# Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia.

Yudy Andrea Quintero

Tesis de maestría presentada para optar al título de Magíster en Ingeniería

Asesor
John Freddy Duitama Muñoz, Doctor en Informática

Universidad de Antioquia
Facultad de Ingeniería
Maestría en Ingeniería
Medellín, Antioquia,
Colombia 2022

Dedicatoria

A mi familia y amigos por su apoyo y acompañamiento.

Agradecimientos

A mi asesor el profesor PhD. John Freddy Duitama Muñoz por su apoyo y comprensión, al Observatorio institucional de la Universidad de Antioquia y a la vicerrectoría de docencia por su apoyo en la consecución de los datos para el desarrollo de esta investigación.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

# SIGLAS, ACRÓNIMOS Y ABREVIATURAS

| | |
|---|---|
| **EDM** | *Educational Data Mining* |
| **ICFES** | Instituto Colombiano para la Evaluación de la Educación |
| **IES** | Instituciones de Educación Superior |
| **LA** | *Learning Analytics* |
| **MEN** | Ministerio de Educación Nacional |
| **ML** | *Machine Learning* |
| **SPADIES** | Sistema para la Prevención de la Deserción de la Educación Superior |
| **RNA** | Redes Neuronales Artificiales |
| **UdeA** | Universidad de Antioquia |
| **XG Boots** | *Xtreme Gradient Boosting* |

# RESUMEN

El fenómeno de la deserción es una problemática que aqueja a la mayoría de las instituciones de educación superior en el país; por esta razón ha sido ampliamente estudiado para determinar sus posibles causas e implementar acciones en pro de su disminución, para lo cual se han usado diferentes técnicas y herramientas del análisis estadístico y computacional. La facultad de ingeniería de la UDEA no es ajena a dicho fenómeno, pues se observa que el nivel de deserción temprana promedio es de alrededor del 40%, por tal motivo se han realizado variadas investigaciones que han apuntado al estudio del fenómeno desde su caracterización e identificación de las causas que lo producen. Con este proyecto de investigación se buscó ir más allá de la caracterización del fenómeno y proponer una herramienta que identifique de manera temprana aquellos estudiantes que están en riesgo de desertar. Para lograr tal propósito se usaron los datos históricos contenidos en las bases de datos de la universidad y la información disponible en la plataforma del ICFES; y aplicando dos técnicas de *machine learning* como son redes neuronales artificiales (RNA) y *Xtreme gradient boosting* (XG Boost) se entrenaron diferentes modelos. Como resultado se construyeron modelos para los tres primeros semestres, pues el propósito fue identificar la deserción temprana. El modelo de primer semestre contiene solo los datos de ingreso y para los modelos de segundo y tercer semestre se incluyeron variables de desempeño académico del estudiante. Los resultados muestran que a medida que el estudiante avanza en su proceso de formación los modelos van logrando un mejor valor en la precisión de la predicción.

***Palabras clave*** — Deserción estudiantil, *Learning analytics*, minería de datos, modelos predictivos.

## ABSTRACT

The phenomenon of desertion is a problem that afflicts most of the higher education institutions in the country, for this reason, it has been widely studied to determine its possible causes and implement actions in favor of its reduction, for which different techniques and tools of statistical and computational analysis have been used. The faculty of engineering is no stranger to this phenomenon since it is observed that the average level of early dropout is around 40%, for this reason, several investigations have been carried out that have aimed to study the phenomenon from its characterization and identification of the causes that produce it. This research project sought to go beyond the characterization of the phenomenon and propose a tool for early identification of those students who are at risk of dropping out. To achieve this purpose, we made use of historical data contained in the university databases and the information available in the ICFES platform, and by applying two machine learning techniques such as artificial neural networks ANN and Xtreme gradient boosting - XGBoost and different models were trained. Models were built for the first three semesters since the purpose was to identify early dropout. The first-semester model contains only enrollment data and for the second and third-semester models, student academic performance variables were included. It was found that as the student advances in his or her education process, the models achieve a better predictive capacity.


*Keywords* — student dropout, learning analytics, data mining, predictive mode

# I. INTRODUCCIÓN

La deserción universitaria es una problemática que aqueja a la mayoría de las instituciones de educación superior a nivel mundial, y trae consigo efectos negativos tanto para el estudiante quien ve sus sueños de tener una carrera truncados, como para sus familias por los gastos económicos y consecuencias emocionales que esto les representa. Igualmente, para la universidad se afecta su desempeño, pues los recursos asignados para un estudiante que no logra culminar sus estudios se desaprovechan y por ende sus finanzas se ven afectadas, adicionalmente, la deserción es un indicador para la medición de la calidad de la educación y de los procesos de gestión administrativos en las instituciones de educación. De igual manera, tales consecuencias negativas se ven reflejadas en el desarrollo socio económico del país, pues este está directamente relacionado con el rendimiento académico de sus estudiantes [1].

Debido a las implicaciones que tal problemática trae consigo su ocurrencia genera gran preocupación en las instituciones de educación superior, por tal razón se ven grandes esfuerzos para estudiar la deserción y analizar las causas que la originan. Tales estudios se han realizado desde varios enfoques y desde diferentes disciplinas como los son: la psicología, la sociología, la pedagogía y la economía [2]; y han logrado construir una base teórica alrededor de dicha problemática y de las variables que influyen en la deserción.

En los recientes años se ha hecho más común el uso de las técnicas de computación aplicadas a bases de datos académicas con el fin de optimizar los procesos de aprendizaje, dando paso a una tendencia conocida como *learning analytics*. Una de las aplicaciones del *learning analytics* ha sido la creación de modelos que estén en capacidad de predecir la deserción. Esta investigación se centró en el uso de *learning analytics* para crear un modelo con la capacidad de predecir la deserción temprana en la Facultad de Ingeniería en la Universidad de Antioquia, la cual para el año 2019 tenía un índice de deserción temprana del 41% [3].

Existen muchos modelos a nivel mundial que se han creado para este propósito, pero por la naturaleza del problema los resultados no son aplicables a cualquier conjunto de estudiantes, pues el fenómeno de la deserción al ser multicausal depende en gran medida de las condiciones del estudiante y del entorno en que este se desenvuelve. Teniendo claro que para proponer un modelo de este tipo se deben incluir la mayor cantidad de variables disponibles que den cuenta de las 4

dimensiones [1], qué son las comúnmente aceptadas para dar explicación al fenómeno de la deserción, como son: las características individuales, socioeconómicas, académicas e institucionales.

Esta investigación tuvo como objetivo principal proponer un modelo basado en *learning analytics* para predecir deserción temprana a partir de la información que se tenía disponible de los estudiantes de la Facultad de Ingeniería de la Universidad de Antioquia; su propósito es poder servir como herramienta para identificar a aquellos estudiantes que se encuentran en riesgo de desertar desde el momento en que se matriculan en la universidad. Para dar cumplimento a este propósito se optó por trabajar una estrategia de modelos incrementales, en la que se crearon modelos predictivos en tres momentos diferentes. El primer momento es cuando el estudiante se matricula para un programa de la facultad, en este caso se contó con la información del formulario de inscripción y para tener una mayor cantidad de variables de cada categoría, tal como lo sugiere la teoría, se extrajeron algunas características disponibles en las bases de datos de la plataforma del ICFES. El segundo momento se presenta cuando el estudiante ha culminado su primer semestre académico y renueva su matrícula para el segundo semestre, en este caso, aparte de la información de ingreso, se tienen variables de desempeño académico de los estudiantes durante el primer semestre. Finalmente, el tercer momento se da cuando el estudiante ha permanecido durante un año en la universidad, en este caso se tiene la información de ingreso y las variables que dan cuenta del desempeño académico durante el primer y segundo semestre.

A partir de una revisión de la literatura y de trabajos previos desarrollados, se identificaron los modelos de *machine learning* que se consideran más adecuados para la tarea específica, en este caso se optó por trabajar con las redes neuronales artificiales y un método de ensamble como el es XGBoots. Se entrenaron los modelos en los tres diferentes momentos con cada conjunto de datos disponibles usando ambas técnicas seleccionadas. A continuación, se configuraron los diferentes parámetros en cada caso y se evaluaron las métricas a fin de seleccionar la técnica y el conjunto de parámetros que mostraron un mejor desempeño a la hora de predecir la deserción temprana. Finalmente se construyó un prototipo de software que usa el modelo entrenado y que recibe los datos de un nuevo estudiante y lo identifica como un posible desertor o no.

Los resultados encontrados mostraron que para el conjunto de datos con el que se contó la técnica XGBoots tuvo mejor desempeño frente a las redes neuronales, adicionalmente se logró un buen resultado para predecir la deserción en el primer momento, esto es, solamente con los datos

de ingreso, pues se obtuvo una precisión en la predicción superior al 70% la cual fue superior en comparación con otros trabajos encontrados con el mismo propósito. También se pudo ver que a medida que se incluyen las variables que dan cuenta del desempeño académico del estudiante el valor de la precisión consigue mejorar.

Este trabajo está estructurado de la siguiente manera: inicialmente se presenta el problema y los objetivos general y específicos. Seguidamente se presenta una contextualización teórica del problema de la deserción y las variables asociadas. A continuación, se muestran varios trabajos relacionados al uso del *learning analytics* en aplicaciones para predecir la deserción, a fin de identificar las técnicas y las variables usadas para dicho propósito. Posteriormente se realiza una descripción del conjunto de datos usados en la investigación. Se describe la metodología usada y finalmente se exponen los resultados y las métricas de evaluación utilizadas para validar los modelos, seguidamente se discuten los resultados y se presentan las conclusiones y las recomendaciones.

# II. PLANTEAMIENTO DEL PROBLEMA

En los últimos años se han logrado avances en materia de educación superior en Colombia. Las cifras del Ministerio de Educación Nacional (MEN) muestran que para el año 2010 sólo 3 de cada 10 jóvenes tenía la posibilidad de acceder a la educación superior, mientras que para el año 2016 esta cifra se aumentó a 5 de cada 10 jóvenes; además, el 60% de los estudiantes en educación superior proviene de hogares de bajos ingresos. Las cifras anteriores evidencian un avance en el acceso y la cobertura; sin embargo, el reto no es solo conseguir que más jóvenes puedan ingresar a la educación superior, sino que estos logren concluir con éxito su plan de estudio [4]. Actualmente, la deserción estudiantil en la educación superior es un problema que afecta a la mayoría de las universidades en toda América Latina, motivo por el cual a la fecha existe una gran cantidad de investigaciones al respecto de este fenómeno en donde se dan cuenta del gran número de estudiantes que no logran culminar satisfactoriamente con sus estudios universitarios y los costos sociales que esto les representa [5].

La universidad de Antioquia no es ajena a este fenómeno y en particular la Facultad de Ingeniería en donde, según el estudio realizado por el grupo de investigación Ingeniería y Sociedad, en el marco del Observatorio de la vida académica, en estudiantes seleccionados de las cohortes entre el 2005 y 2013 se encontró que la deserción acumulada en 10 semestres fue del 52% para las cohortes estudiadas y el 18% de los que desertaron lo hicieron en el primer semestre. En particular, al observar una cohorte concreta, proyectada a graduarse en 10 semestres, se observó que solo el 24,6 % lo logra en este tiempo, mientras que el 37,6% tardó más de lo previsto, el 30,5 % no logró culminar sus estudios, los restantes se graduaron antes este tiempo, debido a que eran estudiantes que ya tenían historia académica previa en otros programas u otras instituciones académicas [6].

Si bien el estudio mencionado anteriormente junto con otros realizados en la universidad y en la facultad [6],[8], [9] y [2] han servido para dar claridad y explicar el problema de la deserción estudiantil en la universidad de Antioquia y en la facultad de ingeniería, estos han abordado el

fenómeno a partir de su ocurrencia para la determinación de sus posibles causas, es decir, se analiza el problema una vez que este ha ocurrido. En la actualidad no se cuenta con una herramienta que a partir de las características y los comportamientos exhibidos por los estudiantes esté en capacidad de determinar el nivel de riesgo que tienen estos de desertar.

Learning Analytics (LA) es un área de investigación y desarrollo que a partir de la recolección, monitoreo, análisis y modelamiento de la información obtenida de los estudiantes y su entorno de aprendizaje busca optimizar los procesos de enseñanza y aprendizaje [10]. Esta área de conocimiento, que surgió en la última década, ofrece diferentes tipos de soporte computacional para monitorear el comportamiento de los estudiantes a fin de hallar patrones y/o tendencias ocultas, además, de identificar correlaciones en los datos de carácter educativo para proveer información de valor a los profesores y administradores acerca de los estudiantes y de los procesos de enseñanza y aprendizaje [11].

La universidad de Antioquia cuenta con una serie de bases de datos en las que se recoge la información de las diferentes dependencias institucionales [12], dicha información puede ser aprovechada usando las técnicas propias del *Learning Analytics* (LA) [11] para extraer la información que subyace en dichos datos. En particular, el LA ofrece un gran potencial para la creación de un modelo que permita diagnosticar el riesgo de deserción estudiantil a partir de los datos obtenidos de tales sistemas de información y que de esta manera la facultad tenga una herramienta efectiva para tomar decisiones correctas en relación con la deserción y la implementación de acciones que faciliten la permanencia de los estudiantes.

La deserción temprana se refiere a los estudiantes de una misma cohorte que solo cursan tres o menos semestres, y ha sido un propósito de gran importancia dentro de la facultad de ingeniería reducir la tasa de deserción temprana, ya que dicha tasa es una de las mayores en comparación con el resto de las dependencias académicas de la universidad [13]. Para el año 2016 la tasa de deserción temprana fue de 31.4% para la universidad, mientras que, para el mismo año el valor de dicha tasa dentro de la facultad fue de 46.1% ; según el plan de acción de la facultad se tiene como meta para el periodo académico del 2022 conseguir una reducción de dicho valor al 36%, tomando como línea base el año 2019, en el cual la deserción temprana fue del 41% [3].

Dada su naturaleza, el fenómeno de la deserción es extremadamente complejo de estudiar, según lo afirma Tinto citado por [7]. Existen una serie de factores que la originan que son de

diferente índole: sociodemográficos, académicos, personales, familiares, entre otros. Estos factores pueden ser cambiantes para cada dependencia académica, institución educativa, región e inclusive para cada momento económico y político del país [14]. Lo anterior indica que los modelos de deserción no son universales y que los resultados obtenidos para un grupo de estudiantes en particular no pueden ser replicables para cualquier grupo de estudiantes, pues estos modelos se deben ajustar de acuerdo con las características de los estudiantes analizados y la información con la que se cuenta; en cada caso se debe determinar cuáles son las variables a incluir y los métodos o técnicas que mejor desempeño tengan a la hora de realizar una predicción; además, dadas las características de la facultad de ingeniería que es la más grande de toda la universidad con alrededor de 8000 estudiantes, es casi equiparable a una universidad local, esta facultad es la tercera con mayor demanda en sus programas, ocupa el tercer puesto en índice de deserción temprana y adicionalmente presenta el índice de graduación más bajo de toda la universidad; razones por las cuales en la facultad de ingeniería la deserción temprana es un problema que requiere de gran atención. Abordar esta problemática nos llevó a la siguiente pregunta de investigación:

¿Qué características debe tener un modelo basado en *Learning analytics* para que esté en capacidad de estimar el riesgo de deserción temprana y presentar alertas tanto para el estudiante, como para el docente en los programas de pregrado presenciales de la facultad de ingeniería de la universidad de Antioquia?

# III. OBJETIVOS

## *A. Objetivo general*

Proponer un modelo que permita predecir la deserción temprana en los programas presenciales de pregrado en la facultad de ingeniería de la U. de A. haciendo uso de los métodos y las técnicas del *learning analitycs*.

## *B. Objetivos específicos*

- Identificar las variables que influyen en la deserción de los estudiantes de la facultad de ingeniería en la universidad de Antioquia a partir de los estudios previamente realizados.
- Realizar el análisis exploratorio de la información disponible en las bases de datos de la UdeA con el fin de construir la cohorte de estudio.
- Explorar los diferentes métodos o técnicas de predicción a fin de encontrar los modelos que ofrezcan mejor desempeño.
- Validar el modelo que ofrezca la posibilidad de predecir la deserción en los estudiantes de la facultad de ingeniería.
- Construir un prototipo del software que permita hacer uso del modelo desarrollado

# IV. MARCO TEÓRICO

## *A. Deserción*

Existen varias definiciones para el concepto de deserción, según Tinto [16] define la deserción como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo. Muchas discusiones se han dado alrededor de la deserción estudiantil y a pesar de que no se tiene una definición consensuada sobre este fenómeno [17] existe una coincidencia en cuanto a que este debe ser explicado por diferentes grupos de variables las cuales en términos generales se han agrupado en 4 categorías principales: socioeconómicas, individuales, institucionales y académicas [7]. Dentro de cada una de estas categorías existen una serie de elementos que interaccionan en el estudiante y que, para cada región, momento económico y político del país son cambiantes [14]; por este motivo se hace complejo el estudio de los factores que influyen en la deserción [18] ya que se debe abarcar no sólo una variedad de perspectivas, sino que también hay que tener en cuenta que existen diferentes tipos de abandono, los cuales se ubican dentro de dos categorías.

La primera categoría establece el abandono en el tiempo, el cual a su vez puede ser precoz, que se presenta cuando el estudiante no alcanza a iniciar su proceso de formación; temprana que se presenta cuando el estudiante deserta en los primeros años de estudio y tardía que indica que el abandono se presenta cuando el estudiante llevaba un avance considerable en su proceso formativo.

La segunda categoría es el tipo de deserción que se considera en el espacio; en esta clasificación se presentan la deserción interna o del programa, la deserción institucional que es cuando el estudiante abandona la institución educativa y finalmente la deserción del sistema en la cual se considera como desertor a aquel que abandona definitivamente el proceso de formación [7],[19] Hay que tener en cuenta que el estudio de la deserción debe realizarse desde diferentes acepciones de acuerdo con el investigador y la situación del ambiente en el que está el estudiante, es por esto por lo que las variables para incluir en el estudio dependen del punto de vista desde el cual se haga el análisis; esto es, individual, institucional y estatal o nacional [7]

### *B. Variables asociadas a la deserción*

El tema de la deserción ha sido ampliamente estudiado y como producto de tales estudios se ha logrado identificar que existen múltiples razones que llevan a un estudiante a tomar la decisión de abandonar su proceso de formación [17]. Si bien los estudios hechos hasta el momento coinciden en que los factores asociados a la deserción no son únicos y que estos dependen del entorno en que se encuentra el estudiante [14], se ha logrado reunir y poner en consenso un conjunto de variables que se hacen comunes a los diferentes grupos de estudiantes. Estos grupos de variables son los tenidos en cuenta por el Ministerio de Educación Nacional para calcular y medir los niveles de deserción [7], dichas variables son mostradas en la TABLA I es de aclarar que estas variables son las más comúnmente utilizadas en diferentes estudios, pero dichas variables no son únicas, pues estas son diversas debido a la pluralidad de los estudiantes y a las características diferenciales de las instituciones de educación superior

TABLA I
RESUMEN DEL GRUPO DE VARIABLES ASOCIADAS A LA DESERCIÓN ESTUDIANTIL

| **Académicas** | **Institucionales** | **Socioeconómicas** | **Individuales** |
|---|---|---|---|
| Orientación socio ocupacional. | Norma académica. | Estrato. | Edad. |
| Tipo de colegio. | Recurso universitario. | Situación laboral e ingreso de los padres. | Género. |
| Rendimiento académico. | Orden público. | Situación laboral del estudiante. | Estado civil |
| Calidad del programa. | Entorno político. | Dependencia económica. | Posición dentro de los hermanos. |
| Pruebas saber. | Nivel de interacción entre los estudiantes y los docentes. | Personas a cargo. | Entorno familiar |
| Resultados del examen de ingreso. | Apoyo académico | Nivel económico de los padres. | Calamidad, problemas de salud. |
| Cualificación docente. | Apoyo psicosocial | Entorno macroeconómico del país. | Integración social. |
| Grado de satisfacción con el programa. | | | Incompatibilidad horaria. |
| | | | Expectativas insatisfechas. |
| | | | Embarazo. |

A continuación, se detallan cada uno de los factores que aparecen en la TABLA I

*Factor académico*

Se refiere a la capacidad intelectual, al compromiso académico y las aspiraciones profesionales que tenga el estudiante, capacidad de adaptación y solución de problemas que posee frente a los retos académicos. Igualmente, está relacionado con objetivos, intereses y proyecto de vida del estudiante. Aspectos que se ven reflejados en el rendimiento alcanzado en el proceso educativo. [21].

