MODELO PREDICTIVO PARA LA DESERCIÓN.

Un modelo de minería de datos para la predicción de la deserción estudiantil en la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander en el periodo 2015-2019.

Juliana Hernández Mosquera

Trabajo de Grado para Optar al título de Ingeniera Industrial

Director

Henry Lamos Díaz

Ph.D en Física - Matemática

Tutor externo

Jorge Esteban Caballero Rodríguez

Ingeniero Industrial – Científico de datos

Universidad Industrial de Santander

Facultad de Ingeniería Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Ingeniería Industrial

Bucaramanga

2024

Dedicatoria

Dedico con todo mi corazón mi tesis a mi novio, por acompañarme a lo largo de los años con tanto amor, sin él no hubiera sido posible este logro; por sus consejos, su paciencia, su comprensión, el ánimo recibido de su parte en los momentos más difíciles y por el maravilloso ser humano que es.

A Mamba por ser el motor de mi vida, el cual me impulsa a seguir adelante día a día, así no cuente con ánimos, por su nobleza y su compañía en donde quiera que me encuentre.

A mi tutor de tesis, por su compromiso con este proyecto, por su disponibilidad y su infinidad de habilidades humanas e intelectuales que hicieron de mí una mejor persona y profesional.

A mi madre y a mi padre por toda la ayuda y cariño recibido durante lo largo de mi vida, son personas increíbles, trabajadoras, berracas y me siento muy afortunada de tenerlos como padres.

A mi familia por siempre apoyarme y quererme a pesar de las diferencias, en especial a mi tía paterna Rosalba que con su forma de ser me ha inspirado a ser una mujer independiente y fuerte.

A mis primas Mari² a las que quiero mucho, atesoro los momentos compartidos y la confianza que hemos construido a lo largo de los años. Porque, aunque la familia no se elija en otra vida quisiera compartir la misma sangre que ustedes.

Agradecimientos

En primer lugar, agradezco profundamente al destino por poner en mi camino a Edin Roberto Canahui González que con su apoyo incondicional durante estos años me ha brindado estabilidad en todos los ámbitos de mi vida y por lo cual he podido cumplir mis objetivos académicos, deportivos y sociales.

Agradezco a mis padres Luz Marina Mosquera y Exelino Hernández González por su apoyo incondicional y por ser tan maravillosos, ellos con su cariño y confianza han labrado mi vida a través del tiempo y me han hecho ser quién soy hoy en día.

Agradezco a mi tutor Jorge Esteban Caballero Rodríguez porque fue la guía más precisa a través de este camino, es un ser humano sumamente valioso en todo aspecto; gracias por cada reunión y cada explicación que contribuyo para conseguir este logro.

Agradezco a mi director de proyecto por la confianza, por brindarme el recurso más valioso que es el tiempo y por su guía precisa en momentos cruciales.

Tabla de contenido

Lista de tablas

Pág.

Lista de figuras

Pág.

Lista de apéndices

Ver apéndices adjuntos y pueden ser consultados en la base de datos de la Biblioteca UIS

Apéndice A. Presupuesto del proyecto.

Apéndice B. Matrícula actual.

Apéndice C. Base de datos Admisiones sin modificaciones.

Apéndice D. Base de datos modificada.

Apéndice E. Código modelo de árbol de decisión.

Apéndice F. Árbol de decisión completo.

Apéndice G. Código modelo de Random Forest.

Apéndice H. Código modelo de Regresión Logística.

Apéndice I. Tabla de contingencia Condición-Colegio.

Apéndice J. Tabla de contingencia Condición-Ocio.

Apéndice K. Tablas de contingencia y coeficientes V Cramer.

Apéndice L. Análisis multivariable Condición.

Apéndice M. Artículo científico con los resultados obtenidos.

MODELO PREDICTIVO PARA LA DESERCIÓN.

Resumen

Título: Un modelo de minería de datos para la predicción de la deserción estudiantil en la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander en el periodo 2015-2019.*

Autor: Juliana Hernández Mosquera**

Palabras Clave: Deserción estudiantil, educación superior, arboles de decisión, estudiantes, minería de datos, modelo predictivo.

Descripción:

La presente investigación tiene como objetivo desarrollar un modelo de minería de datos para predecir la deserción estudiantil de la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander durante el periodo comprendido entre 2015 y 2019. Para lograr este objetivo, se realizó una revisión bibliográfica de la literatura existente sobre la deserción estudiantil en la educación superior, identificando las variables socioeconómicas y demográficas que afectan este fenómeno. Posteriormente, se solicitó a la Dirección de Admisiones y Registro Académico de la universidad los datos de las características de los estudiantes de la facultad, que se analizaron y procesaron utilizando técnicas de ciencia de datos y software especializado mediante distintos lenguajes de programación como Python y R. Se definieron varios modelos predictivos de deserción y se seleccionó el que mejor se ajustó a los datos para desarrollar un modelo predictivo específico para la Facultad de Ingeniería Fisicomecánicas. Los resultados obtenidos se socializarán a través de un artículo académico y el libro final del proyecto de grado, con el objetivo de mejorar la retención estudiantil dentro de la universidad. Este proyecto contribuirá al mejoramiento de las políticas académicas y a la implementación de medidas preventivas que contribuyan a reducir la deserción estudiantil en la facultad.

---

* Trabajo de Grado

** Facultad de Ingeniería Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: Ph. D. Henry Lamos Diaz. Tutor: Ingeniero Industrial Jorge Esteban Caballero Rodríguez.

MODELO PREDICTIVO PARA LA DESERCIÓN.

Abstract

Title: A data mining model for the prediction of student desertion in the Faculty of Physical-Mechanical Engineering of the Universidad Industrial de Santander in the period 2015-2019.*

Author: Juliana Hernández Mosquera**

Key Words: Student desertion, higher education, decision trees, students, data mining, predictive model.

Description

The aim of this research is to develop a data mining model to predict student dropout in the Faculty of Physicomechanical Engineering at the Universidad Industrial de Santander during the period between 2015 and 2019. To achieve this goal, a literature review of existing research on student dropout in higher education was conducted, identifying the socio-economic and demographic variables that affect this phenomenon. Subsequently, the university's Admissions and Academic Registry Department was requested to provide data on the characteristics of students in the faculty, which were analyzed and processed using data science techniques and specialized software using different programming languages such as Python and R. Several predictive models of dropout were defined, and the one that best fit the data was selected to develop a specific predictive model for the Faculty of Physicomechanical Engineering. The results obtained will be disseminated through an academic article and the final project book, with the aim of improving student retention within the university. This project will contribute to the improvement of academic policies and the implementation of preventive measures that will help to reduce student dropout in the faculty.

---

* Bachelor Thesis

**Faculty of Physical-Mechanical Engineering. School of Industrial and Business Studies. Director: Ph. D. Henry Lamos Diaz. Tutor: Industrial Engineer Jorge Esteban Caballero Rodríguez.

Introducción

““Aprender es como remar contra corriente: en cuanto se deja, se retrocede.”. – Edward Benjamin Britten.

La deserción en la educación superior es un fenómeno que afecta directamente el desarrollo social, económico, cultural y político de un país; a nivel mundial los investigadores intentan identificar las causas de la deserción, aunque se ha llegado a hallazgos importantes para las instituciones de educación superior, aún no existe una explicación que pueda aplicarse a todos los casos y en ocasiones las investigaciones son inconclusas, esto se debe en parte a la diversidad cultural y de condiciones de cada región.

La educación superior es un privilegio al que muy pocos colombianos tienen acceso, aun así, existen altos índices de abandono a los estudios técnicos, tecnológicos y profesionales; el gobierno nacional con el fin de comprender este fenómeno ha creado instituciones y programas especializados, como SPADIES, ICETEX, entre otras; para dar apoyo a los estudiantes, sin embargo, los esfuerzos del estado no han alcanzado los resultados esperados.

El intento de encontrar una solución a este problema en la Universidad Industrial de Santander impulsa a realizar la presente investigación; la deserción estudiantil en Santander puede causar efectos negativos en la calidad de vida de los santandereanos, encontrando baja mano de obra calificada, aumento en la delincuencia e incluso aumento de enfermedades mentales tales como la depresión.

Esta investigación propone encontrar un modelo predictivo para la detección temprana del riesgo de abandono escolar en estudiantes de la Universidad Industrial de Santander (UIS), con el fin de disminuir las brechas de desigualdad e inequidad causadas por la no culminación de los

estudios de pregrado; trabajando con información sobre las características socioeconómicas y sociodemográficas de estudiantes de la Facultad de Ingeniería Fisicomecánicas de la UIS en el rango de tiempo de 2015 a 2019; la información reposa en las bases de datos de la Dirección de Admisiones y Registro académico, considerada una fuente de información secundaria.

Inicialmente se realiza una revisión literaria dentro de las bases de datos de la Biblioteca de la UIS, para encontrar referentes dentro del tema investigativo; seguido se hace el análisis exploratorio de la información para organizar, limpiar, depurar y asignar variables a los datos correspondientes; así mismo encontrar posibles correlaciones entre las variables, que sirvan de punto de inicio en la elección del software a utilizar y el método.

El problema de investigación es un problema de clasificación, se intenta dar una respuesta binaria sobre la posible deserción de un estudiante; después del procesamiento de los datos se prueban modelos predictivos existentes como árbol de decisión, regresión logística, Naive Bayes, entre otros; se usan diferentes métricas de exactitud y precisión para la predicción de las clases y así poder elegir el modelo que se implementará en el presente trabajo. Las estadísticas nos ayudarán a elegir el modelo que tenga mayor precisión con respecto a los demás, para realizar la evaluación de los modelos, estudiando minuciosamente la matriz de confusión, la exactitud, la precisión, la sensibilidad y la puntuación F1; que son las características de confiabilidad de un modelo.

El modelo propuesto servirá para encontrar la relación entre las variables de un estudiante y su estado en la universidad. Se espera encontrar un modelo que se ajuste a los datos y tenga un buen margen de confiabilidad para así recomendar acciones correctivas a la dependencia encargada de la disminución de la deserción estudiantil en la UIS.

Tabla de cumplimiento de objetivos.

Tabla 1

Cumplimiento de objetivos

Objetivo	Cumplimiento
Realizar una revisión de literatura vía web de la deserción estudiantil para encontrar hallazgos relevantes a la investigación por medio de bases de datos de la biblioteca de la Universidad Industrial de Santander.	Capítulo 3.
Identificar las variables socioeconómicas y demográficas que afecten la deserción estudiantil en la educación superior para comprender su naturaleza y clasificación.	Capítulo 3.2.
Obtener los datos de las características de los estudiantes de la Dirección de Admisiones y Registro Académico por medio de la Facultad de Ingeniería Fisicomecánicas.	Capítulo 8.
Realizar un análisis preliminar de los datos para encontrar primeras hipótesis de relación entre las variables.	Capítulo 8.1.2.
Utilizar ciencia de datos para procesar la información de las características estudiantiles mediante software especializado.	Capítulo 9.
Definir modelos predictivos de deserción para aplicarlos a los datos.	Capítulo 9.2.
Elegir el modelo que se ajuste mejor a los datos y sirva como modelo predictivo de la deserción en el caso específico.	Capítulo 9.3.
Socializar los resultados de la investigación para mejorar la retención estudiantil dentro de la Universidad Industrial de Santander mediante	Apéndice M.

un artículo académico de carácter publicable y el libro final del proyecto de grado.	

1. Objetivos

1.1. Objetivo general

Proponer un modelo matemático predictivo para la deserción estudiantil en la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander.

1.2.Objetivos específicos

- ✓ Realizar una revisión de literatura vía web de la deserción estudiantil para encontrar hallazgos relevantes a la investigación por medio de bases de datos de la biblioteca de la Universidad Industrial de Santander.
- ✓ Identificar las variables socioeconómicas y demográficas que afecten la deserción estudiantil en la educación superior para comprender su naturaleza y clasificación.
- ✓ Obtener los datos de las características de los estudiantes de la Dirección de Admisiones y Registro Académico por medio de la Facultad de Ingeniería Fisicomecánicas.
- ✓ Realizar un análisis preliminar de los datos para encontrar primeras hipótesis de relación entre las variables.
- ✓ Utilizar ciencia de datos para procesar la información de las características estudiantiles mediante software especializado.
- ✓ Definir modelos predictivos de deserción para aplicarlos a los datos.
- ✓ Elegir el modelo que se ajuste mejor a los datos y sirva como modelo predictivo de la deserción en el caso específico.
- ✓ Socializar los resultados de la investigación para mejorar la retención estudiantil dentro de la Universidad Industrial de Santander mediante un artículo académico de carácter publicable y el libro final del proyecto de grado.

2. Revisión de la literatura

2.1.Análisis bibliométrico

Para el análisis bibliométrico se elige trabajar con la base de datos Scopus a partir de la ecuación inicial de búsqueda:

(((DESERTION AND HIGHER AND EDUCATION) OR (DROPOUT AND HIGHER AND EDUCATION AND UNDERGRADUATE)))

La ecuación de búsqueda arrojo un total de 343 documentos, con ellos se realizó la síntesis de información, extraída de los resúmenes; para encontrar palabras que se deben incluir o excluir de la ecuación de búsqueda para encontrar los artículos que se relacionan directamente con la investigación.

Para la ecuación de búsqueda se excluyeron las siguientes palabras:

COVID & PANDEMIC: Teniendo en cuenta que el periodo al que pertenecen los datos es de 2015 a 2019; estamos en un tiempo prepandemia, la naturaleza de los datos a utilizar es de normalidad académica y presencialidad.

Y se incluye DATA MINING, dado que el modelo resultante se hará por medio de este método estadístico.

La ecuación de búsqueda definitiva que resulta después del análisis exploratorio de información es la siguiente:

(((DESERTION AND HIGHER AND EDUCATION) OR (DROPOUT AND HIGHER AND EDUCATION AND UNDERGRADUATE) AND (DATA MINING) AND NOT (COVID OR PANDEMIC))

Se obtienen un total de 34 documentos relacionados a la investigación; que al revisar dentro de los filtros se encuentra la siguiente información:

Figura 1

Cantidad de producción científica por país (2022)

Nota: Adaptado de Scopus.

Lo que indica que el tema es relevante para el proyecto de grado, dada la producción científica del país y el continente.

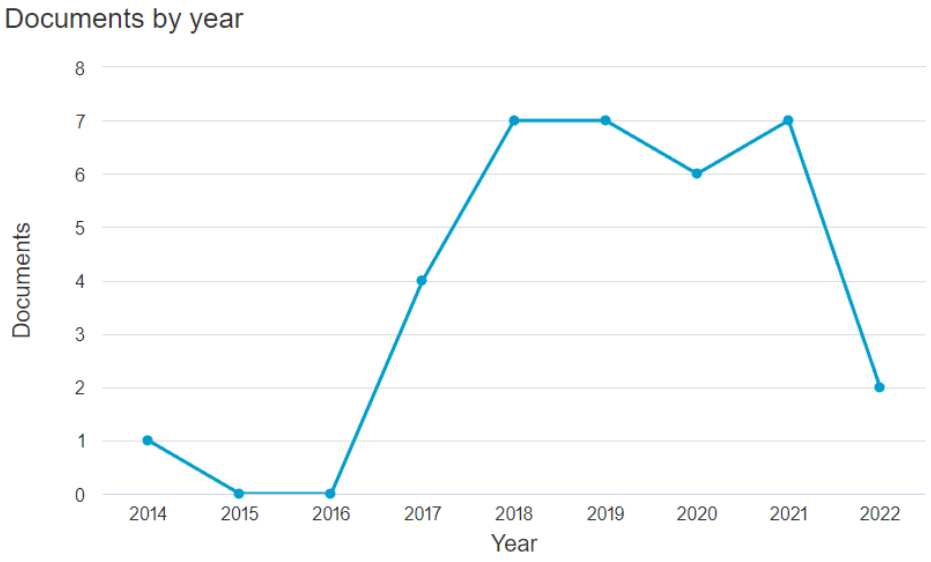
Para obtener el análisis bibliométrico de la revisión de literatura en la base de datos Scopus, se utilizaron las herramientas propias de la base de datos y también el software VOSviewer para complementar el análisis.

2.1.1. Publicaciones por año:

Las publicaciones por año se refieren al nivel de producción científica realizada en un año específico; observando la figura 2, es evidente el aumento en la investigación del tema al pasar los años; la producción científica en diez años ha incrementado uno a siete veces según la figura. Para el año en curso aún existen documentos en espera para publicación e investigaciones en curso.

Figura 2

Cantidad de producción científica por año (2022)



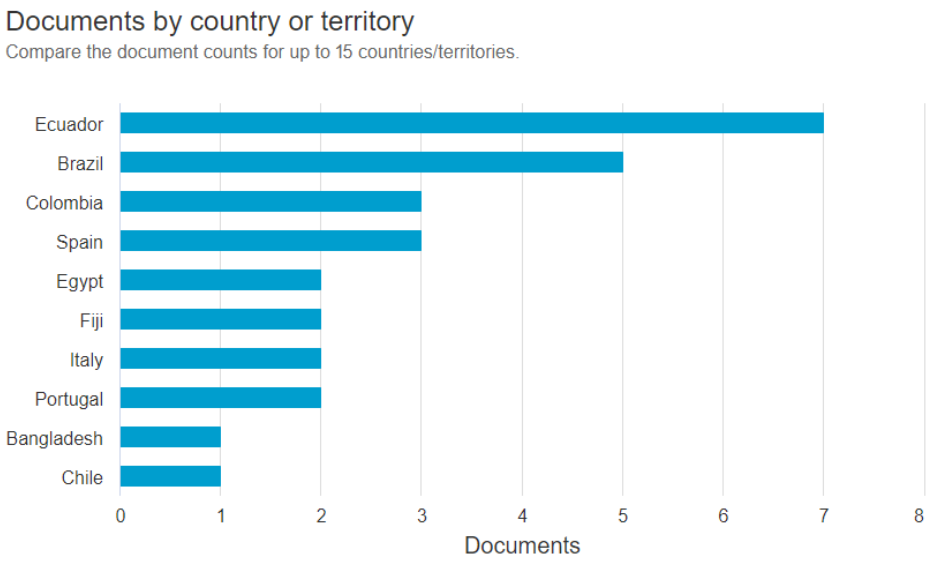
Nota: Adaptado de Scopus.

2.1.2. Países / regiones

El análisis realizado muestra que Ecuador, Brasil y Colombia son los mayores productores de documentos científicos del tema a investigar; con más del 44 % del total de producción científica.

Figura 3

Principales países de producción científica (2022)



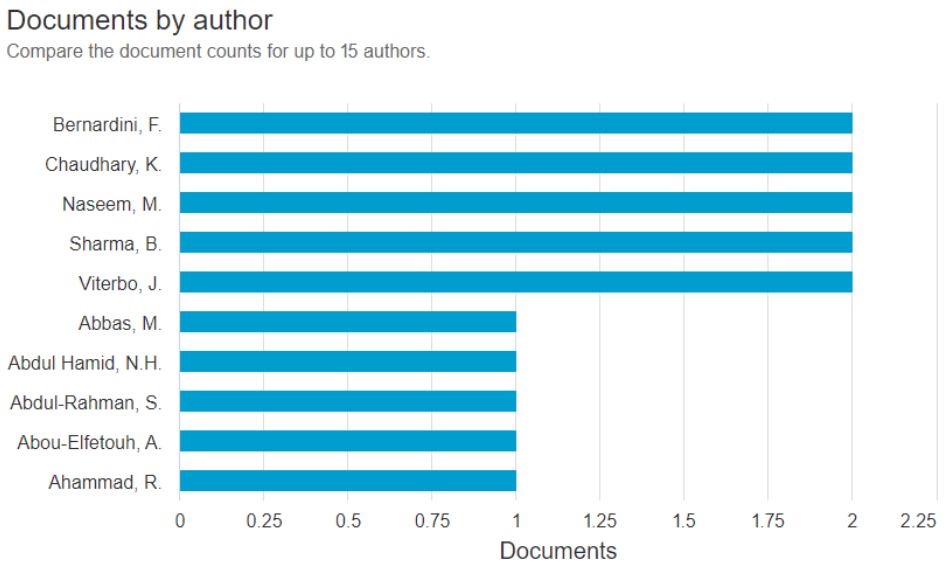
Nota: Adaptado de Scopus.

2.1.3. Autores

Con respecto a los autores se encuentra que la producción científica está distribuida en dos grupos y el máximo de producciones por autor es de 2 documentos.

