DETECCIÓN DE ALERTAS TEMPRANAS PARA LA PREVENCIÓN DE LA DESERCIÓN ESTUDIANTIL EN UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR A PARTIR DE UN MODELO DE CLASIFICACIÓN Y SU PREDICCIÓN POR MEDIO DE TÉCNICAS DE MACHINE LEARNING

DETECTION OF EARLY WARNINGS FOR THE PREVENTION OF STUDENT DROPOUT IN A HIGHER EDUCATION INSTITUTION FROM A CLASSIFICATION MODEL AND ITS PREDICTION THROUGH MACHINE LEARNING TECHNIQUES

William Castiblanco Vargas[1]
Lida Rubiela Fonseca Gómez[2]
Wilmer Pineda-Ríos[3]

Resumen

El abandono estudiantil es uno de los principales problemas en la educación en general y toma especial interés en las instituciones de educación superior. Este artículo aborda la detención de las alertas tempranas para la prevención de la deserción estudiantil universitaria con la aplicación del modelo probabilístico de la regresión logística y su predicción por medio de técnicas de Machine Learning (aprendizaje supervisado); como los árboles de decisiones, Bagging: Boostrap Aggregation y clasificador de bosques aleatorios. Para el estudio se seleccionó una muestra de estudiantes que se encontraban estudiando en un periodo lectivo los cuales pertenecen a todos los programas de la Universidad. La aplicación del modelo de clasificación logístico permitió conocer los principales causantes por las cuales la población estudiantil es potencial desertor de la universidad. Estos resultados permiten a la institución universitaria generar políticas para un plan de retención y permanencia estudiantil que repercutan de manera específica y redireccionando los recursos hacia objetivos claros con el propósito de mitigar la deserción.

Palabras claves: Deserción, Regresión Logística, Balanceo, Aprendizaje Supervisado, Técnicas de Machine Learning, Árbol de Decisión, Baggin: Boostrap Aggregation, Clasificador de Bosques Aleatorios

Abstract

The student dropout is one of the main problems in the general education and takes special interest in higher education institutions. This article addresses the detection of early warnings for the prevention of university student dropout with the application of the probabilistic model of logistic regression and its prediction through Machine Learning techniques (supervised learning); such as

Recepción: Septiembre de 2021 / Evaluación: Octubre de 2021 / Aprobado: Noviembre de 2021

[1]Estudiante de Maestría en Estadística Aplicada. Especialista en Estadística Aplicada. Ingeniero Mecánico de la Universidad de los Andes. Director de Estadística. william.castiblanco@usantotomas.edu.co https://orcid.org/0000-0002-7866-9164.

[2] Estudiante de Doctorado en Educación Universidad Pedagógica, Docente Estadística, Universidad Santo Tomás, Fundación Universitaria Los Libertadores. Magíster en Educación, Especialista en Diseño de Ambientes de Aprendizaje https://orcid.org/0000-0002-3597-728X. lidafonseca@usantotomas.edu.co

[3] Estudiante de Doctorado en Estadística Universidad Nacional De Colombia. Magister en matemáticas de la Universidad Nacional de Colombia. Conocimientos en Diseño de Experimentos, Modelamiento Estadístico, Procesos Estocásticos y Datos Funcionales. Experiencia como Consultor Estadístico en proyectos de investigación académicos. wpinedar@unal.edu.co

decision trees, Baggin: Boostrap Aggregation and random forest classifier. For the study, a sample of students was selected who were studying in an academic period who are belonging to all the programs of the University. The application of the logistic classification model allowed us to know the main causes for which the student population is a potential university dropout. These results allow the university institution to generate policies for a student retention and permanence plan that have a specific impact and redirect resources towards clear objectives in order to mitigate desertion.

Key words: Dropout, Logistic Regression, Balaced Method Learning Superviced, Machine Learning tecniques, Tree Classifier, Baggin: Boostrap Aggregation, Random Forest classifier.

Introducción

Deserción

Una de las bases por las cuales se mide el desarrollo social y económico de una nación es por la capacidad que tiene su sociedad de progresar y proponer posibles soluciones a los desafíos que se presentan tanto a nivel social como económico. Ahora bien, para lograr estos retos la sociedad debe estar preparada tanto intelectual como anímicamente. Esto solo se logra si un gobierno invierte en educación en su propio pueblo. Durante las últimas décadas, los países más desarrollados son los que le apuestan a la formación de su población con un alto nivel de importancia. Esto trae consigo la posibilidad de abrir las brechas de oportunidades para el desarrollo, una interacción o matrimonio con el sector productivo que permite incrementar el desarrollo económico de la industria y una mano de obra altamente calificada necesaria para mejorar las soluciones que tanto adolece el sector productivo. Este factor, no se ve muy reflejado en nuestros países en vía de desarrollo. Es así como el gobierno nacional en cabeza del Ministerio de Educación Nacional (MEN) ha impulsado una discusión nacional, que ha posibilitado la reflexión y la construcción de un proyecto educativo coherente con las exigencias que le impone una sociedad globalizada, especialmente en materia laboral y productiva (VELEZ, 2008).

Deserción en las universidades

Es importante comprender la deserción como un fenómeno que afecta tanto al estudiante, la universidad, como la nación misma. En primer lugar, al mismo estudiante; debido a que el retiro de sus estudios le atrasa el desarrollo económico tanto personal como de su familia. En segundo lugar, se encuentra la universidad; ya que la salida temprana del estudiante causa disminución de los ingresos de la organización y la no graduación causa la pérdida de un fuerte relacionamiento con el sector industrial y en un tercer lugar, se encuentra la Nación; debido a que como se dijo antes, el potencial económico de una Nación se mide por el desarrollo social e intelectual de su población (desarrollos investigativos e industrial, los cuales se ven afectados de manera directa) y esto solo se logra con la profesionalización de sus estudiantes en las diferentes disciplinas. Adicionalmente, la deserción involucra diferentes aspectos de acuerdo al área de conocimiento, por nivel de formación, por origen institucional (públicas y privadas), entre otros. (VELEZ, 2008).

Es importante comprender el fenómeno de la deserción desde su definición; para el Ministerio de Educación Nacional (entidad nacional de supervisión y vigilancia de la calidad de la educación); se entiende la deserción como aquella en la cual el estudiante se ausenta, bien sea de un programa, institución o del sistema de educación nacional. La deserción anual se define como aquella en donde el estudiante se ausenta por dos periodos consecutivos y no se gradúa y la deserción por cohorte cuando se realiza el seguimiento al estudiante desde su primer semestre de

ingreso hasta cuando ha dejado el claustro universitario por retiro en los subsiguientes semestres de su ingreso.

