datos. Los datos faltantes pueden presentar diferentes patrones:

- Completamente al azar: la probabilidad de que un valor falte es la misma para todas las observaciones. En este caso, la ausencia de datos no está relacionada con ninguna variable observada o no observada. Este patrón es el más deseable, ya que no introduce sesgo sistemático en los datos.
- Al azar: la probabilidad de que falte un valor está relacionada con otras variables observadas, este patrón es común en encuestas donde la falta de respuesta puede depender de otras variables observadas (por ejemplo, se acaba el tiempo para responder la encuesta o el sujeto decide abandonarla por alguna circunstancia).
- No al azar: la probabilidad de que falte un valor está relacionada con el valor real que falta, incluso después de considerar otras variables observadas. Este es el patrón más problemático, ya que la falta de datos puede estar influenciada por información no observada, lo que podría sesgar los resultados si no se maneja adecuadamente (por ejemplo, se consulta la orientación sexual y sobre todo si no es anónima conllevaría a un sesgamiento).

Los métodos de imputación utilizan la información disponible en el conjunto de datos para asignar un valor plausible al dato faltante y pretendiendo no alterar la distribución de los datos, las asociaciones de las variables y que los datos imputados sean consistentes con su conjunto. Entre los métodos de imputación más conocidos se encuentran los de regresión, donantes, vecinos más cercanos, clustering y algoritmos iterativos.

Los modelos de regresión se aplican principalmente en casos con una baja tasa de datos faltantes que conduzcan a su estimación. Algunas de las técnicas comunes incluyen regresión lineal, regresión múltiple y métodos más avanzados como regresión robusta o regresión con técnicas de aprendizaje automático.

El método de "donantes", por su parte permite la asignación de un valor plausible al dato faltante tomando de otro individuo o registro el valor existente. Este registro o individuo comparte características similares al del registro o individuo con el dato faltante. En donantes se encuentran los métodos de "Hot-deck aleatorio" y "Hot-deck secuencial", el primero selecciona al azar el dato plausible de un conjunto de individuos con las mismas características auxiliares que presenta el receptor (o individuo con dato faltante); el secuencial imputa el dato faltante tomando el último valor disponible previo del individuo que tenga las mismas características del receptor.

"Vecinos más cercanos" define una distancia sobre las variables independientes de posibles donantes y toma como valor plausible el del vecino más cercano o en algunos casos la mediana de estos valores e imputa el dato faltante en la variable independiente correspondiente del receptor.

Finalmente, los métodos de imputación por algoritmos iterativos responden como tratamientos complejos en ausencias considerables de datos para los cuales los métodos previamente descritos no funcionarían de la manera esperada por el investigador y según el contexto de los datos y del problema de análisis. Los algoritmos iterativos generan "m" valores plausibles para cada valor faltante. Uno de estos es el algoritmo "EM" o esperanza de maximización que iterativamente calcula las medias de un conjunto de observaciones hasta encontrar un valor de convergencia, el cual puede ser considerado por el investigador como posible reemplazo del valor faltante. Resulta en un método más elaborado a la simple imputación con la media o mediana de los datos, no obstante, este método presenta algunas desventajas asociadas a la sensibilidad frente a datos atípicos y presenta convergencia lenta en conjunto de datos grandes y complejos.

Un famoso algortimo de imputación iterativa es "MICE" (Multiple Imputation by Chained Equations) el cual imputa valores faltantes en varias etapas, una variable a la vez y mediante ecuaciones encadenadas. Su funcionamiento se resume de la siguiente manera:

1. **Inicialización**: inicia con la imputación de valores faltantes utilizando un método inicial, como la imputación media o la imputación por regresión simple, proporcionando una base para comenzar el proceso de imputación iterativa.

2. **Iteración**: MICE realiza varias iteraciones, y en cada iteración, se imputa un conjunto de valores faltantes para una variable específica. Este proceso se repite para cada variable con valores faltantes.

3. **Ecuaciones encadenadas**: En cada iteración, MICE utiliza modelos estadísticos para imputar los valores faltantes de una variable dada. Estos modelos pueden ser regresiones, modelos lineales, modelos de regresión logística u otros que se ajustan a la naturaleza de los datos. Se construye entonces un modelo para cada variable con valores faltantes, condicionado a las demás variables en el conjunto de datos.

4. **Actualización de las imputaciones**: Después de imputar valores para todas las variables una vez, se actualizan las imputaciones y se repite el proceso en iteraciones adicionales. Este proceso de iteración y actualización continúa hasta que las imputaciones convergen y se estabilizan.

5. **Combina las imputaciones**: Una vez completadas todas las iteraciones, MICE produce múltiples conjuntos de datos imputados, cada uno de los cuales refleja una posible imputación de los valores faltantes. Estos conjuntos se combinan para proporcionar una estimación global de la variabilidad debida a la imputación.

MICE es flexible y puede manejar diferentes tipos de variables y distribuciones. Además, es capaz de manejar patrones de datos faltantes tanto completamente al azar como no al azar.

Puerta (2002) resume los criterios para considerar y seleccionar la técnica adecuada para imputar, así mismo, los pasos para realizar la imputación y los ánalisis de evaluación de esta. Iniciando con los criterios de consideración de técnicas refiere:

- **Tasas de no respuesta y exactitud necesaria**: cuando el porcentaje de no respuesta es alto en una base de datos, se considera que no hay confiabilidad en los resultados que se obtengan con el análisis de esta base.
- **Tipo de variable a imputar**: si es continua (tener en cuenta el intervalo) y si es cualitativa, las categorías de las variables.
- **Parámetros que se desean estimar**: si se pretende conocer sólo valores como la media, se pueden aplicar métodos sencillos como imputación con la media o moda, sin embargo, puede haber subestimación de la varianza. En caso de que se requiera mantener la distribución de las variables y sus asociaciones entre las distintas variables, se deben emplear métodos más elaborados aplicando imputación de todas las variables faltantes.
- **Información auxiliar disponible**: con ella podemos deducir información de los valores faltantes o hallar grupos homogéneos respecto a una variable auxiliar que se encuentre altamente correlacionada con la variable a imputar, y de esta manera encontrar un donante adecuado que sea similar al registro receptor.

A partir de lo anterior, establece cinco pasos para imputar:

- **Paso 1**: recopilar y validar toda la información auxiliar disponible que pueda ser de ayuda para la imputación.
- **Paso 2**: estudiar el patrón de los datos faltantes. Posteriormente se observa si hay un gran número de registros que simultáneamente tienen no respuesta en un conjunto de variables.
- **Paso 3**: se seleccionan varios métodos de imputación posibles y se contrastan los resultados.
- **Paso 4**: se calculan las varianzas para los distintos métodos de imputación seleccionados con el objetivo de obtener estimaciones con el mínimo sesgo y la mejor precisión.
- **Paso 5**: se concluye a partir de los resultados obtenidos.

Finalmente Puerta propone algunas validaciones de las imputaciones realizadas, así, denomina "precisión de la predicción": debe resultar un valor imputado que sea lo más cercano posible al valor real. La "precisión de distribución": la imputación debe preservar la distribución de los valores reales y conlcluye indicando que la imputación debe producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales. (Goicoechea 2002)

Del Turco (2022) comparó la imputación de valores con diferentes métodos (media, hot-deck, K-NN, por regresión e imputación múltiple con el algoritmo MICE) a un conjunto de 8 bases de datos relacionadas con el sector salud y disponibles en el repositorio digital de la Universidad de California. Los conjuntos de datos imputados son sometidos a análisis a través de un modelo de regresión lineal, un modelo de vecinos más cercanos (para clasificación) y un modelo de máquina de soporte vectorial (para clasificación). En dichos modelos la autora decide comparar los criterios: raiz del error cuadrátivo medio (RMSE) y el coeficiente de determinación o R cuadrado. Así, para el modelo de regresión lineal encuentra que el método Hot-Deck y la imputación múltiple con MICE logran mejorar el RMSE y el R cuadrado de los modelos respecto al conjunto de datos sin imputación y sobre el cual se hizo eliminación de registros con datos faltantes. Por su parte, con K-NN y la imputación por regresión se evidencia una nula o incluso desmejora en ambos criterios y, la imputación por media no presenta desmejora en ningún criterio.

Para la máquina de soporte vectorial encuentra que la imputación por MICE y K-NN presentan mejoras en RMSE y R cuadrado respecto de los datos sin imputación. Por su parte Hot-Deck y la imputación por media no muestran desmejora en los criterios. Por su parte para el modelo de vecinos más cercanos se presenta el mismo escenario que en máquina de soporte vectorial, así, MICE y K-NN evidencian los mejores resultados, mientras Hot-deck y la imputación por la media no presentan cambios para el RMSE

y el R cuadrado de los modelos. La autora concluye que a partir de las características de cada variable a imputar se deberá seleccionar el método más conveniente para esta y no generalizarlo al conjunto de datos, así mismo, indica que los resultados de su análisis no permiten afirmar de manera categórica que los métodos más sencillos y los más elaborados presenten los mejores resultados en la imputación, de manera que, su combinación o procesos intermedios pueden resultar factibles. (Turco 2021)

2.3.2. Técnicas de balanceo de clases o muestras

Algunas bases de datos contienen muestras no lo suficientemente representativas para la aplicación de algortimos de clasificación y, es este desbalance el que ocasiona que los modelos aplicados no logren los mejores resultados. Es así como, se construyeron técnicas estadísticas para mejorar el balance de los datos y que esto no influya negativamente en el entrenamiento de los modelos y por tanto en los resultados y su interpretación.

Los métodos para tratar el desbalanceamiento pueden contemplar la amplificación de la clase minoritaria o reducir la mayoritaria, sin embargo, cada una tiene implicaciones diferentes de acuerdo con el contexto de los datos y del problema a analizar, así mismo, se tienen métodos específicos en el contexto del machine learning.

Inicialmente el "subsampling" o submuestreo busca reducir el grupo mayoritario a partir del muestreo aleatorio simple o mediante la técnica de vecinos más cercanos (KNN); estos tipos de submuestreo, si bien permiten conservar el "comportamiento" de los datos en la clase mayoritaria, reducir sensiblemente esta (al contemplar el tamaño de la clase minoritaria) puede conllevar a la pérdida de información que sea determinante en las estimaciones de los modelos a aplicar y, por tanto, se subestime el efecto de alguna variable explicativa o en el peor de los escenarios descartar variables que no hayan resultado con significancia estadística dado el submuestreo aplicado.

El "oversampling" o sobre muestreo, por otro lado, consiste en la amplificación de casos de la clase minoritaria de manera directa (duplicación) o mediante la creación de muestras sintéticas más elaboradas, lo anterior, permitiría mejorar la proporcionalidad de la clase mayoritaria y minoritaria, sin embargo, la amplificación directa sería considerada para situaciones en las que la diferencia de clases no sea tan marcada y busque reforzar resultados adecuados ya obtenidos en previos análisis. Así, en la literarura se identifican técnicas como el SMOTE, SMOTE-ENN y ADASYN las cuales generan datos similares (sintéticos) a los originales para equilibrar las proporciones de clases y de esa manera los modelos a utilizar tendrán un mejor desempeño y por consiguiente conclusiones robustas.

SMOTE (Synthetic Minority Oversampling Technique) utiliza la técnica de interpolación buscando registros de la clase minoritaria y estableciendo distancias (como la euclidiana) entre estos. Luego, entre esas distancias crea las muestras sintéticas asignando valores plausibles sobre las características explicativas con la interpolación (la estimación de estos valores se logra con modelos lineales, polinómicos, KNN, entre otras).(Abella 2021)

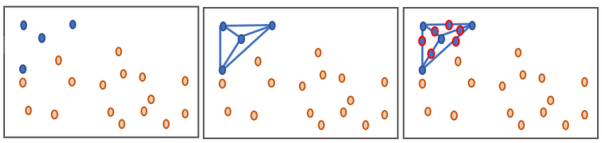


Figura 6: Funcionamiento SMOTE. Fuente: datascienciecampus.github.io

SMOTE-ENN por su parte complementa posteriormente SMOTE con la eliminación de datos atípicos o ruidosos a través de vecinos cercanos y ADASYN (Adaptative Synthetic Sampling) se enfoca en generar ejemplos sintéticos en las regiones donde los valores de las características explicativas son menos identificables para la clase minoritaria, es decir, es aplicable preferiblemente en bases de datos heterogéneas y de alto desbalance, sin embargo, SMOTE resulta en la técnica más común de uso para el tratamiento de

muestras desbalanceadas y sobre la cual profundizaremos a continuación.

Ventura et al (2023) utilizaron SMOTE como técnica de oversampling para aplicar un modelo de árboles de decisión y predecir el éxito del telemarketing bancario en Perú. Inicialmente la clase minoritaria (éxitos del telemarketing medido como la respuesta de los clientes a la campaña ejecutada) representaba poco más del 10 % de la clase mayoritaria. Así, los autores deciden elevar esta cifra al 80 % y dividen el estos nuevos datos en 80 % para entrenamiento del modelo y 20 % para testeo, encontrando que el modelo de árbol de decisión logra en diferentes métricas de evaludación (precisión, recall y F1 score) una tasa por encima del 80 % (el mejor caso correspondió a la precisión con una tasa del 86 % de aciertos).

Blagus y Lusa analizaron la conveniencia de utilizar SMOTE en bases de datos de diferentes tamaños y con desbalance de clases. En principio describen los supuestos teóricos de la técnica:

1. $$E(SMOTE) = E(X) \tag{1}$$
2. $$(Var(X_{smote}) = \frac{2}{3} Var(X) \tag{2}$$
3. SMOTE genera correlación entre los valores de las muestras sintéticas, pero no entre los valores de las variables independientes asociadas a cada valor sintético creado.
4. SMOTE modifica las distancias euclidianas de la clase minoritaria. Particularmente, estas distancias son menores en las muestras sintéticas que entre las muestras originales de la clase minoritaria.

Cada uno de los anteriores supuestos para los autores presentan consideraciones que se deben tener en cuenta frente al uso de SMOTE en modelos de clasfificación. Así, los autores sugieren que, si bien no hay un cambio en el valor esperado de la clase minoritaria, la varianza aumentada por SMOTE es más pequeña que la varianza de la clase minoritaria original, esto implica que, aunque el valor esperado es similar, hay menos dispersión en los valores de la clase minoritaria después de aplicar SMOTE, es decir, en el contexto de modelos predictivos, una menor varianza lleva a modelos más precisos y estables. Modelos con menor varianza tienden a ser menos sensibles a pequeñas variaciones en los datos de entrenamiento, lo que puede resultar en predicciones más consistentes. (Blagus-Lusa 2013)

Sin embargo, los autores indican que SMOTE puede tener un impacto negativo en clasificadores que utilizan varianzas específicas de clase (como el Análisis Discriminante Cuadrático), ya que la varianza más pequeña después de SMOTE puede llevar a estimaciones sesgadas. El análisis discriminante cuadrático es una técnica estadística utilizada para clasificar observaciones en grupos o categorías diferentes. Aunque SMOTE puede tener un impacto limitado en ciertos tipos de clasificadores, su efecto en modelos que dependen de estadísticas específicas, como varianzas de clases, podría necesitar consideraciones adicionales debido a las modificaciones introducidas por SMOTE en la distribución de las muestras.