*Factor institucional*

Reúnen las características de la institución universitaria, los servicios estudiantiles que ofrece la institución, los indicadores de docencia, enseñanza y calidad, la infraestructura y las experiencias de los estudiantes en el aula de clase [5].

*Factor socioeconómico*

Involucra aspectos como la influencia, relación y apoyo del grupo familiar. Se incluyen variables como número de miembros en la familia, niveles educativos de los padres y hermanos, capacidad económica del núcleo familiar y la disponibilidad de recursos que garantizan la permanencia del estudiante, satisfacción de necesidades básicas de alimento, transporte, etc. Por último, se refiere también al ambiente familiar, conflictos y relaciones entre los miembros de la familia del estudiante [20].

*Factor individual*

Estas características están asociadas con la personalidad del estudiante, así como también sus habilidades y hábitos de comportamiento y de estudio [5] y persistencia en el alcance de metas; además de la historia personal y las percepciones del estudiante en lo relacionado con su vida

universitaria y proyecto de vida. Igualmente, se incluyen en esta dimensión las actividades cotidianas del estudiante asociadas con su vida laboral, deportiva y artística.

### *C. Deserción en el contexto de la ingeniería*

En los últimos años la demanda de profesionales en la rama de ciencia, tecnología, ingeniería y matemática -conocida como STEM por sus siglas en inglés- ha aumentado debido al creciente desarrollo tecnológico y sus aplicaciones. Sin embargo, se observa que la deserción en estas carreras es alta [22], a nivel nacional históricamente se han posicionado como programas críticos en cuanto a deserción las ingenierías y las ciencias exactas, con promedios de graduación de cerca de 50% [2]. En un estudio realizado por [23], se indica que en Colombia entre 45 y 52 % de los estudiantes que ingresa a un programa de ingeniería lo abandonan antes de concluir sus estudios, identificando como posibles causas del abandono: la dificultad para adaptarse a la educación superior, una inadecuada orientación profesional, inconvenientes económicos y los bajos niveles de comprensión y competencia de la educación previa. En el caso concreto de la facultad de ingeniería de la universidad de Antioquia, un estudio realizado por el observatorio de la vida académica en la facultad [6] arrojó hallazgos importantes a mencionar. En primer lugar, está el hecho de que al ser una universidad pública ingresan en su gran mayoría estudiantes de estratos medios y bajos con las dificultades propias de su condición social y poca preparación previa lo que se traduce en problemas como: bajo rendimiento académico, repetición de cursos, prolongación de la carrera, conflictos académicos y deserción. Dichos problemas presentan indicadores que afectan significativamente el proceso educativo y si bien se han identificado diversos factores que determinan estos fenómenos, es necesario realizar un seguimiento constante y detallado de estos.

Con relación a la deserción, el mismo estudio que fue realizado con estudiantes pertenecientes a las cohortes de 2005-2 a 2013-1, indica que la deserción promedio acumulada fue del 52% y el 18% de los estudiantes analizados desertaron en el primer semestre. Se resalta el hecho que los estudiantes que desertaron lo hicieron primero por razones académicas y luego por motivos económicos; hasta el cuarto semestre la principal causa de deserción (43%) obedeció a factores académicos y a partir de ese semestre los factores económicos comenzaron a predominar. Materias como matemáticas, físicas y cálculos representan barreras en la vida académica de los estudiantes

de la facultad, pues se encontró que en promedio el 54 % de los estudiantes no aprobaban estos cursos lo cual significa un fracaso académico en la fundamentación científica para la ingeniería, convirtiéndose en una causal de deserción y de prolongación de la carrera. También se mencionan como causa de deserción la carga académica y el alto número de créditos de algunos programas, además, del hecho de que no hay una política de unificación de créditos en la facultad; algunos programas están recargados en los primeros semestres de cursos básicos dificultando para el estudiante identificar una relación con la especialidad de la carrera elegida. Una vocación no clara y la poca preparación que obtienen en la educación secundaria los hace vulnerables a enfrentar tales asignaturas exigiéndoles una mayor preparación y dedicación horaria.

### *D. Minería de datos (Data mining)*

La minería de datos usa herramientas de análisis de datos para descubrir patrones y relaciones válidas previamente desconocidos en un gran conjunto de datos. La minería es un proceso de extracción de información oculta, previamente desconocida y que es potencialmente útil de grandes bases de datos [24]. El objetivo de los esfuerzos en minería de datos normalmente está orientado a crear un modelo descriptivo o un modelo predictivo. Los modelos descriptivos tienen como propósito presentar la información en forma concisa encontrando patrones en los datos y comprendiendo las relaciones que se presentan entre las variables asociadas a dichos datos [25]. Por otro lado, los modelos predictivos consisten en aplicar modelos estadísticos o técnicas de aprendizaje de máquinas a un conjunto de datos con el propósito de extraer nuevas o futuras observaciones [26].

### *E. Learning analytics and Educational Data Mining.*

Con el creciente uso de tecnologías para el Big Data y su expansión a las diferentes áreas del conocimiento surgen nuevos campos de investigación y desarrollo, en particular en el área de la educación y relacionados al tema, aparecen dos nuevos conceptos que comparten intereses comunes como son: *Educational Data Mining* y *Learning Analytics* [27].

El *Educational Data Mining* (EDM) está relacionado con el desarrollo, investigación y aplicación de métodos computacionales para detectar patrones en grandes conjuntos de datos de educación [27].

*Learning analitycs* (LA) se define como: "La medida, recolección, análisis y reporte de datos provenientes de los estudiantes y sus contextos, con el propósito de entenderlos y optimizar el aprendizaje en el ambiente en el cual ocurre" [28]. El campo del LA está influenciado por un amplio rango de disciplinas como son: la educación, la psicología, filosofía, sociología, lingüística, ciencias del aprendizaje, estadística, inteligencia computacional y *machine learning*; sin embargo, las dos disciplinas más dominantes en este tópico de investigación son las ciencias de la computación y la educación [29]. LA es un área que emerge en las últimas décadas y combina los datos de educación, con modelos predictivos y descriptivos, análisis estadístico con el propósito de generar conocimientos para los estudiantes, docentes y administrativos [11].

Dado que el LA está orientado a convertir grandes cantidades de datos de su estado original (estructurado, no estructurados, semi estructurados) en información útil, se nutre de diferentes instrumentos analíticos de otras disciplinas tales como EDM, el *machine learning* y la estadística clásica [30]. Desde su surgimiento LA ha sido empleado sobre diferentes conjuntos de datos académicos para la creación de diferentes aplicaciones como por ejemplo: sistema de percepción de felicidad de los estudiantes, sistemas de recomendación de contenidos académicos, clasificación de actividades y detección del riesgo de deserción [10].

# V. ESTADO DEL ARTE

Con el fin de definir una estrategia a seguir para la solución del problema se analizaron los trabajos que se han realizado al respecto, con el propósito de identificar metodologías utilizadas, variables incluidas, métodos de *machine learning* usados y los resultados obtenidos.

Al realizar la búsqueda se evidenció que las técnicas del *learning analitycs* han sido usadas en la predicción de la deserción y determinación del desempeño de los estudiantes; en la literatura se encuentran varios trabajos en los cuales se pueden ver los esfuerzos para construir un modelo capaz de determinar de manera oportuna la deserción estudiantil.

La búsqueda se realizó en las bases de datos: Scopus y Google Scholar. Utilizando los términos de búsqueda mostrados en la TABLA II**.**

TABLA II
RESULTADOS DE LA REVISIÓN BIBLIOGRÁFICA

| Ecuación de búsqueda | Número de resultados |
|---|---|
| “Learning analytics” AND “Higher education” | 417 |
| ” Learning analytics” AND ”Dropout" | 71 |
| “Learning analytics” AND “Higher education” AND “Review” | 54 |
| “Learning analytics” AND “Systematic review” | 31 |
| "Learning analytics" and "dropout" | 28 |
| “Learning analytics” AND” Dropout” AND “Higher education" | 18 |
| "Educational data mining" and "Higher education" AND "Dropout" | 14 |
| "Educational data mining" AND "Early dropout" | 5 |

Se incluyeron los trabajos del 2016 en adelante, con las palabras “*Learning analytics*” AND ”*Dropout*” AND "*Higher education*" se lograron 18 resultados, de los cuales se extrajeron los que hubieran sido citados, obteniendo 8 trabajos. Y se incluyeron 3 trabajos más, dedicados

exclusivamente a la predicción de la deserción temprana. Los trabajos seleccionados se pueden observar a continuación.

Los trabajos se separaron de acuerdo con el tipo de modalidad, inicialmente se presentan tres trabajos desarrollados en la modalidad virtual ver TABLA III, esto debido a que en los programas virtuales los niveles de deserción son considerablemente altos. El trabajo presentado en

[31] evalúa el uso de algoritmos genéticos para encontrar la combinación de híper parámetros que logre el mejor desempeño de los diferentes modelos tradicionalmente usados en la predicción de la deserción; contrastándola con las técnicas de malla de búsqueda (grid-search), el ejercicio consistió en entrenar los diferentes modelos: Árboles de decisión (DT), *random forest* (RF), Redes neuronales (MLP), regresión logística (LG) y AdaBoost (ADA); optimizando los diferentes híper parámetros de cada modelo, a fin de lograr el mejor desempeño. Inicialmente la optimización la realizaron con malla de búsqueda y posteriormente lo hicieron con un algoritmo genético encontrado que, al hallar la mejor combinación de híper parámetros con el algoritmo se logró un mejor funcionamiento de los modelos. Usaron como métrica de validación el AUC y encontraron que la mejor técnica fue *random forest* (RF) con un valor de AUC de 77,52. Para el desarrollo de los modelos los datos usados consintieron en la información disponible en la plataforma virtual que dan cuenta de la actividad del estudiante en el curso.

Por su lado, [32] proponen el uso regresión logística para la creación de un modelo capaz de perfilar a los estudiantes de cursos virtuales que desertan y a partir del tal perfil identificar aquellos nuevos estudiantes que estén en riesgo de abandonar sus estudios; este análisis incluyó un pequeño conjunto de variables sociodemográficas de los estudiantes como por ejemplo: origen étnico, edad, género, entre otras) e información de interacción con la plataforma virtual logrando una precisión de 81,8 %.

Igualmente, [33] propone un modelo para predecir deserción en un curso de modalidad virtual de diferentes programas de formación; usando técnicas estadísticas como análisis de factores y regresión logística, donde se incluyeron variables relacionadas a la interacción con la plataforma virtual como por ejemplo, tiempo de actividad en línea, participación en foros, mensajes a compañeros y/o tutores, entre otras, logrando una precisión en la predicción de 67,4%.

TABLA III
RESUMEN DE TRABAJOS RELACIONADOS EN MODALIDAD VIRTUAL

| **Proyecto** | **Propósito** | **Método utilizado** | **Población** | **Variables** | **Resultados** |
|---|---|---|---|---|---|
| A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course (Brasil) [31] | Usar algoritmos genéticos para optimizar los híper parámetros de diferentes modelos de machine learning | Árboles de decisión (DT), *Random forest* (RF), Redes neuronales (MLP), regresión logística (LG) y *AdaBoost* (ADA) | 752 estudiantes de un curso técnico a distancia. | Datos de la actividad del estudiante con el curso virtual. Participación en chats, visualización y participación en foros, visualización de recursos y visualización del curso | Validando el AUC<br>DT (0.677), RF (0.77), MLP (0.73), LG (0.69), ADA (0.75) |
| Analyze and Predict Student Dropout from Online Programs. (EEUU) [32] | Determinación de las características de estudiantes que desertan y construir un modelo que identifique posibles desertores con base en dichas características | Regresión logística | 1.211 estudiantes de educación virtual de una universidad pública. Matriculados en el periodo comprendido entre 2014 y 2016 | Origen étnico, género, edad, estado de tiempo (medio tiempo o tiempo completo), puntaje de clasificación, promedio acumulativo y promedio del curso. | Precisión 81,8% |
| Learning Analytics and Scholar Dropout: A Predictive Model. (Colombia) [11] | Modelo de predicción de deserción en cursos de educación virtual | Análisis de factores y regresión logística | 240 estudiantes de diferentes programas (Ingenierías, derecho, comunicación, administración y educación) Matriculados en el segundo periodo académico de 2016 | Edad, género, nivel socioeconómico, tiempo en línea, número de entradas, de participaciones en foros, réplicas en foros, mensajes a compañeros, mensajes a proceso, tareas completadas, recursos descargados. | Precisión 67,4% |

En la TABLA IV se resumen algunos trabajos desarrollados para predecir deserción en modalidad presencial. En el primer trabajo [34] la técnica utilizada es árboles de decisión; el cual tiene como propósito la implementación de un modelo para clasificar el riesgo de deserción usando variables exclusivamente cuantitativas asociadas a las notas obtenidas y los cursos aprobados por los estudiantes, la población usada, fueron 1844 estudiantes de ingeniería de una universidad pública, el modelo demostró tener mejor desempeño para identificar a los estudiantes con más probabilidades de graduarse que aquellos tienen más probabilidad de desertar 93,1% y 67,3% respectivamente. El segundo modelo [35], trabaja con información cualitativa asociada a los hábitos comportamentales de los estudiantes, las variables utilizadas fueron respuestas a encuestas aplicadas a los estudiantes, la técnica usada fue árboles de decisión y logró una precisión para predecir la deserción de 97,95 %. A pesar de que en este trabajo el porcentaje de precisión logrado es bastante alto, este incluye variables que se obtiene de formularios que los estudiantes deben responder acerca de posibles comportamientos adictivos como por ejemplo las drogas, el alcohol, los videojuego, entre otras; lo cual hace que la veracidad de la predicción depende que los estudiantes sean honestos a la hora de responder el cuestionario.

En [36] se presenta un modelo para predecir deserción haciendo uso de una base de datos de 32.000 estudiantes; una de las más grandes utilizadas para este propósito en relación con los otros trabajos analizados; las técnicas utilizadas fueron regresión logística, *k-Nearest Neighbors*, y árboles aleatorios, usando variables tanto de tipo cualitativo como cualitativo, se incluyeron características sociodemográficas y puntajes obtenidos en pruebas de estado y las pruebas de admisión a la institución. De las técnicas utilizadas en dicho estudio la que mejor desempeño mostró a la hora de predecir la deserción fue la regresión logística; sin embargo, el valor de precisión logrado fue de tan sólo 66.59%, por lo cual los mismos autores proponen utilizar otras técnicas con una mejor capacidad predictiva como son las redes neuronales y las máquinas de soporte vectorial.

El reporte [37] presenta un estudio de caso donde se analizan datos educativos haciendo uso de diferentes técnicas para predecir deserción en estudiantes de ingeniería en modalidad presencial. El estudio consistió en comparar diferentes técnicas: árboles de decisión, regresión logística y *Naive Bayes*; encontrando que el mejor desempeño se logró usando árboles de decisión con un 94% en el valor del AUC que fue la métrica seleccionada para la validación del modelo.

TABLA IV
RESUMEN DE TRABAJOS RELACIONADOS EN MODALIDAD PRESENCIAL

| Proyecto | Propósito | Método utilizado | Población | Variables | Resultados |
|---|---|---|---|---|---|
| Using academic analytics to predict dropout risk in engineering courses. (Portugal) [34] | Implementación de un modelo para clasificar el riesgo de deserción | Árboles de decisión (*Decision tree*) | 1.844 estudiantes de diferentes ingenierías de una universidad pública de los cuales 1099 se graduaron y 745 desertaron | Edad de ingreso, número de asignaturas aprobadas, número de intentos para aprobar, notas obtenidas en los cursos aprobados | 93,1 % de precisión para identificar graduados y 67,3% los desertores. |
| Decision Trees for the Early Identification of University Students at Risk of Desertion. (Ecuador) [35] | Predecir la deserción de los estudiantes universitarios basados en hábitos comportamentales | Árboles de decisión | 3.162 estudiantes de ingenierías en una universidad pública. Pertenecientes a las cohortes entre 2012 y 2017 | Respuestas a preguntas como: Adicción a redes sociales, a las drogas, al alcohol, al celular, a los juegos, a los videojuegos, a las compras, apego emocional, edad | Logró una precisión para predecir la deserción de 97,95 % |
| Applying Data Mining Techniques to Predict Student Dropout: A Case Study. (Colombia) [37] | Determinar las características claves de la deserción académica | Árboles de decisión, Regresión logística y Naive Bayes | 802 estudiantes del programa de Informática de una universidad privada. Estudiantes que iniciaron su proceso de formación en 2004 y que para el 2010 ya habían finalizados sus estudios | Información de admisión, incluida información demográfica mínima (género, fecha de nacimiento, estado civil). La fecha de graduación y el programa académico. calificaciones de cursos y el promedio académico acumulativo | Árboles de decisión 0.94%, Logistic Regression 0.92% y Naive Bayes 0.87%. |
| Predicting Student Dropout in Higher Education (EEUU) [36] | Crear un modelo de deserción estudiantil sólo con datos de inscripción | Regresión logística (RL), k-nearest neighbors (KNN), y árboles aleatorios (RF) | 32,500 estudiantes de diferentes programas de formación. Que se matricularon por primera vez entre 1998 y 2006 | Información demográfica (raza, género, fecha de nacimiento, estado de residente, hispano o no), escolaridad previa, puntajes obtenidos en las pruebas SAT ACT | RF: 62,24%<br>RL: 66,59%<br>KNN: 64,60% |

Los trabajos presentados en la TABLA V se enfocaron en predecir la deserción temprana para lo cual se crearon modelos en diferentes momentos. Un momento inicial, cuando solo se cuenta con los datos de inscripción del estudiante, y dos momentos más cuando se tiene resultados de desempeño académico del primer y segundo semestre. En [38] utilizaron redes neuronales (RNA) y árboles de decisión (DT) para conducir un experimento en el que se clasificaba los estudiantes en tres categorías posibles: promoción, repetición o deserción, encontrando que cuando se cuenta solo con datos de preinscripción los árboles de decisión tuvieron una precisión en la clasificación de 59,87% mientras que la red neuronal tuvo una precisión de 60,53%; para el mismo trabajo cuando se contó con la información del primer semestre académico se obtuvo una mejora en la precisión de la clasificación la cual fue 85,08% para DT y de 84,86% para la RNA, finalmente contando con los datos del primer año se lograron precisiones de 89,91% y 96,49 % para DT y RNA respectivamente.

En el trabajo [39] presentado en Alemania se utilizaron diferentes técnicas y se realizaron varios experimentos en los que inicialmente se cuenta solo con datos básicos de los estudiantes, en un segundo experimento, al conjunto de datos básicos se incluyen respuestas de los estudiantes a una evaluación inicial en la que se pretende conocer la personalidad del estudiante y sus motivaciones. Un tercer experimento consistió en trabajar con una base de datos que contiene los datos básicos y adicionalmente respuestas a encuestas aplicadas por los consejeros estudiantiles sobre la satisfacción del estudiante con el rendimiento académico esperado y sobre sus probabilidades de graduarse o desertar. Finalmente, un cuarto experimento consistió en trabajar con los datos básicos del estudiante y las variables que dan cuenta del desempeño académico; esto es; calificaciones y puntos créditos. Las técnicas utilizadas fueron regresión OLS [1] y un método de ensamble como lo es el LightGBM [2].

---

[1] Ordinary Least Squares o método de mínimos cuadrados ordinarios. Es un método utilizado para encontrar los parámetros de un regresor lineal.

[2] LightGBM es un método de ensamble de alto rendimiento y rapidez, basado en árboles de decisión desarrollado por Microsoft https://github.com/Microsoft/LightGBM

Con relación a los resultados del estudio se pudo ver que en los cuatro experimentos y para ambas técnicas usadas, cuando se contaba solo con datos básicos la precisión no fue superior del 67%; a medida que se incluyen más variables la precisión fue mejorando, lográndose el mejor desempeño en el experimento 2 con el LightGBM donde se logró un 80% de precisión con relación a la regresión OLS que tuvo una precisión de 79% en el mismo experimento.