La necesidad y los costos elevados que ocasiona la profesionalización de un estudiante, así como las dificultades tanto económicas, sociales, culturales, familiares y académicas que tienen para obtener un título profesional ha conducido al abandono temprano del alma mater, trayendo consigo el ausentismo y posteriormente la deserción. (RICO, 2019) Colombia es uno de los países de Latinoamérica que se encuentra interesada en disminuir estos indicadores y para ello ha implementado políticas a nivel Nacional para enfrentar este fenómeno. Según SPADIES del MEN la tasa deserción en las Instituciones de Educación Superior a nivel Nacional se encuentra en el 12,5 %, tasa que involucra tanto las universidades públicas como privadas.[4]

Delimitación del problema

Existen innumerables estudios que abordan la problemática de la deserción desde diferentes disciplinas y perspectivas aportando a la búsqueda de alternativas de solución con la identificación de factores que la afectan de manera directa e indirectamente. Es así como Castaño et al., (2004) han identificado una serie de factores y los clasifica en determinadas variables. Entre estos factores se encuentran las personales, las académicas, socioeconómicas y las institucionales.

Para la explicación del fenómeno de la deserción inicialmente se implementaron diversas metodologías estadísticas de corte estático y descriptivo, tales como modelos de regresión logit, probit y análisis discriminante, entre otros. Luego se encaminó a la aplicación de modelos de duración, también llamados análisis de supervivencia; el cual captan la variabilidad de las variables involucradas a través del tiempo hasta el momento de ocurrencia del hecho. (GIOVAGNOLY, 2002). Se llevó a cabo otro estudio en donde se realizó un modelo de duración, en él se encontró según los resultados de la supervivencia y función de riesgo, para los estudiantes de algunas carreras específicas analizadas, la existencia de mayor probabilidad de supervivencia en el primer semestre y fue disminuyendo en los semestres sucesivos. (DÍAZ J, 2009).

Para la detección de patrones de bajo rendimiento académico de los estudiantes hay modelos que utilizaron técnicas de minería de datos para descubrir patrones en grandes volúmenes de datos. Estas técnicas permitieron con análisis descriptivo identificar las variables más influyentes y establecer el modelo de clasificación para el bajo rendimiento académico de los estudiantes. (ULLOA G, 2020).

Para la predicción de la deserción estudiantil hay estudios que utilizaron modelos como la regresión Logística, la cual permitió identificar los patrones como los factores académicos de los estudiantes para aumentar la probabilidad de deserción. Específicamente, utilizó técnicas de Machine Learning y Minería de Datos. (PRIETO A, 2015).

Otros modelos utilizaron las técnicas de Machine Learning para la identificación y predicción de estudiantes en riesgos de deserción académica, tales como árbol de decisión y Random Forest y a través del cálculo de las métricas seleccionó el modelo más influyente. Sin embargo, los cálculos de la matriz de riesgos arrojaron valores discretos entre el 25% y el 67%, en donde el Random Forest clasificó al 79% de los desertores de manera correcta y al restante 21% los clasifica como no desertores. (GONZÁLEZ S, 2021).

Finalmente, otro estudió realizó la comparación de diferentes técnicas de minería de datos para la identificación de indicios de deserción estudiantil, a partir del desempeño académico en cada semestre de la carrera, el cual por medio de la comparación del área AUC se obtuvo una mejor

[4] SPADIES: Sistema para la prevención de la deserción en las Instituciones de educación superior

área en la aplicación del modelo de Random Forest durante los primeros semestres, acentuándose en el tercer semestre, sin embargo, este estudio trabajó con pocos datos de un programa en específico. (PÉREZ G, 2020)

Propósito

Bajo este panorama el presente artículo tiene como propósito generar estrategias para la detección de alertas tempranas para la mitigación de la deserción estudiantil en una institución universitaria por medio de la aplicación de un modelo de clasificación y realizar su predicción a través del uso de técnicas de Machine Learning como Árbol de Decisión, Bagging: Boostrap Aggregation y Clasificador de Bosques Aleatorios.

Organización

En la estructura del artículo se encuentra la primera sección; la cual presenta los preliminares de la problemática de la deserción y estudios realizados sobre el tema en específico, una segunda sección está compuesta por la metodología impartida en el estudio, una tercera sección la conforma los resultados encontrados con la aplicación de cada uno de los modelos aplicados y finalmente, se cuenta con la cuarta sección la cual contiene las conclusiones.

Metodología

El desarrollo de esta investigación inicia con la recolección de la información que lleva a un análisis de variables que determinan la caracterización de la población estudiantil, consecuentemente, con la aplicación de los modelos de Machine Learning para la predicción de la deserción conduce a que el método de esta investigación se fundamente en lo inductivo, de tal manera que se basa en el uso del razonamiento para llegar a conclusiones. Estos razonamientos, a su vez parten de hechos particulares que son aceptados como ciertos y así llegar a ellas, en donde su aplicación es de carácter general. (BERNAL, 2010). Adicionalmente, el estudio es descriptivo debido al uso de determinadas variables que explican el fenómeno de la deserción y que con base en ellas determinan las características más importantes para la predicción de esta. Es así, como los estudios descriptivos buscan especificar las propiedades importantes de personas, grupos, comunidades o cualquier otro fenómeno que sea sometido a análisis, los mide y los resultados le sirven para describir el fenómeno de interés. (HERNÁNDEZ, 2014). Finalmente, esta investigación es no experimental debido a que la información obtenida de los estudiantes no ha sido manipulada deliberadamente. Es decir, se trata de un estudio en el que no se hace variar en forma intencional las variables independientes para ver su efecto sobre otra u otras variables, y corresponde a un estudio transversal o transeccional, por cuanto los datos obtenidos de la población estudiantil fueron tomados en un momento específico de espacio (Institución de Educación superior) y en un tiempo específico (1° semestre del periodo académico del 2021) (HERNÁNDEZ, 2014, pág. 152).

La encuesta

Se construyó un cuestionario que involucró un grupo de 41 preguntas que respondieron a la clasificación en los cuatro grandes aspectos (individuales, socioeconómicas, académicas e institucionales) que son la influencia para la toma de decisión por parte del estudiante de dejar el claustro universitario. [Castaño, et al. 2004]. Dicho instrumento fue validado por un grupo de psicólogos expertos quienes analizaron la pertinencia de las preguntas de acuerdo con las características de la población objeto de estudio y teniendo presente la misión de la universidad.

Adicionalmente, se realizó una prueba piloto con un grupo de estudiantes con el propósito de validar el instrumento respecto a la claridad en las preguntas y a la objetividad del mismo. Este instrumento fue dirigido a la población de los 16700 estudiantes activos de los cuales se seleccionó una muestra de 4987, con la aplicación del método de muestreo aleatorio simple. Primero, se seleccionó las muestras proporcionales a los tamaños de los programas que se encuentran activos en la universidad (estratificación), luego, teniendo presente la proporcionalidad de los tamaños de los semestres se seleccionó la muestra y finalmente, por el método de conglomerado se seleccionaron los integrantes de los cursos quienes diligenciaron la encuesta.