Por otra parte, los autores resaltan que SMOTE puede presentar problemas para clasificadores que asumen independencia entre muestras. Específicamente, mencionan que esto puede ser un problema para clasificadores como la regresión logística penalizada o los métodos de análisis discriminante. El razonamiento detrás de esto es que SMOTE al introducir muestras sintéticas en el conjunto de datos para abordar el desbalance de clases, estas muestras sintéticas se generan mediante interpolación entre muestras existentes de la clase minoritaria. Esta introducción de "dependencia" entre las muestras sintéticas y las originales podría violar la suposición de independencia, por tanto, si los datos de interés no son independientes, SMOTE profundizaría la dependencia causando a los modelos un error de sesgamiento en sus resultados.

Finalmente, partiendo de los supuestos, los autores sugieren que la selección de variables para ajustar un modelo debe ser previa a la aplicación de SMOTE, teniendo en cuenta que, la creación de las muestras sintéticas con las consideraciones anteriormente mencionadas puede conllevar al sesgamiento en la selección de las variables por una "amplificación ficticia" de la correlación entre las variables explicativas y la de interés aun cuando esto no sea necesariamente cierto.

Blagus y Lusa probaron el impacto de SMOTE ajustando modelos de clasificación como vecinos más cercanos, bosques aleatorios, regresión logística con regularización lasso y con regularización Ridge, análisis discriminante lineal y cuadrático y el algoritmo de clasificación por medoides, utilizando tres bases de datos con información de marcadores genéticos para la detección del cáncer (Sotiriou con 7.650 variables, 99 observaciones y un desbalance del 40 %; Pittman con 12.625 variables, 158 observaciones y un desbalance del 30 %; Miller con 22.283 variables, 247 observaciones y un desbalance del 14 %), las bases tendrían entonces una dimensión de 757.000, 1'994.000 y 5'500.000 datos respectivamente.

Sometieron los modelos para la predicción del cáncer para cada base: sin ajustes de balance, con "undersampling" mediante muestreo aleatorio simple y con "oversampling" por SMOTE y concluyeron que la mayoría de los modelos mostraron ser sensibles al desbalance de clases, incluso cuando este desbalance era moderado, lo que refuerza la pertinencia de aplicar técnicas de balanceo para precisar la predicción.

El "undersampling" resultó en la técnica que generó el desempeño relativamente más bajo en todos los modelos, los autores lo aducen a la pérdida de información que pudo ser determinante para los modelos. Por su parte el oversampling por SMOTE presentó resultados disímiles entre los modelos y entre las bases de datos. Así, con la base de Sotiriou los autores resaltaron una mejora sustancial en términos de desempeño general (tomando como referencia el criterio AUC y G-mean) para casi todos los modelos, específicamente, Random Forest, Máquina de soporte vectorial, modelos logísticos con penalización y vecinos más cercanos, no obstante, resaltan los autores que el desempeño mejoraba sustancialmente cuando con SMOTE se elevaba la clase minoritaria al 40 %, y marginalmente o inclusive nulo cuando se elevó al 80 % y 100 %, lo que quiere decir que una tasa razonable de oversampling puede resultar más que suficiente para alcanzar un desempeño aceptable de los modelos de clasificación.

Finalmente, para las bases de datos de Pittman y Miller, a pesar de contar con un tamaño diferente, en la mayoría de los modelos SMOTE no tuvo mejor desempeño frente a no haber realizado balance (especialmente en la base de Miller), o inclusive en algunos modelos con undersampling su desempeño era más consistente (CART, másquinas de soporte vectorial y modelos logísticos con penalización). Otros escenarios destacables fueron los ligeramente más altos desempeños en modelos sin balanceo frente a SMOTE (vecinos más cercanos, análisis discriminante lineal y máquinas de soporte vectorial).

Concluyen Blagus y Lusa que SMOTE en términos generales aporta en la reducción del sesgo de clasficación de los modelos, no obstante, esto se da en bases de datos de baja dimensionalidad (hasta cientos de miles), ya que, con respecto a las bases de datos de alta dimensionalidad (millones de datos) la técnica no resulta efectiva.

SMOTE-ENN consiste en la combinación de dos métodos: SMOTE y ENN (vecinos más cercanos ajustados). En el primer paso, SMOTE se utiliza para generar instancias sintéticas en la clase minoritaria, de manera que, en el segundo paso, ENN se aplica para eliminar instancias ruidosas de ambas clases, lo cual ayuda a mejorar la calidad del conjunto de datos y a evitar el sobreajuste del modelo, es decir que, al equilibrar las clases mediante la generación de instancias sintéticas y la eliminación de instancias ruidosas, se mejora la capacidad de generalización del modelo de clasificación.

ENN proviene del concepto de "vecinos más cercanos" (la técnica de clusterización o clasificación comúmente conocida como K-NN) y en la cual se evalúa la similitud o distancia entre puntos de datos en un espacio dimensional, por tanto, las instancias similares serán agrupadas en el conjunto de "vecinos" con los cuales comparten características. Ahora bien, con ENN las instancias que no logren ser agrupadas serán descartadas del proceso de oversampling.

Al igual que SMOTE, SMOTE-ENN aborda tanto el sobreajuste como la falta de representación de la clase minoritaria en conjuntos de datos desequilibrados, no obstante, la eliminación de instancias mediante ENN puede resultar en una pérdida de información, especialmente si las instancias eliminadas son representativas en la variabilidad de los datos, así mismo, puede introducir cierta complejidad computacional adicional debido a la combinación de estos dos métodos.

Muntasir et al (2022), utilizaron SMOTE-ENN para medir el desempeño de diferentes tipos de clasificadores en la detección temprana de fallas cardíacas. La base de datos empleada contiene 299 registros

de pacientes de un hospital Pakistaní con 13 variables de análisis como la edad, la presión sanguínea, enfermedades de base, entre otras características o diagnósticos clínicos. Los autores decidieron aplicar bosques aleatorios, regresión logística, Naive Bayes Gaussiano, árboles de decisión, vecinos más cercanos (K-NN) y máquinas de soporte vectorial.

Los autores establecen tres escenarios: ajuste de modelos con datos originales y escalamiento de datos, escalamiento de datos y SMOTE-ENN y, por último, escalamiento de datos más SMOTE-ENN más optimización de hiperparámetros. Sobre la optimización, se implementa la técnica denominada "Grid search cross validation" consistente en la búsqueda de combinaciones de valores para los hiperparámetros de cada modelo pretendiendo encontrar los mejores valores según el objetivo planteado (si se desea priorizar la sensibilidad, la especificidad, la precisión, etc). Con lo anterior, los resultados (con Accuracy) en el primer escenario ubica en primer lugar a los bosques aleatorios (0.80), seguido por la regresión logística (0.80) y los árboles de decisión (0.73), los demás modelos obtienen un valor promedio de 0.51. En el segundo escenario, todas las métricas (AUC, f1-score, accuracy y recall) mejoran en todos los modelos, aunque, de nuevo en el destacan (con Accuracy) la regresión logística (0.92), bosques aleatorios (0.90) y árboles de decisión (0.90). Finalmente, en el tercer escenario y utilizando la misma métrica, hubo una ligera reducción en el desempeño de los modelos, así, bosques aleatorios obtienen un puntaje de Accuracy de 0.87, seguido de árboles de decisión con 0.86 y vecinos más cercanos con 0.78.

Los autores concluyen que en cuatro de las cinco métricas y en todos los escenarios los bosques aleatorios presentaron el mejor desempeño, destacando de manera particular que este se logra debido a la implementación de SMOTE-ENN y la optimización de los hiperparámetros del modelo.

3. Aplicación estadística del riesgo de deserción en "Jóvenes a la U"

Para cumplir con los objetivos del presente documento y a partir del análisis de la literatura especializada, se requirió a la Agencia Distrital para la Educación superior, la Ciencia y la Tecnología-Atenea (entidad gestora del programa) la base de datos de los beneficiarios de "Jóvenes a la U", anonimizada, con toda la caracterización individual sociodemográfica, socioeconómica y académica disponible (que incluyera la variable deserción).

Con ello, la entidad remitió una base de datos en formato Excel constituida por 20.230 registros (beneficiarios desde el 2021-2 al 2023-1) y con las siguientes variables: Convocatoria (cohorte de ingreso del beneficiario al programa (1 a 4): sexo (mujer/hombre); código SNIES y nombre del programa académico; área de conocimiento del programa académico; código SNIES y nombre de la institución de educación superior; sector de la IES (oficial y no oficial); edad del beneficiario (obtenida al momento de la remisión de la base); estrato socioeconómico (1 al 6); grupo/comunidad étnica; discapacidad (si o no); nivel SISBEN (A,B,C,D); localidad de residencia; puntaje global de la prueba Saber 11 y puntajes específicos en lectura crítica y matemáticas de la misma prueba; deserción (si o no, con corte de información al generarse la base de datos). Por su parte, la variable de interés (desertor) evidencia 237 registros, siendo el 1.17 % de beneficiarios con esta condición. Partiendo de lo anterior, a priori existe un desbalance entre la variable de interés (deserción) en el universo de beneficiarios, así mismo, la base presenta valores faltantes.

A continuación se realiza el preprocesamiento de la base de datos y se implementan las técnicas descritas en el marco conceptual para dar solución a estos desafíos estadísticos, de tal manera que, se logre suficiencia numérica y calidad de información para lograr los objetivos del proyecto.

3.0.1. Tratamiento de datos faltantes

De la base de datos entregada se identificaron los siguientes porcentajes de datos faltantes para aquellas variables que así lo presentan:

Variable	Porcentaje de datos faltantes
Estrato	54.7 %
Discapacidad	28.6 %
Punt. Lectura	27.5 %
Punt. Matemáticas	23.6 %
Nivel SISBEN	20.4 %
Grupo étnico	14.9 %
Punt. Saber 11	14.9 %
Sector IES	1.2 %
IES	1.2 %
SNIES IES	1.2 %
SNIES Programa	0.9 %
Programa	0.9 %
Localidad	0.9 %

Tabla 1: Porcentaje de datos faltantes.

Como primera medida aquellos registros con datos faltantes asociados a variables categóricas con dificultad de imputación por su naturaleza, como el código SNIES de la universidad, del programa académico y sus nombres, así como, la localidad de residencia del beneficiario son eliminadas (estos valores representaban menos del 2 % del conjunto de los datos). En línea con lo anterior, como característica socioeconómica más objetiva, el sistema de potenciales beneficiarios de programas sociales "SISBEN" se tendrá en cuenta primordialmente sobre el estrato socioeconómico de la vivienda, de tal manera que, esta última variable no será incluida en el análisis de deserción.

Con respecto al nivel SISBEN, se presenta una disyuntiva técnica para su imputación, por un lado, no se cuenta con variables de características individuales de los beneficiarios que funcionen como "proxy" del nivel socioeconómico de estos, adicionalmente, las existentes presentan también valores faltantes y el estrato socioeconómico (con faltantes también) no es una variable adecuada de medición del nivel socioeconómico dada la metodología con la que este indicador fue concebido y calculado.

Las técnicas de imputación analizadas previamente nos refieren a procesos sencillos como la imputación por moda o el uso de técnicas más avanzadas, sin embargo, en este último escenario se requiere de variables correlaciondas con el SISBEN que permitan estimarlo. Así las cosas, inicialmente se procede a completar los datos faltantes utilizando la moda (dato más frecuente) de esta variable dentro de la localidad de residencia del beneficiario, partiendo del supuesto sobre el cual los beneficiarios al ser población principalmente vulnerable comparten similares condiciones socioeconómicas de sus hogares, a diferencia del estrato de la vivienda, el cual "iguala" por la zona de ubicación de la ciudad un conjunto de viviendas sin tener en cuenta las condiciones del hogar. Aunque es una decisión controversial, es válida dentro del mecanismo estadístico, por lo que, se pide al lector aguardar al posterior análisis y uso de las técnicas estadísticas en donde se comprobará si esta imputación es adecuada.

Por otro lado, para las variables "Discapacidad" y "Grupo étnico" se identifica que para los desertores no se presentan valores faltantes y más del 97 % de estos no tiene discapacidad y 91 % no pertenece a alguna comunidad étnica, de tal manera que, a partir de estas proporciones se recurre a la imputacion por la moda de estas variables en el resto de la población para completar así sus valores faltantes al no afectar la caracterización de los desertores.

Por su parte, las variables "Puntaje Saber 11","Puntaje lectura crítica" y "Puntaje matemáticas", presentan el siguiente patrón de datos faltantes:

No	Puntaje Saber 11	Puntaje Matemáticas	Puntaje Lectura	Patrones faltantes
13.015				0
233				1
3179				2
1609				1
571				2
811				3

Tabla 2: Patrón de datos faltantes

La tabla anterior nos muestra un patrón general de datos faltantes, es decir, no se presentan de manera sistemática en todos los registros, su ausencia puede deberse a un error de generación de los datos, ya que, el puntaje Saber 11 es no solo un requisito mínimo de participación, sino que sus resultados son parte del proceso de selección de los aspirantes, así, se presume estar ante un mecanismo de datos faltantes al azar.

De manera particular 811 registros presentan valores faltantes en las tres variables y a 3.750 les hace falta por lo menos dos de estas. Así, como ya se exploró en los métodos para el manejo de datos faltantes, se presentan tres alternativas para su imputación, la primera es la simple, mediante la estimación y uso de la media o la mediana (según se considere conveniente) no obstante, y como se observará más adelante, los datos asociados a desempeños académicos tienen a presentar una distribución normal, por lo que, resulta deseable conservarla en la mayor medida posible y, por tanto, la imputación simple de estas tres variables puede ocasionar un sesgamiento en la estimación de estas medidas de tendencia central y afectar la distribución de los datos con una curtosis leptocúrtica. Entre sus principales implicaciones estará una afectación en los pesos que los modelos a implementar le darán a las variables, por lo que, no resulta apropiado este método de imputación.

Dada la compleja tasa de datos faltantes, la segunda y tercera alternativa invita al uso de técnicas más avanzadas que eviten el alteramiento de la distribución de los datos al máximo posible atendiendo a las consideraciones de Goicoechea (2002). Con lo anterior, a partir de la literatura explorada se considera pertinente realizar dos tipos imputaciones y someterlas a análisis en los modelos a implementar. Se tiene entonces, los modelos iterativos, que como se explicó en el marco conceptual permiten estimar varias funciones ajustadas a la distribución de los datos de las variables y predicen el valor faltante repitiendo el proceso en varias oportunidades (iterativamente), logrando así, un valor plausible (por convergencia o fin de la iteración) al comportamiento esperado de la variable.

Para la tercera alternativa y siguiendo los resultados de Del Turco (2022) se aplicará la imputación con "Hot-Deck" involucrando como técnica implícita los vecinos más cercanos (si bien existen otras como la distancia euclidiana entre registros similares, los vecinos es la más común para Hot-Deck). Se recuerda que este método funciona a través de la donación del valor plausible por parte de un grupo de individuos con similares características al que presenta el valor faltante, esto y al igual que en el algortimo iterativo pueden permitir la conservación de la distribución de los datos y sus medidas de tendencia en general, reduciendo el impacto de la imputación en los modelos a implementar más adelante.