Figura 4

Principales autores (2022)



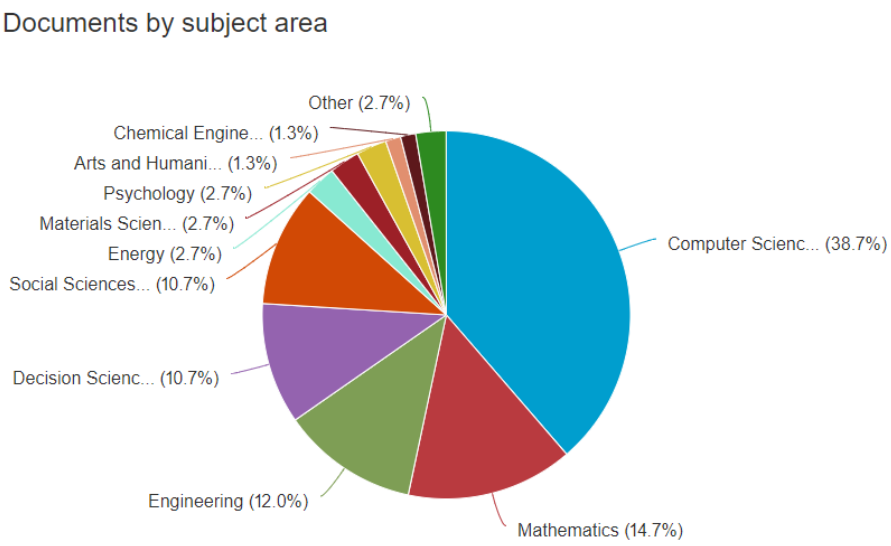
Nota: Adaptado de Scopus.

2.1.4. Áreas de investigación

Las áreas de conocimiento que más han aportado al tema de investigación son la ciencia computacional, las matemáticas y la ingeniería, con un porcentaje de representación del 38,7 %, 14,7 % y 12 % respectivamente. Lo que determina el nivel de relevancia del tema investigativo para la ingeniería industrial, la minería de datos y el análisis de datos.

Figura 5

Producción científica por área de investigación (2022)



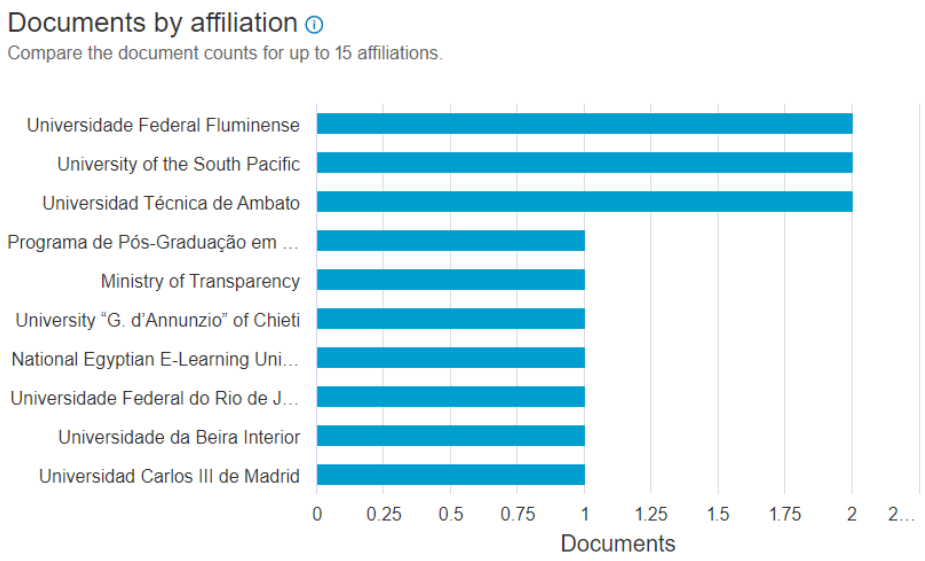
Nota: Adaptado de Scopus.

2.1.5. Instituciones

En el análisis de instituciones generadoras de producción científica se puede encontrar que dentro las 10 principales se encuentran universidades; relacionadas directamente con el tema investigativo.

Figura 6

Instituciones que realizaron la producción científica (2022)



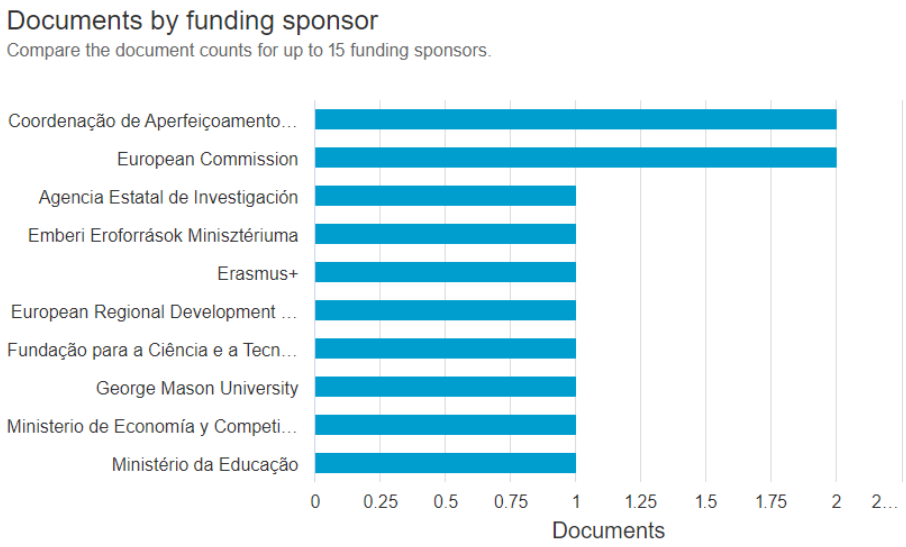
Nota: Adaptado de Scopus.

2.1.6. Documentos por entes financiadores

Las entidades financiadoras de producción científica del tema de investigación en su mayoría son entidades gubernamentales; por ello se deduce que los organismos estatales tienen interés en determinar las causas de la deserción en la educación superior.

Figura 7

Organizaciones financiadoras de investigaciones en el tema (2022)



Nota: Adaptado de Scopus.

2.1.7. Palabras clave

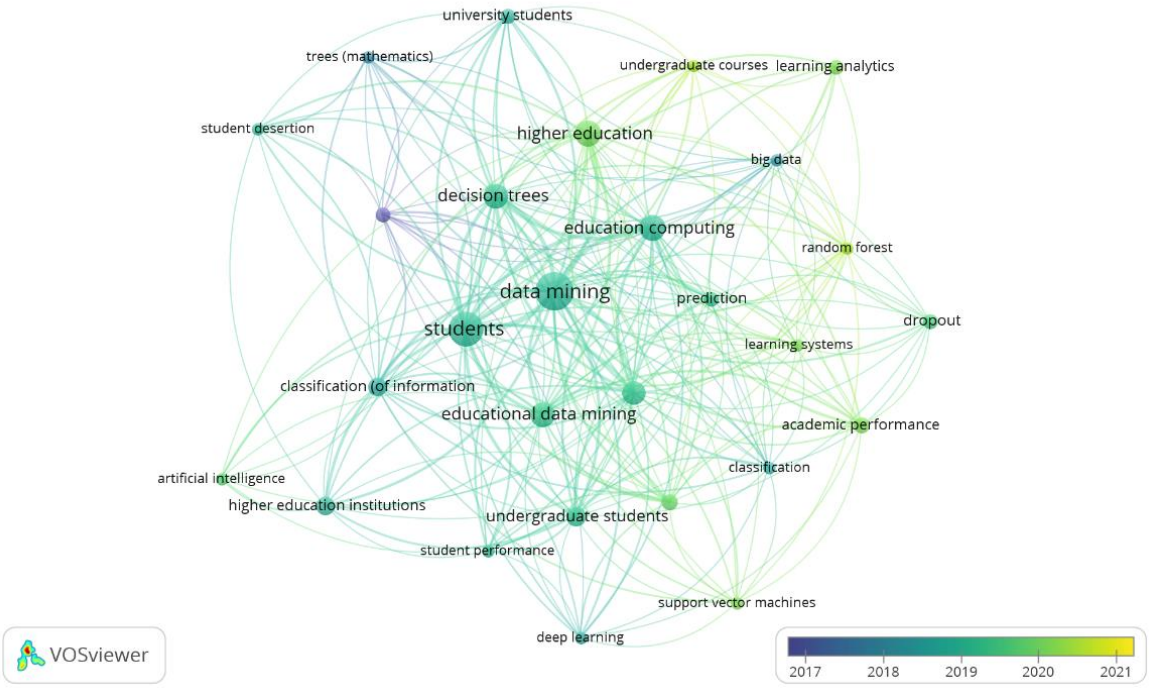
Para el análisis de palabras clave, se utilizó un archivo Thesaurus para agrupar palabras que estaban repetidas por su singular y plural, entre otros casos; por ejemplo, student/students.

El gráfico de palabras clave incluye la relevancia de las palabras según el año de publicación del documento; es importante resaltar que el transcurso del tiempo marca diferencias dentro de la producción científica, vemos que para los últimos años ya las investigaciones tratan de bosques aleatorios, big data, predicción, entre otros.

Esto es positivo y se relaciona directamente con el uso de la tecnología en la educación para apoyar procesos educativos y analizar problemas tales como la deserción estudiantil universitaria.

Figura 8

Palabras clave y su comportamiento en el tiempo (2022)



Nota: Adaptado de VOSviewer.

2.1.8. Autoría y citaciones

A continuación, se mostrará la tabla que contiene el nombre de los autores, el número de documentos que ha producido y las citaciones que le han realizado.

Teniendo a 5 autores con la misma cantidad de documentos científicos pero distinto número de veces en que han sido citado.

Tabla 2

Autores, número de documentos producidos y citaciones

Autor	**Documentos**	**Citaciones**
Bernardini F.	2	11
Chaudhary K.	2	7

Naseem M.	2	7
Sharma B.	2	7
Viterbo J.	2	11

Nota. Elaboración propia, datos extraídos de *VOSviewer*.

2.2.Análisis preliminar de la literatura

Dentro del análisis preliminar de la literatura se encuentran 5 ejes fundamentales dentro del tema de investigación:

1. Definición de deserción.
2. ¿Por qué la deserción es un tema relevante en investigación?
3. Cifras colombianas de la deserción, costos económicos, sociales, entre otros.
4. Determinantes de la deserción.
5. Análisis de datos en la deserción en las instituciones de educación superior.

Según Tinto (1975) la deserción se define como el procedimiento que lleva a cabo un estudiante universitario cuando abandona voluntaria o forzosamente sus estudios, debido a la influencia negativa o positiva de factores internos o externos; esta definición de deserción estudiantil aún tiene vigencia en la actualidad.

Así mismo el autor resalta que la deserción puede estar ligada a las metas individuales del estudiante, las cuales tienen influencia de los procesos sociales e intelectuales que experimente el estudiante a lo largo del desarrollo de su programa académico. Sólo en algunos de los casos las deserciones son causadas por bajo rendimiento académico, puesto que la mayoría son voluntarias. Se ha encontrado que los estudiantes que las desertan pueden tener un mayor rendimiento académico que los estudiantes que continúan con sus estudios; las razones que esgrimen diversos

autores con respecto a esto es que el estudiante que abandona sus estudios lo hace debido a la baja integración personal con los ambientes intelectual y social de la comunidad institucional, que suele darse en etapas tempranas del programa académico (Tinto, 1975).

Por otro lado, el autor estudia la deserción desde el ámbito institucional; contemplando el escenario de las instituciones de educación superior privadas, donde la deserción es una inestabilidad económica en las instituciones puesto que el pago de los estudiantes es la principal fuente de ingresos de estas; con respecto a las IES públicas, la deserción constituye presupuestos insuficientes que pueden entorpecer las actividades misionales de estas (Tinto,1975).

En las primeras investigaciones realizadas sobre el tema; la deserción para los órganos estatales podían ser traslados entre instituciones estatales (universidades públicas) que no representa un abandono sino una transición; cuando los estudiantes dejan sus estudios en una institución de carácter público para migrar a una institución privada o una institución fuera del ámbito estatal, para el gobierno en sí es una deserción (Pascarella, E. T. & Terenzini, P., 1977).

Para concluir la revisión de los predecesores del tema, es significativo resaltar la importancia de estas investigaciones que dieron paso a las primeras hipótesis sobre la deserción en la educación superior y sus posibles causas.

En Colombia la desigualdad afecta la escolaridad desde la niñez, esto se ve reflejado en los índices de pobreza multidimensional donde en zonas urbanas es de 12,3 % y en zonas rurales del 34,5 %; según la Organización para la Cooperación y el Desarrollo Económico (OCDE) sólo el 9 % de las familias más pobres tienen educación superior, mientras que el 53 % son de familias con ingresos altos; la brecha entre el acceso a un colegio privado demuestra la desigualdad en la educación, donde el 61 % de las familias de mayores ingresos puede poner a sus hijos en colegios

privados contra un 2 % de las familias de menores ingresos. En nuestro país únicamente el 30 % de los jóvenes que culminan sus estudios secundarios pueden acceder a la educación superior, además, el 10,4 % en pregrado y 22,2 % en programas técnicos y tecnológicos pueden culminar sus estudios (Barbosa-Camargo, MI *et al*., 2021).

Todo esto conlleva a que los costos sociales y económicos que causa la deserción lleven al gobierno a hacer programas que intenten retener a los estudiantes en las instituciones de educación superior.

Según Barbosa-Camargo *et al*, MI (2021); los determinantes de la deserción se dividen en cinco aspectos:

- ✓ Antecedentes socioeconómicos
- ✓ Características individuales o personales
- ✓ Acceso a sistemas de apoyo
- ✓ Elementos contextuales y geográficos
- ✓ Características de los estudios

Múltiples estudios mostraron que los estudiantes que tienen padres con un mayor nivel educativo tienen alta probabilidad de concluir sus estudios, dado que los padres pueden brindar más apoyo emocional, psicológico y económico a sus hijos; también resalta que los estudiantes de entornos socioeconómicos vulnerables eligen programas de corta duración, como los son programas tecnológicos y técnicos.

Cabe resaltar que el nivel socioeconómico y educativo de los padres influye en el inicio de la etapa universitaria; posteriormente, el tipo de institución y la admisión a su primera opción juegan un papel crucial (Mineducacion, 2014).

MODELO PREDICTIVO PARA LA DESERCIÓN.

Según la literatura, los posibles factores individuales que causan la deserción son: género, nivel de conocimiento en el bachillerato y rendimiento académico (tanto previo como el obtenido durante el programa de educación superior); se encontraron diferencias entre los factores individuales que conllevan a la deserción según el tipo de institución si es pública o privada, para instituciones públicas específicamente el nivel educativo de la escuela secundaria donde culmina los estudios un estudiante es importante; las escuelas secundarias dan las bases pedagógicas a los estudiantes de nuevo ingreso, por ello si los conocimientos y la motivación de un estudiante son bajos, tendrá mayor dificultad para adaptarse al ámbito universitario.

El desempeño académico en el primer año universitario también influye en las motivaciones individuales de los estudiantes, esto hace que persistan en sus estudios o deserten, esto es algo que las universidades detectaron y crearon programas de acompañamiento en los primeros semestres; la Universidad Industrial de Santander específicamente tiene estrategias como el Sistema de Excelencia Académica (SEA) que apoyan a los estudiantes desde diversos programas como SEA-MIDAS que acompaña académicamente a los estudiantes en los primeros niveles de las carreras de ingenierías y ciencias básicas principalmente.

Los entes gubernamentales han creados instituciones de ayuda económica que en su mayoría están dirigidos a alumnos con padres que tienen acceso restringido a créditos. En Colombia existen varias instituciones que proveen estos acompañamientos económicos desde prestamos hasta subsidios; el ICETEX es el principal ente financiador de estudios de educación superior en Colombia pero no es un secreto la mala fama que se ha generado por sus políticas de funcionamiento; dentro de las tesis a estudiar en el marco de antecedentes hay una que estudia el acompañamiento del programa Jóvenes en Acción del Departamento para la Prosperidad Social y la relación que tiene con la reducción de la deserción estudiantil.

MODELO PREDICTIVO PARA LA DESERCIÓN.

De los tres primeros componentes surgen las siguientes hipótesis:

Hipótesis 1a (H1a). *La deserción está directamente relacionada con la existencia de inequidades e inversamente relacionada con los programas y políticas destinados a compensarlas.*

Hipótesis 1b (H1b). *Las personas con entornos socioeconómicos más vulnerables tienen más probabilidades de abandonar los estudios que las personas con entornos más desfavorecidos.*

Hipótesis 1c (H1c). *Cuanto mayor sea la capacidad académica del individuo en la escuela secundaria, menor será la probabilidad de abandonar los estudios terciarios.*

Hipótesis 1d (H1d). *La implementación de programas de ayuda promueve una disminución de las tasas de deserción.*

Dentro de los elementos contextuales y geográficos que afectan la deserción, se incluyen aspectos como la cercanía a la institución, el tamaño de la localidad, la densidad poblacional y si es un sector rural o urbano; las diferencias por ubicación entre las instituciones de educación superior y los estudiantes marcan no sólo un choque cultural que existe entre regiones de Colombia sino que también se ven inmersos otros escenarios como traslado de residencia, costos de manutención en un lugar que probablemente sea más costoso; de lo anterior se llega a la hipótesis numero dos:

Hipótesis 2 (H2). *La existencia de desigualdades regionales influye en las tasas de abandono de los estudios terciarios*.

El último elemento expuesto por los autores son las características de los estudios; el tipo de programa académico del estudiante es un factor importante dentro de la deserción, múltiples estudios de entidades públicas y privadas encontraron que los estudiantes de CTIM (Ciencia,

tecnología, ingeniería y matemáticas) tienen mayor riesgo a desertar en sus estudios que estudiantes de humanidades y carreras sociales. El mercado laboral asociado al programa académico especifico también es un determinante para la deserción, porque a los estudiantes les interesa su futuro laboral.

De lo anterior surge la tercera hipótesis:

Hipótesis 3 (H3). *Existen diferencias institucionales por ciclo de programa y campo de estudio que afectan las tasas de deserción*.

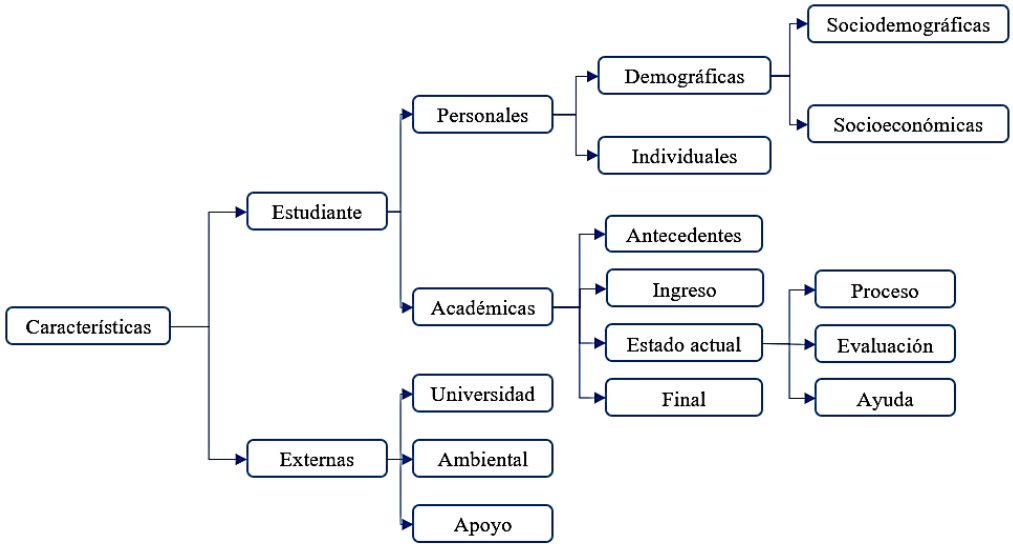
Cabe resaltar que las hipótesis anteriormente mencionadas vienen de la revisión bibliográfica y de literatura de la investigación, no están asociadas ni son las hipótesis del proyecto en curso.

El análisis de datos es una herramienta que permite a las organizaciones anticipar posibles escenarios negativos dentro de su funcionamiento. Por el lado de las IES se cuenta con gran cantidad de información usada por las autoridades académicas para gestionar la información de las características de los estudiantes, mejorar sus programas de prevención de la deserción y predecir posibles resultados dentro de sus actividades misionales. La Universidad Industrial de Santander dentro de su misión institucional “busca el fortalecimiento de una sociedad democrática, participativa, deliberativa y pluralista, con justicia y equidad social, comprometida con la preservación del medio ambiente y el buen vivir” aunque implícitamente no está la retención estudiantil es algo que va sujeto a su misión, la equidad social comprende los derechos y deberes que tiene una persona, sin importar su origen, etnia, estrato socioeconómico, raza, ni antecedentes. Por lo mismo a ser una institución de carácter oficial, cubre las necesidades de educación superior del departamento de Santander y en general del nororiente colombiano.