Diseño Muestral

La construcción del diseño muestral se caracterizó por la existencia de un marco muestral constituido por los programas ofertados por la universidad, adicionalmente, cada uno de ellos cuenta con un número de grupos, los cuales los estudiantes fueron seleccionados de acuerdo a las listas pertenecientes a cada grupo. Esto generó un muestreo en dos etapas, en la primera etapa se empleó un muestreo estratificado, siendo el estrato uno; el tipo de programa (15 profesionales y17 tecnológicos), un segundo estrato lo conformó los 2154 grupos distribuidos en todos los programas de la universidad. La segunda etapa del diseño muestral empleo un muestreo aleatorio simple proporcional al tamaño para elegir los grupos de cada programa; originando 332 grupos quienes fueron los seleccionados para ser encuestados, el cual tuvo un nivel de confianza del 95 %, un margen de error del 5%, una probabilidad de éxito y fracaso del 50% y un factor de expansión de 6.5. Para la selección de cada grupo se empleó el método de coordinado negativo. (LOHR, 2000) A continuación, se presenta la ecuación del tamaño total de la muestra:

$$n = \frac{\sum_{i=1}^{l} N_i P_i Q_i}{NE + \frac{1}{N}\sum_{i=1}^{l} N_i P_i Q_i}$$

Siendo $E = \frac{d^2}{Z_{1-\alpha/2}^{\ 2}}$

y el tamaño de cada estrato está definido por:

$$n_i = n\left[\frac{N_i}{\sum_{i=1}^{l} N_i}\right] = n\left[\frac{N_i}{N}\right] = n(W_i)$$

Siendo la unidad muestral del diseño los grupos que constituyen la unidad básica de la investigación y la unidad Estadística son los estudiantes encuestados en cada uno de los grupos de los diferentes programas. Este diseño muestral generó un total de 4985 estudiantes encuestados.

Técnicas Estadísticas

Modelo de clasificación

Al conjunto de la base de resultados se le aplicó el modelo de regresión logística; el cual hace parte del grupo de los modelos lineales generalizados (GLMs) (HARDIN, 2012), este modelo consiste en poseer una variable dependiente caracterizada por ser dicotómica (0 - 1), la cual para el estudio hace referencia a la variable que representa al estudiante que es potencial (representado como 1) o no potencial desertor (representado como 0) y por el otro lado, un conjunto de variables explicativas que se encuentran representadas por el conjunto de variables que son tenidas en cuenta para la explicación del estudio. El propósito de los modelos generalizados es especificar la relación entre la variable respuesta y el conjunto determinado de covariables. De esta maneara, la variable respuesta es vista como una realización de una variable aleatoria. (MADSEN, 2011)

Los GLMs tradicionalmente se encuentran formulados dentro de la estructura de distribuciones de la familia exponencial; la cual matemáticamente se encuentra representada por: (HARDIN, 2012)

$$f_y(y{:}\,\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

En donde:
θ: Es el parámetro canónico (natural) de localización. Este parámetro relaciona a las medias.
ϕ: Es el parámetro de escala. Relaciona a las varianzas.

La función de probabilidad binomial se expresa así:

$$f(y{:}\,k,p) = \binom{k}{y} p^y (1-p)^{1-y}$$

La distribución Bernoulli o la distribución de probabilidad de respuesta binaria es la forma simplificada de la distribución binomial (parámetro $k = 1$ y la variable respuesta se restringe a los valores $\{0,1\}$). La función de probabilidad de respuesta binaria se representa así:

$$f(y{:}\,p) = p^y (1-p)^{1-y}$$

En la forma de la familia exponencial, la distribución de Bernoulli se representa como (HARDIN, 2012):

$$f_y(y{:}\,p) = exp\left\{yln\left(\frac{p}{1-p}\right) + ln(1-p)\right\}$$

La forma canónica comúnmente se refiere al modelo logit, el cual puede ser vista como la regresión logit o regresión logística, sus parámetros se definen así:

$$\theta = ln\left(\frac{p}{1-p}\right) = \eta = g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$$

El modelo de regresión logístico binario se define como el logaritmo de la probabilidad del éxito respecto a la probabilidad del no éxito; $ln(Odds)$:

$$logit(P) = Log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k$$

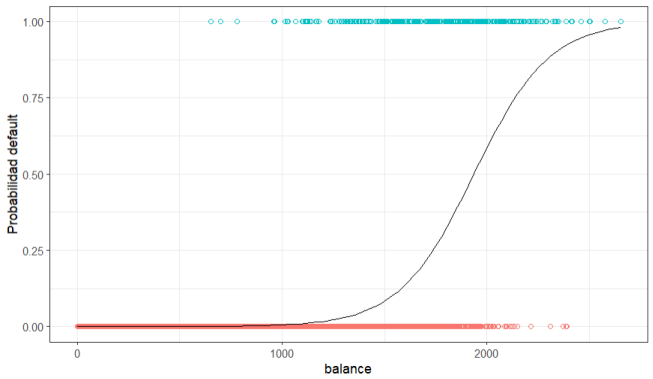
La función inversa de la función $logit(P)$ daría la función sigmoide:

$$logit^{-1}(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

En donde x es la combinación lineal: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
Luego la probabilidad de un evento que sea éxito, sería:

$$P(y=1) = P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k} + 1}$$

Su representación gráfica es:



Fuente: Elaboración propia, basado en la base de datos tomada de https://rpubs.com/Joaquin_AR/229736

Este modelo, puede modelar la probabilidad de que ocurra un evento partiendo de variables. Además, estima la probabilidad de que un evento ocurra para una observación al azar versus la probabilidad de que no ocurra (Odds), predice el efecto de una serie de variables en una variable categórica binaria y clasifica observaciones a través de la estimación de la probabilidad de que se encuentre en una categoría determinada.

Árbol de decisión

Se encuentra clasificado dentro de los modelos de predicción, se fundamenta en el aprendizaje inductivo partiendo de observaciones lógicas, el cual tiene un comportamiento en sus predicciones basadas en reglas que permiten la representación y categorización de las condiciones que se suscitan de forma repetitiva para la solución del problema. Su representación gráfica se hace por medio de un conjunto de nodos, hojas y ramas (BARRIENTOS M, 2009). Este método de clasificación no almacena los datos de entrenamiento en su totalidad, utiliza los datos de entrenamiento para construir una estructura de estratificación o de árbol que divida recursivamente el espacio de entrenamiento en regiones que tengan etiquetas o características similares. Tiene la característica de realizar una partición recursiva en donde la variable dependiente puede ser continua, discreta o categórica y las variables predictoras pueden ser continuas, discretas o categóricas y finalmente, tiene la facilidad de poseer el esquema de interpretabilidad. (RAMASUBRAMANIAN, 2019)

Bagging: Boostrap Aggregation

El método Bagging, conocido como Bootstrap Aggregation; es una técnica de aprendizaje de ensamblaje que de alguna manera ayuda a mejorar el desempeño y la exactitud en los algoritmos con que cuenta el aprendizaje de máquinas, es usado tanto en los modelos de regresión como de clasificación (entre ellos en los algoritmos de clasificación). Una característica importante que tiene el Bagging es que evita el sobre ajuste de los datos. (HAROON, 2017). Como el modelo de árbol de decisión cuenta con la desventaja de tener mucha varianza en la estimación, lo que se busca es reducir esta varianza, adicionalmente, se pierde interpretabilidad con el esquema. El método

Bagging, busca ajustar varios modelos los cuales son independientes, en donde se promedia sus predicciones de tal manera que se obtenga un modelo con una varianza de estimación baja (HAROON, 2017). Se buscan las muestras de bootstrap; el cual se define como un método de remuestreo; en donde en lugar de muestrear en distribuciones de poblaciones que no se conocen en la práctica, remuestrea sobre los datos existentes en donde los toman de forma aleatoria y son independientes (FARAWAY, 2005). Luego, se ajusta un árbol a estos datos, se selecciona un conjunto de variables para su predicción y estas mismas variables se implementan en todas las muestras seleccionadas.