A continuación, se evidencian los resultados de la aplicación del algoritmo "IterativeImputer" en Python con un máximo de 10 iteraciones, las cuales se determinaron esperando una convergencia no tan rápida y por tanto, la generación de un valor más plausible:

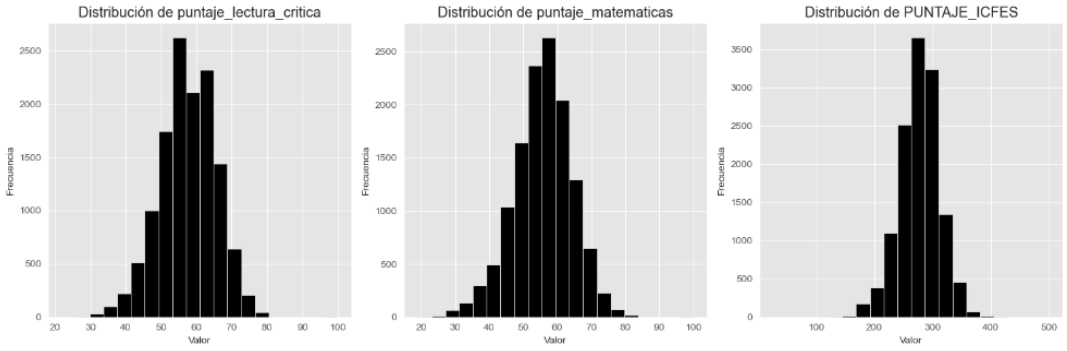


Figura 7: Distribución de variables sin imputación (omitiendo valores faltantes)

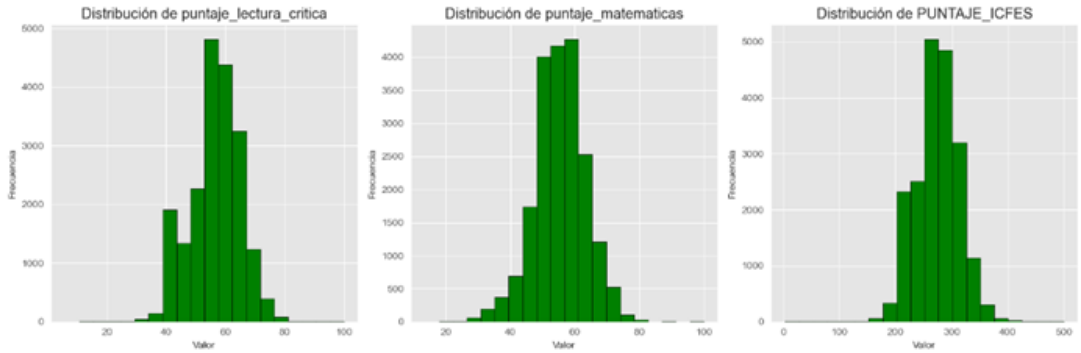


Figura 8: Distribución de variables imputadas con "Iterative Imputer"

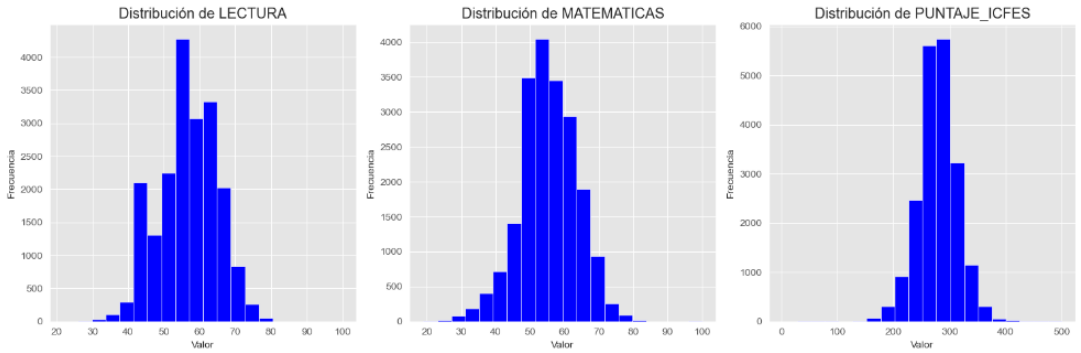


Figura 9: Distribución de variables imputadas con "Hot-Deck"

Variable	**Medida original**	**"Iterative Imputer"**	**"Hot-Deck"**
Puntaje Saber 11	Media=280.03 -Std= 37.98	Media=274.98 - Std= 38.67	Media=279.06 - Std= 35.81
Puntaje Lectura	Media = 56.03 - Std= 8.67	Media=56.6 - Std= 8.53	Media=56.5 - Std= 8.1
Puntaje Matemáticas	Media= 54.79 - Std= 8.54	Media=55.32 - Std= 8.04	Media= 55.36 - Std= 8.24

Tabla 3: Resultados de imputación.

Como se observa en los gráficos, se presenta una ligera diferencia en la distribución de los datos generados por el algoritmo iterativo y Hot-Deck, sin embargo, existe visualmente una mejor coincidencia entre los

datos originales que omiten los valores faltantes y la distribución con Hot-Deck. Lo anterior se refleja en la tabla 3 con las medidas de tendencia central, en donde, Hot-Deck es la técnica de imputación que menos dista de la media de los datos originales y principalmente las desviaciones estándar son inferiores inclusive a los datos originales. Una hipótesis que surge de este ejercicio es que los modelos a implementar tendrán un mejor desempeño con los datos imputados a través de Hot-Deck aunque no se espera una diferencia determinante. En este punto, ambos tipos de imputación resultaron aceptables, no obstante, Hot-Deck requirió menos tiempo de ejecución (pocos segundos de diferencia).

3.0.2. Análisis descriptivo

Una vez completada la base de datos con las imputaciones respectivas, se identifican las siguientes características para las y los beneficiarios de "Jóvenes a la U":

1. El 61 % de beneficiarios son mujeres y 39 % hombres.

2. Las instituciones de educación superior con mayor número de beneficiarios son: Universidad Antonio Nariño con 6.6 %, Universidad Pedagógica Nacional (5.98 %), Corporación Uniminuto con 5.82 %, Universidad ECCI con 5.38 % y la Universidad Colegio Mayor de Cundinamarca con 4.75 %. En contraste, las de menor cantidad son: Escuela Colombiana de Rehabilitación (0.16 %), Universidad de Pamplona (0.17 %) y la Corporación Universitaria TEINCO con 0.2 %.

3. La edad promedio de los beneficiaros es 19 años, con un mínimo de 15 años y un máximo de 29.

4. El 75 % de los beneficiarios cursan sus estudios en instituciones privadas. El 77 % de las mujeres se encuentran en este tipo de instituciones en contraste con el 71 % de los hombres.

5. El 66.3 % de los hombres cursa un programa universitario, el 19.7 % un programa tecnológico y 13.9 % un programa técnico profesional, la participación de las mujeres por su parte corresponde a 71.7 %, 17.5 % y 10.8 % respectivamente.

6. El 45 % de las mujeres se encuentra en pobreza extrema (nivel A) Vs el 44 % de los hombres. Por su parte, 1.2 % de las mujeres es "no pobre (nivel D)" Vs 1.6 % de los hombres de acuerdo con la clasificación de niveles del Sistema de Identificación de Potenciales Beneficiarios de programas sociales SISBEN.

7. El puntaje global promedio de las pruebas Saber 11 corresponde a 280 puntos para hombres y 271 para mujeres. Respecto al puntaje de la prueba específica de lectura crítica, se evidencia un puntaje promedio de 57 puntos para los hombres y 56 para mujeres. Finalmente, el desempeño en matemáticas favorece ligeramente a los hombres con un puntaje promedio de 56 Vs 54 puntos de las mujeres.

8. Los beneficiarios residen principalmente en las localidades de: Bosa (13.84 %), Ciudad Bolívar (12.7 %), Kennedy (12.4 %) y Suba (10.6 %) las cuales corresponden a las zonas de mayor densidad poblacional de Bogotá.

9. Mientras el 54 % de los hombres cursa un programa relacionado con las ingenierías, arquitectura y afines, el 27 % de las mujeres lo hace. En comparación 23 % de las mujeres cursa una carrera relacionada con la economía, administración y afines, 14 % en ciencias sociales y 13 % en ciencias de la educación, en contraste con el 12 %, 6.7 % y 10.7 % de los hombres respectivamente.

3.0.3. Análisis exploratorio de los desertores del Programa.

De acuerdo con la literatura los hombres son quienes más desertan de la educación superior, sin embargo, en la primera exploración de "Jóvenes a la U" son las mujeres la principal población desertora.

Figura 10: Deserción por sexo

Aunado a lo anterior, otras características resaltan que aproximadamente el 3 % de los desertores tiene discapacidad, el 9 % pertenece a una comunidad étnica (negra, afrodescendiente, raizal o palenquera e indígenas). Por su parte, las localidades de residencia de los desertores corresponden de manera natural a aquellas de procedencia de los beneficiarios:

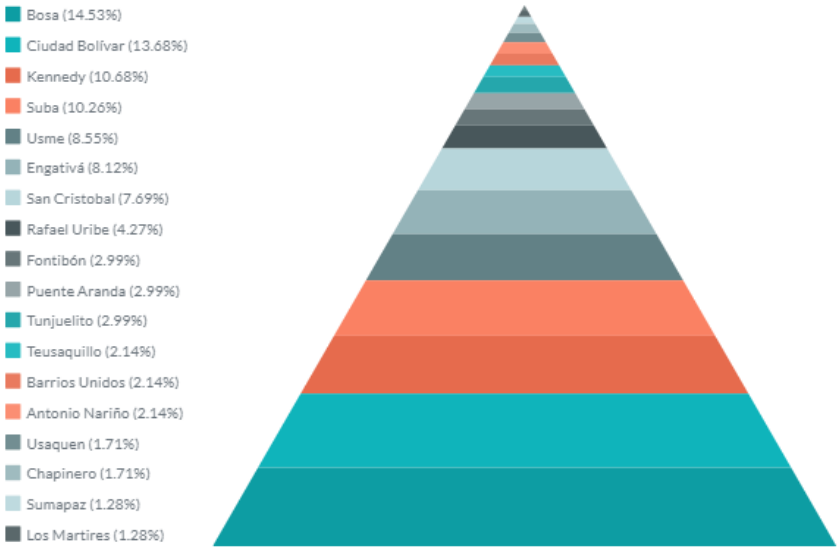


Figura 11: Localidad de residencia de desertores

Sobre las características académicas, la prueba de estado Saber 11 como indicador de los conocimientos adquiridos por los jóvenes en su colegio, permiten identificar algunas características a priori de los beneficiarios desertores:

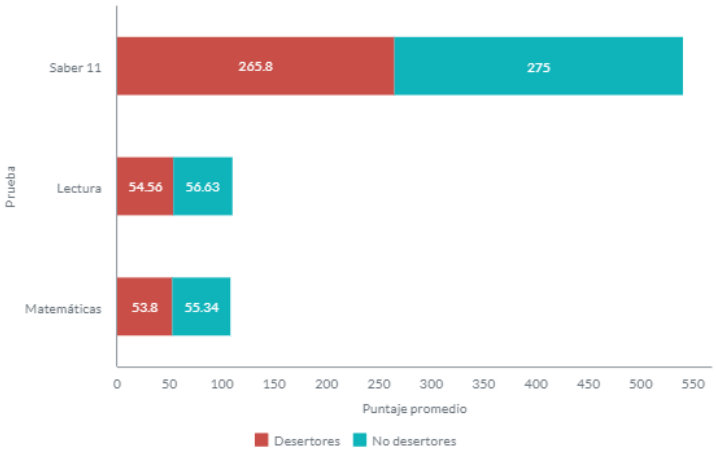


Figura 12: Desempeño promedio Saber 11

En las tres variables (puntaje general, en lectura y en matemáticas) los desertores tienen un desempeño ligeramente inferior frente a los no desertores, sin embargo, el puntaje general muestra una mayor diferencia. No obstante, una diferencia no tan evidente entre ambos grupos puede llevar a que los modelos no consideren determinante los resultados de esta prueba frente al riesgo de deserción. Si bien la literatura menciona el desempeño académico como factor principal, en esta oportunidad se cuenta solamente con esta variable como proxy académica.

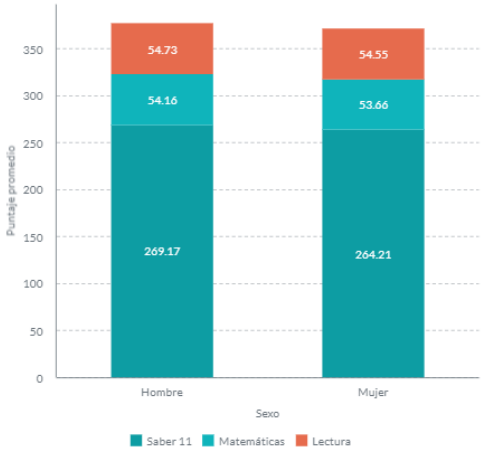


Figura 13: Desempeño promedio Saber 11 desertores

Si se analizan los desertores por sexo, de nuevo la mayor diferencia figura en el puntaje general de la prueba, las pruebas específicas no denotan mayor diferencia entre hombres y mujeres desertores.

Por otra parte, el nivel SISBEN muestra que los desertores se ubican mayoritariamente en el nivel B (pobreza moderada) y A (pobreza extrema), en una menor medida el nivel C (vulnerable) y el D (no pobre, no vulnerable):

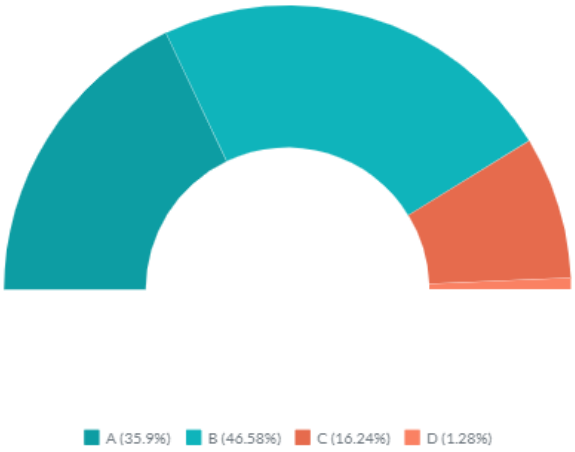


Figura 14: Desertores por nivel SISBEN

Por otro lado, los beneficiarios que cursan programas pertenecientes a áreas del conocimiento con un importante componente de matemáticas al parecer tienen un mayor riesgo de deserción como se muestra a continuación:

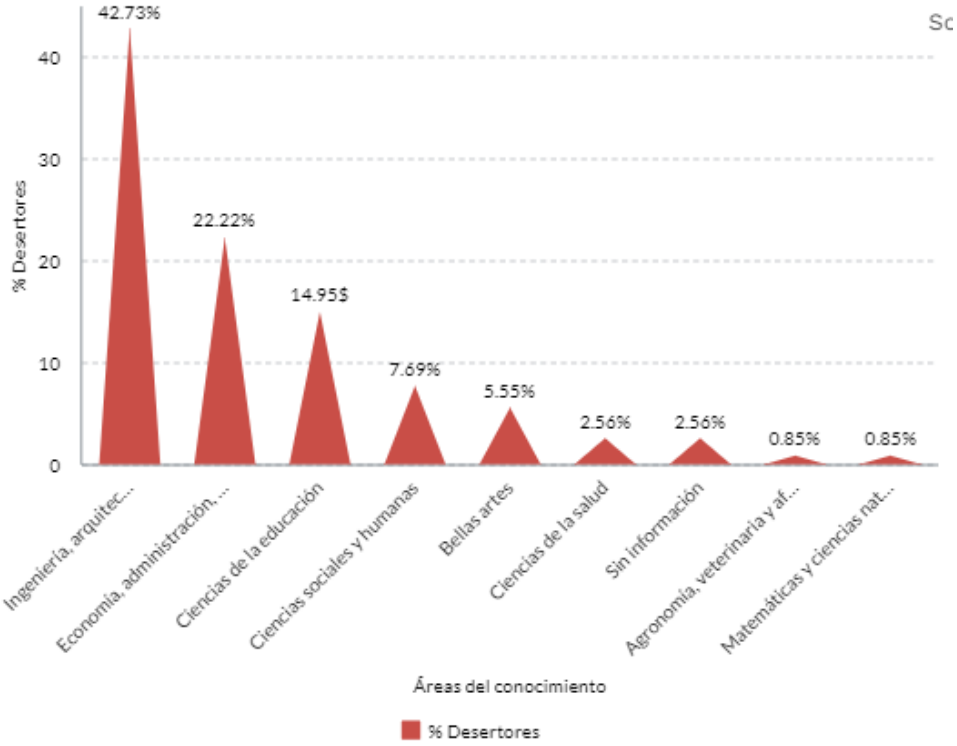


Figura 15: Proporción de desertores por área de conocimiento

3.0.4. Balance de clases: desertores y no desertores

Como se identificó, la base de datos contiene 237 desertores los cuales representan menos del 2 % del universo de beneficiarios. Esta situación conlleva la necesidad de realizar un balance de clases toda vez que la desproporción no permitirá a los modelos realizar una identificación apropiada de las características de los desertores (sobreajuste) y, la sensibilidad, especificidad, precisión y demás métricas que miden el desempeño de los modelos no arrojarán resultados aceptables.