Finalmente, en [40] se presenta un trabajo en el que se utilizaron tres técnicas para predecir deserción: Análisis Lineal Discriminante (LDA), Máquina de Soporte Vectorial (SVM) y Random Forest (RF). Se entrenaron con tres conjuntos diferentes de datos, el primero incluye datos básicos del estudiante al momento de la inscripción, el segundo incluye información básica y adicionalmente, si el estudiante tiene requerimientos adicionales de aprendizaje, esto es, al momento de la admisión se le realiza un examen y dependiendo del resultado se le sugiere al estudiante tomar unos cursos específicos el primer año. El tercer conjunto de datos incluye los datos básicos más la información de los créditos cursados. Los resultados obtenidos fueron: para el primer conjunto de datos LDA tuvo una precisión de 62%, SVM (62%) y RF(56%), en el segundo caso los resultados presentaron una mejora: LDA (75%) , SVM (81%) y RF (63%), finalmente, con el tercer conjunto de datos los resultados fueron más cercanos y mejores, con respecto a los dos casos anteriores, para las tres técnicas usadas: LDA(85%), SVM(87%) y RF(87%).

TABLA V
RESUMEN DE TRABAJOS RELACIONADOS ENFOCADOS EN DESERCIÓN TEMPRANA

| Proyecto | Propósito | Método utilizado | Población | Variables | Resultados |
|---|---|---|---|---|---|
| Predicting Computer Engineering students' dropout in Cuban Higher Education with pre-enrollment and early performance data, (Cuba) [38] | Predecir la deserción después del primer año. | Árboles de decisión y redes neuronales. | 456 matriculados de diferentes provincias de Cuba, matriculados en ingeniería de sistemas | Género, provincia, puntaje en prueba de admisión, notas de los cursos de matemáticas | Inscripción: DT: 59,87% RNA: 60,53 % 1er Sem: DT: 85,08% RNA: 84,86% 1er año: DT: 89,91% RNA: 96,49% |
| Early Identification of College Dropouts Using Machine-Learning. (Alemania) [39] | Evaluar la viabilidad de un sistema de alerta temprana de deserción. | *LightGBM*, regresión OLS | Estudiantes que iniciaron los estudios en el semestre de verano 2010-2011 de una universidad alemana. | Género, edad, condición de migración, puntaje colegio, colegio, educación de los padres, programa, | Inscripción: LightGBM 67% OLS 65% Evaluación 1: LightGBM 74% OLS 71% Evaluación 2: LightGBM 80% OLS 79% |

| | | | | | |
|---|---|---|---|---|---|
| Student Dropout Prediction. (Francia) [40] | Predecir la deserción temprana de un estudiante de primer año. | Linear Discriminant Analysis (LDA), *Support Vector Machine* (SVM) y *Random Forest* (RF) | 15000 estudiantes de diferentes programas del periodo académico 2016-2017 | Edad, género, colegio, puntaje final del colegio, necesidad de créditos adicionales el primer año, escuela (Facultad), créditos aprobados el primer año. | Inscripción: LDA: 62% SVM: 62% RF: 56% Créditos +: LDA: 75% SVM: 81% RF: 63% 1er sem: LDA: 85% SVM: 87% RF: 87% |

Como resultado de analizar todos los trabajos reseñados se tienen las siguientes observaciones:

En cuanto a las técnicas utilizadas se observa que los árboles de decisión son la técnica más utilizada, lo cual puede ser debido a su simplicidad.

En lo que respecta a los datos y las variables utilizadas se destaca lo siguiente:

- En los cursos de modalidad virtual la información usada está relacionada con la actividad del estudiante en el curso, lo cual se debe a la facilidad de obtener dicha información de la plataforma de estudio.
- Las bases de datos en su mayoría no contienen gran volumen de estudiantes analizados.
- Algunos de los trabajos presentados en esta revisión incluyeron datos que son respuestas de encuestas realizadas a los estudiantes, logrando una precisión alta. Sin embargo, como lo exponen los mismos autores estas respuestas pueden introducir un poco de subjetividad al modelo, pues depende de que los estudiantes respondan con sinceridad a preguntas que en algunos casos pueden resultar sensibles.
- No todos los trabajos logran incluir variables de todas las categorías que menciona la literatura como determinantes de la deserción lo cual se debe a que no se encuentran disponibles en las diferentes instituciones de educación dónde se realizaron los estudios.

Con relación al desempeño se evidenció lo siguiente:

- Si bien la técnica más comúnmente usada fue árboles de decisión la precisión lograda por estos no fue superior al 68%, solo en un caso fue superior (97%), sin embargo, como se ya mencionó en este caso se incluyeron variables que fueron respuestas de los estudiantes sobre sus hábitos comportamentales, lo cual le pudo haber incluido subjetividad al modelo debido a la sensibilidad de las preguntas.
- En los modelos enfocados a deserción temprana cuando se cuenta solo con los datos de ingreso, ninguna técnica logró una precisión superior del 67%, no obstante este es el momento más crítico pues como se evidenció en los trabajos realizados en la facultad la mayoría de los estudiantes que están desertado de manera temprana lo hacen en el primer año, por tanto sería ideal contar un modelo que pueda preceder la deserción desde el momento, no obstante debido a la poca información que se tiene del estudiante en ese momento se hace complejo que el modelo pueda aprender a identificar de manera adecuada el perfil de un posible desertor a partir de solo los datos de ingreso a la universidad, lo cual es consecuente con la teoría pues esta establece que al ser un fenómeno multicausal se debe contar con la mayor cantidad de características de las 4 cuatro categorías que describen el fenómeno de la deserción para poderlo explicar de manera más precisa.

# VI DESCRIPCIÓN DE LOS DATOS

En este capítulo se describen los datos utilizados en el desarrollo del proyecto. El primer conjunto de datos se obtuvo de los sistemas de información institucionales de la Universidad de Antioquia, MARES y MOISÉS y se contó además con el apoyo de personal de la vicerrectoría de docencia de la universidad para hacer el cruce de información entre dichas bases de datos y la entrega del conjunto de variables asociadas a los estudiantes de la facultad. Es de aclarar que la información suministrada fue anónima, no se tuvo acceso a la información personal de los estudiantes con el fin de dar cumplimiento a las políticas de privacidad y seguridad de la información de la universidad. Para reconocer a cada estudiante solo se contó con el número de documento y se le asignó un código único de identificación a cada uno de ellos. Adicionalmente, se dispuso de otro conjunto de variables suministrado desde el proyecto denominado: "**Implementación de la unidad de analítica y estudios universitarios: observatorio institucional**". Si bien ambos conjuntos acceden a las mismas bases de datos previamente mencionadas, fue necesario contar con ambas para obtener un conjunto de datos más completo.

Con el fin de lograr tener la mayor cantidad de variables posibles de las 4 categorías que describen la deserción, se optó por hacer uso de las bases de datos que ofrece el ICFES. Los tres conjuntos de datos se cruzaron para lograr un conjunto de datos final con el cual se realizó el entrenamiento de los modelos.

## *A. Base de datos Universidad de Antioquia*

El Sistema de Información de Inscripción y Selección de Estudiantes, MOISES registra la información de los estudiantes durante el proceso de inscripción y admisión y el Sistema de Matrícula y Registro MARES almacena la historia académica de los estudiantes e incluye los cursos que matrícula cada semestre, así como las respectivas notas, el promedio obtenido en el semestre, el promedio acumulado, las materias aprobadas, reprobadas y canceladas, entre otros [12]. Al cruzar los conjuntos de datos de la universidad se obtuvo la información de los estudiantes de la facultad de ingeniería desde el año 1996 hasta el 2019.

### *B. Base de datos ICFES*

El ICFES cuenta con la base de datos pública DataIcfes que almacena la información asociada a los estudiantes que aplican a los diferentes exámenes, los resultados de desempeño de estos en las pruebas, su información socioeconómica, la institución de educación media en la cual están matriculado, entre otros.

El examen SABER 11, anteriormente llamado Examen ICFES, es una evaluación estandarizada que tiene como propósito medir el desarrollo de las competencias de los estudiantes que están a punto de terminar la educación media [47]; se realiza de manera semestral y sus resultados han sido útiles para ayudar a las instituciones de educación superior en diferentes procesos. El 70% de las IES del país utilizan los resultados del examen SABER 11, de una u otra forma, para sus procesos de selección. Esta prueba ha demostrado ser útil en la prevención de la deserción, pues se ha utilizado como indicador para calcular la probabilidad de que un estudiante que ingresa a un programa de educación superior lo abandone antes de terminarlo [48]. Es el caso de la herramienta implementada por el MEN, llamada Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES), la cual sirve para hacer seguimiento de las cifras de deserción de los estudiantes de educación superior [49].

La información disponible en el repositorio relacionada con el examen SABER 11 corresponde a los resultados de los exámenes aplicados a partir del año 2000 hasta la actualidad. Es de anotar, que durante este periodo el examen ha presentado varias modificaciones en su estructura, presentando 3 modelos de exámenes diferentes, donde la puntuación, las áreas evaluadas y los cuestionarios de información del estudiante, presentan diferencias entre los modelos. Como se observa en la TABLA VI, para el periodo 1 se evaluaron 8 áreas y el idioma se seleccionaba entre 3 opciones, para el periodo 2 se evaluaron 7 áreas y para el periodo 3 se evaluaron solo 5 áreas y el idioma extranjero es el inglés.

TABLA VI
CAMBIOS DE ESTRUCTURA EXAMEN SABER 11

| **Período 1 2000-1 a 2005-2** | **Período 2 2006-1 a 2014-1** | **Período 3 2014-II en adelante** |
|---|---|---|
| Lenguaje (L)<br>Filosofía (Fi) | Lenguaje (L)<br>Filosofía (Fi) | Lectura crítica (LC) |
| Matemática (M) | Matemática (M) | Matemática (M)<br>(Incluye Razonamiento cuantitativo) |
| Física (F)<br>Química (Q)<br>Biología (B) | Física (F)<br>Química (Q)<br>Biología (B) | Ciencias naturales (CN) |
| Historia (H)<br>Geografía (G) | Ciencias sociales (CS)<br>(Historia y geografía) | Ciencias sociales y ciudadanos<br>(Incluye competencias ciudadanas) (CS) |
| Idioma (I)<br>(inglés, francés o alemán) | Idioma- inglés (I)<br>(inglés, francés o alemán) | Inglés (I) |

Por lo expuesto anteriormente, a pesar de contar con los puntajes individuales obtenidos en las diferentes áreas evaluadas estos no se incluyeron; pues de acuerdo a lo que indica el ICFES estos puntajes son comparables dentro del mismo período pero no entre períodos, esto es por ejemplo, los puntajes obtenidos en la prueba de matemáticas por un estudiantes en el periodo 1 no son comparables con los resultados de la prueba de matemáticas de un estudiante que la presentó en el periodo 2, debido a que la cantidad de preguntas y la puntuación de esta prueba cambió para cada periodo. En este caso para no descartar por completo el uso de estas variables se optó por trabajar con el índice de puntaje global tal como lo proponen [50]. En donde se calcula el promedio ponderado de los puntajes en las diferentes pruebas, tal como lo hace el ICFES en los reportes de los últimos años donde informa el puntaje global, el cual corresponde al índice global multiplicado por 5.

El índice de puntaje global se calcula para cada periodo de la siguiente manera:

Para el período 1:

$$IG = \frac{P_B + P_Q + P_F + P_H + P_G + P_{Fi} + 3P_L + 3P_M + P_I}{13} \quad (1)$$

Para el período 2:

$$IG = \frac{P_B + P_Q + P_F + 2P_{CS} + P_{Fi} + 3P_L + 3P_M + P_I}{13} \quad (2)$$

Para el período 3:

$$IG = \frac{3P_{LC} + 3P_M + 3P_{CN} + 3P_{CS} + P_I}{13} \quad (3)$$

En donde la P representa el puntaje y el subíndice hace referencia a la prueba evaluada, por lo tanto $P_B$ indica puntaje en la prueba de biología, $P_Q$ se refiere al puntaje de la prueba de química y así sucesivamente, de acuerdo a la letra que identifica cada área según lo visto en la TABLA VI

*C. Base de datos final*

Finalmente, se realizó el cruce de las bases de datos a partir de los campos comunes existentes en la base de datos UDEA y la del ICFES, estos campos fueron: fecha de nacimiento, género, año de titulación, colegio de egreso y municipio de origen.

En la TABLA VII se puede observar la descripción de las variables que se usaron, agrupadas en la respectiva categoría a la que pertenecen, de acuerdo con lo establecido por la teoría del estudio de la deserción que indica que se debe tener en cuenta variables de tipo individual, socioeconómico, académicas e institucionales [14],[7]. También se observa que la variable objetivo, denominada DESERTOR, adquiere dos posibles valores indicando si el estudiante fue o no etiquetado como desertor temprano, esto es, que estuvo en la universidad por 3 o menos semestres. Es de aclarar que algunos estudiantes desertaron en semestres posteriores los cuales se catalogan como desertores tardíos y por tanto no se etiquetaron como desertores para este caso de estudio.

Estas variables son las que se lograron tener en el momento de ingreso de los estudiante, contiene las variables del formulario de registro y las variables extraídas de la plataforma del

ICFES y se utilizaron para entrenar el modelo en el primer momento. Al conjunto completo de las variables se les denominó variables de ingreso.

TABLA VII
VARIABLES DE INGRESO

| Categoría | Campo | Descripción | Tipo Variable | Valores Posibles |
|---|---|---|---|---|
| Variables individuales | Cédula | Identificación del estudiante | Texto | |
| | Edad | Edad de ingreso a la universidad | Numérica | |
| | Género | Género del estudiante | Categórica | F - Femenino<br>M. Masculino |
| | Estadocivil | Estado civil del estudiante. | Categórica | Soltero, Unión libre, casado, separado, viudo |
| | Num_Hermanos | Número de hermanos del estudiante | Numérica | 0, 1, 2, 3... |
| | Tiempo_Ingreso | Tiempo en años que demora para ingresar a la universidad una vez que terminó el colegio. | Numérica | |
| | Discapacidad | Indica si el estudiante posee alguna discapacidad | Categórica | S -Si<br>N- No |
| Variables socioeconómicas | Lee_Escribe_Padre | El padre del estudiante sabe leer y escribir | Categórica | S -Si<br>N- No |
| | Lee_Escribe_Madre | La madre del estudiante sabe leer y escribir | Categórica | S -Si<br>N- No |
| | Educa_Padre | Nivel de educación del padre | Categórica | Postgrado, Profesional, Técnica, Secundaria, Primaria, No sabe, Ninguno |
| | Educa_Madre | Nivel de educación de la madre | Categórica | Igual que la anterior |

| | | | |
|---|---|---|---|
| Ocupa_Padre | Ocupación del padre | Categórica | Profesional dependiente<br>Pensionado<br>Operario-obrero<br>Pequeño empresario<br>Trabajador independiente<br>Administrador-Gerentes Otra Ocupación,<br>Profesional independiente,<br>Empresario<br>No aplica, Hogar |
| Ocupa_Madre | Ocupación de la madre | Categórica | Igual que la anterior |
| Icfes_Trabaja | Estudiante trabajaba al momento de presentar el ICFES | Categórica | S -Si<br>N- No |
| Fami_Computador | En la casa del estudiante tiene computador | Categórica | S -Si<br>N- No |
| Nro_Hijos | Número de hijos | Numérica | |
| Personas_Cargo | Número de personas a cargo | Numérica | |
| Estrato | Estrato socio económico | Numérica | De 1 a 6 ó 0 si vive en una zona donde no hay estratificación económica |
| Región | Región de Antioquia de origen del estudiante | Categórica | Oriente<br>Valle de aburra<br>Norte<br>Urabá Suroeste Bajo cauca Occidente<br>Nordeste<br>Magdalena medio<br>Otra (fuera de Antioquia) |
| Ciudad_Intermedia | Indica si el estudiante proviene de una ciudad intermedia | Numérica | 1-Si<br>0 - No |
| Ciudad_Capital | Indica si el estudiante proviene de una ciudad capital | Numérica | 1-Si<br>0 - No |

| | | | | |
|---|---|---|---|---|
| | Trabajador | Indica si el estudiante trabaja al momento de ingresar a la universidad. | Categórica | S -Si<br>N- No |
| Variables Académicas | Puntaje_Com_Lect | Puntaje en el examen de admisión en la prueba de competencia lectora | Numérica | |
| | Puntaje_Raz_Log | Puntaje en el examen de admisión en la prueba de razonamiento lógico | Numérica | |
| | Indice_global | Índice de puntaje global calculado para el tipo de examen presentado. | Numérica | |
| | Tipo_Cole | Carácter del colegio | Categórica | Oficial<br>No oficial |
| | Ingreso_Beca | Indica si el estudio ingresó a la universidad por cualquiera de las opciones de becas: Andrés Bello, Fundación María Cano, Mejor Bachiller, Deportista | Categórica | 1-Si<br>0 - No |
| | Ingreso_Especial | Indica si la admisión fue por una condición especial (INDIGENA, LEY1084A, LEY1084B, LEY1084C, NEGRITUD). | Categórica | 1-Si<br>0 - No |
| | Exp_Previa | Indica si el estudiante tiene experiencia previa en la educación superior, esto puede ser por: Cambio de modalidad, Educación flexible, Movilidad Transferencia, Reingreso sin examen, Cambio de programa | Categórica | 1-Si<br>0 - No |
| | Cole_Categoria | Categoría del colegio del que egresó el | Categórica | A+, A, B, C, D |

| | | | | |
|---|---|---|---|---|
| | | estudiante según la clasificación otorgada por el ICFES | | |
| | Credmatr_Semest1 | Cantidad de créditos matriculados en el semestre 1 | Numérica | |
| Variables institucionales | Programa | Programa al que se encuentra matriculado el estudiante | Categórica | Industrial, Materiales Sistemas, Sanitaria, Civil, Química, Electrónica, Eléctrica, Ambiental, Bioingeniería, Telecomunicaciones, Mecánica |
| Objetivo | Desertor | Objetivo variable. Indica si el estudiante es desertor temprano. La condición de deserción temprana se da cuando el estudiante abandona la universidad en los tres primeros semestres. | Categórica | 1-Si<br>0 - No |

Una vez que el estudiante logra terminar el primer semestre se tienen resultados de su desempeño académico, igualmente si logra terminar el segundo semestre. Estas nuevas variables se le adicionaron a las variables de ingreso para conformar otros dos grupos de variables para entrenar los modelos de los otros dos momentos.

En las TABLA VIII **y** TABLA IX se pueden observar las variables adicionales incluidas en los modelos para el semestre 1 (modelo 1) y semestre 2 (modelo 2) respectivamente, como se puede evidenciar estas variables son netamente de tipo académicas.

TABLA VIII
VARIABLES ACADÉMICAS DEL PRIMER SEMESTRE

| **Categoría** | **Campo** | **Descripción** | **Tipo Variable** | **Valores posibles** |
|---|---|---|---|---|
| | PROM_SEMEST1 | Promedio semestre 1 | Numérica | Entre 0 y 5.0 |

| | | | | |
|---|---|---|---|---|
| Variables académicas | PROMPROG_SEMEST1 | Promedio programa semestre 1 | Numérica | Entre 0 y 5.0 |
| | CREDAPROB_SEMEST1 | Cantidad de créditos aprobados en el semestre 1 | Numérica | |
| | CREDREPROB_SEMEST1 | Cantidad de créditos reprobados en el semestre 1 | Numérica | |
| | CREDCANCEL_SEMEST1 | Cantidad de créditos cancelados en el semestre 1 | Numérica | |
| | SITUACIÓN_STRE1 | Situación del estudiante al finalizar el primer semestre | Categórica | Normal Periodo de prueba |

TABLA IX
VARIABLES ACADÉMICAS DEL SEGUNDO SEMESTRE

| Categoría | Campo | Descripción | Tipo Variable | Valores posibles |
|---|---|---|---|---|
| Variables académicas | PROM_SEMEST2 | Promedio semestre 2 | Numérica | Entre 0 y 5.0 |
| | PROMPROG_SEMEST2 | Promedio programa semestre 2 | Numérica | Entre 0 y 5.0 |
| | CREDAPROB_SEMEST2 | Cantidad de créditos aprobados en el semestre 2 | Numérica | |
| | CREDREPROB_SEMEST2 | Cantidad de créditos reprobados en el semestre 2 | Numérica | |
| | CREDCANCEL_SEMEST2 | Cantidad de créditos cancelados en el semestre 2 | Numérica | |
| | CREDMATR_SEMEST2 | Cantidad de créditos matriculados en el semestre 2 | Numérica | |
| | SITUACIÓN_STRE2 | Situación del estudiante al finalizar el segundo semestre | Categórica | Normal Periodo de prueba |

*D. Preparación de los datos*

El conjunto de datos resultante fue de 15.187 estudiantes que ingresaron a la universidad en los períodos académicos comprendidos entre años 2000-2 a 2017-2 (equivalentes a 35 semestres) y pertenecientes a los diferentes programas de ingeniería ofertados por la facultad en la sede principal de la universidad. Si bien se tenían datos de la universidad desde el 1996 y hasta 2019, para realizar el cruce con los datos del ICFES se trabajó a partir del 2000; y debido a que la

deserción temprana se cuenta en los 3 primeros semestres, solo se incluyeron los estudiantes que ingresaron hasta el semestre 2017-2, pues los que ingresaron en los años 2018 y 2019 aún no se conoce su comportamiento con relación a la deserción temprana al no existir la información. Para el conjunto de datos resultante la tasa de deserción temprana fue de alrededor del 39%, recordando que se contó como desertor temprano a aquel estudiante que estuvo en la universidad solo por 3 semestres o menos.