Random Forest Classifier

Random Forest Classification es un tipo del método de Bagging, este método de ensamblaje se caracteriza por tomar una cantidad de muestras de tamaños específico, que sean independientes e idénticamente distribuidas al conjunto de datos que se está analizando. Para la clasificación, cada árbol en el Random Forest da una especie de voto para la más repetitiva clase, la salida del clasificador está determinado por la mayoría de votos del árbol. (ULLOA G, 2020) Esto conduce a que en la medida que se tomen muchas réplicas de muestras de la base original; se induce a una correlación positiva alta en las predicciones y esto puede conducir a un sobreajuste. Para evitar esto, realizando una decorrelación en el bootstrap puede disminuir la correlación y por ende disminuir la varianza alcanzando mejores predicciones y así evitar el sobreajuste de los datos, esto recibe el nombre de la técnica de Random Forest (BISONG, 2019). El procedimiento se caracteriza en construir los árboles con diferentes variables para realizar la predicción independientemente del conjunto de datos de cada árbol que está generando la información, esto evita conformar la correlación positiva inducida por el método de Bagging; esto es lo que conduce al método de decorrelación. Finalmente, el método lo ensambla dependiendo del método si es clasificación o regresión y por el método de clasificación escoge la moda

Balanceo

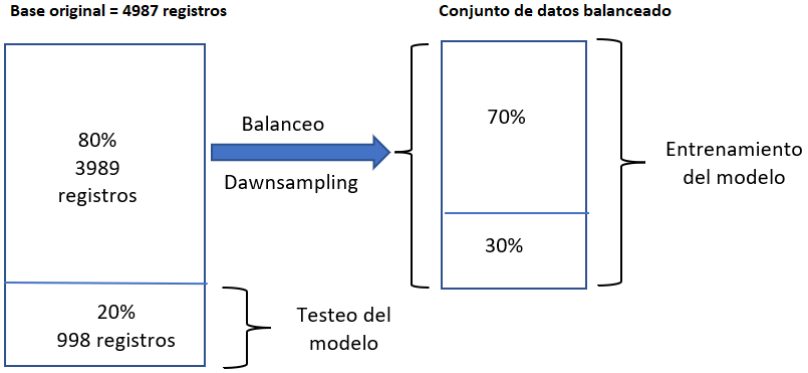
Cuando se trabaja con modelos de clasificación se busca que la variable predictora se encuentre balanceada, es decir que exista un equilibrio entre las categorías de la variable predictora, de lo contrario se conduce a encontrar predicciones sesgadas y los cálculos de las métricas serán engañosos. (NITESH V, 2002). Para el caso en particular, la variable predictora se encuentra categorizada en dos grupos, el grupo de los estudiantes que se consideran potenciales desertores (clase minoritaria con 354 casos que corresponden al 7.6%) y el de los no potenciales desertores (clase mayoritaria con 4633 casos que equivalen al 92.4%). El no balanceo conduce a un posible sesgo al momento de predecir, debido a que tendería a realizar mayor predicción sobre los estudiantes que no son potenciales desertores y una menor predicción sobre los potenciales desertores y esto no es lo que se pretende en el estudio, debido a que el interés se centra sobre la detección en la mayoría de los estudiantes que sean potenciales desertores.

Para el desarrollo del estudio, primero se dividió la base en dos grupos, a saber: un conjunto de datos que corresponde al 80% de la información que recibe el nombre de base de entrenamiento (train) y la otra que corresponde al restante 20% que recibe el nombre de base de prueba (test). La partición se realiza con el propósito de entrenar el modelo con el conjunto de datos de entrenamiento y luego se testea o comprueba el modelo con el conjunto de datos de prueba. Es así como, con la base del 80% se balancea los datos y se entrena el modelo y con el restante 20% se testea el modelo.

Entre los métodos de balaceo se encuentra el Dawnsampling, el cual consiste en disminuir la clase mayoritaria y lograr balancearla con la clase minoritaria. (VIDHYA, 2016). Es así, como se toma el conjunto de datos del 80% y se le aplica el método de balanceo Dawnsampling, al cual se le aplica el entrenamiento del modelo, para luego realizar el testeo con la base de prueba o test. Finalmente, al grupo test del 20% se le aplicó el modelo de clasificación para realizar su predicción y de igual forma se calcularon sus métricas para comparar los resultados con los de la base balanceada.

Para la aplicación del método de balanceo dawnsampling, la base de los 3989 registros se encuentra clasificada con 3703 "No potenciales desertores" y 286 "Potencial desertores". De esta manera se buscó obtener el 70% de estudiantes "NO Potenciales desertores" y un 30% de la clase "Potenciales desertores", (se busca que el 30% corresponda a los 286). Luego para obtener el 100% de los registros se requiere multiplicar los 286 registros "*Potencial desertores*" por una constante (3,332), esto equivale a 952 registros que corresponden al 100%. De esta manera, el 70% corresponde a los 666 registros (Potencial no desertor) y los 286 corresponden al 30% (Potencial desertores). Finalmente, se unen las dos bases y se mezclan para obtener la nueva base con los 952 registros y las correspondiente 41 variables.