La revisión de la literatura sobre el balance de clases evidenció varios procesos, entre ellos, el sub muestreo como técnica de reducción de la clase mayoritaria a través del muestreo aleatorio simple, no obstante, si bien es una técnica aceptable en comparación con la eliminación directa de una parte del conjunto de datos, se presentará pérdida de información en todo caso (como lo corroboraron Blagus y Lusa), así mismo, existen técnicas destacables que lejos de disminuir el conjunto de datos lo pueden ampliar, especialmente sobre la clase minoritaria. Dicho eso, se considera apropiado el sobre muestreo y para el cual se identificaron dos técnicas comunmente utilizadas y con un enfoque diferenciador que pueden contribuir al balance de las clases de la base de datos.

Se tiene así SMOTE y SMOTE-ENN. La primera realiza un balanceo con la generación de muestras sintéticas por interpolación entre los registros de la clase minoritaria y en dicho espacio interpolado crea los nuevos registros con sus respectivos valores en sus variables independientes. Por su parte SMOTE-ENN realizando el mismo proceso, complementa con la eliminación de datos atípicos por KNN (Vecinos más cercanos) en el cual los datos que no sean clasificados en alguno de los grupos creados son descartados del conjunto de datos. Esto último presenta una ventaja comparativa frente a SMOTE para modelos con sensbilidad por datos atípicos.

Ahora bien, el porcentaje apropiado de sobre muestreo para la clase minoritaria no es una medida reglada de acuerdo con la desproporción de clases. Blagus y Lusa en su análisis utilizan tres tasas diferentes que son sometidas en diferentes modelos y se analiza su desempeño. Resaltan que este desempeño era muy similar con una tasa de sobre muestreo del 40 %, 80 % y 100 %. Ventura et al utilizan una tasa del 80 % y Mustasir et al lo contemplan al 100 %.

Teniendo en cuenta lo anterior y atendiendo al alto desbalance de datos que presenta la base de beneficiarios, una tasa de sobremuestreo muy alta para este tipo de desbalances puede inducir un sobreajuste del modelo provocado por la creación de muestras sintéticas "sobreajustadas", es decir, estas muestras tendrían valores en sus variables independientes desafiantemente cercanas a los datos originales, de manera que, en el análisis de este trabajo se considera ser conservadores con la tasa de sobre muestreo y utilizar una proporción del 10 % que, aunque baja en el contexto del sobre muestreo pretende prevenir el sobreajuste inducido en los modelos a implementar, además, pasar de una tasa de desertores del 1.2 % al 10 % representa un salto cuantitativo relativamente amplio. Se busca tambien invitar al debate académico sobre el abuso de las técnicas de oversampling para lograr desempeños (también subjetivos) aceptables de los modelos como sucedió en la revisión de los ejemplos de modelos de clasificación. Así, el impacto de esta tasa se evidenciará en las métricas de desempeño de los modelos a implementar en este trabajo. Adicionalmente, se utilizará esta tasa con SMOTE y SMOTE-ENN para comprender el impacto que ambas técnicas presentan en los modelos y decantar por la mejor en contextos de alto desbalance y heterogeneidad de los datos.

3.0.5. Identificación de factores de deserción usando técnicas del machine learning

De acuerdo con la información anterior se identifica que las y los beneficiarios de "Jóvenes a la U" tienen un mayor o menor riesgo de deserción en función de factores como el sexo, su desempeño general en la prueba Saber 11, el nivel socioeconómico SISBEN y el área de conocimiento del programa que cursan en la universidad.

Por su parte, la literatura referencia la edad y factores institucionales que en la base de datos se complementó con información auxiliar mediante la obtención de la acreditación institucional de alta calidad de la universidad, la acreditación de alta calidad del programa académico (ambas otorgadas por el Ministe-

rio de Educación Nacional) la modalidad de los programas (presencial, virtual y a distancia) y el sector de la institución (oficial o no oficial). Esta información se extrajo del Sistema Nacional de Información de Educación Superior SNIES del Ministerio de Educación Nacional.

Atendiendo así al procesamiento y análisis de la información previa, se procede con la aplicación de las técnicas de machine learning para identificar los factores y el riesgo de deserción de los beneficiarios del programa. Se propone como se realizó en otros trabajos el uso de un modelo logístico, de árboles de decisión y de bosque aleatorio y se propone el clasificador de gradiente acelerado por su uso de árboles de decisión y una red neuronal como modelos complementarios en el análisis.

Previo a esta aplicación y para la selección de las variables se hace uso de la regresión Lasso (Least Absolute Shrinkage and Selection Operator), la cual es una técnica de "regularización" utilizada en modelos de regresión lineal que ayuda a prevenir el sobreajuste de los modelos y, por tanto, permite seleccionar las variables más importantes. La regresión utiliza un parámetro de penalización que fuerza los coeficientes del modelo a ser "cero", de manera que, aquellos coeficientes (de las variables independientes) que sean cero en esta regularización, serán descartadas por el modelo. Adicionalmente, serán tenidos en cuenta los criterios "WoE" (Weight of evidence o peso de evidencia) que es una medida estadística que se utiliza para evaluar la fuerza de la relación entre una variable predictora y una variable objetivo binaria. Ayuda a determinar cómo una variable predictora afecta a la probabilidad de que ocurra un evento y, el "Information value" el cual sigue el mismo objetivo que el WoE pero utiliza este indicador como parte de su cálculo haciéndolo más "ácido" en el contexto de la selección de variables. Finalmente, se ajusta un modelo logístico con la base de datos original para comprender el alcance del desbalance de las clases de cara a los demás modelos con la imputación y balance realizados.

A continuación se presentan las variables no descartadas por Lasso para la base orginal, para la base con imputación SMOTE (10 %) y con SMOTE-ENN (10 %):

Lasso base original	**Lasso + SMOTE**	**Lasso + SMOTE-ENN**
Edad	Edad	Edad
Matemáticas		
Lectura		
Puntaje Saber 11	Puntaje Puntaje Saber 11	Puntaje Saber 11
Sexo mujer	Sexo mujer	Sexo mujer
AC Ciencias de la salud	AC Ciencias de la salud	AC Ciencias de la salud
AC Ciencias sociales y humanas	AC Ciencias sociales y humanas	AC Ciencias sociales y humanas
AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines
AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines
AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales
AC Sin información	AC Sin información	AC Sin información
SECTOR IES Privado	SECTOR IES Privado	SECTOR IES Privado
GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO
DISCAPACIDAD SI	DISCAPACIDAD SI	DISCAPACIDAD SI
SISBEN B	SISBEN B	SISBEN B
SISBEN C	SISBEN C	SISBEN C
IES ACREDITADA SI	IES ACREDITADA SI	IES ACREDITADA SI
PROGRAMA ACREDITADO SI		
MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual

Tabla 5: Resultados regresión Lasso.

La regresión descarta en los modelos sobre muestreados las variables: matemáticas, lectura y el programa acreditado de alta calidad; por su parte, desde la base original se descartó las modalidades presencial y a distancia de los programas. Se realizó un ejercicio con SMOTE y SMOTE-ENN al 25 % sin evidenciarse descartarmientos adicionales, no obstante, los coeficientes de la regresión si se fortalecian lo que indica que un mayor sobremuestreo en efecto puede inducir sobreajuste en los modelos. A continuación, se observan los resultados del WoE y el IV:

WoE base original	**WoE + SMOTE**	**Woe + SMOTE-ENN**
Edad	Edad	Edad
Sexo mujer	Sexo mujer	Sexo mujer
AC Ciencias de la salud	AC Ciencias de la salud	AC Ciencias de la salud
AC Ciencias sociales y humanas	AC Ciencias sociales y humanas	AC Ciencias sociales y humanas
AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines
AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines
AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales
AC Sin información	AC Sin información	AC Sin información
SECTOR IES Privado	SECTOR IES Privado	
GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO
DISCAPACIDAD SI		
SISBEN B	SISBEN B	SISBEN B
SISBEN C	SISBEN C	SISBEN C
IES ACREDITADA SI	IES ACREDITADA SI	IES ACREDITADA SI
PROGRAMA ACREDITADO SI		
MODALIDAD PROGRAMA presencial	MODALIDAD PROGRAMA presencial	MODALIDAD PROGRAMA presencial
MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual

Tabla 7: Resultados WoE.

IV base original	**IV + SMOTE**	**IV + SMOTE-ENN**
Edad	Edad	Edad
Sexo mujer		
AC Ciencias de la salud	AC Ciencias de la salud	
AC Ciencias sociales y humanas	AC Ciencias sociales y humanas	AC Ciencias sociales y humanas
AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines	AC Economía, administración, contaduría y afines
AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines	AC Ingeniería, arquitectura, urbanismo y afines
AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales	AC Matemáticas y ciencias naturales
AC Sin información		
SECTOR IES Privado	SECTOR IES Privado	
GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO	GRUPO ETNICO NINGUNO
DISCAPACIDAD SI		
SISBEN B	SISBEN B	SISBEN B
SISBEN C	SISBEN C	SISBEN C
IES ACREDITADA SI	IES ACREDITADA SI	IES ACREDITADA SI
PROGRAMA ACREDITADO SI		
MODALIDAD PROGRAMA presencial	MODALIDAD PROGRAMA presencial	MODALIDAD PROGRAMA presencial
MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual	MODALIDAD PROGRAMA Virtual

Tabla 9: Resultados Information Value.

El peso de evidencia (WoE) y el "information value" presentan resultados similares. No descartan todas las variables pero difieren en la variable sexo mujer y el área de conocimiento sin información, en donde el information value al ser más "ácido" en su estimación descarta variables que al parecer no aportan de ninguna manera a la correlación con la variable de interés.

De acuerdo con la regresión Lasso, el peso de evidencia y el information value se ajustarán los diferentes modelos con las variables consideradas en estas selecciones. Previo a este proceso se presenta a continuación un modelo logístico con la base de datos sin ajustes para complementar el análisis de selección de variables.

Estimación del Modelo logístico base

Posterior a la transformación de las variables categóricas en "dummies" se ajustó un modelo logístico al conjunto de datos originales, y en el que las variables que resultaron significativas con un nivel de confianza del 95 % son:

Variable	Coeficiente
Edad	0.0990
Sexo mujer	0.4762
Puntaje Saber 11	-0.0056
Sector IES privado	-0.6458
Grupo étnico Si	-4.4908
Grupo étnico ninguno	-5.0674
Discapacidad	1.8542
SISBEN B	0.3827
IES acreditada	-0.4491
Programa acreditado	0.5025
Modalidad virtual	0.6239

Tabla 10: Resultados modelo logístico con base original.

Es preciso indicar que en este punto no se considera la interpretación de los coeficientes en razón al posible sobreajuste. Así, a priori, la edad, ser mujer, tener discapacidad, pertenecer al nivel socioeconómico SISBEN B y estudiar un programa acreditado de alta calidad y en modalidad virtual, incrementa el riesgo de deserción. Resalta que pertenecer a una comunidad étnica reduce el riesgo. Sin embargo, ya se vislumbra el sesgamiento en la capacidad predictiva del modelo dado que las variables significativas son aquellas que en el análisis descriptivo presentaban en mayor medida los desertores. Por su parte, las demás variables no presentan ningún tipo de influencia aun cuando teóricamente si lo puedan tener (por ejemplo, los puntajes de las pruebas Saber 11 y los demás niveles socioeconómicos). Así, para analizar el impacto de la imputación de datos y el balance de clases, se ajustó un segundo modelo para predicción con datos de testeo del 20 % obteniéndose las siguientes métricas de desempeño:

Métrica de evaluación	Valor
Sensibilidad	56.41 %
Especificidad	58.70 %
Accuracy	58.68 %
Precisión	1.33 %
AUC	63.25

Tabla 11: Métricas de evaluación modelo logístico base.

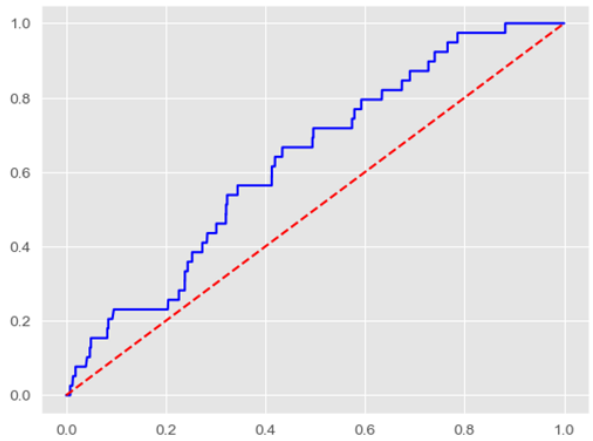


Figura 16: AUC modelo logístico base

Las métricas muestran falencias en la correcta identificación de desertores y no desertores, esta situación implica, por un lado, la aplicación errónea de medidas que prevengan este riesgo a poblaciones que no lo necesitan y por tanto incrementándolo en aquellos potenciales desertores que fueron clasificados como no desertores, y por otro lado, generando costos operacionales, gastos administrativos, financieros y humanos innecesarios para las instituciones involucradas. En promedio el modelo predeciría adecudamente al 57 % de los beneficiarios, es decir, comete el error de discriminación de la misma manera en ambas clases de interés y la precisión tan baja evidencia una alta tasa de falsos positivos evidenciados en aquellos beneficiarios que no son identificados por el modelo (sensibilidad y especificidad).