Según de Oliveira *et al.* después de hacer una revisión sistémica de literatura encontraron que las características que influyen mayormente en la deserción dentro de la educación superior son las siguientes:

Figura 9

Jerarquía de categorías y subcategorías de las características de deserción (2022)



Nota. Traducido por el autor, tomado de *Big Data Cogn. Comput*.

De acuerdo con los autores, el analista de datos y la institución que quiera realizar el trabajo puede tomar las características mostradas en la figura 8 y trabajar cada característica de forma individual o hacer conjunto a su libre elección; aunque en la mayoría de los estudios se utilizan datos personales y académicos de los estudiantes (De Oliveira, C.F. *et al*, 2021).

3. Planteamiento del Problema

La educación superior es considerada el eje fundamental para el progreso de las naciones. El nivel de desarrollo tecnológico, la producción científica e intelectual en los diferentes campos

de la ciencia se relaciona con la alta calidad que tenga la educación de un país. Por ello los países con mejor sistema educativo logra retener a los estudiantes y aumentar la tasa de profesionales con título universitario según Villanueva (2019).

He ahí donde los gobiernos fijan esfuerzos para intervenir en los principales problemas que afectan su sistema de educación superior; dentro de estos problemas se encuentra la deserción estudiantil; revisando los porcentajes de deserción estudiantil, en la educación superior de Colombia se visualiza que se mantiene en un rango entre 8.19 % y 9.89 % en programas universitarios en el periodo del 2010 al 2018 según el Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) (2020).

> SPADIES es un sistema de información especializado para el análisis de la permanencia en la educación superior colombiana a partir del seguimiento a la deserción estudiantil, que consolida y clasifica la información para facilitar el acompañamiento a las condiciones que desestimulan la continuidad en el sistema educativo (Mineducación, 2002).

La Universidad Industrial de Santander desarrollo desde la Vicerrectoría Académica el programa SEA (Sistema de excelencia académica) que cuenta con distintos programas de apoyo en cuatro momentos:

Momento 1. Antes del ingreso a la educación superior.

En esta etapa SEA se encarga de la divulgación de la oferta académica y de articular la universidad con la educación media.

Momento 2. En la transición a la educación superior.

Se da la caracterización del estudiante, ofrece cursos introductorios de matemáticas y lectura universitaria.

Momento 3. Durante la trayectoria académica.

Acompaña al estudiante en los ámbitos académicos, cognitivos, socioeconómicos, salud y ofrece clubes de lectura.

Momento 4. Transición a la vida laboral.

Aún persisten problemas estructurales en el SEA de la UIS; a pesar de que ofrece la Prueba Saber UIS, charlas preparatorias Saber Pro y Saber TyT, acompañamiento en el trabajo de grado y talleres de preparación para la vida laboral.

Al revisar la literatura y consultar múltiples investigaciones se refleja que los programas ofrecidos por SEA se limitan a estar disponibles y no buscan encontrar la raíz y brindar apoyo oportuno a los estudiantes en riesgo.

Teniendo en cuenta la información anteriormente mencionada, el tema de la deserción en la educación superior es de alto interés para el ámbito investigativo; por ello se realiza la presente investigación, específicamente en la facultad de Ingenierías Fisicomecánicas de la Universidad Industrial de Santander, utilizando datos suministrados por la Dirección de Admisiones y Registro Académico.

4. Resultados esperados

Modelo predictivo para prevenir la deserción de un estudiante de la Facultad de Ingeniería Fisicomecánicas de la Universidad Industrial de Santander, según sus características socioeconómicas y demográficas.

MODELO PREDICTIVO PARA LA DESERCIÓN.

Artículo de carácter publicable con los resultados de la investigación realizada.

Documento final con los resultados de la investigación.

5. Marco de referencia

5.1.Marco de antecedentes

Modelo de deserción estudiantil para la población de pregrado de una institución de educación superior privada en Bogotá – Colombia. – Universidad de Bogotá Jorge Tadeo Lozano.

El objetivo de la tesis es encontrar dos modelos predictivos para la deserción de los estudiantes de la Universidad de Bogotá Jorge Tadeo Lozano, por medio de aprendizaje automático supervisado; eligió características de los estudiantes que reposaban en el CRM (Customer relationship management o Gestión de Relaciones con el Cliente) de la universidad: Sexo (Masculino, Femenino), edad del estudiante, ingresos de los padres, nivel educación de la madre, resultado prueba SABER 11, número de compañeros de colegio con los que ingresa a la misma Universidad, número de hermanos, créditos inscritos para su primer semestre, entre otros; la hipótesis inicial es que entre mayor sea la cantidad de datos que se ingresen al modelo, mayor será su precisión de predecir la deserción. En la universidad al ingreso de un estudiante nuevo se le asigna un puntaje “SPADIES” que determina el grado de riesgo que tiene de abandonar los estudios; inicialmente realiza una revisión en la web sobre modelos predictivos existentes, encuentra modelos de árboles de decisión, bosques aleatorios, XGBOOST, Naive Bayes y regresión logística. Para hacer el análisis exploratorio de los datos utiliza Power BI de Microsoft para generar las primeras gráficas sobre el comportamiento de los datos; para identificar la correlación entre variables trabajo con los métodos de Pearson y Spearman, los cuales son comúnmente usados para análisis de datos con distribuciones normales y casos de datos extremos.

Para el procesamiento de datos utilizó Google Colab con lenguaje de Python y asignó codificación binaria a los datos para facilitar su análisis. Los datos fueron balanceados para mayor precisión en los modelos y se propuso que los modelos debiesen estar por lo menos encima del 85 % de precisión; en el modelo 1 se utilizaron la totalidad de los datos para ser analizados bajo árboles de decisión, bosques aleatorios, XGBOOST, Naive Bayes y regresión logística que arrojaron un porcentaje de precisión del 97,49 %, 97,88 %, 84,75 %, 47,52 % y 70 % respectivamente. Para el modelo 2 se usaron los datos de los dos últimos periodos académicos de los estudiantes analizados con árboles de decisión, bosques aleatorios, XGBOOST, Naive Bayes y regresión logística que arrojaron un porcentaje de precisión del 99,04 %, 99,68 %, 97,74 %, 46,13 % y 84,37 % respectivamente. Por lo tanto, el autor realizando la evaluación de modelos concluye que el modelo 2 muestra mejores resultados a nivel general, con mayor porcentaje de éxito en la predicción (Rodríguez, 2022). El trabajo de grado está directamente relacionado con la investigación presente, en él se encuentran distintas herramientas, software, modelos y recursos que sirven de guía directa para la realización del análisis de datos y creación del modelo predictivo.

Evaluación de impacto de jóvenes en acción como política para la reducción de la deserción; caso Universidad Industrial de Santander. – Universidad Industrial de Santander.

El objetivo del trabajo de grado es medir el impacto que genera el programa de política pública Jóvenes en acción sobre la deserción en los estudiantes de la Universidad Industrial de Santander, para aportar al debate sobre si el factor económico es el principal causante de la deserción; Jóvenes en Acción (JEA) es un programa del Departamento para la Prosperidad Social, que apoya a jóvenes en condición de vulnerabilidad con entregas monetarias condicionadas. Se tomo una muestra de 1618 estudiantes, datos suministrados por la Dirección de Admisiones y Registro Académico de la UIS, desagregados en 6 cohortes; los datos se compararon con la tasa

de deserción por programa, la deserción por IES por función del semestre, la matrícula, la cohorte y el género, esperando encontrar una relación entre el porcentaje de estudiantes beneficiarios del programa y la disminución de la deserción. Posteriormente se analizaron los datos con dos modelos de regresión de COX, la diferencia entre los dos modelos fue el uso de las características, al modelo dos le agregaron datos de costo de matrícula semestral y el sexo, dando como resultado mayor precisión; para terminar el autor encontró que si existe una relación entre la deserción en las IES y la pertenencia a JEA (Durán, 2020). Este trabajo en específico deja conclusiones que aportan valor a las hipótesis de la investigación donde la deserción se ve directamente relacionada con el nivel económico de un estudiante de educación superior, específicamente de la UIS.

Prevalencia de problemas de salud mental en los estudiantes UIS en riesgo de deserción por bajo rendimiento académico. – Universidad Industrial de Santander.

El objetivo del autor era validar si los problemas de salud mental y el consumo de sustancias psicoactivas tienen relación con el bajo rendimiento académico que puede llevar a un estudiante a quedar por fuera de la universidad (PFU); la muestra corresponde a estudiantes encuartados en el semestre 2016-2 que en total son 129 estudiantes de todas las carreras, de estos se eligen al azar 80 estudiantes para aplicar entrevistas semiestructuradas y pruebas de tamizaje para determinar si tienen un diagnóstico de salud mental. Los estudiantes seleccionados a participar en el estudio se contactaron por medio de las escuelas a las que pertenecían y eran libres de negarse a participar en el estudio; para concluir el autor encuentra que el trastorno Depresivo Mayor es más frecuente dentro de los estudiantes con un 27,53 % (Durán, 2017). La deserción es un tema que causa preocupación en muchas áreas de investigación, desde la medicina hasta la ingeniería; es valioso contar con tantas fuentes de información diversas que abordan el mismo problema complejo desde perspectivas diferentes.

5.2.Marco teórico

Se consultaron trabajos relacionados con modelos predictivos existentes, para conocer estudios previos que sirvan como base para la investigación presente.

Según Ópazo, D. *et al.* (2021) hay 3 enfoques principales dentro del campo de la investigación de la deserción en la educación superior, depende de la perspectiva en la que sea abordado el problema; distintos análisis se han realizado desde perspectivas económicas y psicológicas, teniendo como fuentes de información datos cuantitativos y cualitativos según el área de investigación. A continuación, se explican las generalidades de los enfoques encontrados dentro de la revisión a la literatura.

Enfoques explicativos

Los modelos de adaptación consideran que la adaptación y la integración social afectan la decisión de abandonar los estudios.

Spady (1970) considera un modelo basado en la teoría del suicidio de Durkheim, siendo la deserción el resultado de un proceso social complejo que contiene características como la formación académica del núcleo familiar, nivel académico previo, el potencial académico, la congruencia normativa, el apoyo social, el desarrollo intelectual, el desempeño educativo, la integración social, la satisfacción y el compromiso institucional.

Tinto (1975) formula un modelo considerando factores como la adaptación del estudiante al entorno universitario, el funcionamiento de la institución de educación superior y el rendimiento académico previo.

MODELO PREDICTIVO PARA LA DESERCIÓN.

En otro enfoque Bean (1982) plantea que la deserción depende de factores netamente individuales y personales que afectan directamente al estudiante, dejando los factores externos afuera de su propuesta.

Para terminar la revisión de los modelos predictivos de la deserción predecesores, Pascarella (1991) aconseja utilizar un modelo mixto donde los factores importantes se relacionan con la calidad de la institución de educación superior, la seguridad en la elección de la carrera o la existencia de becas.

Los primeros modelos que surgieron de los años setenta a los noventa, dieron paso a investigaciones donde se analizaban distintos conjuntos de características y/o factores que conllevan a la deserción; Cabrera *et al.* (2006) sugieren agrupar los modelos teóricos explicativos en cuatro grupos diferentes:

1. El modelo de adaptación, que describe la falta de integración de un individuo en el contexto universitario.

2. El modelo estructural, la estructura universitaria, incluyendo la política, la económica y la social, que influye en la deserción de los estudiantes.

3. El modelo económico, que describe la elección del estudiante de una forma alternativa de invertir tiempo, energía y recursos que podría ofrecer mayores beneficios en el futuro.

4. el modelo psicopedagógico, que abarca una mezcla de diferentes factores de los modelos adaptativo y estructural, más otras dimensiones de carácter psicoeducativo, como las estrategias de aprendizaje.

Para terminar el primer grupo de enfoques, es importante examinar otras investigaciones basadas en entrevistas y análisis manuales; Broc (2011-2018) concluyó que las variables de rendimiento académico previo predicen mejor el rendimiento universitario; por otro lado, Bejarano *et al.* (2017) haciendo uso de encuestas a jóvenes desertores, encontraron que las razones de la deserción en su mayoría eran por influencias socioeconómicas e individuales.

Entre otras investigaciones se llega a la conclusión de que el factor que más influye en la deserción es el acceso a financiamiento para su sostenimiento y matricula, por ellos se sugiere a los servicios de bienestar universitario ser más proactivos en este aspecto (Sinchi & Ceballos, 2018).

Por último, el estudio de Quintero (2016), establece que las causas de deserción de un estudiante no deben aplicarse a otro, puesto que los motivos de los dos pueden ser distintos aun teniendo características sociodemográficas y socioeconómicas similares.

Enfoques predictivos

Dentro de los enfoques de aprendizaje automático las instituciones han utilizado los datos que poseen para crear valor a través de herramientas de aprendizaje automático, van desde predicciones hasta análisis de variables a través de modelos interpretativos (Ópazo, D. *et al.*, 2021). A continuación, se detallan los modelos de enfoques predictivos.

Árboles de decisión:

Dentro de las investigaciones, los árboles de decisión han demostrado alto porcentaje de efectividad en modelos predictivos para la deserción; en la investigación de Kumar *et al.* (2012) compara procesos de entrenamiento para arboles de decisión, llegando a un árbol con 82,8 % de precisión; mientras Heredia *et al.* (2015) en su investigación sobre la deserción en la Universidad

Simón Bolívar, usa el mismo método de entrenamiento para árboles de decisión y explica que son un modelo adecuado, aunque su trabajo no llega a la conclusión de cuáles son las características que más influyen en la deserción. Por lo tanto, se recomienda utilizar árboles de decisión con resultados de optimización de parámetros para obtener mayor precisión dentro de la investigación (Heredia *et al,* 2015).

Regresión Logística:

Una regresión logística es un modelo de probabilidad al que a cada variable se le asocia un parámetro que muestra su relevancia (Cox, 1958); los estudios revisados que usan regresión logística en sus análisis de datos, definen variables que consideran que influyen en la deserción y aportan información básica de los estudiantes, para hallar las variables que más afectan el modelo predictivo; Santelices *et al*. (2013) concluyeron que la deserción está relacionada con el nivel socioeconómico, rendimiento académico previo, puntaje en la prueba de ingreso a la universidad, becas académicas y créditos económicos.

En un estudio con datos de la Universidad Bernardo O'Higgins de Chile, Matheu *et al.* (2018) tomaron 17 características tanto individuales como externas de los estudiantes y analizaron por medio de regresión logística; el resultado de la investigación fue que 7 variables eran las que más influían en el abandono de los estudios: sexo, horario de estudio (diurno, vespertino o nocturno), grupo etario, escuela de procedencia, convivencia con la familia, puntaje en la prueba de ingreso a la universidad y ocupación del padre; el puntaje de la prueba de admisión es la característica con más peso dentro del estudio.

Naive Bayes (Bayesiana ingenua):

Naive Bayes es un modelo probabilístico establecido desde el teorema de Bayes; el teorema de Bayes describe que si conocemos información sobre la ocurrencia de un evento A siempre que haya ocurrido un evento B, necesitamos saber si es posible inferir la ocurrencia del evento B siempre que haya ocurrido A (Soto, 2011).

Kumar y Pal (2011) utilizaron el modelo de Naive Bayes con datos de la Universidad Dr. RML Awadh de la India, llegaron al resultado de que existen factores que están altamente correlacionados con la deserción, entre ellos el rendimiento académico previo, el lugar de residencia, el idioma de enseñanza (clases mixtas en idioma nativo e inglés, o solo en inglés), la educación de la madre, los hábitos de los estudiantes, ingreso familiar anual, y estado familiar del estudiante.

Mientras que Hegde y Prageeth (2018), usaron datos de la Escuela de Artes y Ciencias de Amrita para predecir la deserción temprana; hallaron que los factores académicos, demográficos, psicológicos y de salud, eran los que más incidían en el fenómeno de la deserción.

K-Vecinos más próximos o K-Nearest Neighbors (KNN):

La técnica de KNN es un algoritmo de aprendizaje basado en instancias, en su funcionamiento se almacenan datos históricos y al clasificarse un nuevo objeto, se extraen los objetos más parecidos en sus características y se clasifica automáticamente (Morales, 2009).

La investigación realizada por Valero *et al*. (2005) utiliza dos modelos de minería de datos, los árboles de decisión y la KNN; los resultados del análisis y procesamiento de datos arrojo que el algoritmo de árboles de decisión tenía el 98,98 % de confiabilidad mientras que el del algoritmo KNN apenas superaba el 70 %. Esto indica que la técnica de KNN no contribuye a la elección de

un modelo predictivo sobre la deserción estudiantil; cabe resaltar que al cambiar los datos puede que cambie sus resultados, por ello no se descarta totalmente esta técnica.

Redes neuronales:

Las redes neuronales artificiales son un método inspirado en la biología capaz de crear modelos predictivos complejos no lineales (Zhang, 2000). Siri (2015) desarrollo un modelo con el método de redes neuronales utilizando datos administrativos, entrevistas telefónicas y encuestas de características personales, de los padres, la ubicación, el rendimiento académico anterior y los puntajes de las pruebas de admisión a la universidad.; los resultados obtenidos tuvieron un margen de precisión entre 65 % y el 84 %; un estudio posterior realizado por el mismo investigador determino que las variables más importantes fueron la educación familiar, el origen escolar, la falta de orientación preuniversitaria, el estudio con amigos y la motivación.

Otro estudio con este tipo de modelos lo realizó Alban y Mauricio (2019) con un perceptrón multicapa que obtuvo una precisión de 96,3 % y usando una función de base radial dio una precisión de 96.8 %; como resultado, las variables que afectan la deserción en este estudio eran si el estudiante tiene hijos, conocimiento en software utilizado en la carrera universitaria, compromiso familiar, adaptación a la universidad, ranking universitario y perspectiva del estudiante sobre su inserción en el mercado laboral.

Bosque aleatorio o Random Forest:

Es un método que construye clasificadores basados en árboles de decisión y que puede expandirse arbitrariamente para aumentar su precisión; inicialmente construye múltiples árboles de decisión, cada uno usando una muestra aleatoria de las variables originales. La etiqueta de clase

de un punto de datos se determina utilizando un esquema de votación ponderada con la clasificación de cada árbol de decisión (Alban & Mauricio, 2019).

Lee y Chung (2019) comparan un árbol de decisiones potencializado con un bosque aleatorio usando datos del Sistema Nacional de Información Educativa (NEIS) de Corea del Sur para predecir la deserción escolar; los autores resaltan que el modelo predictivo es aplicable en otros contextos, donde la sensibilidad permanece constante pero el valor predictivo aumenta según el porcentaje de deserción en donde se vaya a aplicar.

Árbol de decisiones de aumento de gradiente:

Se desarrolla un paradigma de potenciación de descenso de gradiente general para expansiones aditivas basado en cualquier criterio de ajuste. Cuando se usa con árboles de decisión, usa árboles de regresión para minimizar el error de la predicción. Un primer árbol predice la probabilidad de que un punto de datos pertenezca a una clase; el siguiente árbol modela el error del primer árbol, lo minimiza y calcula un nuevo error, que es la nueva entrada para un nuevo árbol de modelado de errores; y generalmente se usa para cursos masivos en línea (Friedman, 2002, p 367–378).

Comparaciones de múltiples modelos de aprendizaje automático:

En otras investigaciones se han comparado los modelos anteriormente descritos para confrontar su efectividad; Delen (2010) comparo árboles de decisión, redes neuronales, máquinas de vectores de soporte y regresión logística, para determinar que una máquina de vectores de soporte ofrece la mayor precisión; el análisis también concluyo que los predictores más importantes son el éxito educativo pasado, el éxito educativo presente y la ayuda financiera.

Fischer (2012) analizó la deserción de las carreras de ingeniería de la Universidad de Las Américas comparando redes neuronales, árboles de decisión y K-mediana con las siguientes variables: puntaje en la prueba de admisión a la universidad, rendimiento académico previo, edad y género; la investigación no tuvo resultados favorables por la poca fiabilidad de los datos.