Gráfica 1. Esquema del método de balanceo Dawnsampling



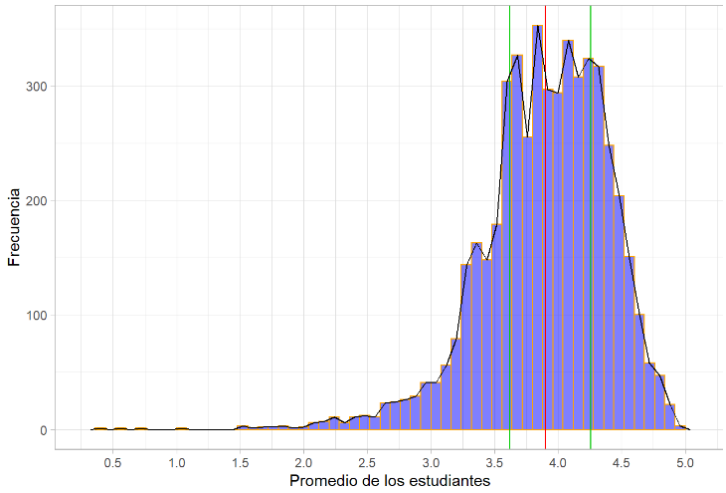
Fuente: Elaboración propia

Resultados

Exploratorios

La caracterización de la población estudiantil de la universidad con relación al aspecto académico se define con un promedio académico de 3.9, el promedio de materias aprobadas es de 18, consecuentemente, el promedio de créditos aprobados es de 57 y el promedio de materias reprobadas es de 2 y un 83% considera no tener buenas bases del bachillerato. Respecto al género se destaca que el 46% son mujeres y el restante 54% son hombres, es una población con un promedio de edad de 24 años; la cual se considera relativamente joven a pesar que la universidad cuenta con una jornada continua (diurna y nocturna).

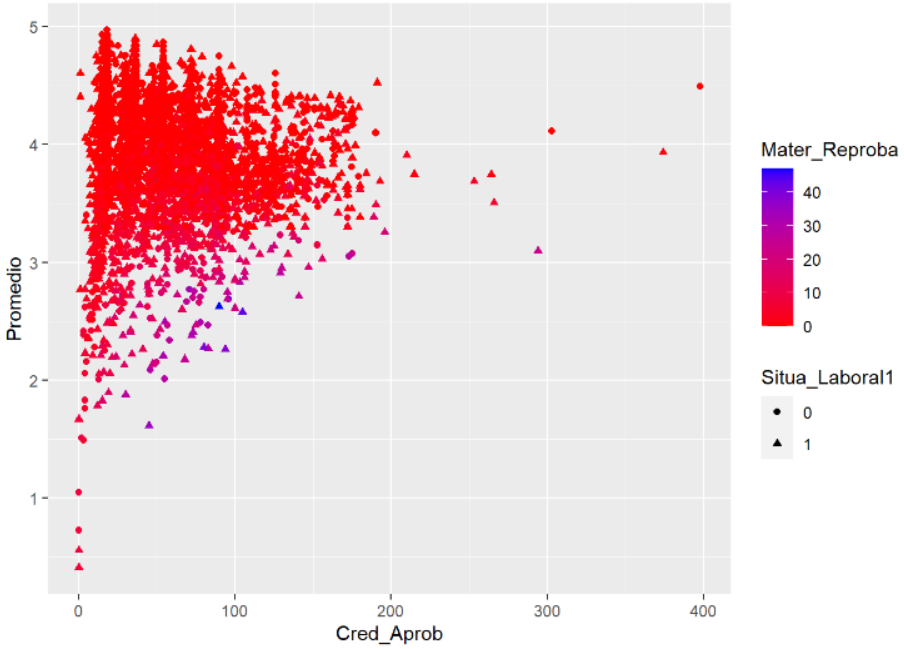
Gráfica 2. Distribución del promedio académico de los estudiantes



Fuente: Elaboración propia

La población estudiantil tiene una caracterización con la existencia de un mayor número de créditos aprobados con promedios superiores a 3.0. Sin embargo, existe un grupo de estudiantes con un buen rango de materias reprobadas (entre 20 y 30) que se encuentran con promedios inferiores a 3.0. Luego, se considera la existencia de un grupo de estudiantes potenciales desertores con estas características. Adicionalmente, se aprecia la existencia de un grupo de estudiantes que estudian y trabajan al mismo tiempo y hacen parte de aquellos que tienen más materias reprobadas (entre 20y 40). Es así como la situación laboral del estudiante influye en la toma de decisión del estudiante a dejar sus estudios superiores.

Gráfica 2. Relación entre el promedio y los créditos aprobados de acuerdo al número de materias reprobadas de los estudiantes



Fuente: Elaboración propia

Adicionalmente, el estado civil de los estudiantes está representado por un 88% de solteros, 9% se encuentra en unión libre y un 3% son casados. Con relación al aspecto socioeconómico; se destaca que el 90% pertenecen a los estratos 2 y 3, un 6% son de estrato 1 y un restante 4% son del 4. Adicionalmente, el 71% cuenta con dependencia económica, un 38% vive en arriendo, un 44% es familiar, un 12% la tiene en hipoteca y solamente un 3% es propia. Además, un 51% no tiene personas a cargo y el restante tiene uno, dos o tres personas a cargo. Finalmente, en relación al aspecto institucional, se destaca que el 51% ha cancelado una o más materias durante su carrera y un 40% ha cancelado el semestre una o más de una vez, además, el 31% ha pensado en retirarse de la universidad, un 83% se siente satisfecho con la carrera y un 57% afirma no tener los recursos económicos para sostenerse en la universidad. La tabla 1 muestra el resumen de algunas de las principales variables contempladas en los cuatro factores determinantes en la deserción estudiantil (CASTAÑO, 2004).

Tabla 1: Variables contempladas en los cuatro aspectos

Factores	Variable	Categorías							
Académicas									
	Promedio académico	3,9							
	Materias aprobadas	18							
	Créditos aprobados	57							
	Materias reprobadas	2							
	Bases de bachillerato	No	83%	Sí	17%				
Individuales									
	Género	Femenino	44%	Masculino	56%				
	Promedio edad	24 años							
	Estado civil	Soltero	88%	Unión libre	9%	Casado	3%	Divorciado	0,30%
Socioeconómicas									
	Estrato	Uno	6%	Dos	45%	Tres	45%	Cuatro	4%
	Dependencia económica	Si	71%	No	29%				
	Tipo de vivienda	Arriendo	38%	Familiar	44%	Hipotecada	16%	Propia	2%
	Personas a cargo	No tiene	56%	Una	25%	Dos	14%	3	5%
Institucionales									
	No. Veces cancelado el semestre	Ninguna	60%	Una	29%	Dos	11%		
	Ha cancelado una o más materias	No	49%	Si	51%				
	Ha pensado en retirarse de la universidad	No	70%	Si	31%				
	Satisfacción con la carrera	Si	83%	No	17%				
	Recursos económicos de sostenimiento en la universidad	Si	47%	No	57%				

Resultados del modelo de Regresión Logística

Para el caso que nos atañe, la variable predictora "Desertor" está representada por aquellos estudiantes que se encuentran con un promedio académico igual o inferior a 3.2; los cuales son considerados como "*potenciales desertores*" y aquellos que tienen un promedio superior a este valor, los cuales se consideran como los "*no potenciales desertores*".