La curva ROC (Característica Operativa del Receptor en español) refuerza la baja capacidad del modelo para predecir las poblaciones de interés. Esta curva permite analizar el equilibrio entre la especificidad y la sensibilidad del modelo. El eje horizontal corresponde a la tasa de falsos positivos y el eje vertical muestra la tasa de verdaderos positivos, por lo que, la linea roja refleja una nula discriminación de ambas clases. Así, el escenario deseable en esta gráfica corresponde a que la curva de predicción (azul) se acerque a la coordenada 1,1, de manera que, la aplicación de modelos de clasificación sobre la base de datos original y desbalanceada no es apropiada en ningún caso.

3.0.6. Tratamiento al desbalance de clases en la base de beneficiarios

Como se identificó previamente, SMOTE y SMOTE-ENN resultan ser las estrategias más utlizadas para el machine learning en la literatura y las más adecuadas para aplicar en este estudio de caso. Observése entonces los resultados que sobre la significancia estadística de las variables tienen estas técnicas ajustando nuevamente un modelo logístico con las variables no descartadas por lasso, WoE y IV, con una tasa de sobremuestreo del 10 % teniendo en cuenta la hipótesis previamente mencionada según la cual una tasa superior generará probablemente sobreajuste en el modelo e indicará significancia estadística en todas las variables. Así mismo, como parte del proceso de prevención de sobreajuste se aplica el modelo bajo la metología "k-fold cross validation" la cual divide el conjunto de datos de entrenamiento en un número (k) determinado de partes (se consideran 5 en este caso) y determina la tasa de exactitud (accuracy) de cada "modelo" en cada división, así, una metríca "accuracy" promedio diferente al "accuracy" de los datos de testeo indicarían que el modelo presenta sobre ajuste y problemas de generalización con nuevos datos:

Variable	**Modelo con SMOTE**	**Modelo con SMOTE-ENN**
Edad	0.2126	0.1882
Saber 11	0.0075	0.0073
Grupo étnico	-3.8124	-3.5213
Grupo étnico ninguno	-3.05	-2.9108
Bellas artes	-3.7994	-3.7312
Ciencias de educación	-3.416	-3.45
Ciencias de la salud	-4.293	-4.336
Ciencias sociales y humanas	-3.562	-3.2
Administración, economía y afines	-3.0511	-2.726
Ingenierías, arquitectura y afines	-2.8185	-2.8177
Ciencias naturales y matemáticas	-6.0672	-4.898
SISBEN B	-0.576	-0.2866
SISBEN C	-1.2938	-1.0820
SISBEN D	-2.435	-2.4168
IES acreditada	-1.22	-2.02
Programa presencial	-1.3168	-1.316
Programa virtual	-2.3733	-2.4991

Tabla 12: Coeficientes modelo logístico con SMOTE y SMOTE-ENN.
*Significancia estadística al 90 % ; **No significancia estadística

Ambas técnicas mejoraron de manera significativa la influencia de nuevas variables como los demás niveles socioeconómicos de SISBEN (C y D), las áreas de conocimiento y la modalidad del programa presencial que había sido descartada por WoE. El nuevo modelo descarta la discapacidad y el ser mujer (aunque WoE no lo hacía). En el análisis realizado por Rodriguez (2012) se concluyó que las mujeres tienen una mayor probabilidad de graduación que los hombres, lo que podría reforzar la idea de una menor deserción de esta población; inclusive es interesante suponer que las labores de cuidado o las responsabilidades que mayoritariamente tienen las mujeres en sus hogares, les lleva a ser más "dedicadas" con el objetivo de titularse. Llama la atención que se mantiene el signo positivo del desempeño en las pruebas Saber 11 que si bien presenta un coeficiente pequeño en el modelo base y en el ajustado con SMOTE, el signo va en contravía de la literatura en la que se espera que un mejor desempeño académico reduzca el riesgo de deserción, sin embargo, como se ha mencionado, "Jóvenes a la U" prioriza población con condiciones de vulnerabilidad, por lo que, esta variable puede verse afectada por las condiciones socioeconómicas de la persona, sus condiciones de salud, factores institucionales u otros como el contexto social de la persona y su familia, es decir, la excelencia académica no garantiza permanecer en los estudios de este Programa.

Nótese así mismo que las imputaciones controversiales del nivel SISBEN (realizado con la moda de la localidad de residencia del beneficiario) resultaron apropiadas, toda vez que, los coeficientes de cada nivel guardan lógica transitiva con lo indicado por la literatura en referencia al nivel socioeconómico de los estudiantes (a mejores condiciones, menor riesgo de deserción) acá se observa que el nivel D (no pobre) tiene menor riesgo que el nivel C (vulnerable) y este sobre el nivel B (pobreza moderada) y estos sobre el nivel de referencia (nivel A).

Se realizó el cálculo del modelo logístico con escala al 20 % y 50 % de sobremuestreo con SMOTE y SMOTE-ENN como prueba de la hipótesis de sobre ajuste inducido por el balance, y todas las variables resultaron significativas en ambas técnicas y sus coeficientes mucho más altos (particularmente con la tasa del 50 %) que los presentados en la tabla anterior, por lo tanto, una tasa del 10 % resulta oportuna para prevenir la sobresignificancia de las variables y en consecuencia sobre ajuste del modelo.

Como refuerzo del análisis de sobreajuste del modelo y la aplicación de K-fold cross validation, se realizó la predicción sobre los datos de prueba (20 % del total de la base), proceso que arroja una ventaja para el modelo con SMOTE-ENN para el cual las métricas de evaluación arrojan resultados superiores (aunque no significativamente). En ese sentido, las muestras sintéticas creadas por las técnicas y la imputación de los datos resultaron adecuadas al no forzar un sobreajuste de los modelos, no obstante, la baja precisión evidencia que el modelo logístico puede no ser el método de clasificación más eficiente ante escenarios conservadores de desbalance, teniendo en cuenta que, se sometió el modelo con sobremuestreo al 50 % (en SMOTE-ENN) y sus metrícas fueron superiores al 85 % indicando un posible sobreajuste del modelo y confirmando la hipótesis planteada previamente.

Criterio de desempeño	**Modelo con SMOTE-ENN**
Sensibilidad	86.12 %
Especificidad	85.9 %
Accuracy	86 %
Precisión	75.26 %
AUC	93.99

Tabla 13: Desempeño modelo logístico SMOTE-ENN al 50 %.

Criterio de desempeño	Modelo con SMOTE	Modelo con SMOTE-ENN
Sensibilidad	81.8 %	83.5 %
Especificidad	82.3 %	83.5 %
Accuracy	82.2 %	83.8 %
Precisión	30.4 %	28.10 %
F1-Score	44.3	42.06
AUC	87.91	88.86

Tabla 14: Desempeño modelo logístico con SMOTE y SMOTE-ENN al 10 %.

Con "K-Fold cross validation" se presentaron los siguientes resultados y que confirman el posible sobreajuste del modelo logístico ante la diferencia de más de 10 puntos de los valores de accuracy en entrenamiento y el correspondiente al testeo, esto significa que el modelo tendría inconvenientes en la generalización de sus predicciones, por tanto, ante datos nuevos y como se evidenció con la precisión, no podría predecir adecuadamente los desertores y los no desertores:

Fold	Accuracy con SMOTE	Accuracy con SMOTE-ENN
Fold 1	93.89 %	95.52 %
Fold 2	93.86 %	95.91 %
Fold 3	94.14 %	95.72 %
Fold 4	94.40 %	95.81 %
Fold 5	93.83	96.24
Accuracy con testeo	82.2 %	83.8 %

Tabla 15: Desempeño modelo logístico con SMOTE y SMOTE-ENN al 10 %.

Ahora bien, los coeficientes del modelo logístico (SMOTE-ENN) permiten inferir que la mayoría de las variables analizadas reducen en diferentes proporciones el riesgo de deserción, particularmente la edad, el desempeño en la prueba Saber 11 y cursar un programa acreditado son las únicas que lo incrementan. Dichos coeficientes en un contexto probabilístico permiten inferir que:

- Por cada año de edad, la probabilidad de desertar se incrementa en 35 %.
- Por cada punto adicional obtenido en la prueba Saber 11, la probabilidad de desertar se incrementa en 0.75 %.
- Ser mujer reduce el riesgo de deserción respecto a un hombre. Ser mujer tiene un riesgo de deserción equivalente al 71.25 % del riesgo de un hombre, es decir, ser mujer reduce el riesgo de deserción en 28.75 %.
- Pertenecer a una comunidad étnica reduce el riesgo de desertar en mayor medida frente a no pertenecer a alguna, en cuyo caso también presenta un impacto negativo en el riesgo de deserción. Se puede inferir de estos resultados que las personas pertenecientes a comunidades afro, indígenas, rrom, afrodescendientes, al igual que la variable mujer, sus vulnerabilidades individuales las lleva a tener un mayor grado de compromiso en procura de una pronta empleabilidad.
- SISBEN B: los beneficiarios en este nivel socioeconómico tienen aproximadamente un 49.71 % del riesgo de deserción en comparación con el nivel SISBEN A, es decir, menor riesgo que este último.
- SISBEN C: los beneficiarios en la categoría nivel SISBEN C tienen aproximadamente un 26.10 % del riesgo de deserción en comparación con el nivel A.
- SISBEN D: los beneficiarios con nivel socioeconómico SISBEN D tienen un riesgo de deserción significativamente menor en comparación con el SISBEN A (0.5 %) así como, con relación a los demás niveles.

- Las áreas del conocimiento arrojan resultados interesantes, se evidencia que quienes estudian programas de las ciencias naturales y matemáticas tienen un riesgo significativamente menor respecto a las demás áreas, esto puede indicar que los beneficiarios que cursan estos programas cuentan con habilidades o preparación adecuada previa a la educación superior, en cambio, quien no sienta afinidad con las matemáticas muy probablemente no se postulará a un programa con un alto componente numérico. Situación similar se presenta en bellas artes. Por su parte, los mayores riesgos relativos de deserción se dan entre las áreas del conocimiento correspondientes a programas de las ciencias de educación y de administración y economía.
- Estudiar en una IES acreditada reduce el riesgo de deserción en 57.8 % respecto a estudiar en una IES no acreditada.
- Estudiar en un programa acreditado incrementa el riesgo de deserción en 14.22 % frente a uno no acreditado, esto se explica en primera medida porque para obtener la acreditación de alta calidad de un programa académico en Colombia, implica que la institución a alcanzado un grado de madurez, reconocimiento y condiciones de calidad que destacan la formación impartida, es decir, la acreditación es equivalente a una mayor exigencia, por lo que, un beneficiario que no logre adaptarse a dicho nivel tendrá naturalmente un mayor riesgo de abandono.
- Estudiar en un programa en modalidad presencial reduce el riesgo de deserción en 29.46 %, por su parte, estudiar en uno en modalidad virtual reduce el riesgo en menor medida (3.01 %) esto último genera una pregunta, ¿las metodologías virtuales aun no logran "atrapar" a los estudiantes?.
- El nivel de formación tecnológico tiene menor riesgo de deserción que los programas universitarios, una explicación de esto recae nuevamente en las condiciones de vulnerabilidad que presente el beneficiario y la decisión tomada al momento de su postulación siguiendo la mejora en sus condiciones de vida. En una periodicidad de estudios más corta, es más rápida la vinculación laboral.

Respecto a los otros análisis de deserción que se han realizado se identifica que los beneficiarios de "Jóvenes a la U" siguen el riesgo de deserción de las variables que la literatura ha identificado como incidentes, sin embargo, como se encontró a través del modelo, un buen desempeño académico no garantiza la permanencia si las condiciones socioeconómicas o individuales del beneficiario pesan más en su decisión de continuar con sus estudios, por ejemplo, una mayor edad o tener hijos lleva a priorizar aspectos laborales sobre los educativos.

Por su parte, los factores institucionales que además de la metodología docente involucra la infraestructura y medidas de acompañamiento que brinden estas instituciones a sus estudiantes, evidencian que la acreditación de estas juega un papel importante en reducir este riesgo. En Colombia las instituciones de educación superior de manera voluntaria se postulan para obtener la acreditación institucional de alta calidad y para ser elegibles a este reconocimiento deben contar con aspectos destacados en sus condiciones de calidad (infraestructura, formación) de bienestar estudiantil, investigación académica, patentes, docentes con formaciones en maestrías o doctorados, entre otros aspectos que implica una importante inversión financiera que mejora las condiciones de los estudiantes.

Los resultados obtenidos permiten inferir que los beneficiarios de "Jóvenes a la U" en general no presentan riesgos de deserción diferentes respecto a estudiantes regulares de la educación superior; los factores identificados en la literatura siguen vigentes. Destaca de manera particular que en esta población y como lo identificó Rodriguez (2012), la vulnerabilidad socioeconómica tiene una mayor incidencia en la deserción y en la probabilidad de graduación aun cuando el beneficiario presente condiciones de excelencia académica a lo largo de su carrera.

Al igual que los análisis de Castro et al (2021) quienes demostraron con el seguimiento que le realizaron a los estudiantes que, a mayor edad de inicio de los estudios, mayor es el riesgo de deserción, situación que también se confirma con el signo que acompaña al coeficiente de edad en el modelo calculado.

A continuación se ajusta un árbol de decisión, un bosque aleatorio, un clasificador de gradiente acelerado y una red neuronal como técnicas que podrían ser implementadas para analizar y predecir el riesgo de

deserción de los beneficiarios logrando establecer estrategias diferenciales según las condiciones de cada uno.

Árbol de decisión

Para la construcción del árbol de decisión se toman las mismas variables del modelo logístico junto con "K-fold cross validation", sin embargo, este modelo conlleva la definición de parámetros como su profundidad (cantidad de ramificaciones), la cantidad de muestras y el peso mínimo en estas para generar divisiones. Así, es importante encontrar fórmulas técnicas que permitan establecer los valores de estos parámetros con el fin de reducir el tiempo que tomaría realizar "ensayo y error" para hayar los valores que mejoren el desempeño del árbol, así mismo, una definición propia de estos valores puede conllevar a un error técnico que afecte consecuentemente los modelos. Con lo anterior, se realiza la optimización de los hiper parámetros mediante la técnica denominada "GridSearch cross validation" de la librería "scikit-learn" en Python, la cual consiste en la postulación de un rango de valores consistentes para cada hiperparámetro del árbol sujeto a una métrica de evaluación objetivo (precisión, sensibilidad, accuracy) y, a través de simulaciones con los datos de entrenamiento en las que combina los diferentes valores propuestos en la "grilla" o cuadrícula buscando el mejor desempeño de la métrica objetivo, la técnica determina los mejores valores.

Esta técnica permite entonces encontrar valores técnicamente óptimos reduciendo el error humano en el ajuste del árbol de decisión. Con lo anterior, y considerándose el uso del "Grid Search Cross Validation" para el árbol considerado, se priorizó como métrica objetivo la sensibilidad del modelo en procura de la identificación adecuada de los desertores, en ese sentido, la técnica arrojó que los mejores hiper parámetros que optimizan la sensibilidad del árbol son:

Hiperparámetro	**SMOTE**	**SMOTE-ENN**
Profundidad	10	10
Muestras mínimas por nodo	5	2
Peso mínimo de muestras para división del nodo	0.0	0.0

Tabla 16: Hiperparámetros de Grid Search Cross Validation.