Para la base de datos final se seleccionaron las características que según la teoría son determinantes a la hora de predecir el riesgo de deserción [14], [15], [2] y que no tuvieran más del 60 % de datos faltantes. Si bien algunos de los cuestionarios del ICFES aportan mucha información sobre la orientación vocacional del estudiante como, por ejemplo: sus expectativas en cuanto a la universidad en la que desea estudiar, la carrera deseada, aspiración salarial, entre otras, en el cruce final se obtuvo más de 70% de datos faltantes para estas variables, razón por la cual no fueron incluidas. Debido a que para el uso de modelos computacionales es determinante contar con una base de datos con información de calidad, es necesario eliminar la mayor cantidad de ruido que puedan generar los datos; por tal razón, a pesar de que en las bases de datos del ICFES se encontraban una gran cantidad de variables que serían de ayuda para poder crear un perfil del estudiante al momento del ingreso no fue posible incluir todas las variables disponibles en estas bases de datos en el análisis ya que algunas de ellas en el cruce final de los datos resultaron con una porción muy alta de datos faltantes. Otro factor problemático que se presentó relacionado con la bases de datos del ICFES fue que a pesar de que en estas están disponibles los puntajes individuales obtenidos en el examen SABER 11 y estos serían útiles para evaluar el desempeño del estudiante en áreas como matemáticas, física e inglés que son de gran importancia en la fundamentación teórica de las ingenierías, esta información corresponde a diferentes modelos de exámenes aplicados durante los diferentes periodos; en consecuencia, no se usaron directamente si no que se utilizaron para calcular el índice de puntaje global, lo anterior debido a que por indicación del ICFES los puntajes individuales de los diferentes modelos de examen no son comparables entre periodos ya que en cada uno se evalúan competencias diferentes.

Para resolver los datos faltantes en las variables de tipo categóricas se utilizó la estrategia de reemplazar por el dato más frecuente y en el caso de las variables de tipo numéricas se reemplazaron por la media. La misma estrategia se utilizó para trabajar los datos atípicos.

Finalmente, debido a que algunas variables numéricas tienen valores significativamente diferentes como es el caso por ejemplo de la edad que toma valores entre 15 y 35 y los puntajes del examen de admisión que pueden tomar valores entre 0 y 100, las variables de tipo numéricas se normalizaron antes de llevarlas a los modelos para el entrenamiento, para esto se dividió el valor de las variables numéricas por el mayor valor del rango de manera que la variable tomó un valor entre 0 y 1.

Las variables numéricas que se normalizaron fueron: edad, número de hijos, número de personas a cargo, número de personas en el hogar, puntajes en las pruebas de admisión, índice de puntaje global obtenido del ICFES, número de créditos matriculados en el semestre 1 y 2, promedios de semestre 1 y 2, cantidad de créditos aprobados, cancelados y reprobados tanto en el semestre 1 como en el semestre 2.

Adicionalmente, teniendo en cuenta que los modelos de clasificación no reciben datos categóricos, estos se codificaron utilizando la representación One-Hot.

Las variables categóricas fueron : género, estado civil, tipo de colegio, categorización del colegio, programa, tipo de admisión, si el estudiante ingresó por beca o por alguna condición especial, región de procedencia, si el estudiante proviene de una ciudad capital o de una ciudad intermedia, nivel de educación de la madre y del padre, ocupación de la madre y del padre, entre otras. Ver TABLA VII

El producto final obtenido en esta parte del trabajo es una base de datos filtrada, depurada y codificada, la cual fue obtenida luego de hacer el cruce de las bases de datos institucionales; y de la información disponible en la plataforma del ICFES.

*E. Análisis exploratorio de los datos*

En la Fig. 1 se observa la cantidad de semestres que permanecen los estudiantes de la facultad matriculados. El gráfico evidencia que hay una porción muy alta de estudiantes que permanecen en la universidad durante solo un semestre, lo cual le da sustento a la realización del trabajo.

De la misma gráfica también se puede observar que hay un grupo considerablemente importante de estudiantes, que duran entre 10 y 14 semestres matriculados; esta cifra es la cantidad de semestres que se tarda un estudiante de la facultad para graduarse [6], [9].

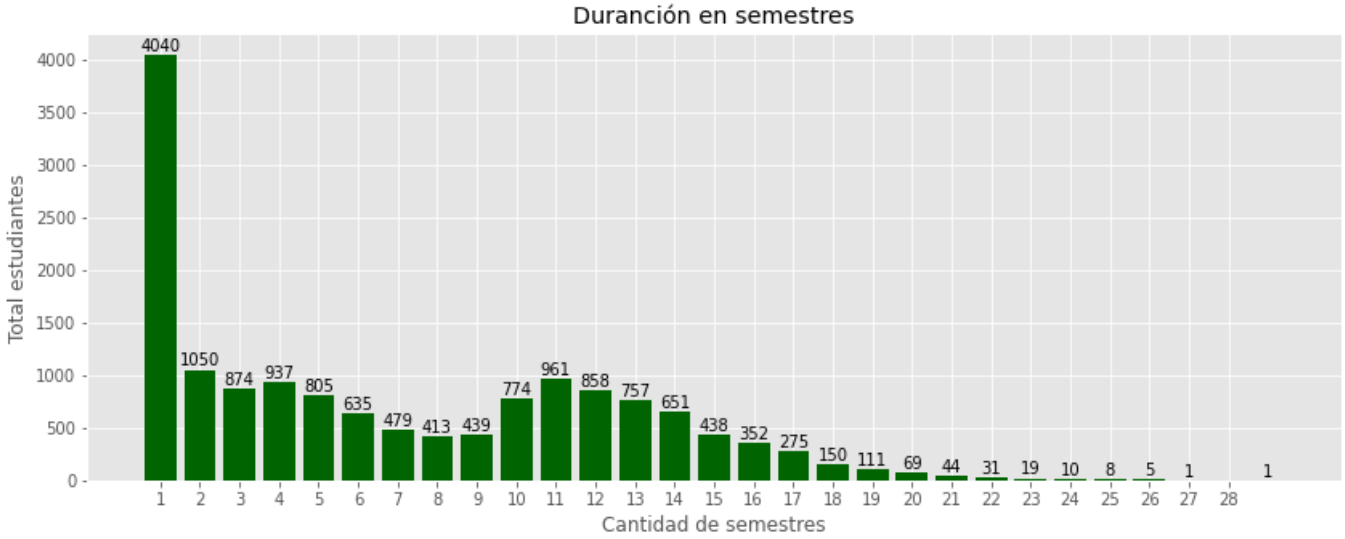



Fig. 1. Duración en semestres de los estudiantes de la facultad de ingeniería años 2000-II a 2017-II

De acuerdo con lo anteriormente mencionado, es de gran importancia entonces poder identificar a los estudiantes que tienen un alto riesgo de desertar desde el momento que realizaron la matrícula del primer semestre, pues como se observa en la Fig. 2 el 67.8% de los estudiantes que desertaron de manera temprana lo hicieron en el primer semestre en muchas ocasiones sin que este haya terminado, razón por la cual se hacía crucial lograr un buen desempeño del modelo 0.

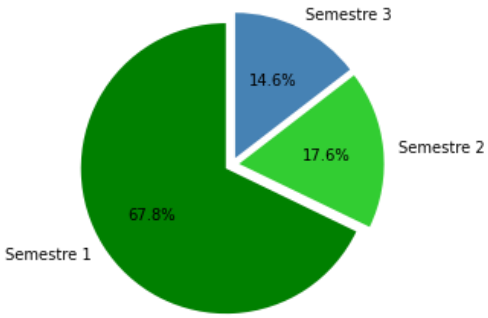



Fig. 2. Porcentajes de deserción temprana por cantidad de semestres

1. *Edad de ingreso*

De la Fig. 3 se puede ver que 11.353 de los 15.187 al momento del ingreso estaban entre los 17 y 19 años, es decir ingresaron a la universidad una vez terminaron el ciclo de educación media. Este atributo presenta un promedio de 18.84 años con una desviación estándar de 2.26 años.

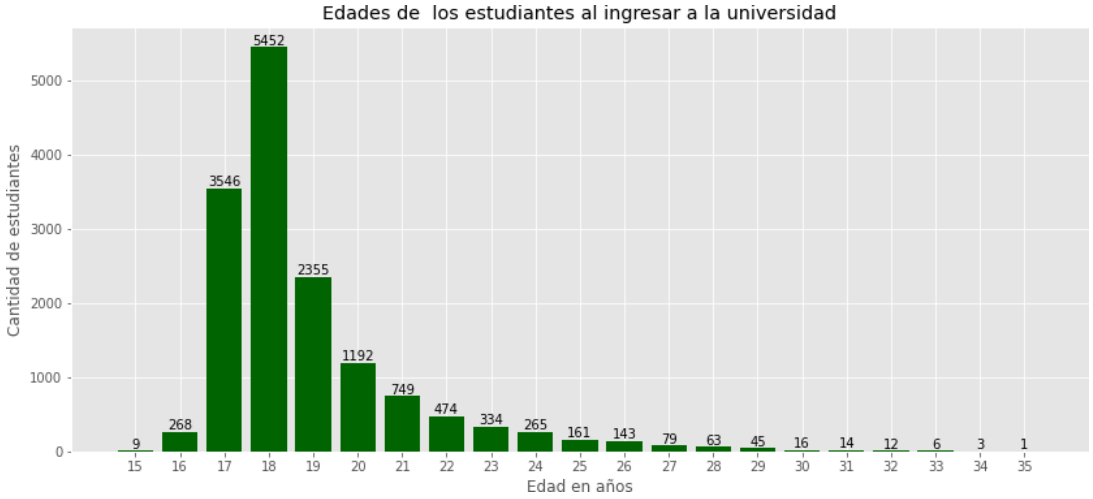



Fig. 3. Edad de ingreso

2. *Estudiantes por programa*

En la Fig. 4 se observa la cantidad de estudiantes matriculados por cada programa donde se evidencia que el programa con mayor número de estudiantes matriculados es ingeniería química, mientras que ingeniería ambiental e ingeniería de telecomunicaciones son las menos demandadas de la facultad.

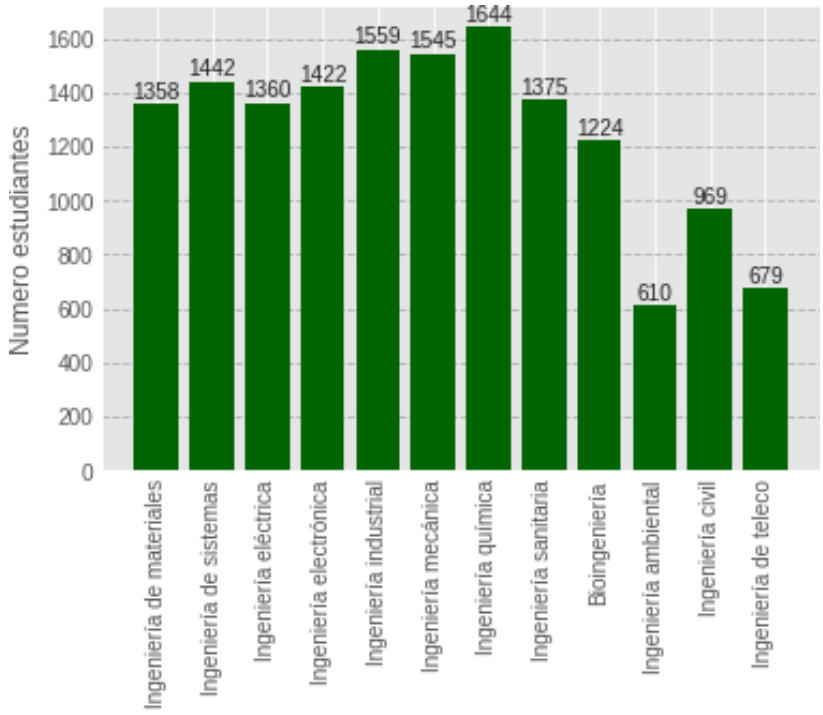



Fig. 4. Cantidad de estudiantes por programa

De la Fig. 5 se puede ver el comportamiento de la deserción temprana por programa, donde se evidencia que la mayor proporción de desertores tempranos se encuentran en el departamento de ingeniería sanitaria con un 45%, seguida por materiales, mecánica y telecomunicaciones, en donde la deserción temprana fue de 43%. Mientras que los programas que presentan una menor tasa de desertores son: ingeniería electrónica e ingeniería química con un valor de 32% y 33% respectivamente.

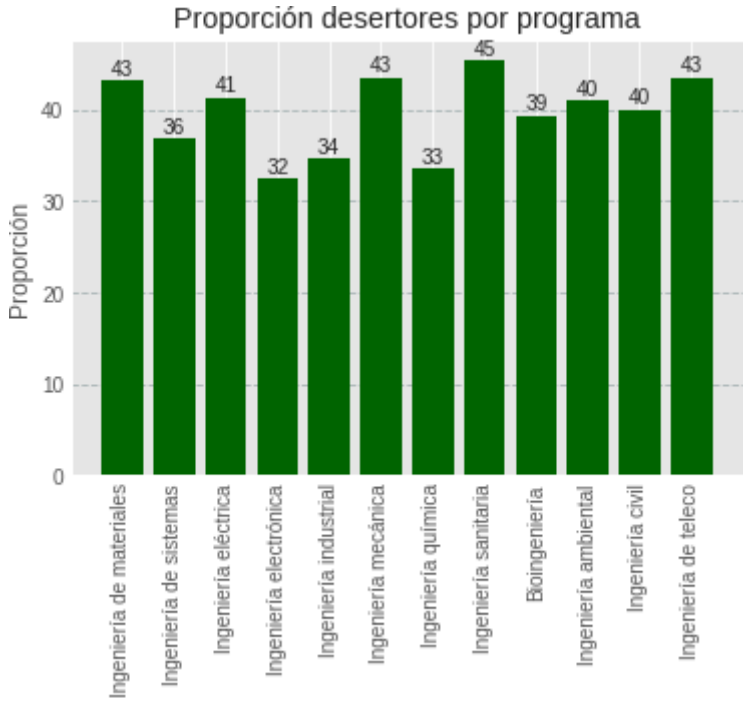



Fig. 5. Proporción de estudiantes desertores por programa

### 3. *Estudiantes por estrato*

En cuanto al estrato socioeconómico de los estudiantes de la facultad de ingeniería se observa que en su gran mayoría pertenecen a los estratos 2 y 3, tal como lo demuestra la Fig. 6.

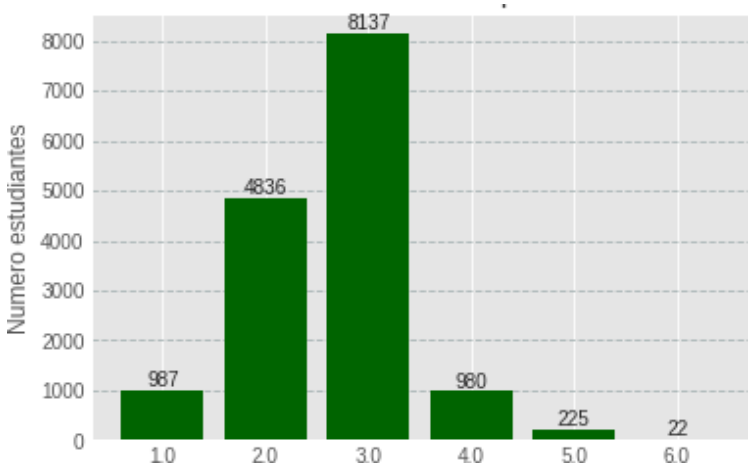


Fig. 6. Cantidad de estudiantes por estrato

En la Fig. 7 se muestra el comportamiento de la deserción temprana por estratos la cual es mayor en el estrato 3 y 6.

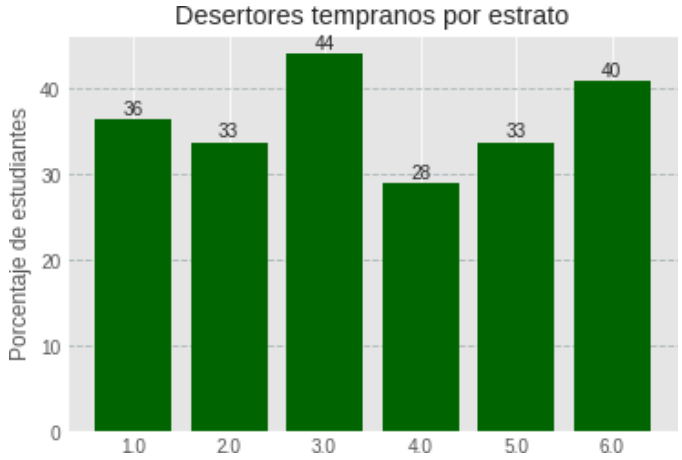



Fig. 7. Proporción de estudiantes desertores por estrato

*4. Estudiantes por género*

Del género de los estudiantes se puede decir que la gran mayoría de estudiantes de la facultad son hombres, los cuales representan el 69,8 % de la población estudiada e igualmente estos presentan la tasa más alta de desertores según lo informan las Fig. 8 y Fig. 9.

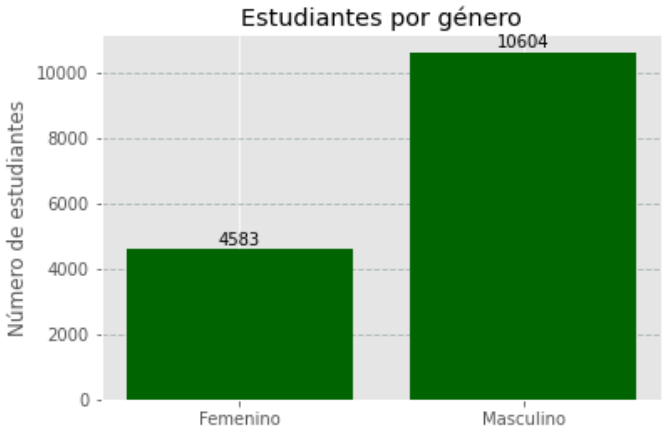



Fig. 8. Cantidad de estudiantes por género

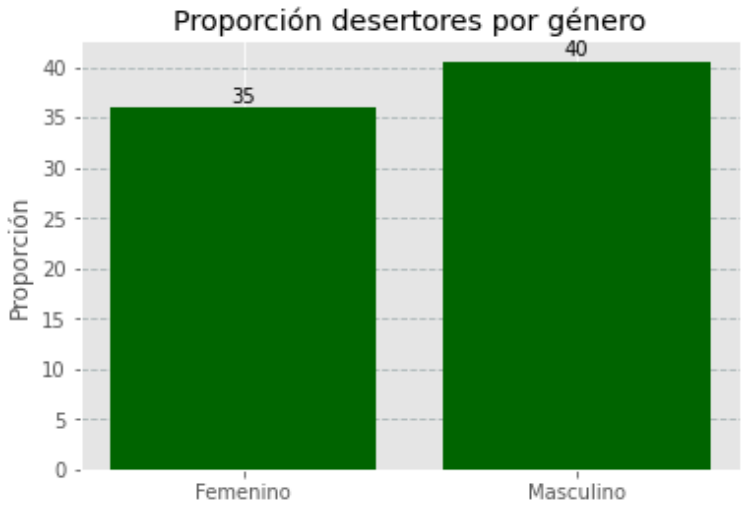



Fig. 9. Proporción de desertores por género

5. *Estudiantes por tipo de colegio*

De acuerdo con lo observado en las Fig. 10 y Fig. 11, cerca del 61% de los estudiantes de la facultad proviene de colegios oficiales, sin embargo, una proporción más alta de los estudiantes de colegios no oficiales son los que desertan de manera temprana.