Eckert y Suenaga (2014) compararon árboles de decisión, Naive Bayes y reglas de asociación, obteniendo el mejor rendimiento con árboles de decisión. El trabajo identificó como variables más importantes el rendimiento académico previo, la procedencia y la edad de ingreso de los estudiantes a la universidad. Adicionalmente, identificó que durante el primer año de la carrera es donde cobra mayor relevancia la contención, el apoyo, la tutoría y todas aquellas actividades que mejoren la situación académica del estudiante.

6. Metodología

La investigación se compone de una revisión de literatura que respalda y valida la relevancia de la investigación; además de un análisis predictivo de datos por medio de Data Mining (procesamiento de datos).

Se propone una metodología mixta que agrupe los dos ámbitos de la investigación. A continuación, se describen las fases de la investigación y las actividades correspondientes en cada una de ellas:

6.1.Fase 1. Definición del problema de interés y búsqueda de información

Según la definición del problema de investigación, se hace una revisión de literatura en alguna base de datos disponible en la biblioteca virtual de la Universidad Industrial de Santander (UIS); el propósito de la revisión fue constatar la relevancia del problema para la comunidad científica. Se presenta a continuación una serie de actividades que se llevaron a cabo:

Actividad 1. Definir las palabras clave para la búsqueda inicial en la base de datos.

Actividad 2. Construir la ecuación de búsqueda inicial.

Actividad 3. Elegir de la base de datos para la revisión bibliográfica.

Actividad 4. Revisar los abstracts para identificar las palabras que se deben excluir o incluir a la ecuación de búsqueda.

Actividad 5. Ejecutar la ecuación de búsqueda en Scopus y aplicar los filtros para reducir el rango de la producción científica, que se ajuste a la investigación.

Actividad 6. Hacer el análisis bibliométrico de la información.

Actividad 7. Analizar la literatura sobre la deserción obtenida de la búsqueda (características de los estudiantes en riesgo, estudios previos de modelos predictivos de deserción, deserción en el ámbito cultural colombiano, entre otros).

Actividad 8. Redactar los hallazgos relevantes encontrados en la revisión bibliográfica.

6.2.Fase 2. Recolección de los datos

Los datos para el análisis se obtuvieron de la Dirección de Admisiones y Registro Académico de la UIS.

Actividad 1. Definir las características a solicitar.

Actividad 2. Solicitar por medio de la facultad de Ingeniería Fisicomecánicas la base de datos a la Dirección de Admisiones y Registro Académico de la UIS.

A continuación, se muestran los campos solicitados que servirán como insumo para la construcción de los modelos de predicción:

- Nombre.
- Código estudiantil.
- Genero.
- Edad.
- Pertenencia a una minoría (a cuál).
- Tipo de colegio de donde se graduó (público o privado).
- Año de graduación del colegio.
- Año de ingreso a la UIS.
- Estrato socioeconómico.
- Residencia núcleo familiar.
- Beneficios ofrecidos por la universidad que tenga el estudiante (comedores, residencia, auxiliaturas).
- Nivel en el plan de estudios.
- Condición actual del estudiante.
- Tiene alguna discapacidad (cuál).
- Tiene hijos.
- Si tiene algún trabajo.
- Promedio ponderado.
- Condicionalidad (qué tipo).
- Si el estudiante deserto o quedo PFU.
- Nacionalidad.
- Departamento y municipio de procedencia.
- Si ha solicitado apoyo a Bienestar Universitario (o algún programa preventivo).

- Motivo de la deserción.
- Cancelaciones de semestre o aplazamientos de semestre (motivo).
- Si tuvo procesos disciplinarios.

6.3.Fase 3. Análisis preliminar y organización de los datos.

Se realiza el proceso de limpieza y transformación de los datos.

Actividad 1. Revisar la estructura y organización de los datos.

Actividad 2. Limpiar los datos para su análisis.

Actividad 3. Organizar los datos y asignar variables a los datos cualitativos.

Actividad 4. Modelar los datos para análisis inicial y las primeras conclusiones.

Actividad 5. Elegir el software para analizar los datos.

6.4.Fase 4. Elaboración del modelo predictivo

Dentro de esta fase se hace el análisis completo de los datos y se validan las hipótesis postuladas, por lo que definen las siguientes actividades:

Actividad 1. Revisar modelos predictivos existentes para la deserción.

Actividad 2. Ajustar los datos a los modelos seleccionados.

Actividad 3. Elegir o crear el modelo predictivo que mejor describa los datos.

Actividad 4. Comprobar las hipótesis planteadas en la fase 1.

6.5.Fase 5. Despliegue del modelo predictivo

Utilizar los resultados para crear informes y métricas, por lo que se plantean las siguientes actividades:

Actividad 1. Validar el modelo predictivo.

Actividad 2. Elegir las métricas a presentar en el libro final.

Actividad 3. Elaborar la presentación de los resultados obtenidos.

6.6.Fase 6. Síntesis de los resultados

Para concluir la investigación se crear los entregables del proyecto bajo las siguientes actividades:

Actividad 1. Elaborar el libro final del proyecto de grado.

Actividad 2. Construir el articulo académico de carácter publicable en el que se presentan los resultados del análisis de datos y el modelo predictivo de deserción.

7. Desarrollo del proyecto

7.1.Indagación de los datos

La base de datos suministrada por la Dirección de Admisiones y Registro Académico de la Universidad Industrial de Santander, contiene datos cualitativos y cuantitativos de los estudiantes que se registraron como activos según el sistema de información utilizado por la universidad; los estudiantes de la base de datos pertenecen a las carreras adjuntas a la Facultad de Ingenierías Fisicomecánicas, a continuación se despliega la información perteneciente a la base de datos para comprender las posibles relaciones entre las características de un estudiante y su situación académica.

7.1.1. Descripción de las variables

Tabla 3

Descripción de variables

Descripción de las variables

Variable	Nombre	Descripción
Condición	Graduado	Estudiante que culmina satisfactoriamente sus estudios de pregrado.
	Excluido por + dos cancel seme	Estudiantes que cancelaron 2 semestres consecutivos.
	Retiro definitivo > 3 periodos	Estudiante que este retirado por más de 3 periodos académicos consecutivos.
	Excluido por vencimiento tiempo	Estudiante que cumplió el tiempo de permanencia para grado.
	PFU	Estudiante queda por fuera de la universidad por bajo rendimiento académico.
	Normal	Estudiante activo, matriculado en algún programa de pregrado de la facultad con normalidad académica.
	Retiro voluntario	Retiro por no renovar matrícula o solicitar cancelación de período, debe solicitar readmisión aún no está PFU
	Cambio de programa	Estudiante cambió de programa de pregrado.
	Cancelación definitiva de matrícula voluntaria	Estudiantes que cancelaron matrícula de forma definitiva y voluntaria
	Condicional primera vez	El estudiante está condicional por primera vez; el promedio del estudiante está entre 2.70 y 3.19, debe subir su promedio ponderado acumulado a 3.2 o más para el próximo semestre.
Programa académico	11	Ingeniería de sistemas
	21	Ingeniería civil
	22	Ingeniería eléctrica
	23	Ingeniería industrial
	24	Ingeniería mecánica
	26	Ingeniería electrónica
	27	Diseño industrial
	Año y periodo académico	Corresponde al año y periodo académico en que el estudiante figura en el sistema de admisiones.
	Año y periodo de condición	Año en el que se reportó la información en admisiones.

Tiempo	Año de ingreso a sede	Indica el año en el que el estudiante ingreso a la Universidad Industrial de Santander en un programa de pregrado asociado a la Facultad de Ingeniería Fisicomecánicas.
	Año de grado	Indica el año en que el estudiante culminó el bachillerato
	Fecha de nacimiento	Fecha de nacimiento del estudiante.
Colegio	Publico	Estudiantes que terminaron el bachillerato en una institución publica
	Privado	Estudiantes que terminaron el bachillerato en una institución privada
Municipio	Área metropolitana	Estudiantes residentes en Bucaramanga, Floridablanca, Girón y Piedecuesta.
	Otros municipios de Santander	Estudiantes residentes en municipios de Santander que no sean los cuatro municipios del área metropolitana
	Municipios lejanos	Estudiantes residentes en municipios que pertenecen a otros departamentos, excluyendo a Santander.
	Desconocido	Estudiantes residentes en municipios desconocidos
Forma de ingreso	Ingreso normal	Estudiantes que se inscribieron con la prueba ICFES y fueron admitidos a la UIS.
	Estudiantes provenientes zonas de difícil acceso o problemas de orden público	Estudiantes que fueron admitidos a la UIS bajo admisiones especiales.
	Víctimas del conflicto armado	Estudiantes que fueron admitidos a la UIS bajo admisiones especiales.
	Comunidades indígenas	Estudiantes que fueron admitidos a la UIS bajo admisiones especiales.
	Intercambio académico	Estudiantes que ingresan a la UIS por intercambio.
	Comunidades afrocolombianas	Estudiantes que fueron admitidos a la UIS bajo admisiones especiales.

Nota. Elaboración propia.

Creación de nuevas variables:

I. **Edad de grado de bachillerato**, utilizando la fecha de nacimiento y la fecha de grado de bachillerato del estudiante.

II. **Tiempo de ingreso a la universidad**, es el tiempo en años que demoro el estudiante desde su grado de bachillerato a ingresar a la universidad.

III. **Tiempo desde ingreso a sede y último reporte**, es el tiempo en años que el estudiante ha tenido una relación académica con la universidad desde su ingreso a alguna sede y su último reporte de admisiones.

IV. **Edad de reporte,** es la edad del estudiante cuando se presentó el último reporte por admisiones según la base de datos.

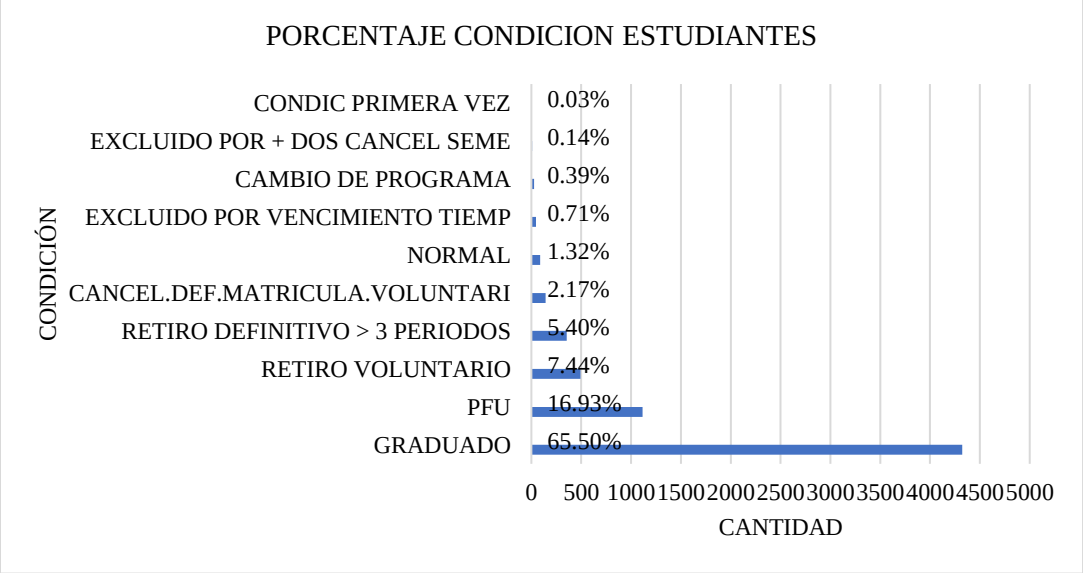
7.1.2. Análisis preliminar de los datos

8.1.2.1. Análisis por características.

Situación académica.

Figura 10

Gráfica porcentual de la condición de los estudiantes

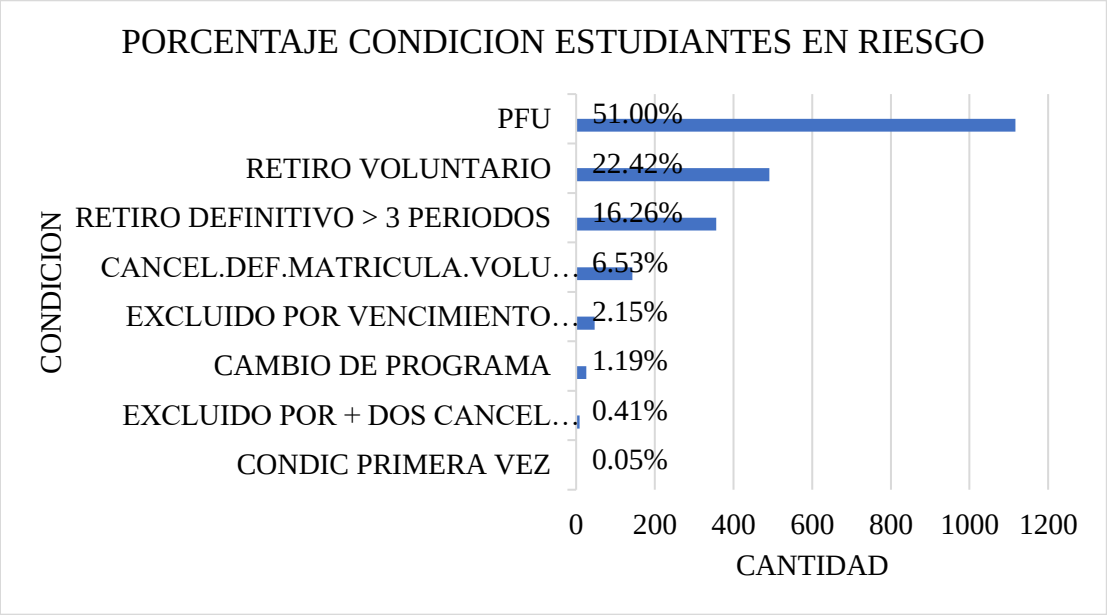


Nota. Elaboración propia.

Se observa a partir de la gráfica que aproximadamente el 33.19 % de los estudiantes de la facultad se encuentra en una situación de riesgo a la deserción o ya ha desertado sus estudios de pregrado, este cálculo se halla agrupando los valores de estudiantes en la condición de PFU, retiro voluntario, retiro definitivo > 3 periodos, cancel.def.matricula.voluntaria, excluido por vencimiento tiemp, cambio de programa, excluido por + dos cancel seme y condic primera vez; esta cifra refleja una diferencia con el promedio de deserción de educación superior universitaria en Colombia que es de 8.684 % entre 2015 y 2019 (SPADIES, 2020).

Figura 11

Gráfica porcentual de la condición de estudiantes en riesgo



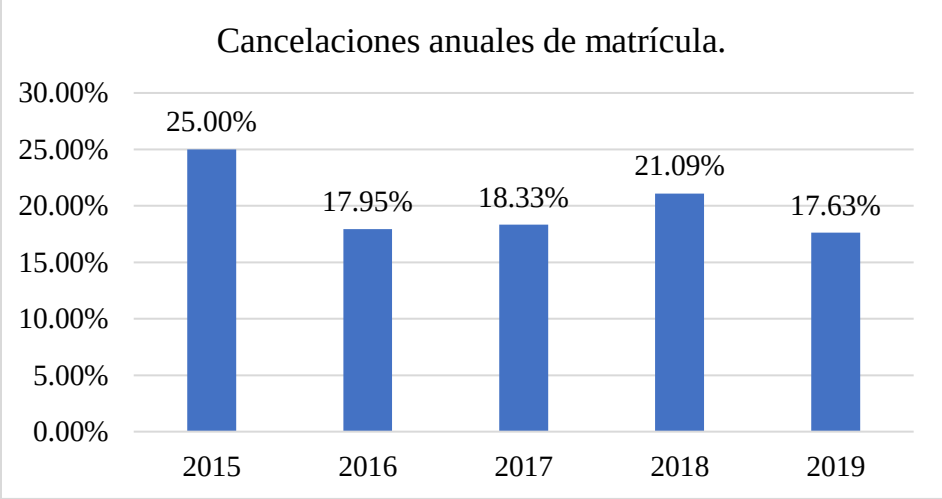
Nota. Elaboración propia.

Excluyendo los estudiantes graduados, en la gráfica anterior se observa que el 96.18 % de los estudiantes registrados en la base de datos se encuentra en riesgo de deserción o ya ha desertado sus estudios universitarios.

Cancelaciones de matrícula.

Figura 12

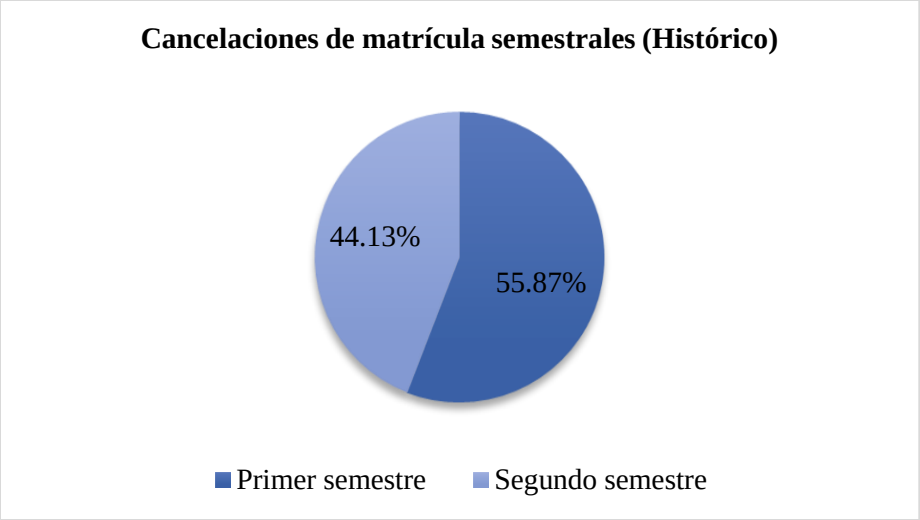
Porcentaje de cancelaciones anual



Nota. Elaboración propia.

Figura 13

Porcentaje de cancelaciones primer y segundo semestre



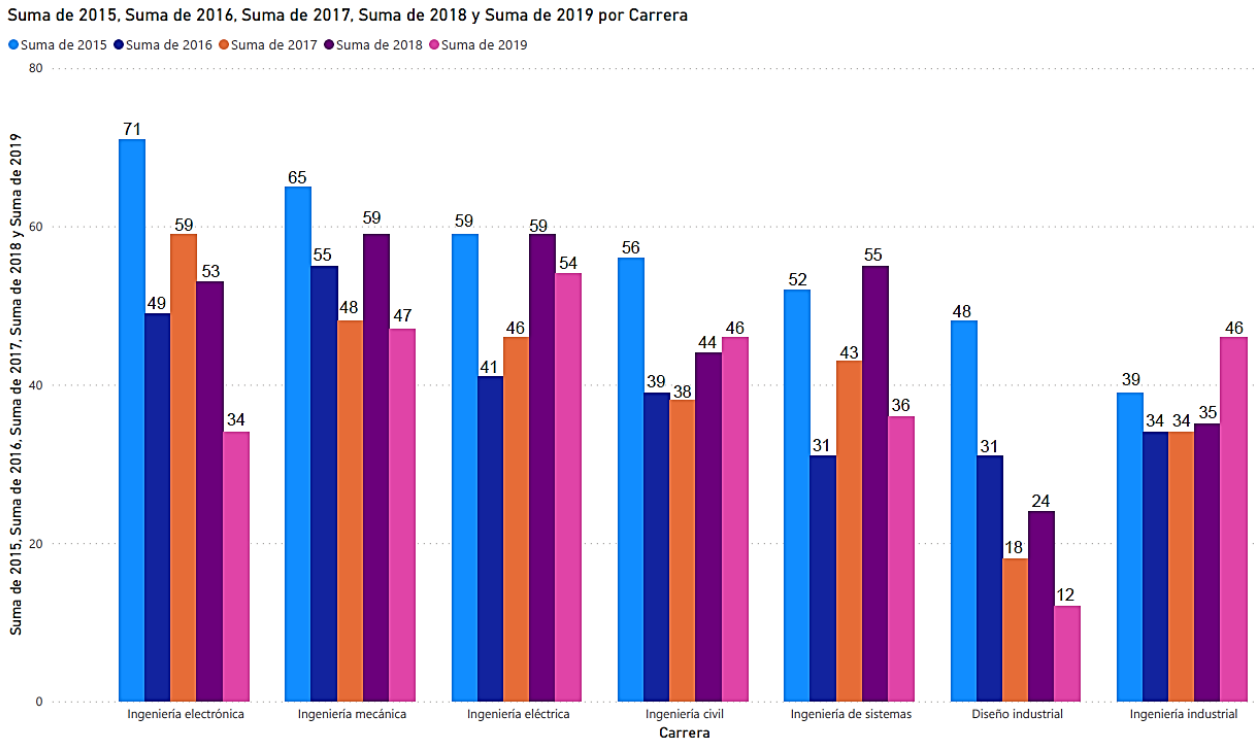
Nota. Elaboración propia.