Los hiperparámetros obtenidos son razonables y ajustados con la dimensión de la base de datos en donde se cuenta con un 10 % de desertores, es decir, el modelo requerirá una identificación más detallada de los registros para realizar su clasificación al no contar con suficientes muestras de la variable de interés o valores determinantes en las variables independientes que le acompañan. Esto se evidencia, por un lado, con la baja cantidad de muestras que requerirá el árbol para tomar una decisión en cada nodo, en donde, SMOTE-ENN resulta en una mayor simplificación; y por otro lado, el no contar con un peso mínimo de las muestras en cada nodo, indica que el árbol no pretende complejizar la clasificación también por la no suficiencia de muestras o valores de las variables independientes que sean determinantes para identificar desertores, sin embargo, la profundidad del árbol (10 niveles en SMOTE y SMOTE-ENN) permite una exploración "lenta" del conjunto de datos de entrenamiento, es decir, se está en un escenario eficiente. Para explicarlo de mejor manera, tómese como ejemplo un conductor de vehículo que se dirige a una ciudad, el conductor precavido tomará la ruta más larga, con menos pendientes y curvas, frente a un conductor no precavido que tomará el camino más corto, con más curvas y más pendientes. Con lo anterior, el árbol de decisión con SMOTE y SMOTE-ENN al 10 % arrojó los siguientes resultados en métricas de desempeño de clasificación:

Métrica de evaluación	Valor SMOTE	Valor SMOTE-ENN
AUC	74.85	82.50
Sensibilidad	50.80 %	66.04 %
Especificidad	98.91 %	98.96 %
Accuracy	94.76 %	96.65 %
Precisión	81.54 %	82.71 %
F1-Score	62.60	73.44

Tabla 17: Desempeño del árbol de decisión.

Los resultados resultan mejores en comparación con el modelo logístico (en precisión, especificidad y F1-SOCRE), sin embargo, se notan diferencias importantes en el impacto que tiene SMOTE y SMOTE-ENN. Particularmente, el segundo permitió al árbol tener una mejor identificación de los desertores (sensibilidad) y esto a su vez se refleja en el F1-SCORE que combina la sensibilidad y la precisión en una sola métrica y el área bajo la curva ROC (AUC).

- ***"Características importantes o Feature importance"***

Esta técnica permite identificar en modelos de aprendizaje automático aquellas variables que son relevantes o representativas en los resultados (o predicción) del modelo estimado, así, esta técnica permite darle interpretabilidad a las variables de modelos que como el árbol de decisión y los que posteriormente se ajustan tienen como propósito la predicción de valores. Ahora bien, las variables con mayor importancia consideradas en ambos modelos son:

Variables SMOTE	Variables SMOTE-ENN
institución acreditada	Puntaje Saber 11
Puntaje Saber 11	Modalidad virtual
Edad	Edad

Tabla 18: Variables con mayor importancia en el árbol.

Por otra parte, ambos árboles le otorgan menor importancia relativa y en la misma proporción a las áreas de conocimiento y los niveles socioeconómicos, auque en estos últimos el modelo otorga ligeramente más peso al nivel B, seguido del C y por último el nivel D, guardando la lógica transitiva que identificó el modelo logístico y confirmando la correcta imputación de esta variable. Así, para el árbol de decisión resulta determinante la edad, el puntaje de la prueba saber 11, si la institución se encuentra acreditada de alta calidad y la modalidad virtual frente al riesgo de deserción de los beneficiarios del Programa, siguiendo los hallazgos de la literatura (factores académicos, institucionales e individuales).

Finalmente, el accuracy promedio de los k-fold realizados con los datos de entreamiento del 94 % con SMOTE y 96 % para SMOTE-ENN en contraste con el accuracy en testeo (94.76 % y 96.65 % respectivamente) evidencian que el árbol de decisión no presentó sobreajuste y por consiguiente, resulta en una técnica apropiada para para el modelamiento o predicción del riesgo de deserción en el escenario conservador y desafiante de una base de datos aun con desbalance de clases, es decir, el árbol de decisión es generalizable para nuevos conjuntos de datos.

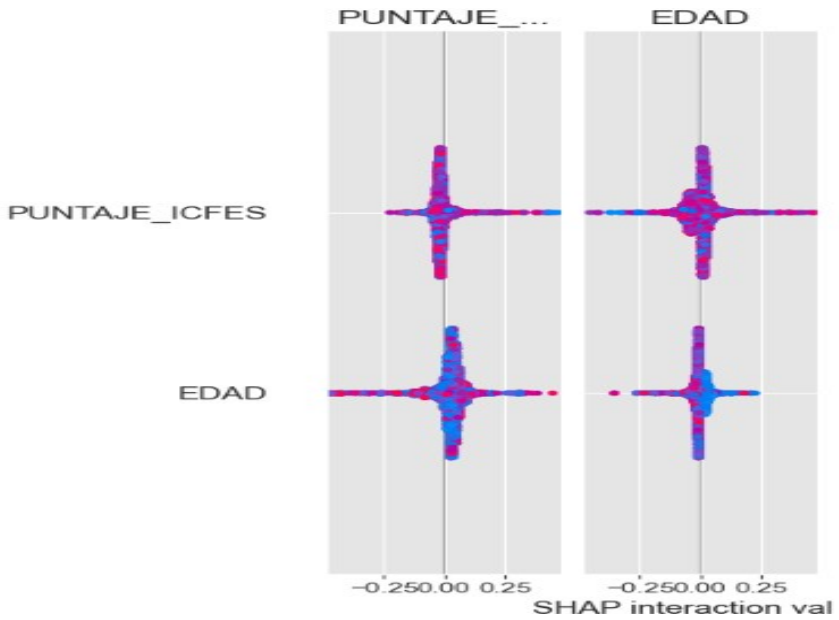
- ***"SHAP (Shapley Additive Explanations)"***

"SHAP" es una metodología que basada en la teoría de juegos pretende mostrar la contribución de cada variable al resultado de la predicción de un modelo de aprendizaje automático. Desde la teoría de juegos, en 1953 el Economista y Matemático Lloyd Shapley propuso un mecanismo para la repartición de ganancias "justa"para individuos que participan de un juego cooperativo, lo anterior, partiendo del

supuesto sobre el que las contribuciones de cada individuo son diferentes y aportan de forma marginal a la ganancia total del grupo, dicha contribución se conoce entonces como el valor "Shapley".

Así, mediante la metodología SHAP se puede medir la contribución que cada variable independiente de un modelo de aprendizaje automático aporta al resultado de una predicción específica, de manera que, por un lado, SHAP podrá evidenciar las características representativas de un modelo de manera general a partir de los resultados estimados por este y también está en la capacidad de generar la contribución marginal de cada variable o característica sobre una observación en particular que se desee observar. Para el caso del árbol de decisión, SHAP identificó las siguientes variables como las más representativas para el modelo:

Figura 17: Valores SHAP, árbol de decisión



En el cuadrante superior izquierdo se observa la interacción del PUNTAJE ICFES (Saber 11) consigo mismo permitiendo analizar su impacto en la deserción. Los círculos de la gráfica corresponden a las observaciones estimadas por el modelo, además, aquellos círculos de color azul corresponde a los valores más bajos del PUNTAJE ICFES y los rojos son el caso opuesto. Los puntos ubicados a la derecha del eje central del cuadrante indican el aporte positivo del puntaje icfes al riesgo de deserción, por lo que, la gráfica permite dilucidar que la mayoría de los puntos (rojos y azules) se concentran ligeramente a la izquierda del eje, guardando coherencia con los resultados del modelo logístico en donde el coeficiente que acompaña la variable es cercano a cero, es decir, el desempeño de esta prueba no prediciría si un beneficiario del programa desertará o no. El lector podrá fijarse que algunos valores altos y bajos de la prueba saber 11 incrementan el riesgo de deserción, siendo las segundas ubicaciones más lógicas, los valores altos deben analizarse a la luz de la interacción con el resto de las variables del modelo.

Por el lado de la edad también se evidencia un resultado similar. Si bien el modelo logístico indicó que la edad incrementa el riesgo de deserción, el gráfico de SHAP evidencia resultados no concluyentes, se observa por ejemplo una concentración de puntos azules a la derecha del gráfico, es decir, menor edad incrementando el riesgo de deserción y puntos rojos (valores de edad altos) reduciendo dicho riesgo.

Ahora, obsérvese la interacción entre estas dos variables, así, en el cuadrante superior derecho se analizan los efectos de la interacción entre un puntaje alto de la prueba Saber 11 y valores bajos de la edad del

beneficiario, así, esta interacción resulta negativa, o lo que es lo mismo, reduce el riesgo de deserción especialmente cuando el puntaje de la prueba Saber 11 es alta y la edad es baja, en contraposición, en el cuadrante inferior izquierdo la misma interacción (analizando valores altos de edad y bajos en Saber 11) muestran el efecto positivo de la interacción sobre el riesgo de deserción, es decir, una mayor edad sumado a un desempeño bajo en la prueba Saber 11 incrementa el riesgo de deserción.

Una conclusión importante de este análisis es que el riesgo de deserción de los beneficiarios y en comparación con la literatura, implica que no es correcto determinar un posible riesgo de deserción observando condiciones o características de manera individual; deberá entenderse para los beneficiarios del programa como la suma de estas.

Bosque Aleatorio:

Como se mencionó previamente, es un algoritmo que genera y combina varios árboles de decisión encontrando características comunes para la clasificación de las observaciones de una manera esperada más robusta y eficiente. Así, se espera que varios árboles tengan mejores resultados (en el bosque aleatorio) que un árbol de decisión individual.

Para la construcción del bosque aleatorio se recurrió también a la optimización de los hiper parámetros mediante la técnica "GridSearch cross validation" y tal como se indicó para el árbol de decisión, se busca la optimización sobre el criterio de sensibilidad. Así, esta búsqueda arrojó los siguientes hiper parámetros para SMOTE y SMOTE-ENN:

- Número de árboles: 50
- Profundidad: 10
- Muestras mínimas por nodo: 1
- Peso mínimo de las muestras: 0.0

A partir de lo anterior, se ajustó el bosque aleatorio son SMOTE y SMOTE-ENN al 10 % y el cual evidencia los siguientes resultados en métricas de desempeño:

Métrica de evaluación	Valor SMOTE	Valor SMOTE-ENN
AUC	95.38	96.51
Sensibilidad	90.10 %	91.41 %
Especificidad	90.16 %	91.48 %
Accuracy	90.15 %	91.48 %
Precisión	46.35 %	44.70 %
F1-Score	61.21	60.04

Tabla 19: Desempeño del Bosque aleatorio.

Los resultados en general respaldan el impacto de SMOTE-ENN en los modelos aplicados, no obstante, la precisión del bosque es peor que el árbol pero mejor frente al modelo logístico. La precisión al indicarnos la tasa de desertores considerados por el modelo sobre los reales desertores muestra una subestimación de estos, lo que no es deseable dentro del contexto de identificación esperado.

- ***"Características importantes o Feature importance"***

Obsérvese a continuación las variables que mayor peso consideró el bosque frente al riesgo de deserción:

Variables SMOTE	Variables SMOTE-ENN
Puntaje Saber 11	Puntaje Saber 11
Institución acreditada	Institución privada
Institución privada	Institución acreditada

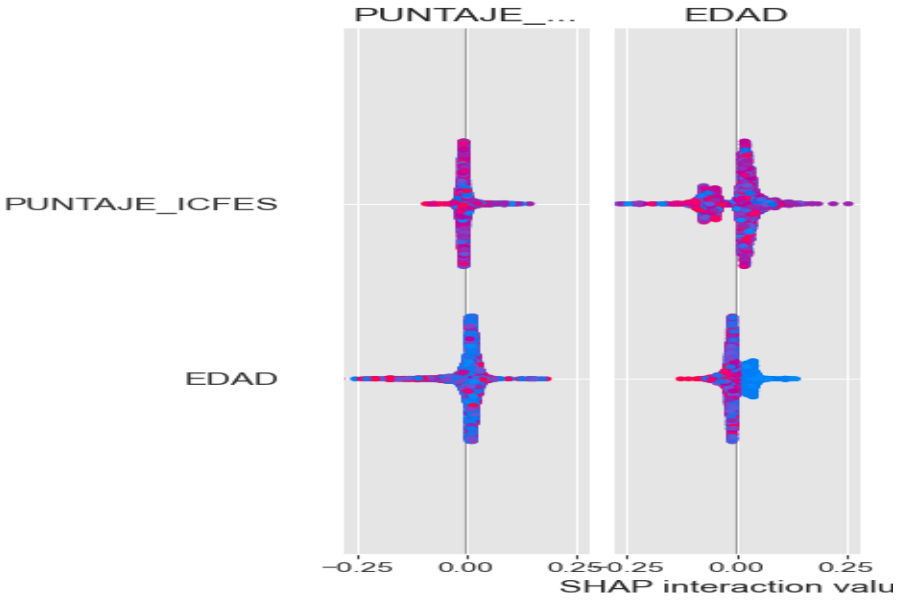
Tabla 20: Variables con mayor importancia para el Bosque.

Por otra parte, ambos modelos al igual que el árbol de decisión le dan relativa menor importancia y en la misma proporción a las áreas de conocimiento y los niveles socioeconómicos, auque en estos últimos el modelo otorga ligeramente más peso al nivel C, seguido del B y por último el nivel D. Así, para el bosque los factores institucionales y académicos resultan más relevantes en el riesgo de deserción de los beneficiarios del Programa.

Finalmente, el accuracy promedio de los k-fold realizados con los datos de entreamiento del 94.2 % con SMOTE y 96.5 % para SMOTE-ENN en contraste con el accuracy en testeo (90 % y 91 % respectivamente) son relativamente similares y se podría suponer la generalización del modelo ante una base de datos con mayor cantidad de desertores, aunque la baja precisión representa un importante obstáculo en esta hipótesis.

- ***"SHAP (Shapley Additive Explanations)"***

Figura 18: Valores SHAP, Bosque aleatorio



Partiendo del hecho que el bosque aleatorio tuvo un desempeño inferior al árbol de decisión, la gráfica muestra resultados con menores efectos. Por una parte, la edad indica que un beneficiario joven tiene un claro mayor riesgo de deserción, sin embargo, su valor shapley es menor que en el gráfico anterior, en cuanto a las pruebas Saber 11 es definitorio que valores altos reducen dicho riesgo aunque el impacto (valor shapley) es menor.

Respecto a la interacción de las variables, sobre el cuadrante superior derecho, el modelo indicó que mayores puntajes en la prueba saber 11 junto con una menor edad, incrementa el riesgo de deserción,

aunque se evidencia un clúster con similares características que muestra lo contrario. Por su parte, en el cuadrante inferior izquierdo se guarda cierta relación con el gráfico del árbol de decisión, en esta oportunidad el efecto de la interacción es más negativo por el comportamiento de los puntos en el eje X, así, la interpretación de la interacción no cambia, una mayor edad y un puntaje bajo en la prueba Saber 11 incrementa el riesgo de deserción y en ciertos beneficiarios se presentan efectos contrarios, por lo tanto, se debe analizar si estos comportamientos atípicos obedecen a beneficiarios con condiciones socioeconómicas o individuales que mitiguen parcialmente el riesgo de deserción, por ejemplo, un beneficiario con 28 años (límite de ingreso al programa) con un relativo bajo desempeño en la prueba saber 11, puede tener un riesgo bajo de deserción si sus condiciones socioeconómicas le permiten mantenerse en los estudios o incluso asumir costos de recuperación de asignaturas perdidas, o, beneficiarios que bajo un supuesto de racionalidad buscan aprovechar la financiación del programa se esmeran por la titulación.