Variables más influyentes en el modelo

Para la selección de las variables más influyentes en función de la variable que representa la deserción estudiantil del modelo, se requiere seleccionar aquellas que mayor influencia tienen en el modelo. Para tal efecto, se aplicó el criterio de AIC. En teoría este criterio se fundamenta en encontrar el conjunto de variables predictoras más influyentes en el modelo; el cual responde al conjunto que tenga un menor AIC (Criterio de información Akaike) y en consecuencia es considerado el mejor modelo Adicionalmente, la aplicación de este criterio de información conduce a evitar la multicolinealidad. (AKAIKE, 1974).

En teoría este criterio se fundamenta en encontrar el conjunto de variables predictoras más influyentes en el modelo; el cual responde al conjunto que tenga un menor AIC (Criterio de información Akaike) y por ende es considerado el mejor modelo. Se utiliza los métodos stepwise o backward, el cual va eliminando una a una las variables de tal manera que irá disminuyendo el AIC. Al utilizar el método backward; se encuentra un AIC = 353.6, el cual corresponde con la selección de las variables más influyentes. La tabla 2 presenta las 15 variables más influyentes en el modelo logístico, entre las que se destacan; el número de materias reprobadas, la metodología de enseñanza, el sexo biológico del estudiante, si cuenta con dependencia económica, motivo de selección la universidad para realizar los estudios, si ha pensado retirarse de la universidad, si el personal directivo y administrativo atienden las necesidades del estudiante, los profesores de la universidad trabajan con profesionalismo y calidad, la metodología de enseñanza, deficiencia en las bases académicas del bachillerato, problemas para culminar el trabajo de grado, obtuvo trabajo, falta de recursos para sostener la universidad, pérdida de auxilios económicos por parte de la universidad y los problemas con créditos directos.

Tabla 2: Modelo logístico: coeficientes con las variables más influyentes (criterio AIC)

Coeficientes:

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept	-1.531.560	0.910413	-1.682	0.092517 .
Mater_Reproba	0.739419	0.068729	10.758	< 2e-16 ***
as.factor(METOD_ENSEÑ)Aceptable	-1.346.028	0.717467	-1.876	0.060644 .
as.factor(METOD_ENSEÑ)Bueno	-0.008581	0.582410	-0.015	0.988244
as.factor(METOD_ENSEÑ)Deficiente	-2.201.142	0.976094	-2.255	0.024130 *
as.factor(Sexo)M	0.643003	0.338686	1.899	0.057628 .
as.factor(Depen_Econ)2	-0.518304	0.332521	-1.559	0.119065
as.factor(Moti_Sele_Univ)2	0.410605	0.472945	0.868	0.385293
as.factor(Moti_Sele_Univ)3	1.739.834	0.669117	2.600	0.009317 **
as.factor(Moti_Sele_Univ)4	0.206217	0.377443	0.546	0.584823
as.factor(Retir_U)1	0.983150	0.382115	2.573	0.010085 *
as.factor(Retir_U)2	-0.464773	0.557933	-0.833	0.404830
as.factor(Pers_Atienden_est)1	-1.305.365	0.579811	-2.251	0.024362 *
as.factor(Prof_Profesional)2	-0.069355	0.390012	-0.178	0.858857

as.factor(Prof_Profesional)3	2.584.254	0.609238	4.242	2.22e-05 ***
as.factor(Prof_Profesional)4	-1.914.125	0.578849	-3.307	0.000944 ***
as.factor(Met_Est)2	-1.081.313	0.421953	-2.563	0.010388 *
as.factor(Met_Est)3	-0.547258	0.966649	-0.566	0.571299
as.factor(Met_Est)4	-2.114.979	0.679046	-3.115	0.001842 **
as.factor(Bases_Bachi)1	-0.621371	0.331559	-1.874	0.060918 .
as.factor(Proble_Trab_Grad)1	-0.723718	0.394753	-1.833	0.066751 .
as.factor(Obtuv_Tra)1	0.838185	0.375570	2.232	0.025630 *
as.factor(Falt_RecSost_U)1	-0.584050	0.349222	-1.672	0.094439 .
as.factor(Perd_AuxEco_U)1	0.896140	0.407380	2.200	0.027824 *
as.factor(Prob_Cred_Dir)1	-0.887174	0.399993	-2.218	0.026557 *
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**				

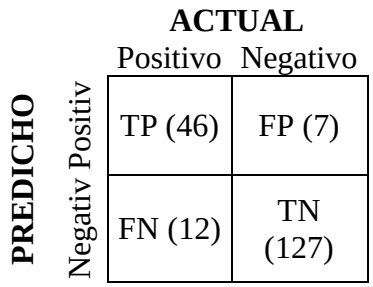
La selección de las variables responde a aquellas que cuenten con un nivel de significancia menor a 0.05 (p value < 0.05). Con estas variables más influyentes en el modelo de regresión logística se calcularon sus métricas.

Predicción del modelo

Se realiza los cálculos de los parámetros del modelo utilizando la base train, la predicción del modelo muestra el cálculo de las siguientes métricas respecto a la matriz de confusión, la cual se muestra en el gráfico 3; con una tasa de clasificación (accuracy) de 90.10%; es decir la tasa de los verdaderos positivo y verdaderos negativos respecto al total de datos. La tasa de verdaderos positivos (sensitivity) de 79.31%; en otras palabras, representa la tasa de los verdaderos positivos respecto a los actuales positivos, y una precisión de 86.79%; el cual representa la tasa de los verdaderos positivos respecto a los positivos predichos.

Accuracy	Error_rate	Sensitivity	Specifity	Precisión	npv
90.10%	9.89%	79.31%	94.77%	86.79%	91.36%

Gráfico 3 Matriz de confusión



Fuente: Elaboración propia

Resultados del modelo de Clasificación por Árbol de decisión

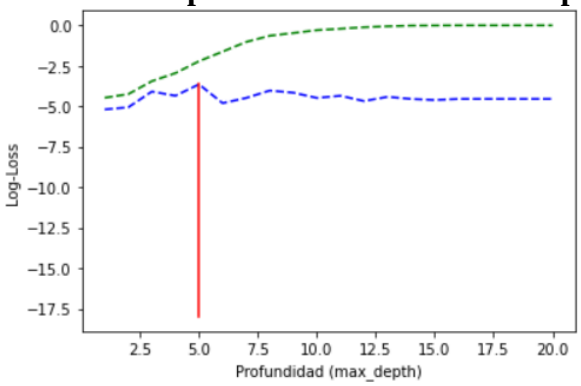
Teniendo presente los parámetros con que cuenta este método, entre los que se tiene la entropía; la cual es una medida de la impureza del conjunto de datos, en donde se define como la medida del desorden o incertidumbre de un grupo de datos. La entropía es un método utilizado para dividir un árbol de decisión en subconjuntos más pequeños, al dividir el árbol, actúa como un valor umbral para un nodo de árbol el cual mide que tan aleatorio son los nodos. Entre las métricas del

modelo se tiene el accuracy con un valor del 85.96%, el área bajo la curva ROC es del 83.25%, la precisión para los potenciales desertores es del 76.05%.