Las cancelaciones han tenido una disminución con respecto al tiempo; el periodo académico donde más se hacen cancelaciones de matrícula es el primero con un 11 % aproximadamente por encima del segundo periodo académico.

Dentro de la base de datos suministrada por la Dirección de Admisiones y Registro Académico se encuentran las cancelaciones de matrícula presentadas en el periodo 2015-2019, en qué semestre académico se presentó y la carrera del estudiante que presentaba la cancelación.

Figura 14

Cancelaciones anuales por programa académico



Nota. Elaboración propia.

Admisiones especiales.

Según el Acuerdo 282 de 2017 del 7 de noviembre, la Universidad Industrial de Santander cuenta con el programa de ingreso a la Universidad de aspirantes por la modalidad de admisiones especiales. El número de estudiantes que ingresaron durante el periodo 2015 – por modalidad de ingreso especial fue de 81 distribuidos de la siguiente manera:

Tabla 4

Cantidad de estudiantes por modalidad de ingreso especial

Tipo de ingreso	Cantidad
COMUNIDADES AFROCOLOMBIANAS	10
COMUNIDADES INDIGENAS	14

VICTIMAS DEL CONFLICTO ARMADO	31
ESTUDIANTES PROVENIENTES ZONAS DE DIFICIL ACCESO O PROBLEMAS ORDEN PUBLICO	26

Nota. Elaboración propia.

En la tabla 5 se muestra el estado de estos 81 estudiantes.

Tabla 5

Condiciones estudiantes con ingreso especial

Condición	Cantidad	Porcentaje	Porcentaje de deserción
PFU	30	37.04%	**93.83%**
RETIRO VOLUNTARIO	27	33.33%	
RETIRO DEFINITIVO > 3 PERIODOS	16	19.75%	
GRADUADO	5	6.17%	
CANCEL.DEF.MATRICULA.VOLUNTARI	3	3.70%	
Total	**81**	**100.00%**	

Nota. Elaboración propia.

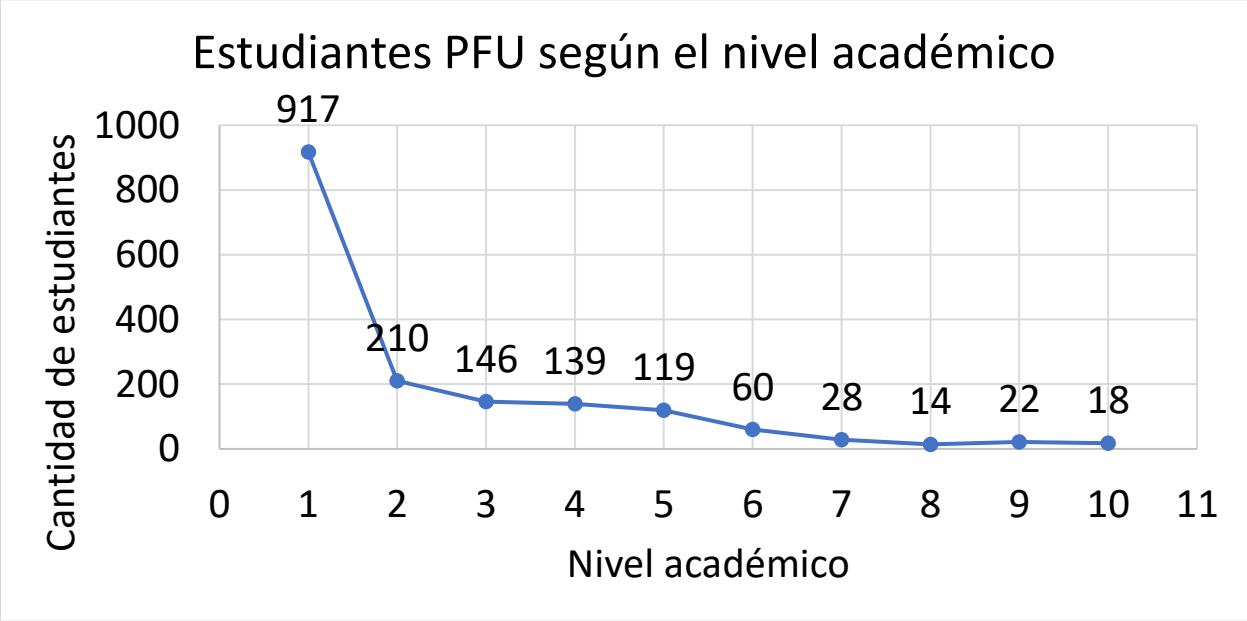
Como se puede observar el 93.83 % de los estudiantes que ingresan a la universidad por modalidad especial se encuentran en deserción estudiantil. Este fenómeno particularmente es preocupante puesto que el ingreso especial es un objetivo institucional, apegándose al inciso segundo del artículo 13 de la constitución política de Colombia establece que "*El estado promoverá las condiciones para que la igualdad sea real y efectiva y adoptará medidas en favor de grupos discriminados y marginados*" (Consejo académico UIS, 2017).

El alto porcentaje de deserción de los estudiantes que ingresan por modalidad especial marca que el objetivo institucional no se está cumpliendo y los estudiantes que vienen de entornos sociales difíciles, de etnias indígenas y afrocolombianas están quedando excluidos de la educación superior.

Para terminar con el análisis preliminar de los datos, tenemos la gráfica de los estudiantes que quedan PFU según el nivel académico en el que estaban como se observa en la figura 15.

Figura 15

Estudiantes PFU según nivel académico



Nota. Elaboración propia.

De la figura anterior se obtiene que el 84.4 % de los estudiantes que quedan por fuera de la universidad (PFU) se encuentran dentro de los primeros cuatro semestres de su programa académico que para ingenierías sería el ciclo básico de la carrera.

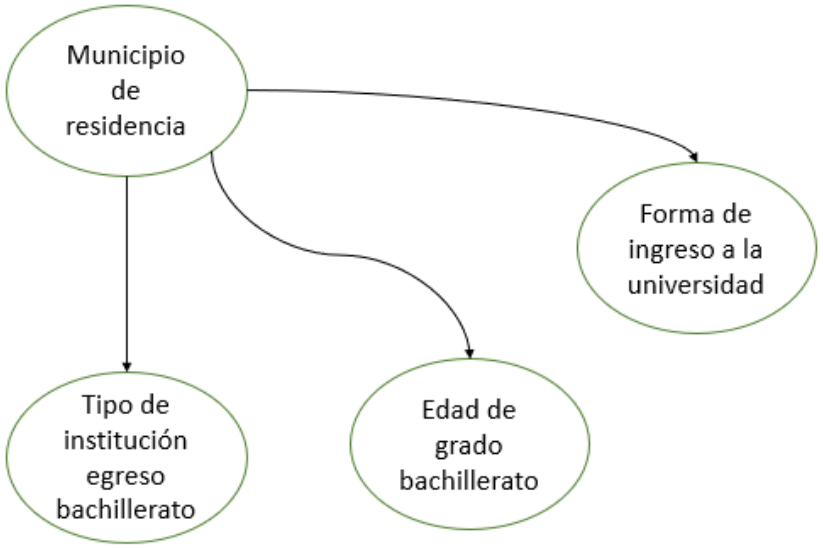
8.1.2.2. Análisis multivariable.

Para todo el análisis realizado se utiliza Jupyter Notebook, este es un proyecto sin ánimo de lucro para desarrollar software de código abierto, estándares abiertos y servicios para computación que contiene múltiples lenguajes de programación (Cabrera y Diaz, 2021); se decide trabajar con el lenguaje de programación Python.

Antes de empezar con el análisis multivariable se hacen los diagramas de relación para tener las primeras hipótesis sobre el comportamiento de las características independientes sobre la variable respuesta, que es la condición del estudiante.

Figura 16

Diagrama de relación caracterización del estudiante, municipio

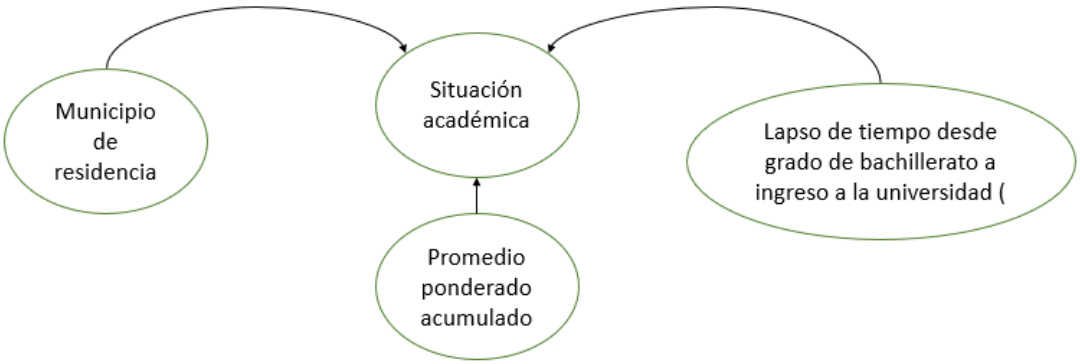


Nota. Elaboración propia.

De la figura anterior se infiere que el municipio de residencia de un estudiante influye las tres características asociadas en la figura; en el tipo de institución egreso de bachillerato puesto que en los municipios pequeños de Colombia es poco común que hayan instituciones de educación media privadas, en la edad de grado de bachillerato dado que en los municipios pequeños y en su mayoría la zona rural de Colombia los estudiantes acceden a la educación en una etapa tardía y finalmente en la forma de ingreso a la universidad dado que las condiciones especiales de ingreso en su mayoría

Figura 17

Diagrama de relación caracterización del estudiante, situación académica



Nota. Elaboración propia.

A continuación, se muestran las tablas de contingencia de las variables condición, tipo de institución de grado de bachillerato (colegio) y tiempo ocioso en años que hay desde que se graduó un estudiante de bachillerato e ingresa a la universidad.

Tabla 6

Tabla de contingencia condición-colegio

CONDICIÓN	PRIVADO	PÚBLICO
CAMBIO DE PROGRAMA	0.38%	0.40%
CANCEL.DEF.MATRICULA.VOLUNTARI	2.37%	2.09%
CONDIC PRIMERA VEZ	0.00%	0.02%
EXCLUIDO POR + DOS CANCEL SEME	0.11%	0.17%
EXCLUIDO POR VENCIMIENTO TIEMP	0.54%	0.78%
GRADUADO	62.57%	66.62%
NORMAL	0.70%	1.56%
PFU	15.89%	17.33%
RETIRO DEFINITIVO > 3 PERIODOS	6.62%	4.91%
RETIRO VOLUNTARIO	10.82%	6.11%

Nota. Elaboración propia.

De la tabla de contingencia condición-colegio se deduce que de la cantidad total de estudiantes que ingresan a la universidad de colegios públicos hay 4.05 % más graduados de los que ingresan de colegios privados; con respecto del retiro voluntario, los estudiantes que son egresados de colegios privados tienen 4.71 % más deserciones por esta condición que los que ingresan de colegios públicos, puede estar sujeto a la facilidad que tienen los estudiantes de migrar a universidades privadas.

Al obtener el coeficiente V de Crammer para la tabla de contingencia de la tabla 6 se obtiene:

```
colegio
0    0.182477
1    0.114179
dtype: float64
```

La interpretación de este coeficiente está sujeta al resultado obtenido; entre 0 y 0,2 indica que no hay asociación, de 0,2 indica una asociación débil, entre 0,2 y 0,6 indica una asociación moderada y entre 0,6 y 1 indica una asociación fuerte (Ojeda *et al.* 2018).

Los resultados de los coeficientes de la condición-colegio indican que no existe una asociación entre la condición que tenga un estudiante y el tipo de colegio del que se haya graduado de bachillerato.

Siguiendo con el análisis multivariable, se presenta la tabla de contingencia de la condición de un estudiante y el tiempo ocioso (tiempo en años desde que se graduó del colegio e ingreso a la UIS).

Tabla 7

Tabla de contingencia condición-tiempo ocioso

CONDICION	0	1	2	3	4	5	6	7	8	9	10
CAMBIO DE PROGRAMA	0.00%	0.41%	0.39%	0.24%	0.78%	0.00%	0.99%	0.00%	0.00%	0.00%	0.00%
CANCEL.DEF MATRICULA VOLUNTARI	2.90%	2.56%	1.55%	1.67%	0.78%	1.28%	0.99%	1.85%	0.00%	0.00%	8.33%
CONDIC PRIMERA VEZ	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EXCLUIDO POR + DOS CANCEL SEME	1.45%	0.02%	0.23%	0.24%	0.39%	0.00%	0.99%	0.00%	0.00%	0.00%	8.33%
EXCLUIDO POR VENCI MIENTO TIEMP	0.00%	0.53%	0.93%	1.44%	0.39%	1.28%	1.98%	0.00%	2.63%	0.00%	0.00%
GRADUADO	55.07 %	64.82 %	71.27 %	64.35 %	63.81 %	69.23 %	52.48 %	59.26 %	63.16 %	44.44 %	8.33%
NORMAL	1.45%	0.94%	1.79%	1.67%	1.17%	3.85%	3.96%	5.56%	0.00%	0.00%	0.00%
PFU	20.29 %	17.71 %	14.13 %	18.42 %	17.51 %	12.82 %	21.78 %	9.26%	13.16 %	22.22 %	8.33%
RETIRO DEFINITIVO > 3 PERIODOS	7.25%	5.60%	4.58%	5.02%	3.11%	4.49%	5.94%	12.96 %	5.26%	7.41%	25.00 %
RETIRO VOLUNTARIO	11.59 %	7.38%	5.12%	6.94%	12.06 %	7.05%	10.89 %	11.11 %	15.79 %	25.93 %	41.67 %

Nota. Elaboración propia.

En la tabla 7 se puede ver completa en el Apéndice J; se observa que el 55.07 % de los estudiantes que se gradúan e ingresan a la UIS el siguiente periodo académico después de culminar su bachillerato logran graduarse de programa académico; el mejor porcentaje de graduados según el tiempo ocioso es el 71.27 % de estudiantes que ingresaron a la UIS 2 años después de haberse graduado del colegio; de todas las columnas de tiempo ocioso desde 0 años de tiempo ocioso a 8 años de tiempo ocioso los porcentajes de graduados varían entre un 0 % a 10 % exceptuando el

tiempo ocioso de 2 años; para el año 9 se observa que el porcentaje de graduados es menor al 50 % de estudiantes y para el tiempo ocioso de 10 años el porcentaje de graduados no logra alcanzar sino un 8.33 %, siendo el tiempo ocioso de 10 años el que tiene peores resultados de deserción, el 91.66 % de los estudiantes que tienen un tiempo ocioso de 10 años deserta los estudios por distintas circunstancias pero la principal es el retiro voluntario con un 41.67 % seguido por el 25 % que no matriculan 3 periodos consecutivos.

Coeficientes V de Cramer para tabla de contingencia Condición-Tiempo ocioso:

```
ocio
 0     1.041230
 1     0.134357
 2     0.240998
 3     0.423042
 4     0.539516
 5     0.692483
 6     0.860618
 7     1.176994
 8     1.403070
 9     1.664521
 10    2.496782
 11    2.398830
 12    4.993564
 13    3.867998
 14    4.993564
 15    4.324553
 16    4.993564
 17    8.649106
 19    8.649106
 20    6.115842
 23    8.649106
 dtype: float64
```

Según la interpretación de los coeficientes se tiene que de 2 años a 4 años de tiempo ocioso que tiene un estudiante hay una asociación moderada entre la condición de un estudiante y esta cantidad de años de tiempo ocioso, cuando un estudiante tiene un tiempo ocioso de 5 o 6 años si hay una asociación fuerte entre la condición y el tiempo ocioso.

MODELO PREDICTIVO PARA LA DESERCIÓN.

Análisis de conglomerados o clustering:

Primero se elige el método de clustering a utilizar, por la naturaleza de los datos se elige el método de KMeans (K-medias) que es un algoritmo de clasificación no supervisada que agrupa datos en *k* grupos basándose en sus características (Zahra, S. *et al.* 2015). Las variables que se usan en la segmentación son: promedio, nivel (la variable ordinal se transforma así:

1) La variable ordinal se trata como una variable de escala. Sea r={1,2,..M} las categorías de la variable.
2) Se "mapea" el rango de cada variable en [0, 1] mediante la transformación

$$z_i = \frac{r_i - 1}{M - 1}$$

Por ejemplo, para el nivel $z_1 = \frac{1-1}{10-1} = 0, z_2 = \frac{2-1}{10-1} = \frac{1}{9}, .. z_{10} = 1.$

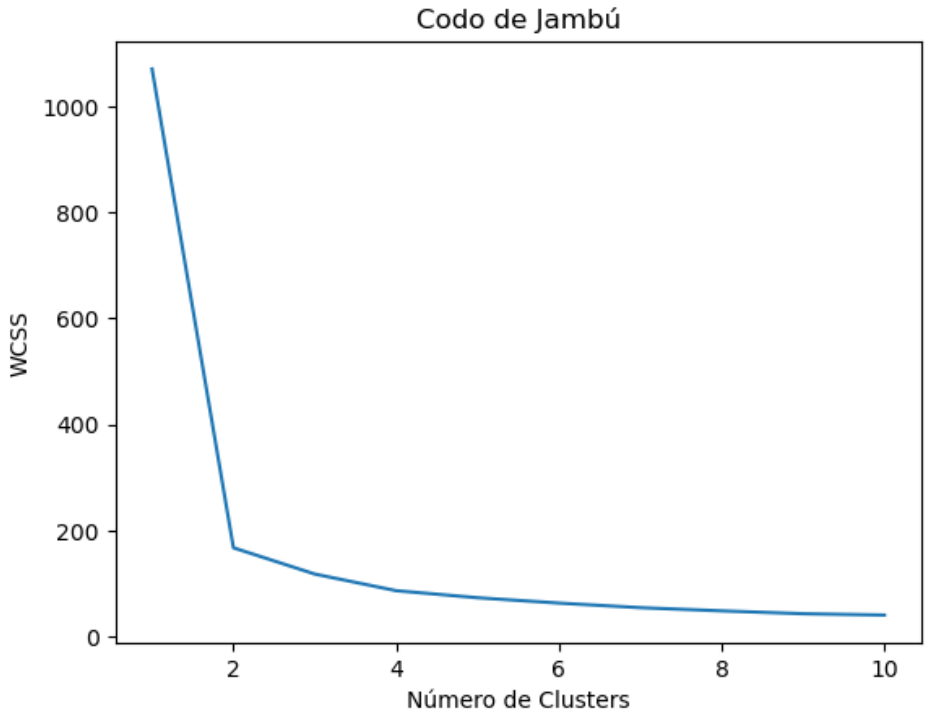
Otras variables que se usan son la edad de grado de bachillerato, graduación-ingreso a sede (tiempo ocioso), tipo de institución y Tiempo activo(años).

El objetivo de la clasificación no supervisada es construir una tipología de los estudiantes que permita elaborar estrategias de apoyo para la disminución de la deserción estudiantil de la facultad de ingeniería fisicomecánicas.

Para hallar el número óptimo de clústeres utilizamos el método codo de Jambú; utilizando el valor de WCSS (suma de los cuadrados) se forma la gráfica con la cantidad de clústeres de cero a diez, en el punto en que el valor de WCSS deja de descender drásticamente se hallara el número de clústeres que se ajustan a los datos, como se observa en la figura 18.