Clasificador por impulso de Gradiente:

Este modelo en particular llama la atención porque realiza un aprendizaje sobre los errores residuales de modelos previos ajustados. Su proceso contempla un primer modelo y cuyos errores se predicen en un segundo modelo y, de manera secuencial, ajusta varios modelos bajo la premisa del aprendizaje del error de predicción de sus predecesores, por lo que, se espera un mejor desempeño del mismo. El gradiente en su configuración automática toma árboles de decisión para su desarrollo secuencial, sin embargo, puede adoptar otro tipo de modelos como la regresión logística, las máquinas de soporte vectorial, los bosques aleatorios e inclusive las redes neuronales, entre otros modelos de clasificación.

Así, teniendo en cuenta al desempeño del árbol de decisión que hasta el momento presenta las mejores métricas por sobre el modelo logístico y el bosque aleatorio, se desarrolla el clasificador y, siguiendo la toma de decisiones técnicas se utiliza la optimización de los hiper parámetros mediante la técnica "GridSearch cross validation". Así, esta búsqueda arrojó los siguientes (mismos) hiper parámetros para SMOTE y SMOTE-ENN:

- Número de estimadores: 200
- Nivel de profundidad: 10
- Tasa de aprendizaje: 0.1

La cruadrícula nos indica que el gradiente tendría el mejor desempeño incorporando de manera secuencial 200 árboles de decisión, con un nivel de profundidad de 10 niveles para cada uno (mismo nivel de análisis que el árbol de decisión ajustado anteriormente) y la tasa de aprendizaje que hace referencia al peso o aporte que dará cada árbol al modelo final que, al ser baja, permite que el gradiente tenga más posibilidades de generalización dado que realizará un proceso de aprendizaje secuencial lento y con contribuciones bajas en cada árbol por lo que ninguno de estos impactará particularmente las predicciones generales del gradiente.

A partir de lo anterior, se ajustó el gradiente son SMOTE y SMOTE-ENN al 10 % reflejando los siguientes resultados en métricas de desempeño:

Métrica de evaluación	Valor con SMOTE	Valor con SMOTE-ENN
AUC	98.45	98.35
Sensibilidad	91.97 %	94.91 %
Especificidad	92 %	95.00 %
Accuracy	92 %	94.99 %
Precisión	52.04 %	55.58 %
F1-Score	61.21	70.10

Tabla 21: Desempeño del Gradiente

Las métricas mejoran respecto del bosque aleatorio, inclusive la sensibilidad, especificidad y accuracy

resultan en las más altas entre los modelos realizados hasta el momento, no obstante, la precisión no es mejor que la del árbol de decisión, así, se enfrenta el gradiente ante una subestimación de los desertores reales, teniendo en cuenta que la cantidad de los clasificados como tal es inferior a la real.

- ***"Características importantes o Feature importance"***

A continuación las tres variables que mayor peso consideró el gradiente en el riesgo de deserción:

Variables SMOTE	**Variables SMOTE-ENN**
Puntaje Saber 11	Puntaje Saber 11
Institución acreditada	Institución privada
Edad	Edad

Tabla 22: Variables con mayor importancia para el Gradiente

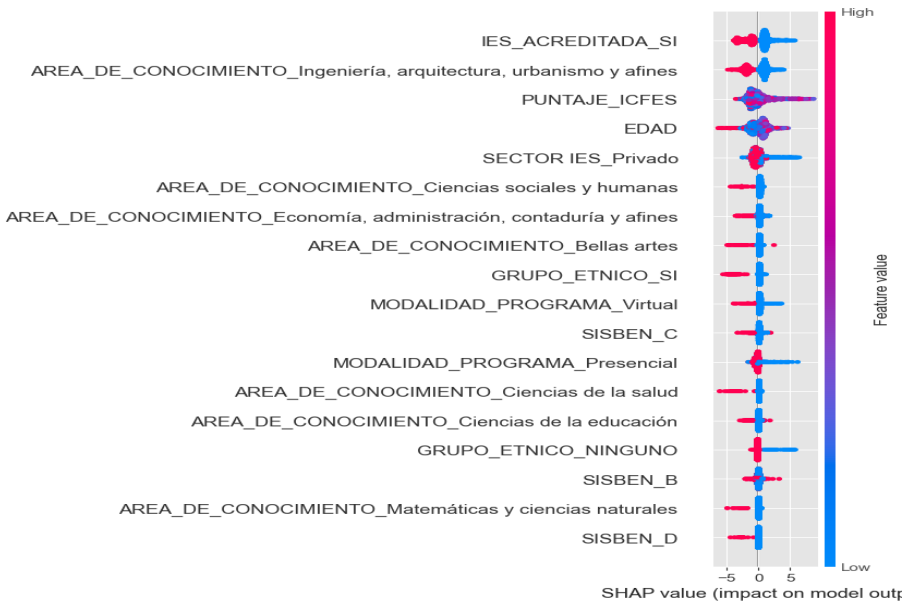
Ambos modelos guardando las consideraciones de los modelos previos, le dan menor peso a las áreas de conocimiento y los niveles socioeconómicos, aunque el gradiente contempla la edad. Sobre esto último, la edad es un factor importante si se tiene en cuenta que las personas a cierta edad se convierten en aportantes económicos de sus hogares, por tanto, el equilibrio entre la vida laboral y la académica representa por sí mismo un desafío para la permanencia en los estudios. Por otro lado, el gradiente retoma la lógica transitiva en los niveles socioeconómicos al otorgar más peso al nivel B, seguido del C y por último el nivel D. Así, nuevamente los factores institucionales y académicos resultan relevantes en el riesgo de deserción de los beneficiarios sumados a factores demográfico, en este caso representado por la edad.

Finalmente, el accuracy promedio de los k-fold realizados con los datos de entreamiento del 97.29 % con SMOTE y 98.44 % para SMOTE-ENN comparados con el accuracy en testeo (92 % y 95 % respectivamente) no refleja una diferencia significativa bajo el contexto de un conjunto de datos desbalanceado, por lo que, con mejores condiciones de estos el modelo presentaría sobresalientes métricas de precisión y se aseguraría su generalización.

En comparación con los modelos previos, la metodología SHAP aplicada al modelo del gradiente destaca todas las variables consideradas. Destaca sobre aquellas variables binarias que, si un beneficiario estudia en una institución acreditada de alta calidad es claro que disminuye el riesgo de deserción, en línea con lo anterior, cursar programas en áreas del conocimiento de las ciencias de la salud, matemáticas y ciencias naturales, ingenierías y bellas artes reducen sustancialmente el riesgo de deserción guardando relación con los resultados obtenidos en el modelo logístico. Obsérvese el coherente impacto del nivel socioeconómico, el nivel D (no pobre) evidencia el menor riesgo de deserción seguido por el nivel C (población vulnerable) y el nivel B (pobreza moderada) que inclusive muestra incrementos del riesgo de deserción para ciertos beneficiarios.

Llama la atención la consideración negativa de la edad sobre el riesgo, difiriendo sustancialmente de los modelos previos (quizás por esa interpretación que el modelo le da a la edad, este no logra una mejor precisión) y frente al resultado de las pruebas Saber 11 (puntaje-icfes) se muestran impactos positivos y negativos sobre el riesgo de deserción sin una tendencia clara, sin embargo, en este modelo se considera un efecto mayor (eje horizontal) en esta variable y en las demás a partir del posible valor shapley que se mueve en un rango de entre -5 (o menor) y 5.

Figura 19: Valores SHAP, Gradiente



Red Neuronal:

La red neuronal con la simulación de neuronas humanas representa un interesante mecanismo de clasificación y el cual no ha sido abordado ampliamente en la literatura de deserción y, al igual que los modelos anteriores, la red neuronal requiere de la definición de los siguientes hiperparámetros básicos:

- **Número de capas ocultas**: en donde se procesa la mayor cantidad de información de entrenamiento. Así, el número de capas ocultas determinará la profundidad y tiempo de operación de la red.
- **Número de neuronas por capa oculta**: establece la cantidad de neuronas dentro de cada capa oculta. Así, un mayor número de neuronas complejizará el modelo y el costo computacional.
- **Función de activación**: corresponde a la función que las neuronas aplicarán al conjunto de datos en cada capa. El resultado se transmite a la siguiente capa y por tanto a las demás neuronas. La funciones de activación más comunes son no lineales, sigmoides, tangentes hiperbólicas y para clasificación multiclase (softmax).

En las redes neuronales se contempla la función de pérdida como hiperparámetro que se utiliza para medir la diferencia entre las predicciones del modelo y los valores reales del conjunto de datos. La elección de esta depende del tipo de problema (regresión, clasificación binaria, clasificación multiclase, etc.). Así, en clasificación binaria se cuenta con la "Entropía cruzada binaria (Binary Cross-Entropy)", la cual mide la discrepancia entre las distribuciones de probabilidad de las etiquetas reales y las predicciones del modelo y que será contemplada en esta red.

Con lo anterior, la optimización de los hiper parámetros mediante la técnica "GridSearch cross validation" para la red arrojó los siguientes resultados para SMOTE y SMOTE-ENN:

- Función de activación: Relu (no lineal)

- cantidad de capas ocultas: 2
- Cantidad de neuronas por capa: 100 y 50 respectivamente

Se adicionó para la capa de salida una función de activación sigmoide partiendo de la variable binaria de interés. Se ajustó la red neuronal con SMOTE y SMOTE-ENN al 10 % con los siguientes resultados en métricas de desempeño:

Métrica de evaluación	**Valor con SMOTE**	**Valor con SMOTE-ENN**
AUC	93.61	93.36
Sensibilidad	85.66 %	88.41 %
Especificidad	85.65 %	88.42 %
Accuracy	85.65 %	88.42 %
Precisión	37.70 %	38.13 %
F1-Score	52.36	53.29

Tabla 23: Desempeño de la red neuronal.

Los resultados en general son aceptables, no obstante, la precisión de las redes es la menor entre todos los modelos lo que no es deseable dentro del contexto de identificación de desertores y, por tanto, en la capacidad de generalización esperada, si bien se tiene en cuenta el desbalance con la que debieron trabajar los modelos. Obsérvese a continuación las variables que mayor peso consideraron las redes frente al riesgo de deserción (se seleccionan las tres variables con mayor peso):

Variables SMOTE	**Variables SMOTE-ENN**
Puntaje Saber 11	Puntaje Saber 11
Edad	Edad
Grupo Étnico si	Grupo Étnico Si

Tabla 24: Variables con mayor importancia para la red.

Ambas, dan relativa menor importancia (y siguiendo los modelos previos) a las áreas de conocimiento y los niveles socioeconómicos SISBEN. Sin embargo, las redes consideran que los factores individuales (edad y grupo étnico) y los académicos resultan influyentes en el riesgo de deserción. Resalta la pertenencia de grupo étnico que en la fase descriptiva de los datos era ínfima con relación al universo de desertores, así, junto con la baja precisión de las redes es posible que este factor individual en realidad sea producto de la limitada capacidad del modelo para extraer información detallada con el desbalance de datos presentado.

El accuracy promedio de los k-fold realizados con los datos de entrenamiento (95.21 % con SMOTE y 95.68 % para SMOTE-ENN) en contraste con el accuracy en testeo (87 % y 88 % respectivamente) es relativamente similar. Las redes entonces son suceptibles al alto desbalance, por lo que, con una mejor cantidad de datos la red neuronal podría mejorar su precisión y generalizar su capacidad predictiva.

3.1. Comparación de los modelos implementados

Como resumen de las métricas de evaluación y variables consideradas por los modelos, se presenta la siguiente tabla resumen en donde se prioriza la mejor precisión de las técnicas de balanceo consideradas, en este caso, SMOTE-ENN:

Métrica de evaluación	Árbol de decisión	Bosque aleatorio	Gradiente	Red Neuronal
AUC	82.50	96.51	98.35	93.36 %
Sensibilidad	66.04 %	88.41 %	94.91 %	88.41 %
Especificidad	98.96 %	88.42 %	95 %	88.42 %
Accuracy	96.65 %	91.48 %	94.99 %	88.42 %
Precisión	82.71 %	44.70 %	55.58 %	38.13 %
F1-Score	73.44	60.04	70.10	53.29

Tabla 25: Comparación de métricas de los modelos.

La selección de la sensibilidad como criterio de interés parte de la utilidad pretendida por los modelos, esto es, la identificación adecuada de los desertores, para que, identificando posteriormente aquellos que presentan dicho riesgo se realicen las acciones respectivas para prevenirlo. Así, bajo este criterio el Gradiente resulta ser el mejor modelo de clasificación. Al respecto, es pertinente reiterar que la estructura de este concibe errores de modelos calculados dentro del algoritmo haciendo que el aprendizaje sea más robusto y esto se ve reflejado en los resultados obtenidos.

Si se comparan los modelos en términos de estabilidad de sus métricas (entendida como la menor diferencia entre la mejor y la peor) el árbol de decisión evidencia estabilidad, incluso el árbol presenta la mejor especificidad (identificación de reales negativos) seguido muy no de lejos por el gradiente. Ahora bien, el F1-SCORE que tiene en cuenta le precisión (predicciones correctas positivas y negativas) y la sensibilidad le otorga una ligera ventaja al árbol por sobre el gradiente dado que los demás modelos no presentan resultados similares. El gradiente presenta un bajo desempeño en la métrica de precisión en comparación con el árbol pero mejor que la red neuronal y el bosque aleatorio, dicha situación se presenta en la medida en que como se evidenció en los valores de la metodología SHAP, el gradiente valora particularmente en sentido contrario el efecto de la edad sobre el riesgo de deserción, el gradiente indica que una mayor edad reduce el riesgo aun cuando el modelo logístico y los valores SHAP el árbol de decisión y el bosque aleatorio permiten intuir que una mayor edad incrementa el riesgo.

Árbol de decisión	Bosque aleatorio	Gradiente	Red Neuronal
Puntaje Saber 11	Puntaje Saber 11	Puntaje Saber 11	Puntaje Saber 11
Modalidad virtual	Institución privada	Institución privada	Edad
Edad	Institución acreditada	Edad	Grupo Étnico

Tabla 26: Comparación variables relevantes para los modelos.

Los cuatro modelos seleccionaron como la variable más importante en la predicción del riesgo: el resultado de la prueba Saber 11, guardando relación con la literatura y (probablemente) con la realidad formativa en las cuales se destaca que un estudiante con deficiencias de aprendizaje heredadas de la educación escolar tendrá inconvenientes para adaptarse a la vida universitaria, enteder las temáticas y por tanto presentarán un alto riesgo de pérdida de asignaturas.

La edad se convierte en una variable común para 3 de los 4 modelos, le sigue la institución de carácetr privado y la modalidad virtual del programa, la acreditación institucional de alta calidad y la pertenencia a un grupo étnico. Así las cosas, retomando el desempeño del gradiente y el árbol (los mejores modelos por métricas) se identifica la prueba saber 11 y la edad como determinantes en el riesgo de deserción y, como se mencionó previamente se considera observar la sumatoria de condiciones y no su individualidad al momento de evaluar el riesgo de deserción y las acciones que se deben tomar, por ejemplo, dos beneficiarios con desempeño bajo en la prueba Saber 11 tendrían el mismo riesgo y reuqerirían estartegias de nivelación académica, sin embargo, si uno de ellos tiene 28 años y el otro 18, al primero le convendría un programa virtual (si aporta económicamente en su hogar y debe trabajar) y si bien los modelos no lo resaltan, las condiciones socioeconómicas (medidas como el nivel SISBEN) se deben tener en cuenta en la caracterización de los beneficiarios, así, el segundo joven puede tener un Nivel SISBEN A o B (pobreza extrema o moderada) y requerirá de otro tipo de apoyos (como los económicos).