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
85.96%	14.04%	79.31%	78.80%	76.05%	91.36%

Con la inclusión de los hiperparámetros en la cual se tiene el máximo número de características (max_features = 15) el cual dice el máximo número de características para cuando se está observando la mejor separación del árbol. La máxima profundidad del árbol (max_depth = 5), el cual expresa la profundidad que se puede tener en los niveles del árbol, como se evidencia en el siguiente gráfico.

Grafico 4. Representación de la máxima profundidad del árbol



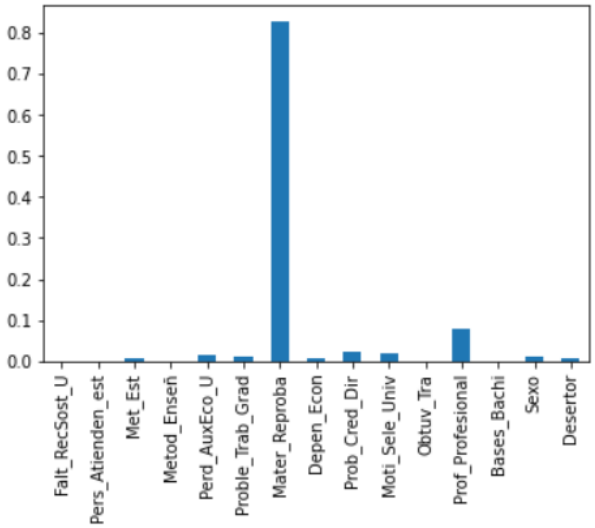
Fuente: Elaboración propia

Con la inclusión de estos hiperparámetros en el modelo, hizo que las métricas como el accuracy mejorara al llegar al 89.91% y el área bajo la curva es de 87.95%. A continuación, se presentan las respectivas métricas.

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
89.91%	10.09%	84.06	92.45%	82.86%	93.04%

Adicionalmente, el modelo presenta cuales de las variables involucradas en él son más influyentes en la predicción, destacándose en primer lugar el número de materias reprobadas, le sigue los profesores de la carrera trabajan con profesionalismo y calidad y las deficiencias en las bases académicas del bachillerato, como se aprecia en la figura 5.

Figura 5. Variables más influyentes en el modelo de Árbol de decisión



Fuente: Elaboración propia

Resultados del modelo Baggin: Boostrap Aggregation

Debido a que el conjunto de datos es igual y como provienen de una misma distribución; se introduce una correlación positiva dentro de la estimación (este es el costo que se tiene que pagar para mejorar la predicción). De los parámetros tomados en consideración se encuentra la utilización del criterio de la entropía, se toma el número de estimadores igual a 50 y el estimador base es el "decisión tree classifier". Se encontró con los cálculos de las métricas de predicción; el accuracy igual 89.91% y un área bajo la curva ROC igual a 87.62%. A continuación, se presentan las métricas respectivas.

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
89.91%	10.09%	86.96%	91.19%	81.08%	94.16%

Ahora con la inclusión de los hiperparámetros para mejorar la predicción del modelo; se tiene utilizando vecinos más cercanos (Bagging with KNN), el cual permite entrenar el clasificador Bagging con KNN como estimador base e inicializando el número de vecinos = 5 y llegando a 9 como el mejor número de vecinos, el número de muestras (max_sample) es 1.0, el máximo número de características (max_features) es 1.0. El modelo presenta sus métricas mejoradas; el accuracy = 90.35% y un AUC-ROC = 88.19. A continuación, se presentan sus métricas respectivas.

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
90.35%	9.65%	86.96%	91.82%	82.19%	94.19%

Resultados de la aplicación del modelo de Random Forest Classifier

Teniendo presente las características del modelo de clasificación de Random Forest y con la aplicación del criterio de la entropía, utilizando por defecto el criterio del mínimo número de muestras requeridas para la separación en cada nodo es igual a 1 y el máximo número de

características que se consideran cuando se desea la mejor separación (max_features), el cual se toma como "auto", se obtuvieron las métricas como el accuracy con el 90.35%, el área bajo la curva AUC-ROC es del 88.81%,

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
90.35%	9.65%	82.61%	93.71%	85.07%	92.55%

En la búsqueda de mejorar la predicción del modelo se incluyeron los hiperparámetros como la máximo profundidad (max_depth) de 19, el número de estimadores (n-estimators) de 100, el mínimo número de muestras separadas (min_samples_split) igual a 2, el mínimo número de muestras de hojas (min-samples_leaf) es de 1. De esta manera se obtuvieron las métricas como el accuracy del 91.23%, con una precisión del 85.51% y una especificidad del 93.71%. A continuación, se presentan las métricas respectivas.

Accuracy	Error_ rate	Sensitivity	Specifity	Precisión	npv
91.23%	8.77%	85.51%	93.71%	85.51%	93.71%

Selección del mejor modelo

De acuerdo con los modelos que presentaron las mejores métricas obtenidas tanto en el modelo de clasificación de regresión logística y en los modelos de Machine Learning aplicados en el estudio se evidenció que el modelo de Random Forest Classifier fue el que presentó las mejores métricas, tuvo el más alto Accuracy con el 91.23%; lo que significa que la tasa de clasificación tanto de potenciales desertores como no potencial desertores es del orden del 91.23%. Por otro lado, de acuerdo a la métrica de la Precisión la cual afirma que del 100% de los que clasificó como potenciales desertores, cuál fue el porcentaje que realmente desertó; se observó que fue más alta la precisión del modelo de la Regresión Logística con el 86.79%, sin embargo, la métrica del Random Forest estuvo muy cerca con el 85.51%. Ahora bien, si se busca identificar la probabilidad con que pudo haber desertado el estudiante (cantidad de verdaderos potenciales desertores sobre la totalidad de potenciales desertores observados), en otras palabras, del total de estudiantes que desertaron, cuál porcentaje está detectando que realmente van a desertar, se observa la métrica de la Sensibilidad y es el modelo de Random Forest que tiene esta métrica más alta con el 85.51%. Adicionalmente, si se observa la métrica de la Especificidad (Specifity); la cual calcula la tasa de los verdaderos negativos (la tasa de los verdaderos No potencial desertores); el modelo de Regresión Logística obtuvo mejor métrica con 94.77% que el de Random Forest. Sin embargo, esto no es el pilar del estudio, debido a que el mayor interés de las universidades se centra en identificar los potenciales desertores y no los otros, los no potenciales desertores. Finalmente, el modelo Random Forest fue el que obtuvo la mayor área bajo la curva AUC-ROC (Receiver Operator Charcteristic) el cual el punto que maximiza los verdaderos positivos y los verdaderos negativos se encuentra con un 90.17%; el cual identifica el umbral de mejor conveniencia.

En consecuencia, teniendo presente los resultados expuestos, se considera que el mejor modelo es el de Random Forest. A continuación, se presenta las métricas de los diferentes modelos.