Figura 18

Gráfica codo de Jambú

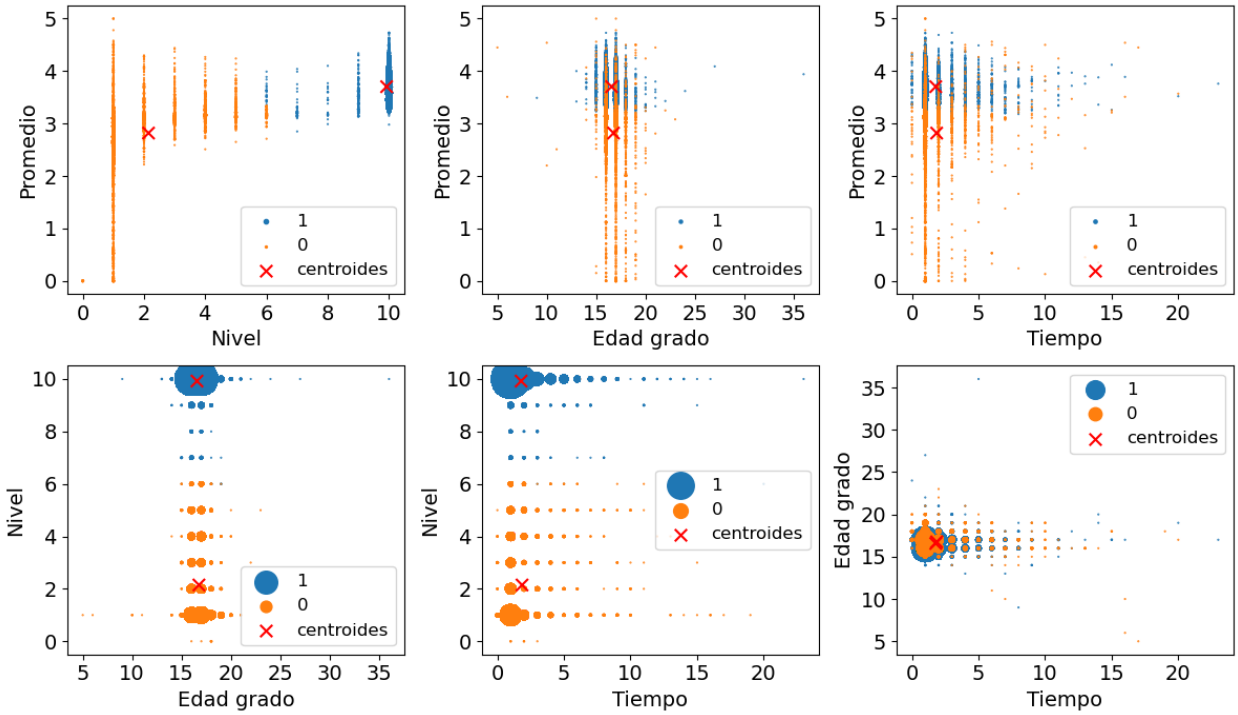


Nota. Elaboración propia.

La figura 18 muestra que el número de clústeres que se forman según la base de datos es dos (2). A continuación, se presenta la figura 19 que contiene el clustering para las variables preseleccionadas:

Figura 19

Gráficas de clústeres y sus centroides



Nota. Elaboración propia.

Conglomerado 1: Naranja.

Conglomerado 2: Azul.

Teniendo en cuenta los dos clústeres y las figuras 18 y 19 podemos deducir que la variable edad de grado y tiempo no agrupan a los individuos en grupos significativos, lo que se puede observar en las coordenadas de los centroides de cada clúster, las variables significativas que pueden dar grupos de individuos con características similares son el promedio y el nivel.

Tabla 8

Coordenadas de los centroides

Centroide	Promedio	Nivel	Edad grado	Tiempo
1	2.818616	2.147557	16.759071	1.854378

2	3.702938	9.921024	16.561438	1.785352

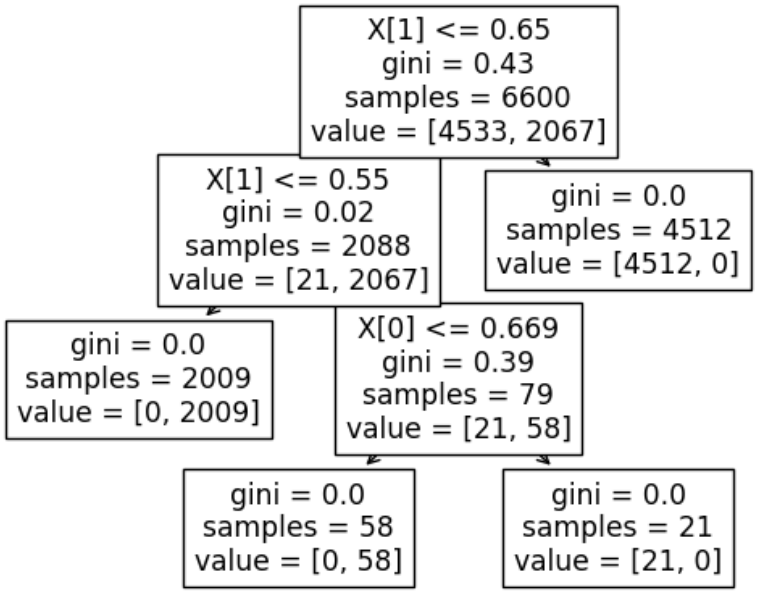
Nota. Elaboración propia

En el conglomerado 1 la coordenada del centroide en la variable promedio es 2.82 y en la variable nivel es 2.15; esto se debe al bajo rendimiento académico reportado en los semestres iniciales, donde la mayoría de los estudiantes desertores quedan PFU por su promedio ponderado, a partir del nivel 6 es poco probable que un estudiante quede PFU por su rendimiento académico, esto marca el conglomerado 2 donde en centroide de la variable promedio es 3.70 y el centroide de la variable nivel es 9.9; los individuos del conglomerado 2 se encuentran es una normalidad académica marcada por su promedio y nivel dentro de la institución.

Seguidamente se construye un logaritmo de aprendizaje automático para clasificar los estudiantes según las variables de los clústeres.

Figura 20

Árbol de decisión variables clústeres



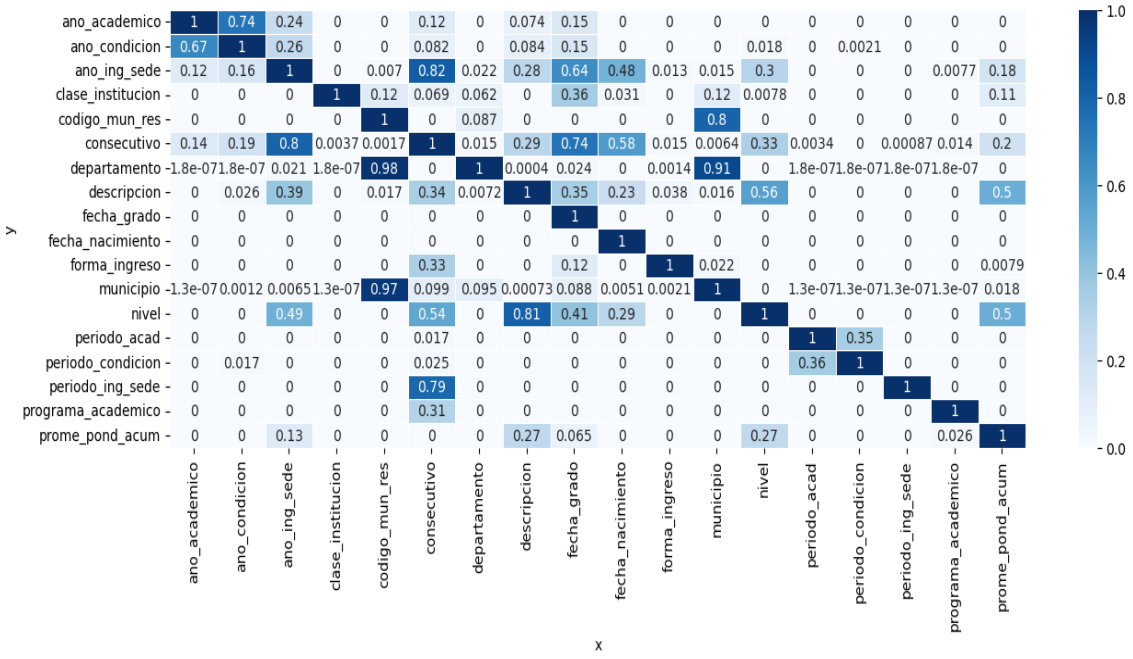
Nota. Elaboración propia.

El árbol de decisión está relacionado al análisis de clústeres, para ello se normalizaron las variables, esto se hace para que las variables como promedio que va de 0.0 a 5.0 no sea comparada con el tiempo activo de un estudiante en la UIS que va de 0 años a 36 años. La normalización ayuda a que los datos usen una escala común sin distorsionar las diferencias entre los intervalos de las características.

Las variables que tienen mayor relación con la condición del estudiante son X0 = Promedio y X1 = Nivel, la primera decisión del árbol se basa en el nivel del estudiante este clasifica los estudiantes en 2 grupos clúster, estudiantes que están en nivel mayor o igual a 6 que por lo general logran culminar los estudios, en este grupo se encuentran 4512 estudiantes; de nivel menor o igual a 5 en los que se encuentran 2088 estudiantes, en este grupo se encuentran las mayores tasas de deserción dado que por bajo nivel rendimiento académico los estudiantes entre primer y segundo semestre quedan PFU.

Figura 21

Matriz de relación base de datos completa

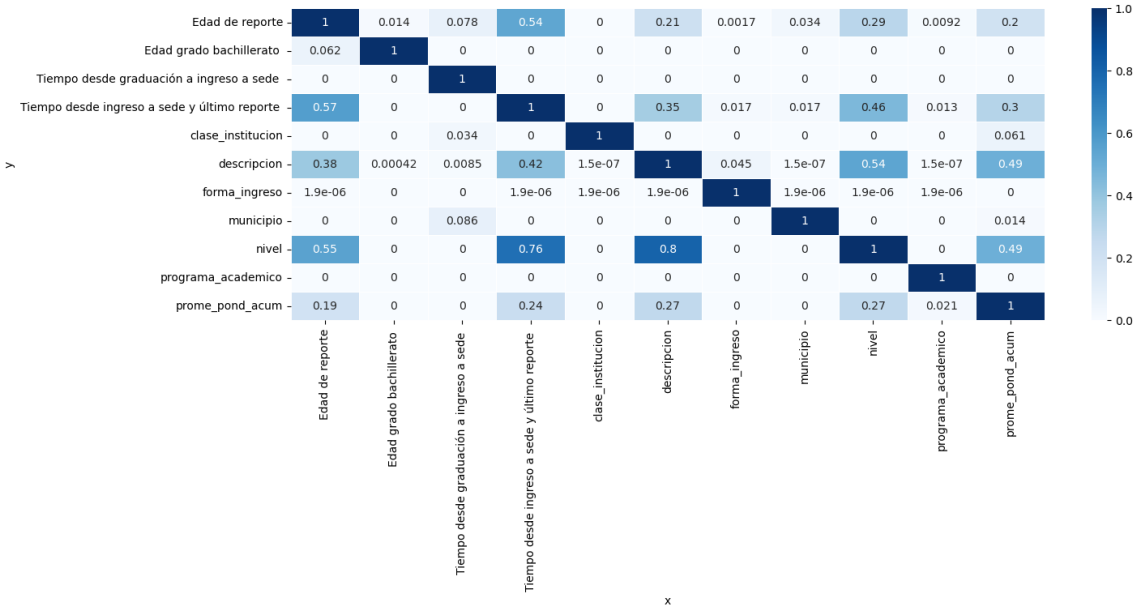


Nota. Elaboración propia.

La matriz de relación muestra la influencia que tienen otras variables sobre una variable específica, para ello se asocia con el porcentaje de influencia de las variables sobre la situación de un estudiante (descripción). De lo anterior se deduce que la variable nivel es la que más influye en la variable respuesta con un porcentaje de 81 %, el promedio ponderado es la segunda variable más influyente en la descripción con un porcentaje de 27 %.

Figura 22

Matriz de relación de las variables seleccionadas



Nota. Elaboración propia.

En el grafico anterior se usan datos de las variables preseleccionadas y transformadas, quitando las que no tienen relación significativa con la variable respuesta.

7.2. Preparación de los datos

Dentro de la preparación de los datos se hacen las respectivas transformaciones según el método del modelo predictivo y el lenguaje de programación utilizado que será Python.

Para la variable respuesta **“Condición del estudiante”** se eligen dos grupos así:

Normalidad académica: estudiantes que tienen condición de graduados y estudiantes activos con normalidad académica; “NORMAL”.

En riesgo académico: estudiantes que tienen retiro voluntario, que cambiaron de programa y estudiantes que se encuentran condicionales por primera vez, estudiantes excluidos por vencimiento de tiempo de permanencia, estudiantes que quedaron PFU, estudiantes excluidos por cancelar más de dos semestres consecutivos, estudiantes que se retiraron y duraron más de tres

semestres sin pedir readmisión y estudiantes que cancelaron definitivamente su matrícula; "RIESGO".

Con respecto a la variable **"Municipio"** que contiene el municipio de residencia del estudiante, se hizo una sustitución inicial donde se dejaron únicamente cuatro grupos así:

1º. **Área metropolitana de Bucaramanga:** estudiantes que pertenecen a Girón, Piedecuesta, Floridablanca y Bucaramanga.
2º. **Otros municipios de Santander:** estudiantes que pertenecen a municipios de Santander ajenos al área metropolitana.
3º. **Municipios lejanos:** estudiantes procedentes de municipios de departamentos diferentes a Santander.
4º. **Desconocidos:** estudiantes que no cuentan con información de procedencia.

Además del agrupamiento de los datos en cuatro grupos, se realizó la transformación de estos en One-Hot Encoding que reemplaza las variables categóricas por variables numéricas; como las variables no son ordinales no sirve el encoding ordinal y se usa este método para evitar la multicolinealidad.

La última variable transformada para el modelado fueron los **"programas académicos"**, los que se definen según la universidad con un código de dos cifras, por ejemplo, Ingeniería Industrial con el código 23. Es importante realizar la transformación para que el algoritmo no interprete los valores numéricos de los programas académicos como un rango de salida sino como grupos de clasificación.

7.3.Modelado

Los modelos se eligieron según la naturaleza de los datos y por la precisión mostrada en otras investigaciones:

- Arboles de decisión.
- Bosques aleatorios o Random Forest.
- Regresión Logística.

7.3.1. Modelado 1 Árbol de Decisión

A continuación, se muestran las variables utilizadas para el modelo árbol de decisión, estas variables no han tenido ninguna transformación.

```
      <class 'pandas.core.frame.DataFrame'>
RangeIndex: 6600 entries, 0 to 6599
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   descripcion                 6600 non-null   object
 1   programa                    6600 non-null   object
 2   prome                       6600 non-null   float64
 3   nivel                       6600 non-null   int64
 4   forma_ingreso               6600 non-null   int64
 5   institucion                 6600 non-null   int64
 6   municipio                   6600 non-null   object
 7   Edad bachillerato           6600 non-null   int64
 8   graduación-ingreso a sede   6600 non-null   int64
 9   Tiempo activo               6600 non-null   int64
 10  Edad de reporte             6600 non-null   int64
dtypes: float64(1), int64(7), object(3)
memory usage: 567.3+ KB
```

Después de realizar la transformación de las variables con One Hot Encoding, únicamente quedaron datos de tipo enteros y decimales; esto se hace para eliminar las variables categóricas.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5280 entries, 3158 to 2732
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
```