Finalmente, es necesario evaluar el impacto que tienen las instituciones privadas en el riesgo de deserción. Para ello se debe tener en cuenta que en Colombia, las instituciones públicas presentan una desfinanciación histórica y, por tanto, el músculo financiero no es comparable con instituciones privadas que podrían destinar recursos para la implementación de medidas que prevengan el riesgo de deserción (apoyos psicosociales, profesionales específicos según el riesgo a tratar, infraestructura física, entrega de incentivos económicos, entre otros). En este punto, es recomendable comparar los planes de bienestar de las instituciones públicas y privadas para que en una armonización de procesos se identifiquen buenas prácticas aplicables desde las segundas hacia las primeras.

4. Conclusiones

1. Las y los beneficiarios de "Jóvenes a la U" presentan factores que influyen en su riesgo de deserción en general no diferentes a los que afectan a un estudiante regular de la educación superior.

2. La excelencia académica de los beneficiarios no necesariamente garantiza su permanencia en los estudios bajo la influencia de factores determinantes como los socioeconómicos y los individuales.

3. La edad es quizás el factor que en mayor medida puede generar un riesgo de deserción temprana, teniendo en cuenta que, un beneficiario con responsabilidades económicas y un tiempo cesante de estudios desde la graduación del colegio tendrá suficinetes condiciones para no adaptarse a las exigencias académicas, la jornada y la modalidad del programa de formación. Es así como, desde el momento de la inscripción deberá guiarse la selección de programas hacia aquellos que mitiguen estos riesgos, promoviéndose en este tipo de población formación virtual y en contra jornada a su trabajo.

4. Los estudiantes de programas presenciales tienen un mayor riesgo de deserción en contraste con programas virtuales. Como se explicó en el análisis exploratorio, la mayoría de beneficiarios proviene de las localidades ubicadas en el sur occidente de la ciudad, mientras las instituciones de educación superior se ubican principalmente en el borde oriental de la ciudad y aunque existe transporte masivo en dicho borde, la distancia desde el hogar al sitio de estudios o inclusive la capacidad económica para asumir los costos de transporte puden ser factores que explican el mayor riesgo de abandono en programas presenciales.

5. Los estudiantes de las ciencias de la salud y las ciencias naturales y matemáticas presentan un menor riesgo de abandono. Esta situación puede obedecer a aspectos vocacionales ya establecidos (no todos gustan de las emergencias médicas y no todos se consideran buenos en matemáticas). Así, este factor puede tener incidencia en la deserción que se presenta en otras áreas del conocimiento.

6. La vulnerabilidad económica evidencia la diametral diferencia de riesgo entre las poblaciones con pobreza moderada, vulnerables y no pobres quienes naturalmente presentan el menor riesgo de deserción. Esto implica que la oportunidad en la entrega del apoyo económico otorgado por el programa resulta en una estrategia trascendental para los beneficiarios.

7. Las estrategias para prevenir el riesgo de deserción deben tener un enfoque trasnversal. Si bien los beneficiarios del Programa tienen una afectación particular por las mayores condiciones de vulnerabilidad, es importante no descartar el acompañamiento vocacional y el fortalecimiento de competencias académicas. Así mismo, los factores individuales no contemplados en este análisis como la distancia del lugar de residencia, la discapacidad e inclusive la salud mental y general de la persona.

8. A partir de los factores identificados los esfuerzos de la Agencia Atenea y de las instituciones de educación superior en materia de prevención del riesgo de deserción podrán partir de la concepción de la vulnerabilidad socioeconómica de los beneficiarios, no necesariamente desde la perspectiva monetaria (que estaría cubierta en parte con la entrega del apoyo económico) sino entendiéndose

como el conjunto integral de condiciones que material y físicamente impiden el adecuado desarrollo de la formación académica.

9. La sumatoria de factores que incrementan el riesgo de deserción debe interpretarse como el primer grupo de beneficiarios sobre quienes deben aplicarse de manera integral medidas que reduzcan el riesgo de deserción y la temporalidad de intervención debe ser también considerada; no es lo mismo promover nivelaciones académicas cuando el beneficiario presenta a su vez vulnerabilidad económica o tiene una edad cercana al límite superior del programa, en este caso, además de la nivelación, debe acelerarse la entrega del apoyo económico y promover que el beneficiario curse un programa virtual o en jornada nocturna si se encuentra laborando.

10. La variabilidad de instituciones de educación superior que hacen parte del programa implica para la Agencia Atenea la necesidad de comprender el alcance que cada una de estas otorga a sus estrategias de prevención del riesgo de deserción, de manera que, se generen espacios entre las instituciones para que se adopten mecanismos y acciones funcionales desde la experiencia de cada una.

11. Subsidios al transporte también complementarían el apoyo económico otorgado por el programa, de tal manera que, si bien los estudiantes pueden ser primeros respondientes de su hogar, el proceso integral de cubrimiento de necesidades permita a este encontrar alternativas al abandono de la formación en virtud del apoyo brindado. Inclusive, en el caso de las madres y padres, resulta determinante contar con guarderías en las instituciones de educación superior para sobrellevar las labores de cuidado que desde la administración de Bogotá se intenta abordar con el programa "Manzanas del cuidado".

12. No se debe dejar de lado el papel de la Agencia Atenea como entidad Distrital. Al respecto, es imperativa la articulación con otras entidades oficiales que aporten en acciones de prevención o atención a las situaciones y condiciones que generan deserción en los beneficiarios, por ejemplo, orientar la atención en salud para la población con discapacidad, promover la oferta de oportunidades de empleo, generar redes para el apoyo jurídico y psicosocial en casos de víctimas de violencia basada en género, y en general, la oferta a servicios oficiales.

13. Las técnicas de machine learning empleadas coincidieron en que los factores institucionales y académicos son los más relevantes en el riesgo de deserción, particularmente reflejaron que la acreditación institucional y el sector de la institución tienen incidencia. Al respecto la acreditación es una distinción otorgada por parte del Ministerio de Educación Nacional y que reconoce condiciones pedagógicas, de infraestuctura, académicas, de bienestar estudiantil, entre otras y que impactan la formación y, por tanto, el perfil de egreso de los estudiantes. Adicionalmente, estos modelos identificaron la edad como factor individual de riesgo y el cual en el modelo logístico se acompañaba de un signo positivo, es decir, a mayor edad, mayor riesgo de deserción.

14. Otros factores individuales como la discapacidad o la pertenencia a algun grupo étnico pueden tener una mayor influencia en la deserción, sin embargo, la poca cantidad de desertores con estas características llevó a la no relevancia de los mismos, de manera que, en el futuro se esperaría que con más desertores se logre identificar si estos factores tienen incidencia determinante. Así mismo, otros factores que no fueron suministrados por la Agencia, como la condición de maternidad/paternidad, información de empleabilidad y constitución del hogar que son considerados en la literatura de deeserción, también pueden tener importante influencia en esta población.

15. Los métodos de imputación para el tratamiento de datos faltantes permiten conservar información en los conjuntos de datos, no solo sus valores absolutos sino las medidas de tendencia central que se ven reflejadas en la distribución de estos y su natural impacto en el modelamiento. Todos los métodos de imputación son igualmente válidos según el contexto del conjunto de datos, así, métodos sencillos como la imputación por media, mediana o la moda pueden aportar de manera significativa y a un costo computacional bajo tal y como se evidención con el nivel SISBEN. Por su parte, métodos más elaborados (donantes, vecinos más cercanos y algoritmos iterativos) solo son necesarios ante escenarios de considerables tasas de valores faltantes y no contar con suficiente

información que permita un modelamiento apropiado (regresiones). Adicionalmente, es consistente el uso de más de un método de imputación si la base de datos contiene información con diferente tipo de estructura, inclusive, bajo una misma estructura (variables continuas) pueden resultar apropiados doferentes métodos de imputación.

16. El desbalance de clases se debe contemplar desde la priorización de técnicas para la amplificación de la clase minoritaria y no recurrir a sub muestreo que conlleve a la pérdida de información. Sin embargo, según el contexto del conjunto de datos, el balanceo a aplicar no puede ser el propuesto por las técnicas por defecto (balance 1:1) y tampoco se debe considerar subjetivamente un balance que garantice desempeños aceptables para los modelos a implementar, toda vez que, como se ha explicado en este documento, una tasa alta de sobremuestreo lleva a que los modelos sesguen su desempeño debido al sobreajuste que se provoca con las muestras sintéticas, las cuales replican los valores en las variables independientes de manera que coincidan con los registros de interés. Así, el modelo contemplará únicamente como ciertos los valores correlacionados con la variable de interés y no tendrá en cuenta en nuevos conjuntos de datos información útil y heterógenea que no se presentó en el conjunto de entrenamiento, es decir, los modelos no serían generalizables.

17. Las técnicas implementadas en este documento resultaron susceptibles ante la imputación de datos y el balance aplicado. Particularmente, los árboles de decisión, el bosque aleatorio y el gradiente consideraron en general el impacto transitorio de los niveles SISBEN y el puntaje Saber 11 siguiendo los patrones encontrados en la literatura. En contraste, sobre el balance de clases se presentaron resultados heterogéneos, mientras el árbol de decisión presentó el mejor desempeño, los demás modelos presentaron limitaciones en su precisión, aunque la sensibilidad, especificidad, accuracy y F1-Score fueron aceptables y buenos en todos los modelos. Así mismo, la poca diferencia entre el promedio de valores accuracy de los k-fold con el valor accuracy de los datos de testeo de todos los modelos indican que estos tienen posibilidades reales de generalización. Se debe tener en cuenta que la tasa de sobremuestreo considerada en este trabajo fue conservadora con el fin de comprender el nivel óptimo de balance.

18. Con la identificación de los factores de deserción en los beneficiarios de "Jóvenes a la U", las técnicas de machine learning consideradas en este trabajo son válidas para la modelación del riesgo y, cada una responderá según el objetivo de interés, por ejemplo, priorizar la sensibilidad y la especificidad lleva a aplicar el bosque aleatorio y el gradiente acelerado. Si interesa la precisión y la exactitud de los modelos, encontramos en los árboles de decisión el mejor modelo.

5. Bibliografía

Referencias

Abella, B. M. (2021), 'Mejora de las predicciones en muestras desbalanceadas', **1**.

Amaya Torrado, Y. K., Barrientos Avendaño, E. & Heredia Vizcaíno, D. J. (2014), 'Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos'.

Andrade-Flores (2018), 'Comparativa entre classification trees, random forest y gradient boosting; en la predicción de la satisfacción laboral en ecuador', **2**.

Arismendy Fuentes, C. A., Morales Parrado, N. L. et al. (2018), 'Modelo de regresión logística como alternativa para medir la probabilidad de deserción temprana en la universidad de los llanos periodo 2015-2, 2018-1'.

Blagus-Lusa (2013), 'Smote for high-dimensional class-imbalanced data', **1**.

Castillo, M. B. & Giraldo, A. M. (2010), 'Los retos de la educación superior en colombia: una reflexión sobre el fenómeno de la deserción universitaria', *Revista Educación en ingeniería* **5**(10), 85–98.

Castro, et, a. (2021), 'Analizar factores demográficos, socioeconómicos y académicos asociados con deserción y graduación, mediante un modelo de riesgos competitivos. se incluyeron 639 estudiantes matriculados en 2009-2010, a quienes se les realizó seguimiento durante 14 periodos académicos'.

Diaz, D. B. & Garzón, L. P. (2013), 'Elementos para la comprensión del fenómeno de la deserción universitaria en colombia. más allá de las mediciones.', *Cuadernos latinoamericanos de Administración* **9**(16), 55–66.

Escarria, A. S. (2010), 'Deserción universitaria en colombia', *Academia y virtualidad* **3**(1), 50–60.

et all, M. (2022), 'A comprehensive investigation of the performances of different machine learning classifiers with smote-enn oversampling technique and hyperparameter optimization for imbalanced heart failure dataset', **1**.

et all, V. (2023), 'Predicción del éxito del telemarketing bancario mediante el uso de árboles de decisión', **4**.

Fandiño Benavides, A. (2022), Modelo de predicción para la deserción en la Fundación Universitaria Konrad Lorenz, PhD thesis, Bogotá DC: Fundación Universitaria Konrad Lorenz, 2022.

Flores-Congacha (2021), 'Factores asociados a la desnutriciÓn crÓnica infantil en ecuador. estudio basado en modelos de regresiÓn y Árboles de clasificaciÓn', **1**.

Friedman, J. (2000), 'Additive logistic regression: A statistical view of boosting', **28**.

Goicoechea, A. P. (2002), 'ImputaciÓn basada en Árboles de clasificaciÓn', **1**.

Gonzalez.et.all (2014), 'Algoritmos de clasificación y redes neuronales en la observación automatizada de registros', **1**.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Kennedy, S. (2005), *The business of lobbying in China*, Harvard University Press.

López, A. V. D. (2004), 'Estrategias para vencer la deserción universitaria', *Educación y educadores* **7**, 177–203.

McCulloch, W. & Pitts, W. (1990), 'A logial calculus of the ideas immanent in nervious activity', *Carnegie Mellon University Journall* **52**(1), 1–17.

McGinnis, D. (1994), 'Predicting snowfall from synoptic circulation: A comparison of linear regression and neural network methodologies', **1**.

MEN (2023), 'Sistema para la prevención de deserción'.

Páramo, G. J., Maya, C. A. C. et al. (1999), 'Deserción estudiantil universitaria. conceptualización', *Revista Universidad EAFIT* **35**(114), 65–78.

Perez Bedia, C. A. & Rojas Segovia, L. E. (2020), 'Diseño de un sistema para predecir la deserción de los alumnos mediante machine learning en la universidad tecnológica del perú'.

Peter Gould, volume=1, y. p. (n.d.), 'Neural computing and the aids pandemic: The case of ohio'.

Rodríguez Rios, C. Y. (2012), 'Diseño de un modelo estocástico usando cadenas de markov para pronosticar la deserción académica de estudiantes de ingeniería. caso: escuela colombiana de ingeniería julio garavito'.

Romero-Carmona-Pozuelo (2021), 'La predicción del fracaso empresarial de las cooperativas españolas. aplicación del algoritmo extreme gradient boosting', **101**.

Rugeles, N. (2022), 'Algoritmos de boosting para modelos de clasificaci´on y regresi´on lineal', **1**.

SDP (2020), 'Plan distrital de desarrollo 2020-2024:'un nuevo contrato social y ambiental para la bogotá del siglo xxi".

Tinto, V. (1975), *Dropout from higher education: A theoretical synthesis of recent research*, Vol. 45, Sage Publications Sage CA: Thousand Oaks, CA.

Turco, L. Y. D. (2021), 'Mejora de la prediccIÓn en tareas de aprendizaje automÁtico supervisado-imputaciÓn de valores perdidos', **1**.