Tabla 3. Métricas de los mejores modelos aplicados

Modelo	TP	TN	FN	FP	Accuracy	Error rate	Sensitive	Specifity	Precisión	npv	Área AUC-ROC
Regresión Logística con AIC = 353	46	127	12	7	90,10%	9,89%	79,31	94,77	86,79%	91,36%	80,00%
Árbol de decisión	58	147	11	12	89,91%	10,09%	84,06%	92,45%	82,86%	93,04%	87,95%
Baggin: Boostrap Aggregation	60	146	9	13	90,35%	9,65%	86,96%	91,82%	82,19%	94,19%	88,19%
Random Forest Classifier	59	149	10	10	91,23%	8,77%	85,51%	93,71%	85,51%	93,71%	90,17%

Conclusiones

El desarrollo de este trabajo ha podido determinar las alertas tempranas para la prevención de la deserción estudiantil a partir del modelo de clasificación Logístico y de predecir la deserción por medio de la aplicación de técnicas de Machine Learning como fue con los modelos de árboles de decisión, Baggin: Boostrap Aggregation y Random Forest.

La aplicación del modelo de clasificación logístico permitió identificar las principales causas por las cuales la población estudiantil es potencial desertor de la universidad. Entre los aspectos académicos se destacaron; el número de materias reprobadas del estudiante durante su carrera, el método de enseñanza que se imparte en el programa, la metodología de estudio impartida en el programa, la calidad y las deficiencias en las bases académicas del bachillerato con que cuenta los estudiantes al momento de ingresar a la universidad y los problemas que se le presentaron con el trabajo de grado. Entre los aspectos socioeconómicos se encontraron; la falta de recursos económicos para el sostenimiento de la universidad, los problemas de obtención de créditos directos con la institución, la pérdida de auxilios económicos que ofrece la universidad, el obtener trabajo mientras se encuentra estudiando y la dependencia económica en que se encuentra el estudiante. Entre los aspectos individuales se destacó; el sexo biológico del estudiante, y entre los aspectos institucionales sobresalió el motivo de selección de la universidad, el personal que atiende al estudiantado, el estudiante ha pensado en retirarse de la universidad y los profesores de la universidad trabajan con profesionalismo y calidad.

El uso de la técnica de clasificación de bosques aleatorios (Random Forest classifier) permitió predecir con buenos niveles de medición las causas de deserción de los estudiantes. Con una exactitud del 91.23% predijo la tasa de los estudiantes potenciales desertores y no desertores respecto a la totalidad de los estudiantes. Adicionalmente, con una precisión del 85,51% predijo que del 100% de los que clasificó como potenciales desertores, el 85.51% fue el porcentaje que realmente desertó. Una especificidad del 93.71% predijo la tasa de los verdaderos negativos (No potenciales desertores) respecto al total de la suma de los falsos positivos y los verdaderos negativos. Finalmente, el área bajo la curva AUC muestra el 90.17% que maximiza los verdaderos positivos y los verdaderos negativos (los verdaderos desertores y los verdaderos no desertores).

Estos resultados permiten a la universidad crear políticas para un plan de retención y permanencia estudiantil, creando políticas claras que afecten de manera específica y redireccionando los recursos hacia objetivos claros con el propósito de mitigar la deserción. Los efectos comenzarán a verse reflejados después de un año que se esté aplicando dichas políticas de manera constante y realizándole el seguimiento y la trazabilidad respectiva a los estudiantes

involucrados en este proceso. Esto conduce al mejoramiento de la calidad de vida de los estudiantes y por ende se verá reflejado en la calidad de la educación con que cuenta la universidad.

Referencias bibliográficas

Akaike, H. (1974). *A new look at the satistical model identification.* EEUU: IEEE transactions on automatic control.

Barrientos M, R. C. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Univrsidad Veracruzana, 9*(2), 19-24. Recuperado el 01 de Junio de 2022, de https://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf

Bernal, C. A. (2010). *Metodología de la Investigación* (Tercera edición ed.). Naucalpan, MEXICO: Pearson Educación de México S.A.

Bisong, E. (2019). *Building Machine Learnig and Deep Learning Models on Google Cloud Platform. A Comprehensive Guide for Beginners.* Ottawa-Canada: Apress Media LLC.

Castaño, E. G. (2004). Deserciín estudiantil universitaria: una aplicación de modelos de duración. *lecturas Económicas*, 39-65.

Díaz J, C. (2009). Factores de Deserción Estudiantil en Ingeniería: Una Aplicación de Modelos de Duración. *Información Tecnológica*, 129-145.

Faraway, J. J. (2005). *Linear Models with R.* Florida, EEUU , United States of Amerca: Chapman & Hall/CRC.

Giovagnoly, P. (2002). Determinantes de la deserción y graduación univrsitaria: una aplicación utilizando modelos de duración.

González S, J. C. (2021). Identidicación y predicción de estudiantes en riesgo de deserción académica por medio de modelos basados en Machine Learning. *Fundación Universitaria Los Libertadores*, 3-16.

Hardin, J. W. (2012). *Genralized Linear Models and Extensions.* Texas, USA: Stata Press, 4905 Lakeway Drive, College Satation.

Haroon, D. (2017). *Python Machine Learning Case studies. Five Case studies for the Data Scientist.* New York EEUU: Springer Science + Business Media.

Hernández, S. R. (2014). *Metodología de la Investigación.* Mexico D.F,: McGRAW-HILL.

Lohr, S. L. (2000). *MUESTREO: Diseño y Análisis.* Madrid España: Thomson Learning.

Madsen, H. T. (2011). *introduction to General and Generalized linear Models.* United Satates of America: Taylor and Francis Group, LLC.

Nitesh V, C. (2002). *Data minig for inbalanced datasets: an overview.* Estados Unidos: Univesity of Notre Dame.

Pérez G, B. R. (2020). Comparción de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *UIS Ingenierías*, 193-204.

Prieto A, C. (2015). Uso de Regresión Logística para predecir deserción estudiantil temprana. *Universidad de los Andes*, 5-28.

Ramasubramanian, K. S. (2019). *Machine Learning Using R.* New Dehi, India: Karthik Ramasubramanian and Abhishek Singh.

Rico, H. D. (2019). Caracterización de la deserción estudiantil en la Universidad Nacional de Colombia sede Medellín.

Ulloa G, N. J. (2020). *Técnicas de minería de datos para detectar patrones de bajo rendimiento académico.* Bogotá, Colombia: Instituto Latinoamericano de Altos Estudios -ILAE-.

Velez, W. C. (2008). Deserción estudiantil en la Educación Superior Colombiana. Elementos para su diagnóstico y tratamiento. Ministerio de Educacioón Nacional. *Educación*, 7-34.

Vidhya, A. (28 de 03 de 2016). *Practical Guide to deal with Imbalanced Classification. Problems in R.* Recuperado el 20 de abril de 2022, de Imbalanced classification problems: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). *Deserción estudiantil universitaria: una aplicación de modelos de duración. Lecturas de economía, (60), 39-65.*