```
 0   prome                   5280 non-null   float64
 1   nivel                   5280 non-null   int64
 2   forma_ingreso           5280 non-null   int64
 3   institucion             5280 non-null   int64
 4   Edad bachillerato       5280 non-null   int64
 5   Tiempo_sabatico         5280 non-null   int64
 6   Tiempo activo           5280 non-null   int64
 7   Edad de reporte         5280 non-null   int64
 8   programa_Civil          5280 non-null   uint8
 9   programa_Diseño         5280 non-null   uint8
 10  programa_Electrica      5280 non-null   uint8
 11  programa_Electronica    5280 non-null   uint8
 12  programa_Industrial     5280 non-null   uint8
 13  programa_Mecanica       5280 non-null   uint8
 14  programa_Sistemas       5280 non-null   uint8
 15  municipio_Area          5280 non-null   uint8
 16  municipio_Desconocidos  5280 non-null   uint8
 17  municipio_MunLejanos    5280 non-null   uint8
 18  municipio_MunSant       5280 non-null   uint8
dtypes: float64(1), int64(7), uint8(11)
memory usage: 428.0 KB
```

Las variables que se presentan en la tabla son:

1. Float64 = Numero decimal.
2. Int64 = Numero entero.
3. Uint8 = Variable binaria.

Donde se observa que el número de columnas de las variables independientes del modelo pasan de ser 10 a 19 por la transformación de las variables de programa académico y municipio de procedencia de los estudiantes.

Para los tres modelos predictivos se utilizaron los datos transformados con One Hot Encoding, teniendo el mismo Data Set (conjunto completo de datos).

Hiperparámetros del modelo árbol de decisión:

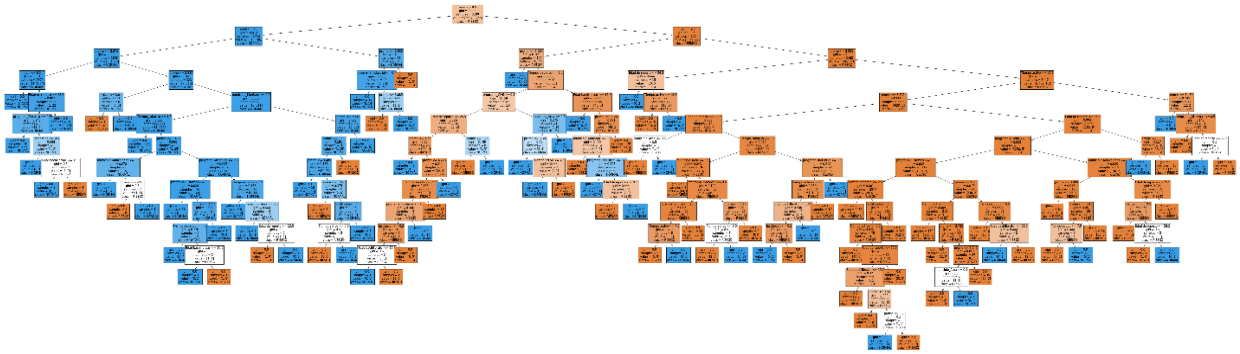
La cantidad máxima de características (max_features): son las variables independientes tomadas por el modelo árbol de decisión, que en esta investigación serán las 10 columnas de los datos transformadas en 19.

La profundidad máxima (max_depth): es el total de los niveles del árbol; este hiperparámetro se dejó abierto a que el algoritmo plasmara el máximo de niveles posibles y eligió 14 niveles de profundidad.

En el modelo se utilizaron el 80 % de los datos (5280) como entrenamiento y el 20 % restante (1320) como prueba de la precisión del árbol de decisión. A continuación, se comparte el árbol de decisión obtenido.

Figura 23

Árbol de Decisión



Nota. Elaboración propia.

En el primer nivel del árbol de decisión (adjunto en los apéndices) el modelo toma como decisión el nivel del estudiante que es una variable categórica, si el nivel del estudiante es mayor o igual a 7.5 ≈ 8 el estudiante estará en el estado normal, entre los datos de entrenamiento (80 %) el árbol de decisión clasifico 3527 estudiantes en estado normal y 1753 en riesgo; la segunda

característica del árbol para definir la situación del estudiante es el promedio, donde si un estudiante tiene un promedio menor o igual a 3.185 ≈ 3.2 el estudiante se encontrará en riesgo.

Otros hallazgos importantes para resaltar son los encontrados en los nodos que toman a consideración el tiempo sabático del estudiante, donde si el tiempo es mayor a dos años el estudiante tiene mayor probabilidad de estar en riesgo.

La tabla de Feature Importance contiene la información de la importancia de las variables independientes sobre el modelo predictivo.

Tabla 9

Importancia de variables árbol de decisión

#	feat	importance
1	nivel	0.958210825
0	prome	0.015733969
6	Tiempo activo	0.005214758
7	Edad de reporte	0.004319351
4	Edad bachillerato	0.00282612
5	Tiempo_sabatico	0.002721737
18	municipio_MunSant	0.002159486
3	institucion	0.001853828
8	programa_Civil	0.001302495
14	programa_Sistemas	0.00100334
12	programa_Industrial	0.000916679
17	municipio_MunLejanos	0.00075077
13	programa_Mecanica	0.000747881
9	programa_Diseño	0.000732308
15	municipio_Area	0.000489814
2	forma_ingreso	0.000426989
11	programa_Electronica	0.000426989
10	programa_Electrica	0.000162662
16	municipio_Desconocidos	0

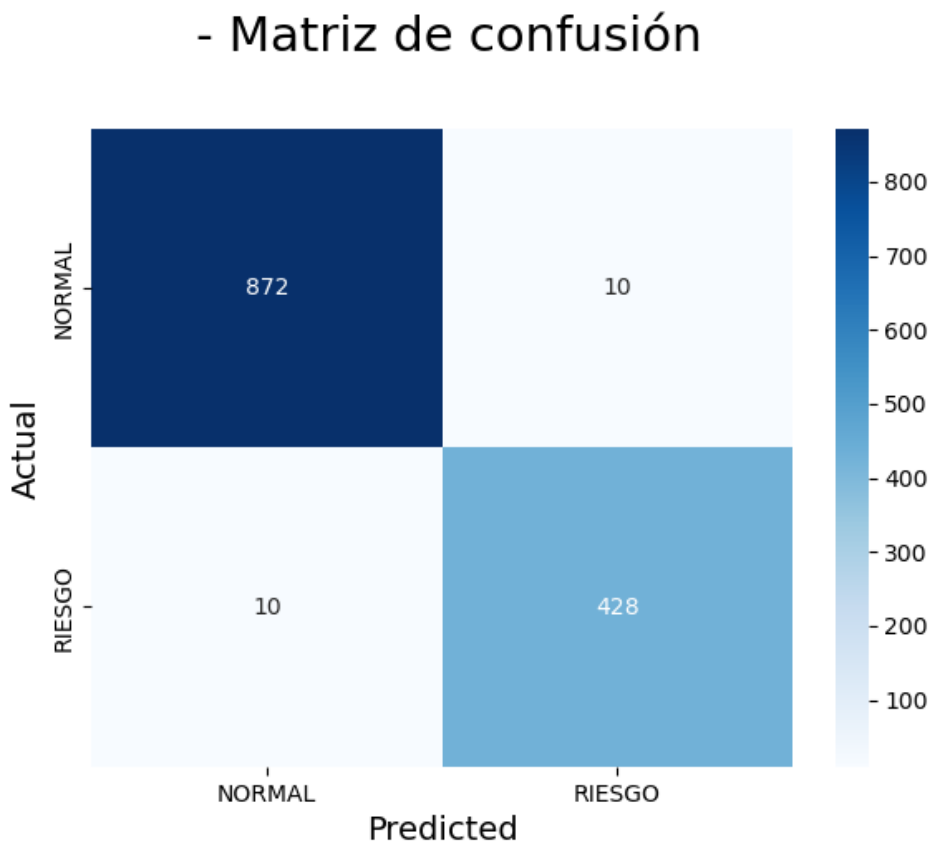
Nota. Elaboración propia.

De la tabla 9 se deduce que la única variable que tiene influencia en la condición de un estudiante es la variable categórica del nivel; esto concuerda con los hallazgos realizados en el análisis multivariable de los clústeres.

Próximamente se presenta la matriz de confusión del modelo.

Figura 24

Matriz de confusión del Árbol de Decisión



Nota. Elaboración propia.

Según la matriz de confusión, el modelo de árbol de decisión categorizo correctamente 872 estudiantes de los 882 que eran estudiantes con normalidad académica y 428 de los 438 que estaban

en riesgo según sus características. Según lo obtenido anteriormente se halla que la precisión global del árbol de decisión es `0.986363` (98.63 %).

7.3.2. Modelado 2 Random Forest.

Hiperparámetros del modelo Random Forest:

Para el modelado con bosques aleatorios o Random Forest también se utilizaron el 80 % de los datos para entrenamiento y el 20 % para prueba del modelo.

Números de estimadores (n_estimators): 250, que son el número de árboles en el bosque aleatorio.

Máximas características (max_features): son las variables independientes tomadas por el modelo árbol de decisión, que en esta investigación serán las 10 columnas de los datos transformadas en 19 columnas.

Mínimo de muestras necesarias (min_samples_leaf): es el número de ejemplo mínimo para que un nodo sea considerado hoja, para este caso se eligió 16.

A continuación, se comparte la tabla de Feature Importance del modelo de bosques aleatorios.

Tabla 10

Importancia de variables Random Forest.

#	feat	importance
1	nivel	0.420042177
6	Tiempo activo	0.245420473
0	prome	0.187216902
7	Edad de reporte	0.12529818
12	programa_Industrial	0.00500875

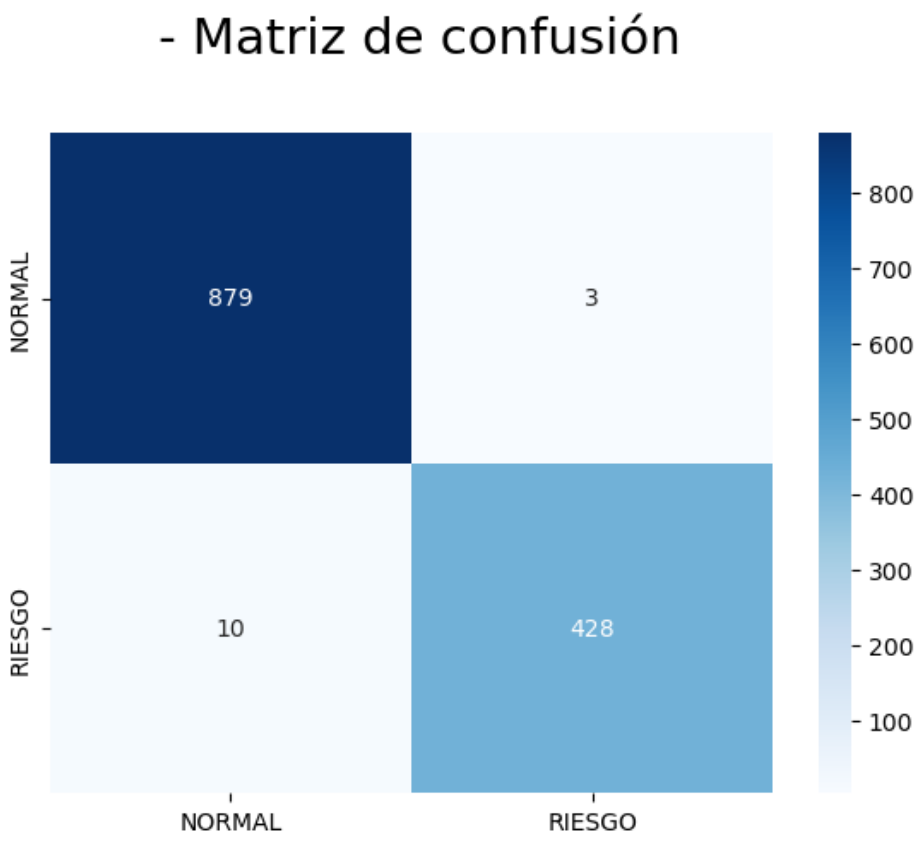
15	municipio_Area	0.002838556
14	programa_Sistemas	0.002668224
5	Tiempo_sabatico	0.002577936
4	Edad bachillerato	0.002068453
2	forma_ingreso	0.001833049
11	programa_Electronica	0.001797051
8	programa_Civil	0.001245939
18	municipio_MunSant	0.000984129
3	institucion	0.000261835
9	programa_Diseño	0.000220717
10	programa_Electrica	0.000205826
13	programa_Mecanica	0.000193911
17	municipio_MunLejanos	0.000083677
16	municipio_Desconocidos	0.000034216

Nota. Elaboración propia.

Observando la importancia de las variables dentro del modelo de Random Forest y al hacer la comparación con el modelo de árbol aleatorio se deduce que, en el primer modelo toma sólo una variable (nivel del estudiante, importancia de 95.82%) para clasificar los estudiantes en riesgo o en normalidad académica, mientras que el modelo de Random Forest tiene cuatro variables que tienen un importancia significativa dentro de la clasificación de los estudiantes (Nivel del estudiante 42 %, Tiempo activo en la universidad 24.54 %, promedio ponderado acumulado 18.72 % y edad de reporte 12.53 %).

Figura 25

Matriz de confusión bosques aleatorios



Nota. Elaboración propia.

Según la matriz de confusión, el modelo de Random Forest categorizo correctamente 879 estudiantes de los 882 que eran estudiantes con normalidad académica y 428 de los 438 que estaban en riesgo según sus características. Según lo obtenido anteriormente se halla que la precisión global del árbol de decisión es `0.9902` (99.02 %). A comparación del primer modelo de árbol aleatorio tuvo una mejora de 0.4 % aproximadamente en la clasificación correcta de los estudiantes en los dos grupos respuesta.

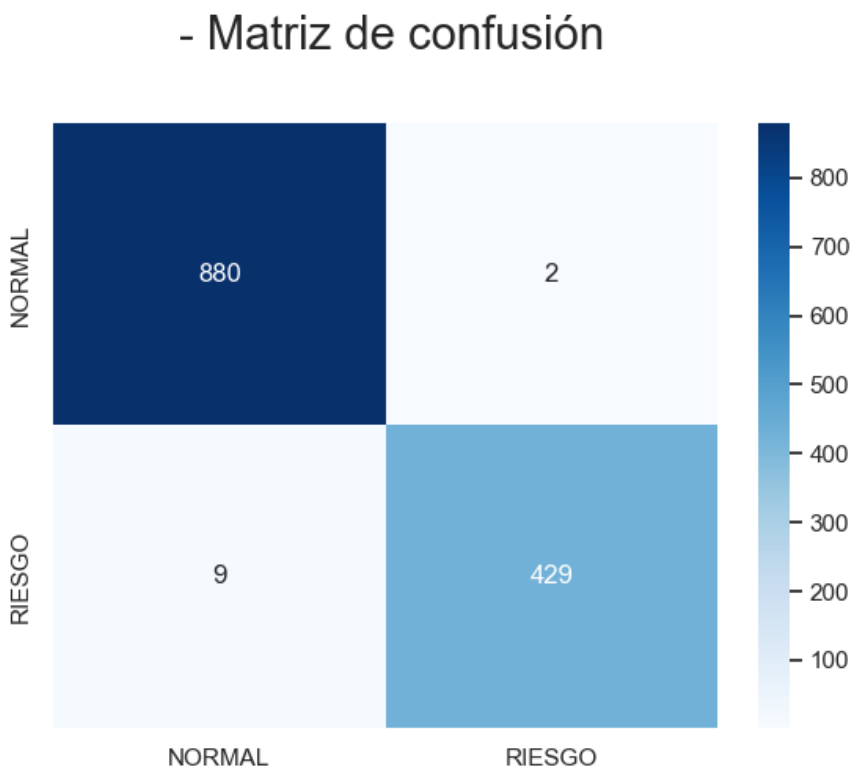
7.3.3. Modelado 3 Regresión Logística.

Hiperparámetros del modelo de Regresión Logística:

Máximas iteraciones (max_iter): Número máximo de iteraciones necesarias para que los solucionadores converjan, que en el caso específico de la investigación es de 5000 iteraciones.

Figura 26

Matriz de confusión Regresión Logística



Nota. Elaboración propia.

En la matriz de confusión de la Regresión Logística clasifico correctamente 880 estudiantes de 882 como estudiantes con normalidad académica y 429 de 438 en riesgo de deserción. Comparado con los dos modelos anteriores, el modelo de Regresión Logística tiene una precisión global de `0.99166` 99 (99.17 %); contra el primer modelo presenta un aumento en la precisión de 0.54 % y contra el segundo modelo, presento un aumento en la precisión global de 0.15 %. Lo

que infiere que el modelo de Regresión Logística presenta mejor rendimiento en su clasificación correcta de los estudiantes.

Tabla 11

Coeficientes de variables Regresión Logística.

#	Feat	p-values	Coef	Odds
1	nivel	0.00E+00	-8.09449	0.00031
0	prome	1.40E-70	-2.09052	0.12362
4	Edad bachillerato	5.78E-02	-0.23579	0.78995
5	Tiempo_sabatico	1.64E-03	-0.0726	0.92997
14	municipio_Desconocidos	1.49E-02	-0.04272	0.95818
15	municipio_MunLejanos	1.09E-02	-0.02933	0.9711
3	institucion	4.33E-02	-0.02319	0.97708
10	programa_Electronica	1.72E-26	-0.00355	0.99646
16	municipio_MunSant	4.75E-23	0	1
2	forma_ingreso	4.78E-31	0	1
9	programa_Electrica	5.93E-04	0.0189	1.01908
12	programa_Mecanica	3.64E-01	0.0309	1.03138
11	programa_Industrial	8.18E-38	0.06646	1.06872
7	Edad de reporte	1.63E-263	0.09988	1.10504
8	programa_Diseño	5.18E-10	0.25771	1.29396
13	programa_Sistemas	1.48E-35	0.33251	1.39446
6	Tiempo activo	0.00E+00	0.51057	1.66624

Nota. Elaboración propia.

En el modelo de Regresión Logística obtenemos los coeficientes de importancia que son distintos a los de las tablas de Feat Importance de los modelos de árbol de decisión y bosques aleatorios, puesto que utiliza una formula donde los coeficientes de las variables son los exponentes de Euler como lo indica la siguiente la siguiente ecuación:

Figura 27

Ecuación de la Regresión Logística

MODELO PREDICTIVO PARA LA DESERCIÓN.

$$p_i = \frac{exp(\beta_0 + \sum_k \beta_k(u_i, v_i))}{1 + exp(\beta_0 + \sum_k \beta_k(u_i, v_i))}$$

Nota: Extraída de la revista Espacios.

La fórmula de la figura 27 se usa para calcular la probabilidad de que un estudiante pertenezca a una de las categorías.

De la tabla de coeficientes se deduce que las variables más significativas para el modelo predictivo y que influyen mayormente en la variable respuesta son dos variables de programa académico (programa de Diseño Industrial y Programa de Ingeniería de Sistemas) y una variable del tiempo activo que lleva un estudiante en la universidad.

Después de la normalización de las variables, el nivel de referencia del programa académico que utiliza el modelo es Ingeniería Civil, el nivel de referencia se ajustan los valores medidos en escalas comparadas a una escala en común que para esta investigación es la carrera de Ingeniería Civil contrastada con los demás programas académicos; se interpreta que un estudiante que pertenezca al programa de Diseño Industrial tiene 29.4 % más de probabilidad de estar en riesgo de deserción que un estudiante de Ingeniería Civil. Igualmente, un estudiante de ingeniería de Sistemas tiene un 39.4 % más de probabilidad de estar en riesgo que un estudiante de Ingeniería Civil.

Por último, la variable de tiempo activo del estudiante indica que los años activos máximos de un pregrado asociado a la Facultad de Ingenierías Fisicomecánicas son 5, sin embargo, hay estudiantes que llegan a tener un tiempo activo mayor a 20 años. Según la tabla de coeficientes del modelo después del quinto año activo de un estudiante, por cada año más que un estudiante está activo disminuye la probabilidad de estar en riesgo de deserción un 66 %. Al contrastar la

información obtenida en la tabla de coeficientes del modelo con lo encontrado en el análisis multivariable hecho antes del modelado, se encuentra coherencia, puesto que los primeros semestres (primero y segundo) son los que cuentan con mayor tasa de estudiantes en riesgo.

7.4. Optimización y validación de los modelos

Para la evaluación de los modelos se utilizó validación cruzada (Cross Validation). La validación cruzada busca reducir el sobreajuste (overfit) de los modelos encontrando los mejores hiperparámetros donde los modelos muestran los mejores resultados. La personalización de hiperparámetros se hace utilizando una grilla de hiperparámetros que usan una misma métrica para cada modelo en particular.

En la validación cruzada de esta investigación se tomó un N (número de grupos de prueba) que en este caso para los tres modelos fue de cinco grupos.

> La validación cruzada se utiliza en investigaciones estadísticas para comprobar la fiabilidad de un grupo de datos de prueba antes de aplicar el análisis a un grupo de datos de entrenamiento, se usa para regular los parámetros, ajustarlos y luego utilizarlos en el grupo desconocido de datos (Castro C. y Proaño A., 2022, p.50).

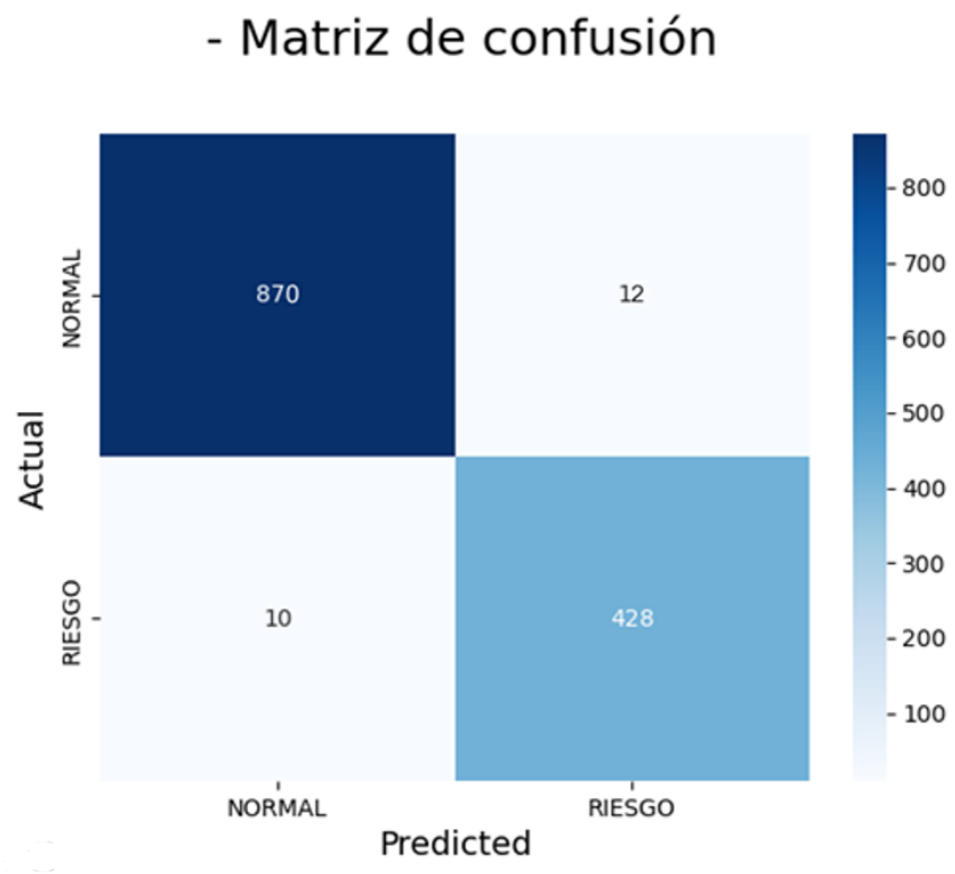
7.4.1. Modelo árbol de decisión.

Para el Cross Validation del modelo de árbol de decisión se variaron sus hiperparámetros para encontrar los que más optimizan la clasificación de los estudiantes en el modelo.

En el hiperparámetro de max_depth [8, 9, 10, 11, 12, 13, 14], cuando el árbol de decisión tiene 11 niveles de profundidad es cuando se optimiza la clasificación en el Cross Validation.

Figura 28

Matriz de confusión Cross Validation del modelo árbol de decisión



Nota. Elaboración propia.

Comparando la matriz de confusión de la Cross Validation del árbol de decisión con la matriz de confusión del árbol de decisión tenemos, que disminuyó su eficiencia de clasificar los estudiantes en un 0.23 %.

7.4.2. Modelo Random Forest

Para el Cross Validation del modelo Random Forest se probaron los siguientes hiperparámetros:

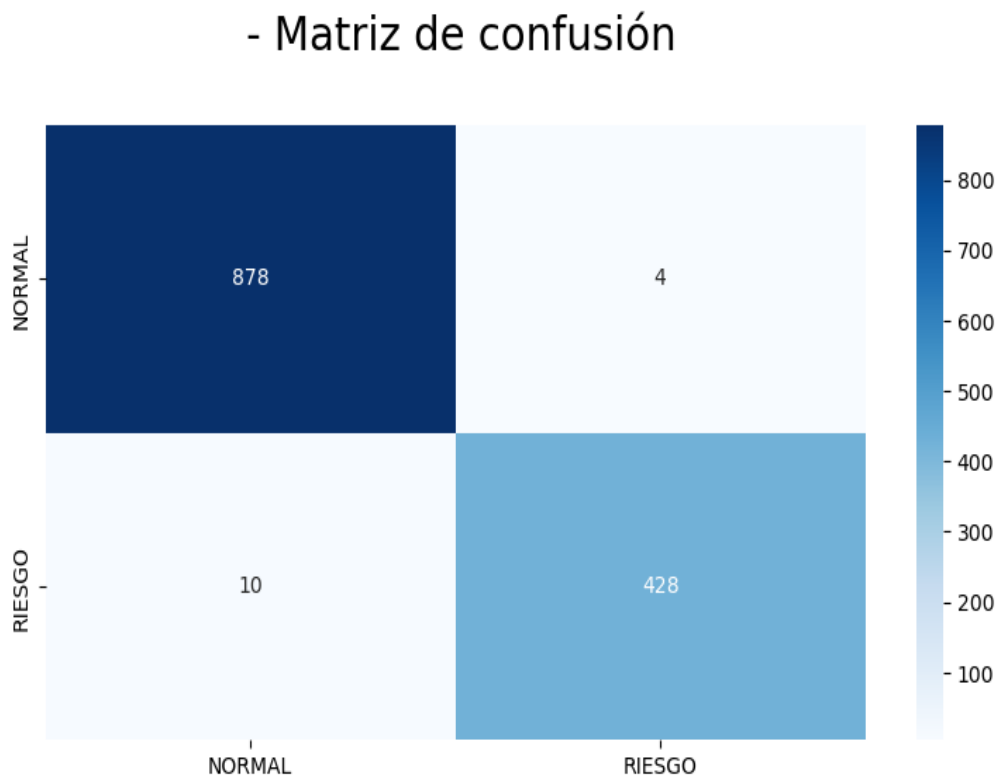
N_estimators: [50, 100, 200, 250, 500, 1000].

Min_samples_leaf: [2, 3, 4, 5, 7, 8, 16, 32, 64, 124].

De los hiperparámetros seleccionados, tenemos que los hiperparámetros que optimizan la clasificación de los estudiantes son n_estimators = 100 y min_samples_leaf = 7. De estos dos hiperparámetros surge la siguiente matriz de confusión.

Figura 29

Matriz de confusión Cross Validation del modelo Random Forest



Nota. Elaboración propia.

Comparando la matriz resultante de la validación cruzada del modelo Random Forest con la matriz del modelado sin optimizar los hiperparámetros, se tiene que disminuyo su eficiencia en la clasificación de los estudiantes en un 0.12 %.

7.4.3. Modelo de Regresión Logística

El solucionador en el Cross Validation del modelo de Regresión Logística es Saga porque es el único que permite hacer Elastic-Net.

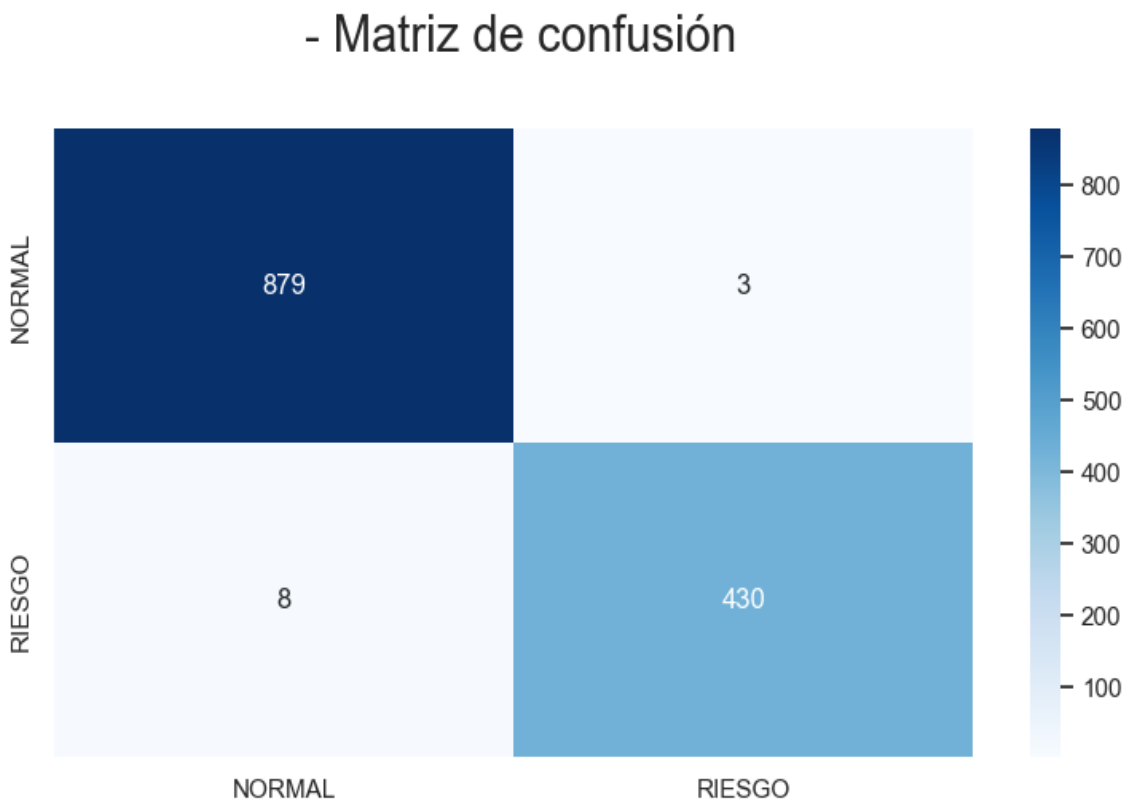
> Elastic-Net es un modelo de regresión que normaliza el vector de coeficientes con las normas L1 y L2. Esto permite generar un modelo en el que solo algunos de los coeficientes sean no nulos, manteniendo las propiedades de regularización de Lasso (L1), de Ridge (L2) o la regularización L1+L2 (Elastic Net). (Barrueco D. 2018).
>
> La regularización incorpora penalizaciones en el ajuste por mínimos cuadrados ordinarios (OLS) con el objetivo de evitar overfitting, reducir varianza, atenuar el efecto de la correlación entre predictores y minimizar la influencia en el modelo de los predictores menos relevantes. Por lo general, aplicando regularización se consigue modelos con mayor poder predictivo (Amat J. 2020).

Overfitting o sobreajuste es un comportamiento de aprendizaje automático no deseado que se produce cuando el modelo de aprendizaje automático proporciona predicciones precisas para los datos de entrenamiento, pero no para los datos nuevos (Horrillo I. y Barrena M. 2008).

En el caso de la Cross Validation del modelo, este uso las propiedades de la regularización de Lasso (L1).

Figura 30