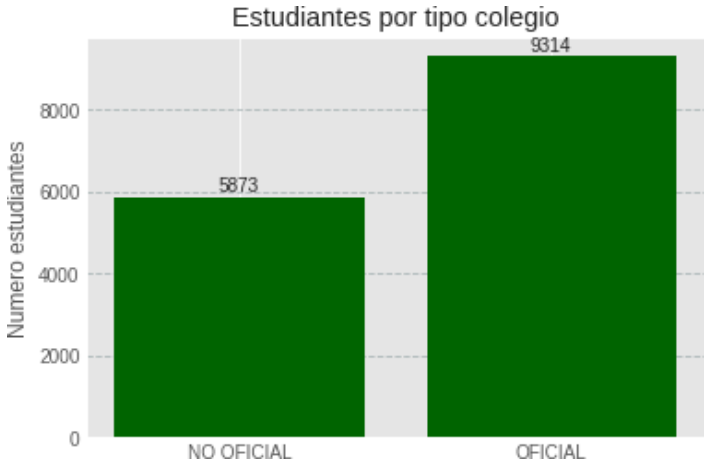



Fig. 10. Cantidad de estudiantes por tipo de colegio


Fig. 11. Proporción de estudiantes desertores por tipo de colegio

### *6. Tipo de admisión*

Existen diferentes modalidades de admisión a la universidad de Antioquia: por examen, reingreso, transferencia, cambio de sede o cambio de programa, entre otros. Como se puede observar en la Fig. 12, la mayoría de los estudiantes, el 89.5% ingresaron por examen.


Fig. 12. Cantidad de estudiantes por tipo de admisión

*7. Región de procedencia*

El atributo región corresponde a la región de Antioquia donde nació el estudiante, se observa una alta concentración de estudiantes pertenecientes al Valle de Aburrá Fig. 13; estos representan aproximadamente el 60% de la totalidad de los estudiantes y 3.266 que provienen de otras regiones fuera de Antioquia, esto es el 21,5%. Aparte del Valle de Aburrá y regiones fuera de Antioquia la tercera región con más estudiantes provenientes de ahí es el Oriente antioqueño que representan el 6,7%. Lo cual se puede deber a su cercanía con el Valle de Aburrá no obstante, para estos estudiantes el desplazarse a la universidad podría generar costos económicos y tiempos adicionales.

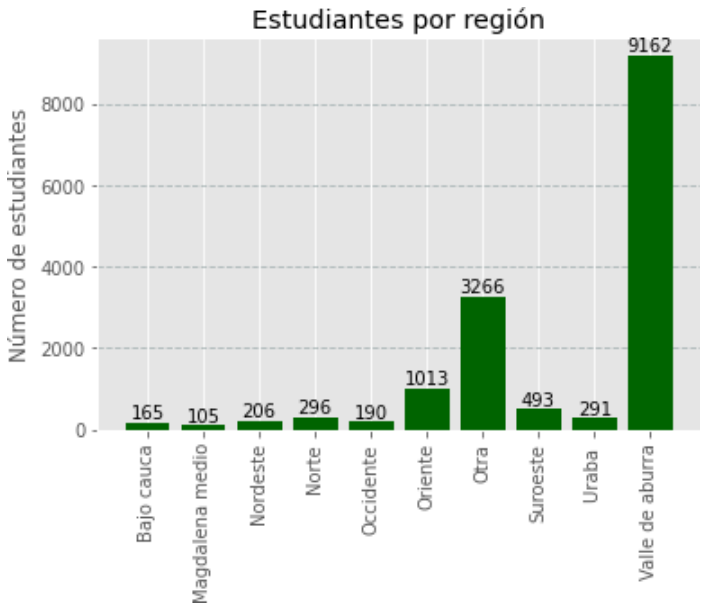



Fig. 13. Cantidad de estudiantes por región

*8. Ocupación de los padres*

En relación con la ocupación de los padres como se observa en Fig. 14 y en Fig. 15 la mayoría de los padres de los estudiantes son trabajadores independientes, mientras que las madres se dedican en su gran mayoría al hogar.

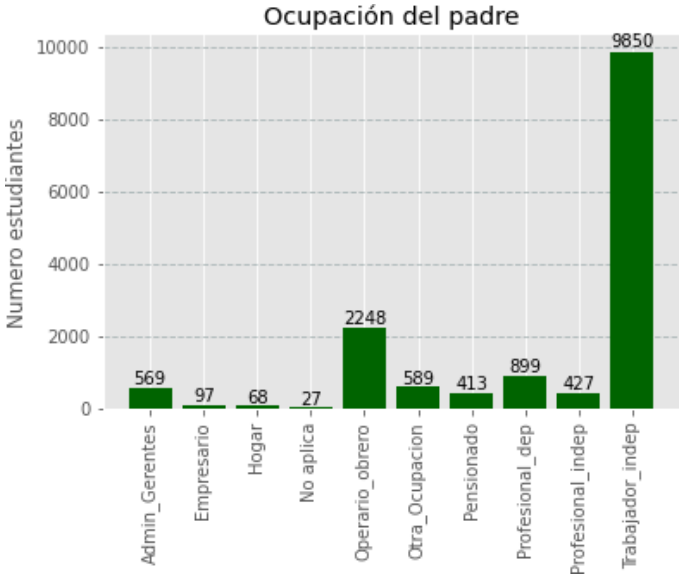



Fig. 14. Número de estudiantes por ocupación del padre

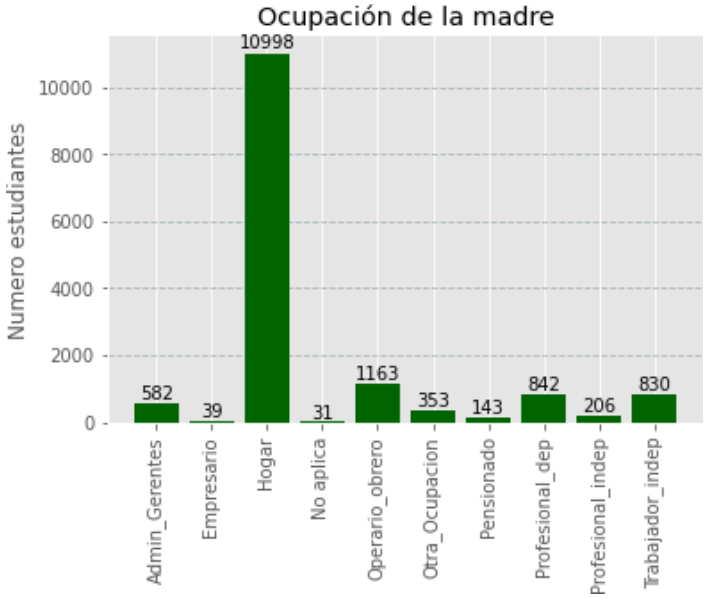



Fig. 15. Número de estudiantes por ocupación de la madre

### *9. Nivel estudio de los padres*

De acuerdo a lo observado en las Fig. 16 y Fig. 17 tanto la madre como el padre posee un nivel máximo de educación secundaria, representando el 69.9% en el caso de la madre y 65,9% en el

padre. Solo el 7,7 %  de las madres poseen un nivel de educación profesional mientras que el 8,7 % de los padres de los estudiantes son profesionales .

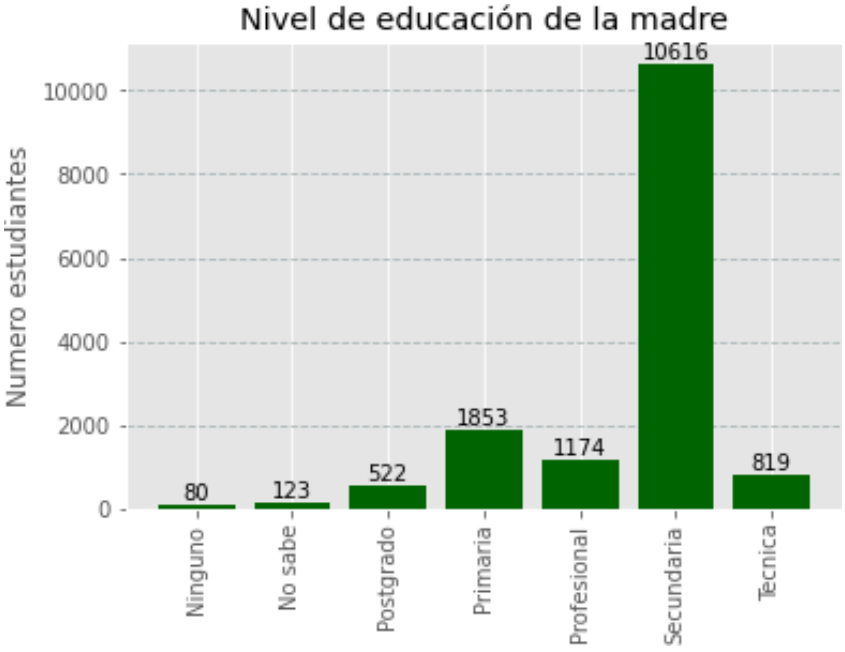



Fig. 16. Número de estudiantes por nivel de estudios de la madre

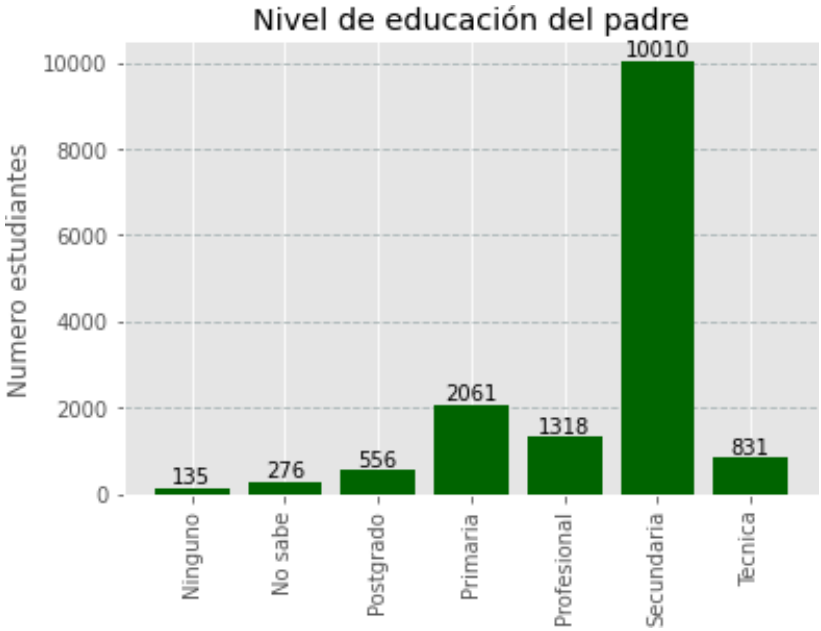



Fig. 17. Número de estudiantes por nivel de estudios del padre

# VII METODOLOGÍA USADA

En la Fig. 18 se presenta el diagrama de flujo de la metodología propuesta para el desarrollo de la investigación.

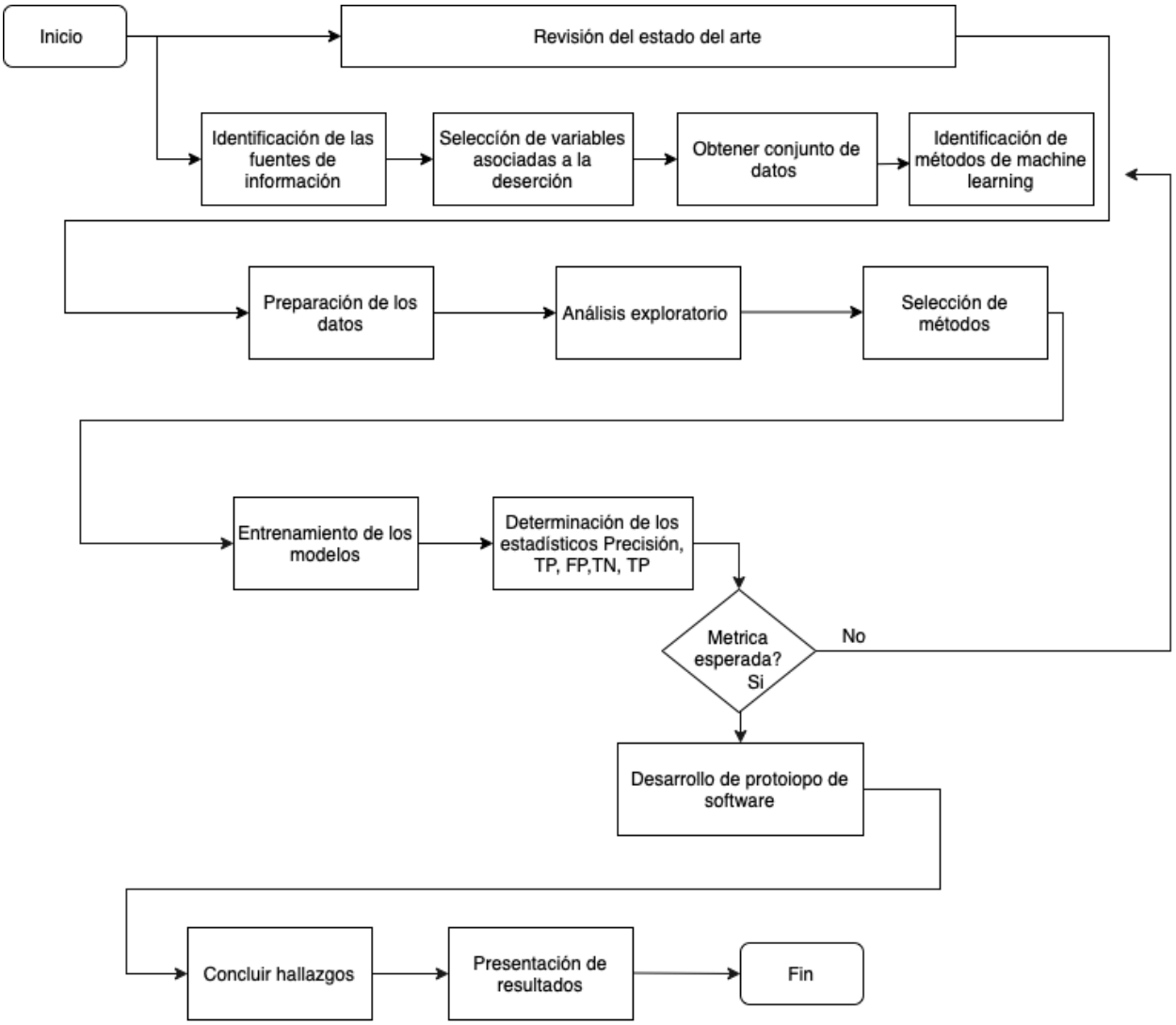



Fig. 18. Diagrama de flujo de la metodología propuesta

La metodología empleada consistió inicialmente en un análisis del estado de arte y del marco teórico para identificar las variables que influyen en la deserción, a la par que se fueron identificando las fuentes de datos que se tenían disponibles para seleccionar las variables que serían incluidas en el estudio. Las fuentes de datos identificadas fueron, las bases de datos institucionales

y la plataforma del ICFES. Se construyó un conjunto de datos procurando incluir la mayor cantidad de variables de cada una de las categorías que, según la teoría deben tenerse en cuenta para dar explicación al fenómeno de deserción, cuidando que los datos tengan una buena calidad, en este caso a pesar de que se contaba con una gran cantidad de variables disponibles especialmente en la base de datos del ICFES se excluyó gran cantidad de estas por tener una porción muy alta de datos faltantes.

De la revisión bibliográfica se identificaron los métodos del *machine learning* más comúnmente utilizados en el estudio de la deserción y el desempeño logrado por estos, también se revisaron los métodos no explorados para este propósito.

Una vez seleccionado el conjunto de datos se pasó a la etapa de preparación de estos, identificando datos faltantes o datos atípicos y validando que las variables se encuentren en el formato correcto para ingresarlos en los modelos.

Se realizó el análisis exploratorio de la información para tener un conocimiento previo del conjunto de estudiantes que serían objeto de estudio.

De los métodos revisados se estableció que los más popularmente usados eran los árboles de decisión, no obstante estos no tenía un buen desempeño satisfactorio en el momento de analizar la deserción temprana cuando se cuenta solo con los datos de ingreso por lo que se optó por trabajar con otros métodos más robustos y entrenar un modelo con árboles de decisión para tenerlo como modelo base.

Una vez seleccionados los métodos se pasó al entrenamiento de los modelos y se fueron ajustando los parámetros de cada uno para lograr un desempeño aceptable de cada modelo. Se evaluaron las métricas tales como precisión, falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos, entre otros, para evaluar la eficiencia de clasificación temprana del o los métodos seleccionados.

Uno vez que se obtuvo un resultado que se consideró aceptable de cada modelo y teniendo en cuenta las limitaciones que se tenía en cuanto a la disponibilidad de la información, se construyó un prototipo de software que pudiera utilizar el modelo previamente entrenado de manera que ingresando los datos de nuevos estudiantes, el software pudiese clasificar (predecir) si el estudiante se encuentra en riesgo de desertar de manera temprana.

Finalmente, se pasó a realizar las conclusiones sobre los hallazgos, redacción del informe y del artículo.

# VIII INVESTIGACIÓN EXPERIMENTAL

Dado que el tipo de deserción que se estudió fue la deserción temprana y de acuerdo con la definición establecida por la UDEA, es la que se da en los 3 primeros semestres [13] se optó por entrenar 3 modelos diferentes: un modelo 0 que se da en el momento en que el estudiante ingresa a la universidad y matricula su primer semestre. Para este modelo se contó con los datos de inscripción en el proceso de matrícula y la información recuperada del ICFES estas variables fueron las denominadas variables de ingreso ver TABLA VII . El modelo 1 se da en el momento en que el estudiante logró permanecer un semestre en la universidad y se matricula para el segundo semestre, en este caso se cuenta con la misma información del modelo 0 (variables de ingreso ) y adicionalmente se cuenta con los resultados de desempeño académico obtenidos durante el primer semestre, tales variables adicionales se observan en la TABLA VIII Finalmente, para el tercero modelo 2, se cuenta con toda la información del modelo 1 y se incluye la información de desempeño académico durante el segundo semestre TABLA IX.

Existen varias técnicas que se han utilizado para el propósito de predecir deserción temprana, tal como se pudo observar en el capítulo de revisión del estado del arte, la técnica más comúnmente utilizada son los árboles de decisión; sin embargo, en el caso en el que solo se cuenta con los datos de ingreso los resultados no lograron ser superiores al 68% de precisión en la predicción, razón por la cual algunos autores proponen utilizar técnicas con mayor capacidad predictiva como son las redes neuronales artificiales [36]. Por tal razón se decidió usar para este propósito las redes neuronales y la técnica del XGBoots que está basada en árboles de decisión pero que ensambla varios árboles de manera secuencial y realiza varias repeticiones mientras se va corrigiendo el error. También se trabajó con árboles de decisión para tenerlo como modelo base y comprobar si las técnicas anteriormente propuestas logran un mejor desempeño frente a los árboles de decisión que es la técnica más popularmente usada.

En esta sesión se presentan los resultados obtenidos durante el entrenamiento de los diferentes modelos con ambas técnicas.

## *Métricas empleadas para el análisis*

Se utilizaron las siguientes métricas de clasificación de evaluación de clasificadores binarios

- Accuracy expresa el porcentaje de individuos que fueron clasificados correctamente.

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

En el caso de estudio particular indica el porcentaje de estudiantes que fueron correctamente clasificados dentro de su respectiva clase: desertor, no desertor.

- Precisión

$$Precision = \frac{TP}{TP + FP}$$

Indica la proporción de estudiantes desertores que fueron identificados correctamente.

- Recall

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

Indica la proporción de estudiantes que fueron clasificados como desertores del total de estudiantes desertores existentes.

El f-score se define como una media armónica de la precisión y el recall. Un valor alto de la medida F indica que tanto la precisión como el recall son razonablemente altas.

$$F1 = 2 * precisión * recall \ / \ (precision + recall)$$

Donde:

TP: True Positive (Verdadero Positivo)

FN: False Negative (Falso Negativo)

FP: False Positive (Falso Positivo)

TN: True Negative (Verdadero Negativo)

AUC ofrece una estimación de qué tan bueno es el modelo para la tarea de clasificación, se expresa como el área bajo la curva (ROC).

De las métricas anteriormente mencionadas fue considerado el Recall como la más importante, puesto que obtener un alto *Recall*, aumenta la probabilidad de calcular de manera acertada los estudiantes desertores el cual es el propósito principal del estudio.

Los modelos se entrenaron en los diferentes momentos usando las técnicas seleccionadas, en cada caso se configuraron los híper parámetros a fin de lograr el mejor desempeño de los modelos frente a las métricas mencionadas.

Se utilizó una herramienta de Sklearn llamada "GridSearchCV", la cual a través de una búsqueda exhaustiva entrega la mejor combinación de parámetros en cuanto a una determinada métrica de rendimiento.

*Modelo base con árboles de decisión*

TABLA X
MALLA DE HÍPER- PARÁMETROS CONFIGURADOS EN LOS ÁRBOLES DE DECISIÓN

| | PARÁMETRO | VALORES |
|---|---|---|
| 1 | *Criterion* | ['gini', 'entropy'] |
| 2 | max_depth | [2, 3, 4, 5, 6, 7, 8, 9] |
| 3 | min_samples_splits | [2, 3, 4] |
| 4 | *min_samples_leaf* | [1, 2, 3, 4, 5, 6, 7, 8, 9,10] |
| 5 | *max_leaf_nodes* | [1, 2, 3, 4, 5, 6, 7, 8, 9,10] |

En la TABLA X se muestran los diferentes conjuntos de parámetros configurados en el modelo de árboles de decisión el cuál será el modelo base, se seleccionó esta técnica ya que es la más popularmente utilizada para este propósito, se tomaron sus resultados como base y se compararon con las otras técnicas seleccionadas con el propósito de buscar una mejora en el desempeño en los diferentes modelos entrenados.

Se probaron las diferentes combinaciones de parámetros buscando la mejor combinación de estos las cuales se pueden observar en la TABLA XI y los resultados obtenidos de los

entrenamientos haciendo uso de tales conjuntos de parámetros en cada caso se pueden observar en la TABLA XII

TABLA XI
CONJUNTO DE HIPERPARAMETROS RESULTANTES PARA ÁRBOLES DE DECISIÓN

| Parámetro | Modelo 0 | Modelo 1 | Modelo 2 |
|---|---|---|---|
| *Criterion* | 'gini' | 'gini' | 'gini' |
| max_depth | 4 | 6 | 6 |
| min_samples_splits | 0.2 | 0.2 | 0.3 |
| *min_samples_leaf* | 5 | 4 | 6 |
| *max_leaf_nodes* | 9 | 2 | 9 |

TABLA XII
RESUMEN DE MÉTRICAS USANDO LA TÉCNICA DE ÁRBOLES DE DECISIÓN

| Modelo | Accuracy_Train | Accuracy_Test | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| Modelo 0 (datos de ingreso) | 0.733706 | 0.732900 | 0.681233 | 0.6453788 | 0.662821 | 0.723834 |
| Modelo 1 Desempeño académico sem1 | 0.853815 | 0.847268 | 0.838272 | 0.783879 | 0.81014 | 0.845630 |
| Modelo 2 Desempeño académico sem1 y sem 2 | 0.895135 | 0.885122 | 0.878069 | 0.834943 | 0.85596 | 0.883838 |

*Resultados de la técnica XGBoots*

En este caso se construyó la malla de parámetros y se entrenaron los diferentes modelos realizando una búsqueda de la mejor combinación de parámetros que ofrecieran un resultado aceptable. En la TABLA XIII se pueden observar los valores de los diferentes parámetros introducidos en la malla.

TABLA XIII
MALLA DE HÍPER- PARÁMETROS CONFIGURADOS EN EL XGBOOTS

| | PARÁMETRO | VALORES |
|---|---|---|
| 1 | *learning_rate* | [ 0.05, 0.1, 0.2, 0.5, 0.8] |
| 2 | *n_estimators* | [ 100, 200, 300, 500] |
| 3 | max_depth | [2,3,4,5] |
| 4 | *min_child_weight* | [0,1,2] |
| 5 | *Reg. gamma* | [0.0, 0.1, 0.2, 0.3, 0.4] |
| 6 | *subsample* | [0.6, 0.7, 0.8, 0.9] |
| 7 | *colsample_bytree* | [0.6, 0.7, 0.8, 0.9] |
| 8 | *reg_alpha* | [1e-5, 1e-2, 0.1, 1] |

Si bien los datos no presenta un alto desbalance pues hay una relación aproximadamente de 40 a 60 de la clase desertora con respecto a la clase no desertora, para caso del XGboots es posible configurar un parámetro denominado scale_pos_weight, el cual controla el balance ponderando las clases y se calcula como:

suma (instancias negativas) / suma (instancias positivas)

En el caso de la base de datos usada (noDesertor/desertor) se calculó así: (5950)/(15187)= 1.55

Y dado que es clasificación binaria el parámetro objetive= ‘binary:logistic’

Una vez ejecutada la búsqueda de parámetros se obtuvo las siguientes combinaciones de parámetros en cada caso ver TABLA XIV

TABLA XIV
CONJUNTO DE HIPERPARAMETROS RESULTANTES PARA EL XGBOOTS

| Parámetro | Modelo 0 | Modelo 1 | Modelo 2 |
|---|---|---|---|
| learning_rate | 0.01 | 0.1 | 0.01 |
| n_estimators | 200 | 100 | 100 |
| max_depth | 3 | 3 | 3 |
| min_child_weight | 5 | 2 | 1 |
| gamma | 0.1 | 0.4 | 0.4 |
| subsample | 0.8 | 0.8 | 0.85 |
| colsample_bytree | 0.8 | 0.8 | 0.8 |
| reg_alpha | 0 | 0.01 | 0.001 |
| objective | binary:logistic | binary:logistic | binary:logistic |
| nthread | 4 | 4 | 4 |
| scale_pos_weight | 1.55 | 1.55 | 1.55 |

Una vez entrenados los diferentes modelos en los tres momentos con los conjuntos de parámetros resultantes en cada caso se obtuvo el conjunto de métricas de la TABLA XV dónde se observa que el para el modelo 0 el cual era el momento más crítico; puesto que es el semestre donde los estudiantes están desertando con mayor frecuencia, se logró un *accuracy* de 74,90% con los datos de test y también se observó que el modelo tiene la capacidad de detectar correctamente el 67,5% de los estudiantes desertores.

TABLA XV
RESUMEN DE MÉTRICAS USANDO LA TÉCNICA XGBOOTS

| Modelo | Accuracy_Train | Accuracy_Test | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| Modelo 0 (datos de ingreso) | 0.75512 | 0.749177 | 0.675214 | 0.676949 | 0.676080 | 0.81151 |
| Modelo 1 Desempeño académico sem1 | 0.891349 | 0.884793 | 0.891612 | 0.801016 | 0.843889 | 0.869544 |
| Modelo 2 Desempeño académico sem1 y sem 2 | 0.916619 | 0.907834 | 0.910811 | 0.856054 | 0.882584 | 0.953702 |

*Variables de importancia*

El algoritmo de XGBoots ofrece la posibilidad de mostrar las variables de importancia, es decir aquellas variables que más aportan información al proceso de clasificación, las cuales se pueden observar en las Fig. 19, Fig. 20 y Fig. 21

En la Fig. 19 se muestran la variables de importancia en el modelo 0, momento de ingreso, donde se puede observar que para este caso la variable que más aporta al proceso de clasificación es CRED_MATR_SEM1 que representa la cantidad de créditos matriculados en el primer semestre lo cual es consecuente con los resultados de otros estudios en la facultad que indican que la carga de créditos en los primeros semestres es determinante en la deserción pues estos corresponden a los cursos básicos de matemáticas y físicas las cuales son materias que representan gran dificultad para los estudiantes por la poca preparación con la que llegan de la educación secundaria.

Por otro lado la Fig. 20 presenta las variables de importancia del modelo 1, donde se observa que variables como CRED_MATR_SEM2 que indica la cantidad de créditos matriculados en el segundo semestre y PROM_SEM1 promedio de carrera obtenido durante el primer aportan en gran medida al modelo de clasificación. Igualmente, que el caso anterior, son consecuentes con los hallazgos de investigaciones previas en los que se menciona que en los primeros semestres el desempeño académico fue determinante a la hora de tomar la decisión de desertar. Caso similar se observa en la Fig. 21 donde se evidencia que las variables más relevantes en el modelo 2, son aquellas que dan cuenta del rendimiento académico del estudiante.

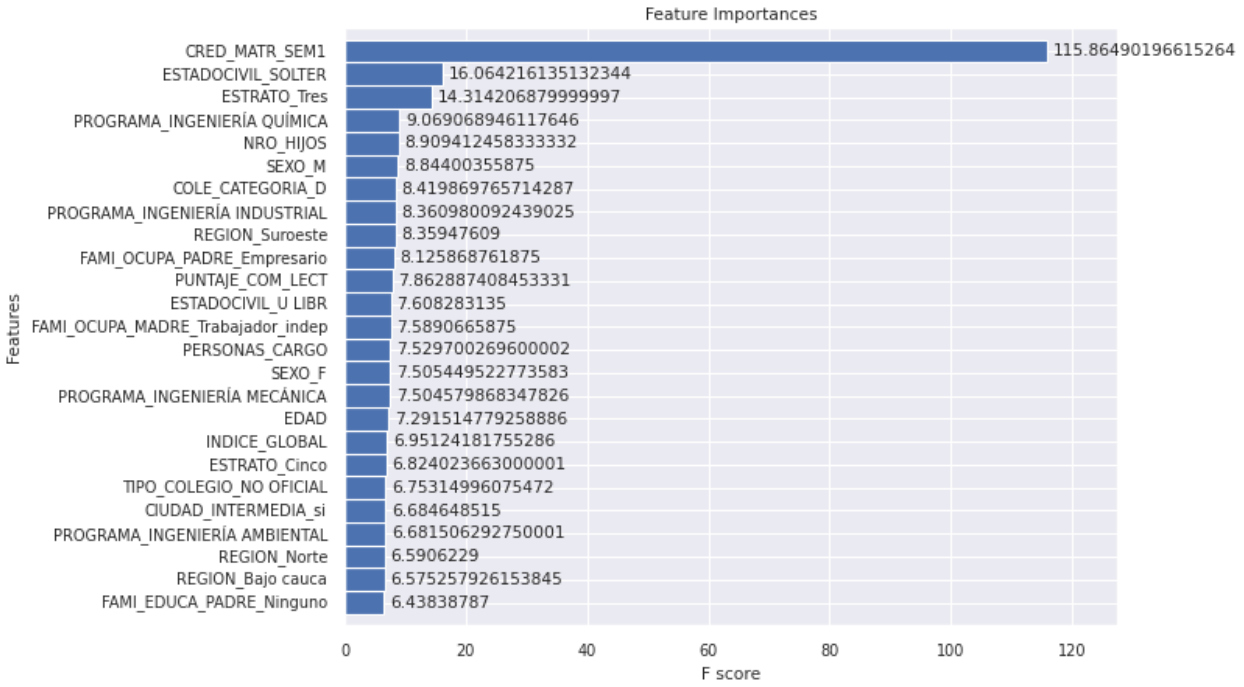



Fig. 19 Variables de importancia para el modelo 0

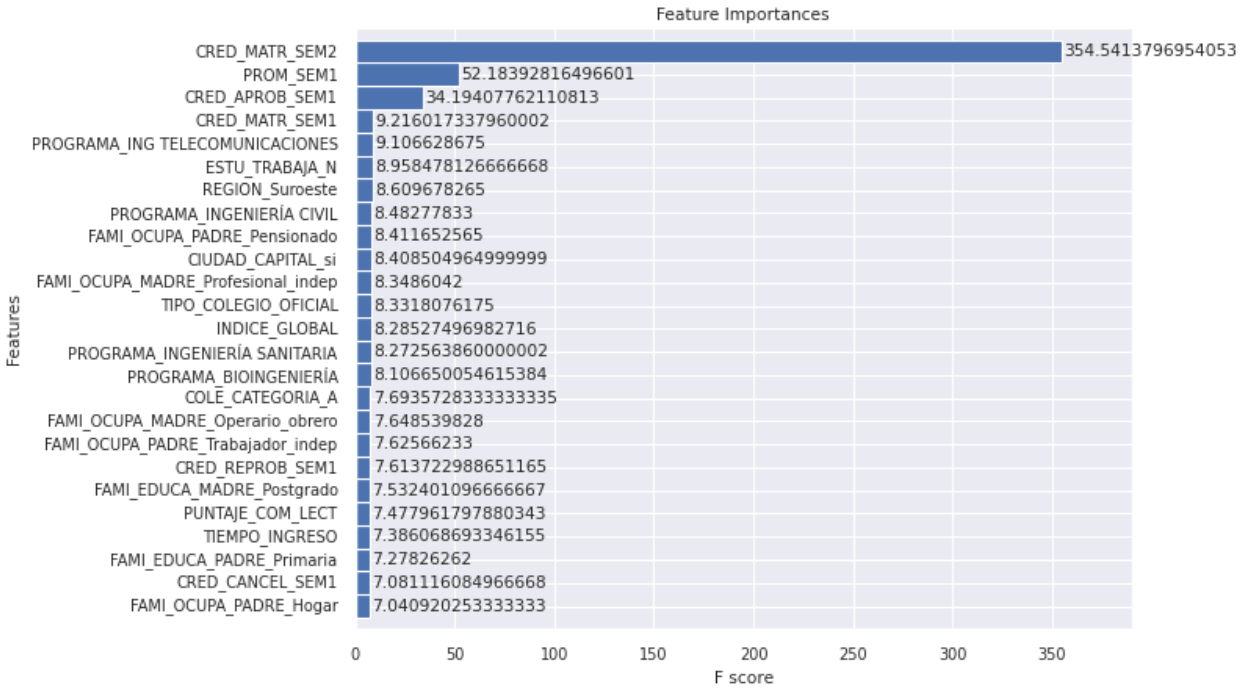



Fig. 20 Variables de importancia para el modelo 1

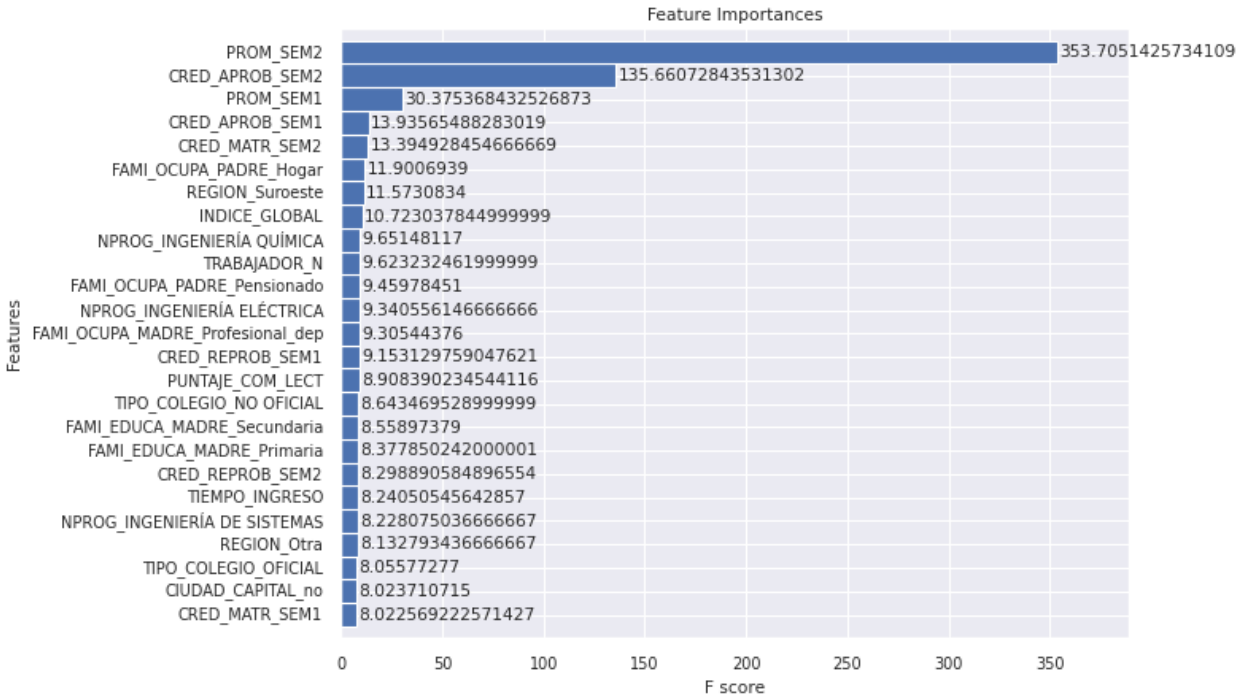



Fig. 21 Variables de importancia para el modelo 2

*Resultados de Técnica RNA*

Igualmente que en el caso previo se construyó una malla de parámetros y se entrenaron los diferentes modelos buscando la mejor combinación de parámetros, es decir, aquellos que ofrecieran un resultado aceptable. Los parámetros configurados para las redes neuronales artificiales se observan en la TABLA XVI.

TABLA XVI
MALLA DE HÍPER PARÁMETROS CONFIGURADOS EN LAS REDES NEURONALES ARTIFICIALES

| | Parámetro | Valores |
|---|---|---|
| 1 | Tasa de aprendizaje | [ 0.001, 0.005, 0.01, 0.05, 0.1] |
| 2 | Tamaño del *batch* | [10, 20, 50] |
| 3 | Optimizador | [adam, SGD, RMSprop] |
| 4 | Dropout | [0, 0.1, 0.2, 0.5, 0.8] |

Las diferentes arquitecturas de las RNAs utilizadas en cada uno de los momentos se pueden observar en la TABLA XVII. Y en la TABLA XVIII se puede observar los diferentes parámetros con los que se entrenó cada red neuronal.

TABLA XVII
ARQUITECTURAS PROPUESTAS PARA LAS RNA

| Modelo | Número de capas | Neuronas por capa |
|---|---|---|
| Modelo 0 | 4 ocultas | [150,60,40,20] |
| Modelo 1 | 5 ocultas | [80,66,52,48, 24] |
| Modelo 2 | 5 ocultas<br>2 capas de dropout | [80,66,52,48, 24] |

TABLA XVIII
CONJUNTOS DE PARÁMETROS DE ENTRENAMIENTO DE LAS RNAS

| Parámetro | Modelo 0 | Modelo 1 | Modelo 2 |
|---|---|---|---|
| learning_rate | 0.01 | 0.1 | 0.01 |
| Tamaño del *batch* | 50 | 30 | 30 |
| epochs | 20 | 20 | 20 |
| optimizer | adam | adam | adam |
| Dropout | 0 | 0 | 0.2 |
| loss | binary_crossentropy | binary_crossentropy | binary_crossentropy |

Los conjuntos de métricas resultantes en los entrenamientos se pueden observar en la TABLA XIX

TABLA XIX
RESUMEN DE MÉTRICAS USANDO LA TÉCNICA RNA

| Modelo | Accuracy_Train | Accuracy_Test | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| Modelo 0 | 0.7818 | 0.73963 | 0.695396 | 0.601185 | 0.64486 | 0.78926 |
| Modelo 1 | 0.904653 | 0.8759 | 0.90216 | 0.763229 | 0.82690 | 0.923223 |
| Modelo 2 | 0.92019 | 0.90413 | 0.91050 | 0.83514 | 0.87119 | 0.949695 |

*Descripción del prototipo de software*

Una vez entrenados los modelo y hecha la selección del que tuvo mejor desempeño se pasó a la construcción de un prototipo de software que permitiera hacer uso del modelo previamente entrenado. Dicho prototipo consistió en un aplicativo web que se compone de un formulario a través del cual se ingresan las variables asociadas a los estudiantes, dichas variables corresponde a las mostradas en la TABLA VII las cuales el usuario que haga uso de la herramienta deberá ingresar seleccionando los valores de los diferentes menús desplegables, esto con el fin de que cada variable esté dentro de los rangos posibles; una vez hecha la captura de los datos se pasa a ejecutar una función que hace uso del modelo previamente entrenado y guardado para generar una predicción sobre el estudiante evaluado y la probabilidad de considerarlo un posible desertor, el sistema mostrará un mensaje en pantalla indicando desertor o no desertor. En Fig. 22 y Fig. 23 se observan las partes del formulario que recibirá los datos de un nuevo estudiante y al presionar el botón de predecir muestra el resultado de la predicción.