Matriz de confusión Cross Validation de Regresión Logística



Nota. Elaboración propia.

Al comparar la matriz de confusión de la Cross Validation de la Regresión Logística con la matriz de confusión de la Regresión Logística se tiene que, la matriz obtuvo el mismo porcentaje de precisión al clasificar los estudiantes.

7.5.Elección del modelo predictivo

Lo que hace el Cross Validation es tomar N grupos de datos que fueron 5, toma el primer grupo (20 %) de datos y utilizarlos para entrenamiento y el restante de datos (80 %) los utiliza para probar el modelo, por eso en los resultados obtenidos en los modelos de árbol de decisión y de Random Forest se evidencia una disminución en la efectividad del modelo al clasificar los estudiantes. Esto puede señalar problemas de overfit de los dos primeros modelos sin aplicar la

Cross Validation, después de aplicarla se eliminan estos problemas y generalizan mejor los modelos.

El modelo de Regresión Logística fue el que mantuvo su nivel de precisión al hacer la comprobación con Cross Validation, por ello es el modelo que mostró mejores resultados fue el de Regresión Logística.

8. Discusión

Contrastando los resultados de la presente investigación con la teoría consultada en la primera fase del proyecto tenemos que:

En el análisis preliminar de los datos, se encontró que los estudiantes que entran a la universidad por alguna de las modalidades especiales de ingreso tienen un alto riesgo de desertar de sus estudios universitarios, esto concuerda con lo encontrado en la literatura donde las personas con entornos socioeconómicos más vulnerables tienen más probabilidades de abandonar los estudios que las personas con entornos más desfavorecidos; como lo son las personas de comunidades afrocolombianas, comunidades indígenas y víctimas del conflicto armado.

Con respecto a las características individuales de un estudiante, la teoría indica que el desempeño académico de un estudiante en sus estudios secundarios está relacionado con el desempeño que tendrá en los estudios de educación superior; los primeros dos modelos señalaron que el nivel y el promedio del estudiante influyen en la condición del mismo. Igualmente, en la revisión bibliográfica realizada al inicio de la investigación se encontró que el desempeño de un estudiante durante el primer año académico de los estudios de educación superior es clave para la deserción de sus estudios; esto se evidencia en el análisis de los datos, donde la deserción estudiantil se presenta principalmente en los dos primeros semestres de la vida universitaria.

9. Conclusiones

El interés de encontrar soluciones que contribuyan a la disminución de la deserción estudiantil es algo que comparten todas las naciones alrededor del mundo, con el constante avance de la tecnología en la predicción de estudiantes en riesgo de deserción y los estudios compartidos por países de Latinoamérica, hace que haya un precedente histórico de información.

Aunque la Universidad Industrial de Santander cuente con programas y áreas enfocadas en disminuir la deserción estudiantil, se evidencia según esta investigación que la tasa de deserción es alta viendo los datos de la Facultad de Ingeniería Fisicomecánicas.

Según los datos presentados en la investigación se observa que la deserción está directamente relacionada con la inequidad y afecta a personas que viven en entornos socioeconómicos vulnerables; aunque la universidad tenga programas de acceso a la educación para personas que pertenecen a estos entornos o grupos vulnerados, no logra retener estos estudiantes lo que conlleva a que realmente el programa no cumpla el objetivo de que la igualdad sea real y efectiva (Acuerdo No. 282 de 2017, Universidad Industrial de Santander).

Teniendo en cuenta lo hallado con respecto a los programas académicos adjuntos a la Facultad de Ingeniería Fisicomecánicas, los programas de ingeniería electrónica, ingeniería eléctrica, ingeniería mecánica e ingeniería de sistemas presentan altas tasas de deserción y de cancelaciones de matrícula. Lo que concuerda con la hipótesis 3 (H3) presentada en el marco teórico.

Así mismo las características que evidencian relación con el riesgo de un estudiante a desertar sus estudios dentro de los modelos analizados son principalmente el promedio, el nivel y el tiempo activo del estudiante; si el promedio ponderado de un estudiante está entre 2.7 y 3.2 lo

pone en riesgo de condicionalidad, si es menor a 2.7 inmediatamente queda PFU (por fuera de la universidad) y se convierte en desertor; con respecto al nivel, dentro del análisis de variables se encontró que el 84.4 % de los estudiantes que quedan PFU se encuentran dentro de los primeros cuatro niveles del programa académico que para los programas de ingenierías estaría comprendido por el ciclo básico, además que entre mayor sea el nivel alcanzado por un estudiante menos es la probabilidad de que este en riesgo de desertar sus estudios, esto se relaciona con el tiempo activo del estudiante.

Finalmente se concluye que el modelo que mejor predice la probabilidad que tiene un estudiante de estar riesgo de desertar sus estudios luego de la optimización por medio de Cross Validation es el modelo de Regresión Logística dado que refleja mejores resultados a la hora del tuneo de sus hiperparámetros y presenta menos sobre ajuste manteniendo la efectividad como predictor.

10. Recomendaciones

Para futuras investigaciones, existen oportunidades de extender la presente investigación, utilizando datos de todos los estudiantes de pregrado de la Universidad Industrial de Santander, solicitando características adicionales que puedan estar relacionadas al riesgo de desertar los estudios.

Incentivar a los estudiantes a investigar sobre la deserción estudiantil, que la universidad y la Dirección de Admisiones y Registro Académico conceda acceso a la información no sensible de los estudiantes para un estudio más profundo sobre la relación de las características de un estudiante con la posible deserción de sus estudios.

Considerar la posibilidad de tener un sistema de acceso abierto a los datos de los estudiantes de pregrado para que en futuras investigaciones se puedan utilizar variables adicionales como el género del estudiante, nivel económico de los padres, nivel educativo máximo alcanzado por los padres, promedio del último grado del colegio, entre otras que pudieran tener una relación con la deserción.

Referencias bibliográficas

Acuerdo No. 282 de 2017. Por el cual se dictan disposiciones sobre el ingreso a la Universidad de aspirantes por la modalidad de Admisiones Especiales. 7 de noviembre de 2017. Consejo académico de la Universidad Industrial de Santander.

Agudelo, G., Franco L., Franco L. E. (2017). Tiempo necesario para rentar una vivienda: una aplicación de regresión logística geográficamente ponderada. *Revista Espacios*. 38(18), 23. https://www.revistaespacios.com/a17v38n18/1738182w3.html

Alban, M.; Mauricio, D. (2019). Neural Networks to Predict Dropout at the Universities. *International Journal of Machine Learning and Computing*. 9 (2), 149–153. http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=84&id=905

Amat, J. (2020). *Regularización Ridge, Lasso y Elastic Net con Python*. Ciencia de datos. https://cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python#:~:text=La%20regularizaci%C3%B3n%20Ridge%20penaliza%20la,que%20estos%20lleguen%20a%20cero.

Barbosa-Camargo, MI; García-Sánchez, A.; Ridao-Carlini, ML. (2021). *Desigualdad y Deserción en la Educación Superior en Colombia. Un análisis multinivel de las diferencias regionales, las instituciones y el campo de estudio*. https://doi.org/10.3390/math9243280

Barrueco, D. (s.f.). *Elastic Net.* Interactive Chaos. https://interactivechaos.com/es/manual/tutorial-de-machine-learning/elastic-net

Bean, J.P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education.* (17), 291–320. https://doi.org/10.1007/BF00977899

Bejarano, L.; Arango, S.; Johana, K.; Durán, H.; Ortiz, C. (2017). Caso de estudio: Caracterización de la deserción estudiantil en la Fundación Universitaria Los Libertadores 2014-1–2016-1. *Tesis Psicológica*. (12), 138–161.

Boshier, R. (1973). Educational participation and Dropout: a Theoretical Model. *Adult Education*. 23 (4), 255-282. https://doi.org/10.1177/074171367302300401

Broc, M. (2011). Voluntad para estudiar, regulación del esfuerzo, gestión eficaz del tiempo y rendimiento académico en alumnos universitarios. *Revista de Investigación Educativa*. (29), 171–185.

Cabrera, E.; Díaz, E. (2021). *Manual de uso de Jupyter Notebook para aplicaciones docentes*. Universidad Complutense de Madrid.

Cabrera, L.; Bethencourt, J.; Álvarez, P.; González, M. (2006). El problema del abandono de los estudios universitarios. [The dropout problem in university study]. *Revista Electrónica de Investigación y Evaluación Educativa*. (12), 171–203. https://www.redalyc.org/pdf/916/91612201.pdf

Camacho M., Montalvo A., Galezo P. (2019). Determinantes de la Deserción estudiantil en estudiantes universitarios. *Panorama Económico*. 27 (1), 134-162. https://revistas.unicartagena.edu.co/index.php/panoramaeconomico/article/download/2621/2198/5703

Castro Martínez, C. D., & Proaño Indacochea, A. G. (2022). *Determinación de autoría de textos en español mediante análisis estilométrico de palabras de uso frecuente y validación cruzada para Machine Learning*. (Tesis de pregrado). Universidad de Guayaquil, Guayaquil, Ecuador.

Cox, D.R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Serie B (Methodol).* 20 (2), 215–232. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1958.tb00292.x

De Oliveira, C.F.; Sobral, S.R.; Ferreira, M.J.; Moreira, F. (2021). How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. *Big Data and Cognitive Computing*. 5 (4), 64. https://doi.org/10.3390/bdcc5040064

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*. 49 (4), 498–506.

Datos Mundial. (2021). *Las 50 economías más grandes del mundo*. https://www.datosmundial.com/economias-mas-grandes.php

Duran, C.; Sánchez, L. (2020). *Evaluación de impacto de jóvenes en acción como política para la reducción de la deserción; caso Universidad Industrial de Santander*. (Tesis de pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia.

Dúran, J. (2017). *Prevalencia de problemas de salud mental en los estudiantes UIS en riesgo de deserción por bajo rendimiento académico.* (Tesis de pregrado). Universidad Industrial de Santander, Bucaramanga, Colombia.

Eckert, K.; Suenaga, R. (2014). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Formación Universitaria.* 8 (5), 3–12.

Fischer, E. (2012). *Modelo Para la Automatización del Proceso de Determinación de Riesgo de Deserción en Alumnos Universitarios*. (Tesis de maestría), Universidad de Chile, Santiago, Chile.

Friedman, J. (2002). Stochastic gradient-boosting. *Computational Statistics & Data Analysis*. 38 (4), 367–378. https://www.sciencedirect.com/science/article/pii/S0167947301000652

Hackman, J. & Dysinger, W. S. (1970). Commitment to College as a Factor in Student Attrition. *American Sociological Association.* 43 (3), 311-324. https://www.jstor.org/stable/2112069?origin=crossref

Heredia, D.; Amaya, Y.; Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. IEEE Lat. Am. Trans. 13, 3127–3134. https://ieeexplore.ieee.org/document/7350068

Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación*. 17, 91-108. https://doi.org/10.31619/caledu.n17.409

Iberdrola S. A. (s.f.). *Análisis predictivo, una manera de adelantarse al futuro de la mano de las nuevas tecnologías.* https://www.iberdrola.com/innovacion/analisis-predictivo

Kumar, S.; Bharadwaj, B.; Pal, S. (2012). Mining Education Data to Predict Student's Retention: A comparative Study). *International Journal of Computer Science and Information Security*. 10, 113–117

Kumar, B.; Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *International Journal of Computer Science and Information Security*. 2, 686–690.

Lee, S.; Chung, J. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences*. 9 (15), 3093. https://doi.org/10.3390/app9153093

Matheu, A.; Ruff, C.; Ruiz, M.; Benites, L.; Morong, G. (2018). Modelo de predicción de la deserción estudiantil de primer año en la Universidad Bernardo O'Higgins. *Educação e Pesquisa*, 44. https://dialnet.unirioja.es/servlet/articulo?codigo=7315140

Morales, E. (2009). *Descubrimiento de conocimiento en bases de datos*. http://ccc.inaoep.mx/~emorales/Cursos/KDD/principal.html

Ojeda, V.; Fernández, J. Isea, R. Gutiérrez, A. Salazar, V. (2018). *Coeficiente V de Cramer (V).* Facultad de humanidades y educación. Universidad Central de Venezuela.

Opazo, D.; Moreno, S.; Álvarez-Miranda, E.; Pereira, J. (2021). Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities. *Mathematics*. 9 (20), 2599. https://doi.org/10.3390/math9202599

Pascarella, E. T. & Terenzini, P. (1977). Patterns of Student-Faculty Informal Interaction Beyond the Classroom and Voluntary Freshman Attrition. *The Journal of Higher Education*, 48 (5), 540-562.

Pascarella, E. T. & Terenzini, P. (1991) *How College Affects Students: Findings and Insights from Twenty Years of Research*; Jossey-Bass Publishers: San Francisco, CA, USA

Páez, L. (2021). *¿Qué es Notion? La mejor app de productividad que organizará tu vida.* https://www.crehana.com/blog/negocios/que-es-notion/

Quintero, I. (2016). *Análisis de las Causas de Deserción Universitaria*. (Tesis de maestría). Universidad Nacional Abierta y a Distancia UNAD, Colombia, Bogotá, Colombia.

Sinchi, E.; Ceballos, G. (2018). Acceso y deserción en las universidades. Alternativas de financiamiento. *Alteridad*. 13, 274–287. https://revistas.ups.edu.ec/index.php/alteridad/article/view/2.2018.10

Siri, A. (2015) Predicting Students' Dropout at University Using Artificial Neural Networks. *Italian Journal of Sociology of Education*. 7, 225–247.

Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior. (2014). *Informes determinantes de la deserción*. https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_Informe_determinantes_desercion.pdf

Soto, G. (2011). El teorema de Bayes. *Revista de Educación Matemática.* 26(3), 3-25. https://revistas.unc.edu.ar/index.php/REM

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1, 64–85. http://www.jstor.org/stable/1084192

System Development Corporation. *Progress Report on the Evaluation of Special Service in Higher Education*. Santa Monica: System Development Corporation, 1981.

Tinto, V. (1975). *Dropout from Higher Education: A Theoretical Synthesis of Recent Research.* 45(1), 89–125. https://doi.org/10.3102/00346543045001089

Valero, S.; Salvador, A.; García, M. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*. 779(73), 33. https://www.academia.edu/download/34203825/e1.pdf

Villanueva, J. A. (2019, 2 abril). *Los mejores sistemas nacionales de educación superior de 2019 – Oficina de Acreditación y Calidad*. Oficina de Acreditación y Calidad, Universidad Nacional de Ingeniería. https://acreditacion-fiis.com/los-mejores-sistemas-nacionales-de-educacion-superior-de-2019/

Zahra, Ghazanfar, M. A., Khalid, A., Azam, M. A., Naeem, U., & Prugel-Bennett, A. (2015). Novel centroid selection approaches for KMeans-clustering based recommender systems. *Information Sciences*, 320, 156–189. https://doi.org/10.1016/j.ins.2015.03.062

Zhang, G. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 30 (4), 451–462. https://doi.org/10.1109/5326.897072