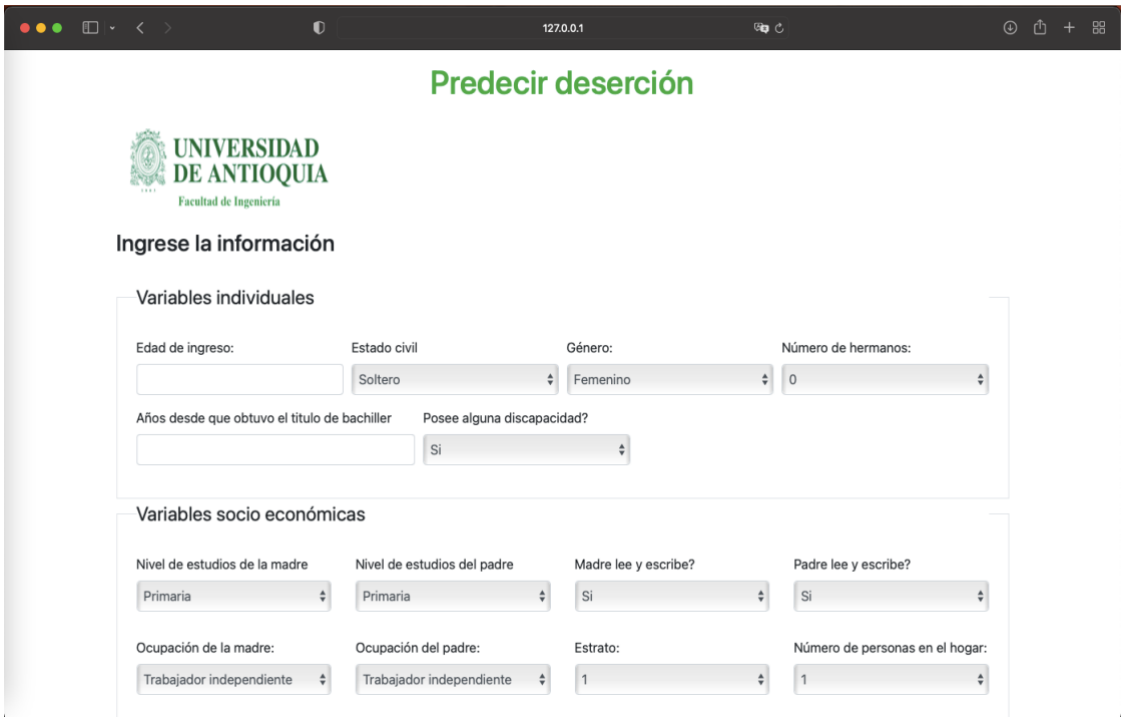



Fig. 22 Formulario del prototipo software parte 1

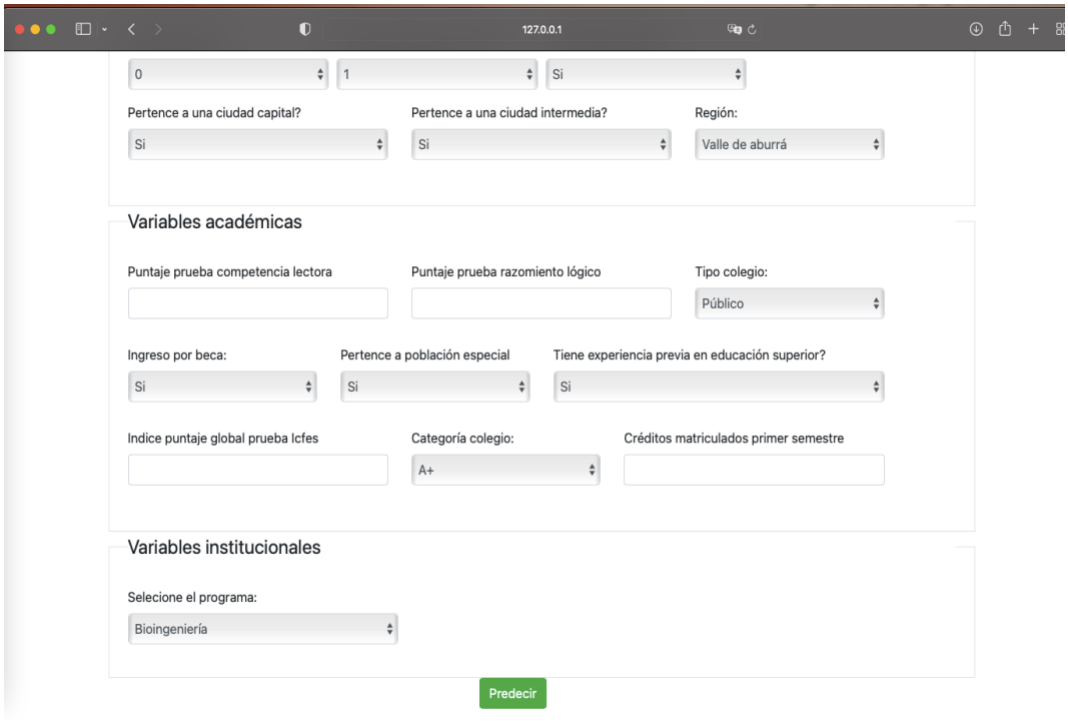



Fig. 23 Formulario del prototipo software parte 2

Inicialmente el prototipo no ofrece la posibilidad de hacer múltiples predicciones, es decir, se tendría que ingresar uno a uno los estudiantes que se desee identificar como posible desertor. Como una mejora se propone continuar desarrollando la herramienta de manera que permita la carga masiva de datos desde un archivo de Excel y retorne un archivo con los estudiantes y su respectiva identificación como desertor o no desertor.

# IX. DISCUSIÓN

De los resultados obtenidos después de entrenar los diferentes modelos presentados, con relación a la métrica *accuracy* (74,91%) se tuvo el mejor desempeño con el XGBoots; aunque el modelo construido con RNA alcanza un *accuracy* muy cercano de 73,96, en ambos casos superar a los árboles de decisión. Analizando las otras métricas, F1- score que mide la capacidad del modelo para identificar los desertores, se observa que para el modelo 0 el mayor F1- score igualmente fue para XGBoots (67,60%). Igualmente, la métrica AUC (81,51%) es superior con XGBoost; lo anterior nos permite concluir que el XGBoots presenta mejor desempeño para identificar la deserción temprana cuando se cuenta con solo los datos de ingreso.

En el caso del modelo 1, si bien en algunos casos las métricas en la RNA son un poco superiores como el caso del precisión y el AUC, se observa que el *accuracy* durante el test y el f1-score fueron mejores en el XGBoost. Bajo las métricas de validación seleccionadas se concluye que en este caso al igual que en el modelo 0, el mejor desempeño logró el XGBoots.

Para todos los casos las métricas del XGBoots fueron superiores que las RNA. Lo anterior conlleva que se concluya que para el conjunto de datos propuesto en este trabajo el mejor desempeño se logró con el XGBoots, pues, si bien los resultados no estuvieron muy alejados entre ambas técnicas, en todos los casos las métricas de XGBoots fueron superiores, sobre todo aquellas que dan cuenta del comportamiento de la técnica con relación a la identificación de la clase "desertor".

Se pudo observar que a medida que el estudiante logra avanzar en los semestres académicos se puede predecir con mayor precisión la posibilidad de desertar. Adicionalmente, se evidencia que después del primer semestre las variables académicas son las que toman mayor capacidad predictiva por lo cual se puede afirmar que durante los primeros semestres el rendimiento académico es determinante a la hora de tomar la decisión de abandonar el proceso educativo, lo cual es consecuente con los estudios realizados en la facultad y presentados en el capítulo de marco teórico, los cuales indican que los estudiantes que desertaron de manera temprana en la facultad lo hicieron en primer lugar por razones académicas y seguidamente por cuestiones económicas. Otro hallazgo que se corrobora es que como se mencionó en otros estudios hechos en la facultad, una de las causa de deserción es la carga de créditos en algunos programas, lo cual se evidencia en las variables de importancia ofrecidas por el XGBoots en donde se observa que cuando se cuenta

solo con los datos de ingreso, la cantidad de créditos matriculados es la variable que tiene más peso en el modelo de clasificación.

Al observar las curvas ROC Fig. 24, Fig. 25 y Fig. 26, obtenido su métrica asociada AUC se puede notar que el desempeño de los modelos tiende a mantenerse estable y tiene un valor aceptable pues en todos los casos es superior al 70% incluso para el modelo 0 que sólo tiene los datos de ingreso lo cual indica que los modelos tienen la capacidad de identificar de forma bastante aceptable los verdaderos casos de deserción en proporción con los casos existentes.

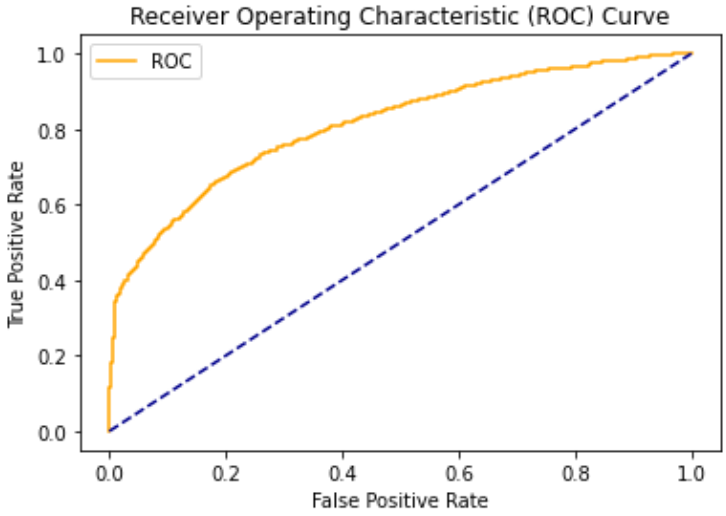



Fig. 24 Curva ROC Modelo 0 - XGBoots

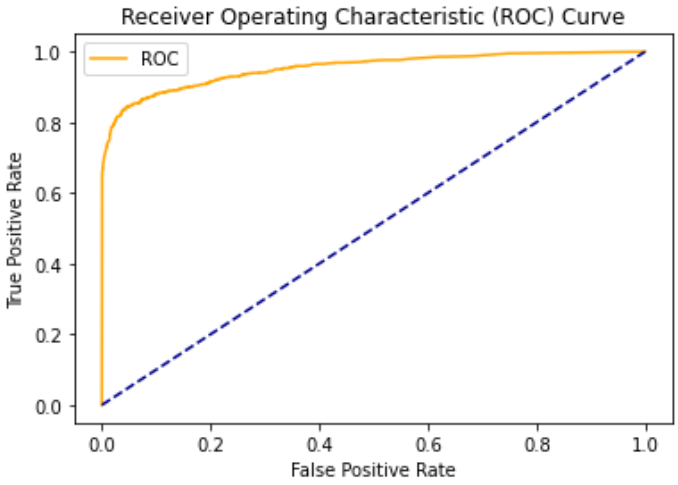



Fig. 25 Curva ROC Modelo 1 - XGBoots

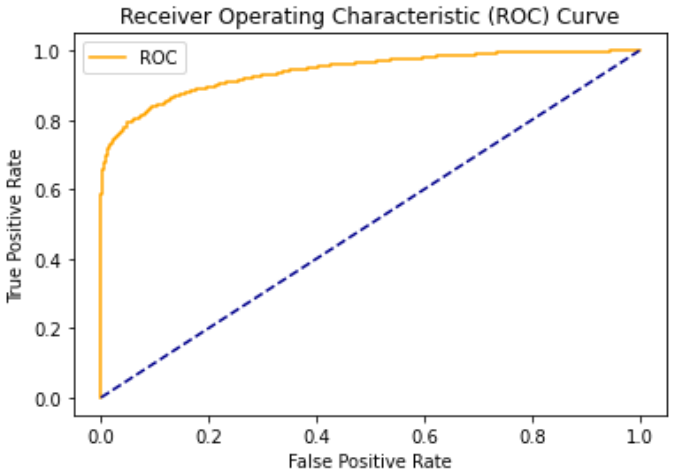



Fig. 26 Curva ROC Modelo 2 - XGBoots

# X CONCLUSIONES Y TRABAJOS FUTUROS

Esta investigación tuvo como objetivo principal usar las técnicas de *learning analytics* para crear un modelo que permita la identificación de riesgo de deserción temprana en los estudiantes de la facultad de ingeniería de la Universidad de Antioquia, usando los datos disponibles en la base de datos institucionales y la información disponible en la plataforma de ICFES. Partiendo del reconocimiento de las variables que explican el fenómeno de la deserción, encontrando que si bien el MEN ha estandarizado un conjunto de variables que determinan la deserción, existen otras variables que según lo establece la teoría son de importancia y se debe considerar su inclusión en los estudios de deserción.

Se identificaron las técnicas del *machine learning* que comúnmente son utilizadas en la tarea de predecir la deserción y se seleccionaron dos técnicas con gran capacidad predictiva, se identificaron los parámetros que consiguieron el mejor desempeño de cada técnica para los datos que se tenía disponibles. Encontrando que con la técnica de XGBoots se pudo lograr un modelo en el momento de ingreso que está en capacidad de clasificar a un estudiante como posible desertor temprano, con un 74,91% de precisión, resultado que se considera aceptable y contrastando con algunos trabajos realizados que utilizan el mismo enfoque de predecir la deserción con solo los datos de ingreso se observa que no lograron una precisión superior al 68%, esto teniendo en cuenta el hecho anteriormente mencionado que no fue posible incluir una gran cantidad de variables que según la teoría son explicativas del fenómeno de la deserción.

Con relación a los resultados obtenidos se evidenció que si bien se consideraron como aceptables, se observa que en cuanto al reconocimiento de la clase de interés (DESERTOR) validando las métricas *recall* y *f1-score* se podría buscar una mejora a fin de que la identificación de los estudiantes en riesgo de desertar sea más precisa, de manera que todas las acciones que desde la facultad y desde bienestar universitario se están llevando a cabo en pro de la permanencia de los estudiante recaigan sobre aquellos que más lo requieran y les puedan brindar el acompañamiento que necesitan para concluir de manera adecuada sus estudios. Tal mejora podría lograrse si fuera posible incluir más variables como por ejemplo aquellas que informen sobre la orientación vocacional puesto que algunos de los estudiantes que ingresan a la universidad desconocen el

programa al cual ingresan o en algunos casos no son admitidos al programa que habían seleccionado en primera opción.

Igualmente se debería incluir variables que den cuenta del ámbito institucional pues para la construcción de este modelo no se lograron incluir debido a que no se contó con su disponibilidad.

Como trabajos futuros se recomienda trabajar con bienestar universitario para obtener información sobre ayudas, acompañamientos y seguimientos ofrecidos a los estudiantes las cuales son variables que hacen parte del ámbito institucional también se recomienda incluir variables sobre el perfil vocacional de los estudiantes que ingresan a la facultad, y conseguir información sobre el apoyo brindado por la familia y los recursos económicos con que cuenta el estudiante para avanzar con su proceso de formación, lo anterior con el objetivo de que se pueda conseguir una mayor cantidad de variables y volver a reentrenar los modelos a fin de buscar una mejora de estos, es de aclarar que dada la naturaleza del fenómeno de la deserción la cual se explica entre otras cosas por el entorno en el que está el estudiante, estos modelos no deben ser inmutables sino que por contrario requieren de una constante actualización de acuerdo a las realidades que rodean al estudiante.

# REFERENCIAS

[1] J. Zárate-Valderrama, N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, N. Bedregal-Alpaca, and V. Cornejo-Aparicio, "Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios," *Ingeniare. Rev. Chil. Ing.*, vol. 29, no. 1, pp. 168–177, Mar. 2021, doi: 10.4067/S0718-33052021000100168.

[2] A. M. Úsuga Ciro, "La deserción estudiantil universitaria: análisis relacional del fenómeno en la Universidad de Antioquia para la cohorte 2009 - I," 2017, Accessed: Aug. 24, 2021. [Online]. Available: http://bibliotecadigital.udea.edu.co/handle/10495/11400.

[3] Facultad de ingeniería, "Plan de acción. Facultad de ingeniería periodo 2020 a 2022." https://www.udea.edu.co/wps/wcm/connect/udea/94fd3b6b-07c0-468e-8ff1-00f0e0deaee5/Documento+Plan+de+Acción+Facultad+de+Ingeniería+2020-2022.pdf?MOD=AJPERES&CVID=nsXw4g0 (accessed Nov. 18, 2021).

[4] MEN (Ministerio de Educación Nacional), "REPORTE SOBRE DESERCIÓN Y GRADUACIÓN EN EDUCACIÓN SUPERIOR AÑO 2016, 2017".https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles - 357549_recurso_5.pdf (accessed Aug. 24, 2021).

[5] S. Cortés-Cáceres, P. Álvarez, M. Llanos, and L. Castillo, "Deserción universitaria: La epidemia que aqueja a los sistemas de educación superior The epidemic that afflicts higher education systems," *REV. Perspect.*, vol. 20, no. 1, pp. 13–25, 2019, doi: 10.33198/rp.v20i1.00017.

[6] A. Valencia Giraldo, L. Mejía Vélez, C. Parra Mesa, E. Castañeda Gómez, R. Mendoza Herrera, and G. Restrepo González, "Vida académica en ingeniería: observar para decidir," *Ing. y Soc.*, vol. 2, no. 10, p. Completo, Oct. 2017, [Online].Available: https://revistas.udea.edu.co/index.php/ingeso/article/view/24677.

[7] C. Guzmán Ruiz *et al.*, "Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos de

prevención," Bogotá- Colombia, 2009. Accessed: Aug. 27, 2021. [Online]. Available: www.mineducacion.gov.co.

[8] Grupo de investigación Ingeniería y Sociedad, "Rendimiento académico de los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia: cohorte 2012-2," *Ing. y Soc.*, pp. 2–10, Sep. 2013, Accessed: Aug. 25, 2021. [Online]. Available: https://revistas.udea.edu.co/index.php/ingeso/article/view/16537.

[9] E. C. Gómez, "Rendimiento académico de los estudiantes en el primer semestre: Facultad de Ingeniería cohortes 2016-1 y 2015-1," *Ing. y Soc.*, pp. 27–33, 2016, Accessed: Aug. 24, 2021. [Online]. Available: https://revistas.udea.edu.co/index.php/ingeso/article/view/327005.

[10] D. K. Mah, "Learning Analytics and Digital Badges: Potential Impact on Student Retention in Higher Education," *Technol. Knowl. Learn.*, vol. 21, no. 3, pp. 285–305, Jun. 2016, doi: 10.1007/s10758-016-9286-8.

[11] V. D. Vera Gil, "Learning Analytics and Scholar Dropout: A Predictive Model Learning Analytics and Scholar Dropout : A Predictive Model," vol. 25, no. January, pp. 1414–1419, 2018, doi: 10.5829/idosi.mejsr.2017.1414.1419.

[12] Universidad de Antioquia, "Sistemas de información institucionales." https://www.udea.edu.co/wps/portal/udea/web/inicio/somos-udea/empleados/informatica-telecomunicaciones/contenido/asmenulateral/sistemas-informacion-institucionales (accessed Aug. 25, 2021).

[13] Universidad de Antioquia, "Plan de acción institucional 2018- 2021," Medellín, Colombia, 2019. [Online]. Available: https://www.udea.edu.co/wps/portal/udea/web/inicio/institucional/direccionamiento- estrategico/plan-accion-institucional.

[14] I. Velasco Quintero, "ANÁLISIS DE LAS CAUSAS DE DESERCIÓN UNIVERSITARIA," Universidad Nacional Abierta y a Distancia UNAD, 2016.

[15] P. I. Giovagnoli, "Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración," *Doc. Trab.*, vol. no. 37, 2002,

Accessed: Aug. 24, 2021. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/3436.

[16] V. Tinto, "Completing College: Rethinking Institutional Action.," *Univ. Chicago Press*, p. 228, Apr. 2012.

[17] J. Areth, E. Jaime, C.-M. Henry, and R. Granobles, "La educación virtual en Colombia: exposición de modelos de deserción.," *Rev. Innovación Educ.*, vol. 7, no. 1, p. 10, 2015.

[18] E. E. Rocío Balmori Méndez, M. DE Teresa La Garza Carranza, and E. Reyes Varela, "EL MODELO DE DESERCIÓN DE TINTO COMO BASE PARA LA PLANEACIÓN INSTITUCIONAL: EL CASO DE DOS INSTITUCIONES DE EDUCACIÓN SUPERIOR

TECNOLÓGICA," Accessed: Nov. 23, 2021. [Online]. Available: www.dgest.gob.mx.

[19] N. E. M. Agudelo and P. J. R. Angulo, "Motivos de deserción estudiantil en programas virtuales de posgrado: revisión de caso y consideraciones desde el mercadeo educativo y el mercadeo relacional para los programas de retención.," *RED. Rev. Educ. a Distancia*, vol., no. 45, pp. 1–23, 2015, Accessed: Aug. 24, 2021. [Online]. Available: https://www.redalyc.org/articulo.oa?id=54738735006.

[20] A. Carvajal, "Factores Cualitativos Que Inciden En La Deserción De La Educación Superior," *Congr. CLABES*, 2014, Accessed: Nov. 28, 2021. [Online]. Available: https://revistas.utp.ac.pa/index.php/clabes/article/view/1050.

[21] C. Atuesta, D. Catherine, D. Gamba, and M. Gisela, "CAUSAS ASOCIADAS A LA DESERCIÓN ESTUDIANTIL Y ESTRATEGIAS DE ACOMPAÑAMIENTO PARA LA PERMANENCIA ESTUDIANTIL," *Conf. Latinoam. sobre el Abandon. en la Educ. Super. CLABES*, 2014.

[22] N. Lázaro Alvarez, Z. Callejas, D. Griol, and M. Durán Benejam, "La deserción estudiantil en educación superior: S.O.S. en carreras de ingeniería informática," *Conf. Latinoam. sobre el Abandon. en la Educ. Super. CLABES*, 2017, [Online].

Available:
http://www.revistas.utp.ac.pa/index.php/clabes/article/view/1674/2410.

[23] V. Barona and C. Fernando, “Experiencias formativas service learning: modelo pedagógico para promover aprendizaje activo y contextualizado desde primer año,” *Congr. CLABES*, 2015, Accessed: Nov. 28, 2021. [Online]. Available: https://revistas.utp.ac.pa/index.php/clabes/article/view/1187.

[24] S. Agarwal, “Data mining: Data mining concepts and techniques,” *Proc. - 2013 Int. Conf. Mach. Intell. Res. Adv. ICMIRA 2013*, pp. 203–207, Oct. 2014, doi: 10.1109/ICMIRA.2013.45.

[25] K. S. Deepashri and A. Kamath, “Survey on Techniques of Data Mining and its Applications,” *Int. J. Emerg. Res. Manag. Technol.*, vol. 9359, no. 2, pp. 198–201, 2017.

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

[27] H. L. Dos Santos, C. Cechinel, J. B. C. Nunes, and X. Ochoa, “An initial review of learning analytics in Latin america,” *12th Lat. Am. Conf. Learn. Objects Technol. LACLO 2017*, vol. 2017-January, pp. 1–9, Nov. 2017, doi: 10.1109/LACLO.2017.8120913.

[28] Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12 (Vol. Part F1305, p. 252). New York, New York, USA: ACM Press. https://doi.org/10.1145/2330601.2330661

[29] Education: A review of UK and international practice Full report,” 2016.

[30] P. R. Castro, “Learning Analytics: una revisión de la literatura,” *Educ. y Educ.*, vol. 20, no. 1, pp. 106–128, Feb. 2017, Accessed: Aug. 25, 2021. [Online]. Available:
https://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/6412.

[31] E. M. Queiroga *et al.*, “A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course,” *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113998.

[32] K. Kang and S. Wang, “Analyze and predict student dropout from online programs,” *ACM Int. Conf. Proceeding Ser.*, pp. 6–12, Mar. 2018, doi: 10.1145/3193077.3193090.

[33] V. Gil Vera, “Learning Analytics and Scholar Dropout: A Predictive Model,” *Middle East J. Sci. Res.*, vol. 25, no. November 2017, pp. 1414–1419, 2017, doi: DOI: 10.5829/idosi.mejsr.2017.1414.1419.

[34] J. Lima, P. Alves, M. Pereira, and Almeida Simone, “Using Academic Analytics to predict dropout risk in engineering courses,” 2018. Accessed: Nov. 19, 2021. [Online]. Available: https://books.google.es/books?hl=es&lr=&id=Jox5DwAAQBAJ&oi=fnd&pg=PA316&dq=Using+Academic+Analytics+to+Predict+Dropout+Risk+in+Engineering+Courses. +In+E CEL+2018+17th+European+Conference+on+e-Learning+(p.+316).&ots=hc6hbqyYW6&sig=s1YZaj5rQcavpy2iUBPfFr-Z3lU#v=onepage&q&f=false.

[35] M. Albán and D. Mauricio, “Decision Trees for the Early Identification of University Students at Risk of Desertion,” *Int. J. Eng. Technol.*, vol. 7, no. 4.44, pp. 51–54, Dec. 2018, doi: 10.14419/ijet.v7i4.44.26862.

[36] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, “Predicting Student Dropout in Higher Education,” June. 2016, Accessed: Aug. 24, 2021. [Online]. Available: https://arxiv.org/abs/1606.06364v4.

*[37]* B. Perez, C. Castellanos, and D. Correal, “Applying Data Mining Techniques to Predict Student Dropout: A Case Study,” *2018 IEEE 1st Colomb. Conf. Appl. Comput. Intell.*

*ColCACI 2018 - Proc.*, Oct. 2018, doi: 10.1109/COLCACI.2018.8484847.

[38] N. L. Alvarez, Z. Callejas, and D. Griol, “Predicting Computer Engineering Students’ Dropout in Cuban Higher Education With Pre-Enrollment and Early Performance Data," J. *Technol. Sci. Educ.*, vol. 10, no. 2, pp. 241–258, 2020, doi: 10.3926/jotse.922.

[39] I. E. Isphording and T. Raabe, "Early Identification of College Dropouts Using Machine- Learning. Conceptual Considerations and an Empirical Example," 2019. doi: 10.5157/NEPS:SC5:11.0.0.

[40] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jul. 2020, vol. 12163 LNAI, pp. 129– 140, doi: 10.1007/978-3-030-52237-7_11.

[41] B. Perez, C. Castellanos, and D. Correal, "Applying Data Mining Techniques to Predict Student Dropout: A Case Study," 2018 IEEE 1st Colomb. Conf. Appl. Comput. Intell. ColCACI 2018 - Proc., Oct. 2018, doi: 10.1109/COLCACI.2018.8484847.

[42] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," Jun. 2016, Accessed: Aug. 24, 2021. [Online]. Available: https://arxiv.org/abs/1606.06364v4.

[43] C. Galafassi, F. F. P. Galafassi, and R. M. Vicari, "Predictive Teaching and Learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10423 LNAI, pp. 549–560, Sep. 2017, doi: 10.1007/978-3-319-65340-2_45.

[44] N. L. Alvarez, Z. Callejas, and D. Griol, "Predicting Computer Engineering Students' Dropout In Cuban Higher Education With Pre-Enrollment and Early Performance Data," *J. Technol. Sci. Educ.*, vol. 10, no. 2, pp. 241–258, 2020, doi: 10.3926/jotse.922.

[45] I. E. Isphording and T. Raabe, "Early Identification of College Dropouts Using Machine-Learning. Conceptual Considerations and an Empirical Example," 2019. doi: 10.5157/NEPS:SC5:11.0.0.

[46] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev. 2018 521*, vol. 52, no. 1, pp. 381–407, Feb. 2018, doi: 10.1007/S10462-018-9620-8.

[47] ICFES, "Documentación del examen Saber 11." https://www.icfes.gov.co/documents/20143/1885630/1.+Documentacion_Saber11.pdf/e72d7e45-7b05-fbee-aed7-c0dfafa25e2f?t=1590543922537 (accessed Aug. 25, 2021).

[48] ICFES, “Sistema Nacional de Evaluación Estandarizada de la Educación. Alineación del examen SABER 11°,” 2013. Accessed: Aug. 26, 2021. [Online]. Available: www.icfes.gov.co.

[49] MEN (Ministerio de Educacion Nacional), “¿Qué es el SPADIES? - Sistemas información.” https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/254648:Que-es-el-SPADIES (accessed Aug. 26, 2021).

[50] W. Acero R., J. F. Sánchez, D. Suárez, and C. F. Téllez, “Modelo de recalificación para la prueba Saber 11,” *Comun. en Estadística*, vol. 9, no. 1, p. 45, 2016, doi: 10.15332/s2027-3355.2016.0001.02.

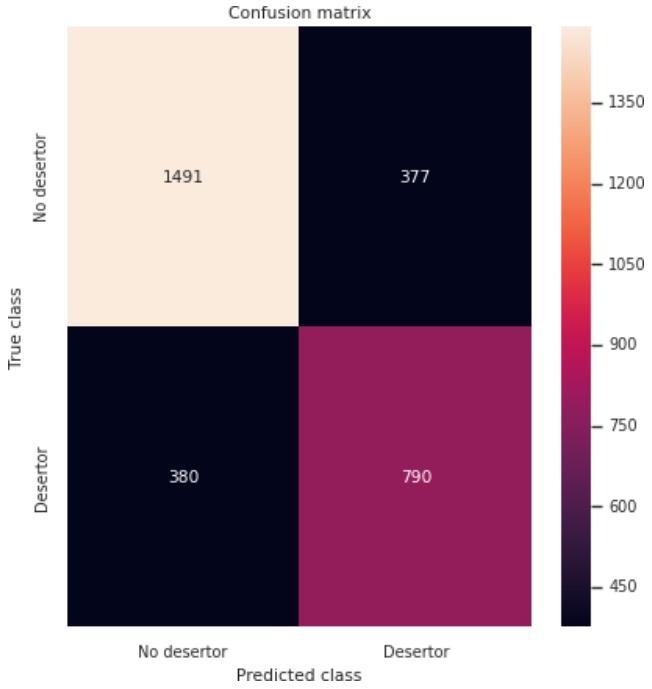
## Anexos

MODELOS ENTRENADOS

1. XGboots modelo 0

```
xgb5 = XGBClassifier(
 learning_rate =0.1,
 n_estimators=100,
 max_depth=3,
 min_child_weight=5,
 gamma=0.1,
 subsample=0.8,
 colsample_bytree=0.8,
 reg_alpha=1,
 objective= 'binary:logistic',
 nthread=4,
 scale_pos_weight=1.55,
 seed=27)
alg, dtrain_predictions,dtrain_predprob, cv_result=modelfit(xgb5, X_train, y_train)
result(alg,y_train,dtrain_predictions,dtrain_predprob,X_test,y_test)
```

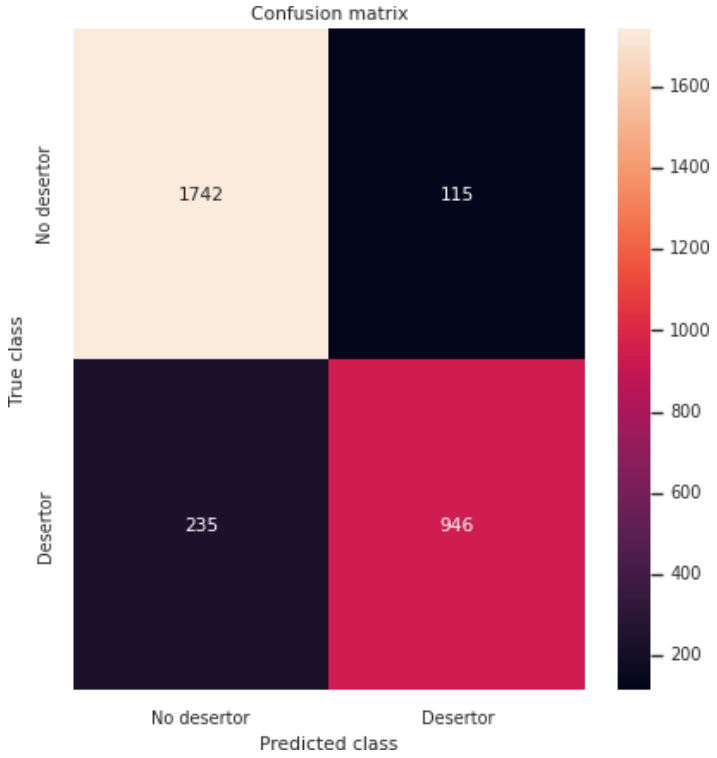
## 1.1. Matriz de confusión



2. XGboots modelo 1

```
xgb4 = XGBClassifier(
 learning_rate =0.1,
 n_estimators=100,
 max_depth=3,
 min_child_weight=2,
 gamma=0.4,
 subsample=0.8,
 colsample_bytree=0.8,
 reg_alpha=0.001,
 objective= 'binary:logistic',
 nthread=-1,
 scale_pos_weight=1.55,
 seed=50)
```

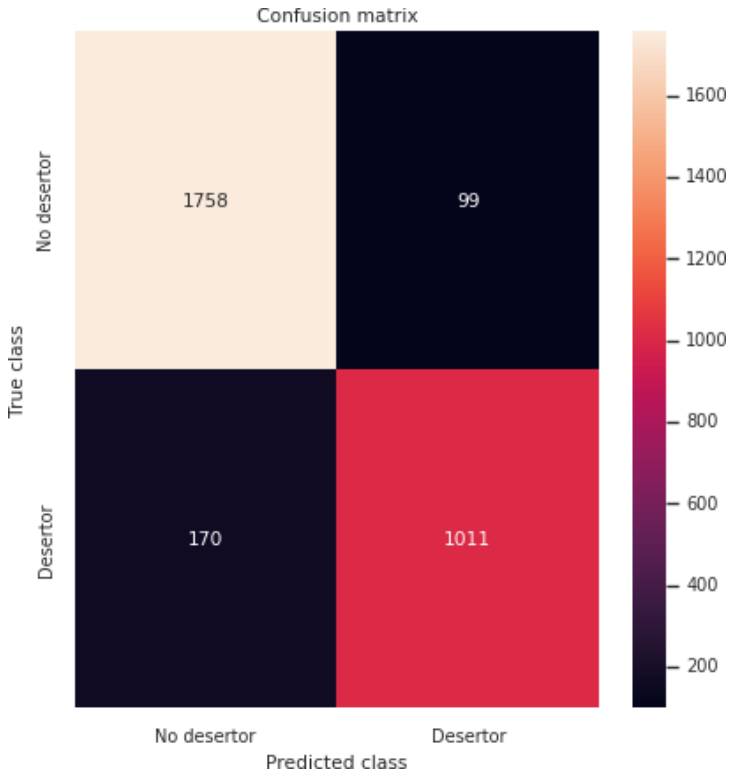
2.1 Matriz de confusión



3. XGboots modelo 2

```
xgb4 = XGBClassifier(
 learning_rate =0.001,
 n_estimators=100,
 max_depth=3,
 min_child_weight=1,
 gamma=0.4,
 subsample=0.85,
 colsample_bytree=0.8,
 reg_alpha=0.001,
 objective= 'binary:logistic',
 nthread=4,
 scale_pos_weight=1.55,
 seed=27)
```

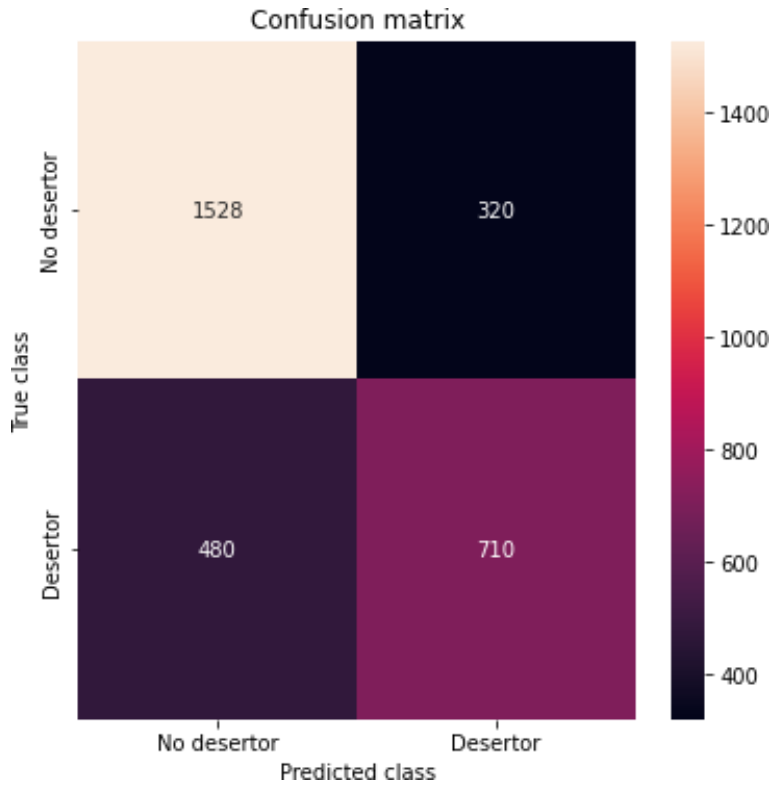
3.1 Matriz de confusión



## 4. RNA modelo 0

```
[ ]  1 # define model
     2 model = Sequential()
     3 model.add(Dense(150, input_dim=102, activation='relu'))
     4 model.add(Dense(60, activation='relu'))
     5 model.add(Dense(40, activation='relu'))
     6 model.add(Dense(20, activation='relu'))
     7 model.add(Dense(1, activation='sigmoid'))
     8
```

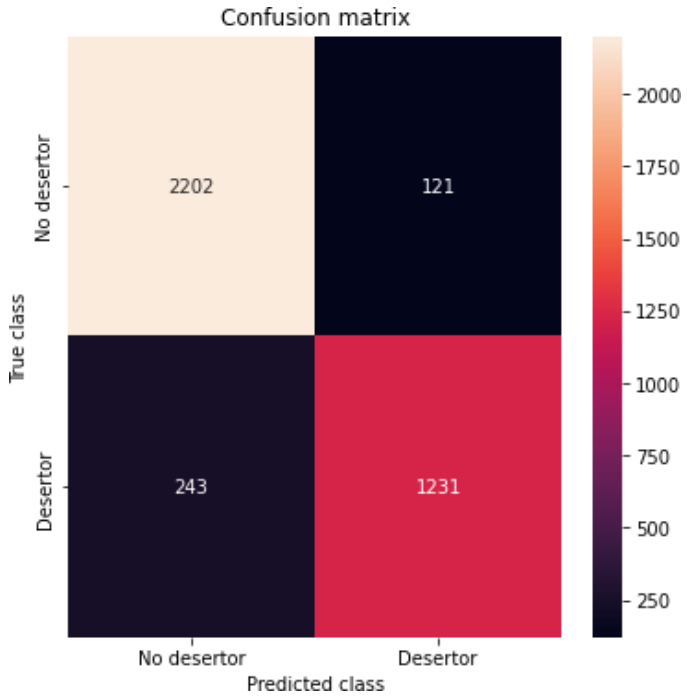
### 4.1 Matriz de confusión



5. RNA modelo 1

```python
def create_larger(optimizer):
  # create model
 model = Sequential()
 model.add(Dense(80, input_dim=107, activation='relu'))
 model.add(Dense(66, activation='relu'))
 model.add(Dense(52, activation='relu'))
 modelC.add(Dropout(rate = 0.2))
 model.add(Dense(48, activation='relu'))
 modelC.add(Dropout(rate = 0.2))
 model.add(Dense(28, activation='relu'))
 model.add(Dense(1, activation='sigmoid'))
  # Compile model
 model.compile(optimizer=optimizer,loss='binary_crossentropy',metrics=['accuracy'])
 return model
```

5.1 Matriz de confusión



## 6. RNA modelo 2

```
modelC = Sequential()
modelC.add(Dense(80, input_dim=113, activation='relu'))
modelC.add(Dense(66, activation='relu'))
modelC.add(Dropout(rate = 0.2))
modelC.add(Dense(48, activation='relu'))
modelC.add(Dense(28, activation='relu'))
modelC.add(Dropout(rate = 0.2))
modelC.add(Dense(1, activation='sigmoid'))
```

### 6.1 Matriz de confusión