Análisis de la predicción temprana de los factores determinantes en la deserción de la educación superior desde el enfoque de Machine Learning.
Un análisis desde el contexto colombiano

Jaminson Enrique Herrera Flórez

Asesor

Andrés Felipe Hernández Giraldo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Nota de Aceptación

_____

Nombre director de Trabajo de Grado

_____ _____

Jurado Jurado

2025

Dedicatoria

Primeramente, a Dios por darme la fortaleza y la sabiduría; a mi madre María de Lourdes Flórez por haberme enseñado valores éticos y morales y el deseo por la superación académica, a mí esposa Marelbis Silvera por el apoyo incondicional y a mis hijos quienes son mi mayor motivación para lograr cada una de las metas propuestas.

Agradecimientos

A mis docentes por su motivación y dedicación y por compartir sus conocimientos.

De Manera especial quiero agradecer a mi asesor Andrés Felipe Hernández Giraldo, por su acompañamiento y su valioso aporte en el desarrollo de este trabajo de grado.

Resumen

La deserción en la educación superior es un problema creciente en Colombia, con tasas que afectan significativamente la calidad educativa y el desarrollo del país. Estudios previos han identificado múltiples factores y variables asociados a este fenómeno, incluyendo dificultades económicas, bajo rendimiento académico, falta de apoyo institucional, problemas de adaptación al entorno universitario y problemas con el entorno familiar entre otros.

Es por ello por lo que los objetivos de esta monografía se enfocan en analizar la deserción universitaria por medio de una revisión sistemática bibliográfica referente a la problemática abordada desde el machine learning y posteriormente clasificar las técnicas y métricas más utilizadas en esas revisiones, identificando limitaciones asociadas con la implementación de esas técnicas en la detección de la deserción universitaria.

Para llevar a cabo esa revisión sistemática de la bibliografía, se implementa una metodología mixta de tipo documental que se obtiene de fuentes secundarias apoyándose en el protocolo PRISMA. La metodología se divide en 3 fases, Exploratoria, Análisis documental y Análisis de resultados.

Los resultados permiten obtener de la revisión sistemática, las métricas y las técnicas más implementadas y las de mayor rendimiento. En estos resultados también se arroja un análisis de los factores y variables más recurrentes dentro de la revisión y por ende los de mayor incidencia en la deserción estudiantil. Posteriormente se específica una serie de factores que limitan la implementación de las técnicas de machine learning al predecir las variables de más peso en la causalidad de la problemática planteada.

Finalmente, las conclusiones reflejan el uso de machine learning como herramienta esencial e imprescindible en la predicción temprana de los factores que determinan la deserción

en la educación superior. Del mismo modo describe el modelo de mayor precisión y la métrica más utilizada para evaluar esas técnicas de machine learning.

***Palabras clave*:** Machine Learning, College Dropouts, Higher Education, Predictions, Data Science.

Abstract

Dropout in higher education is a growing problem in Colombia, with rates that significantly affect educational quality and the country's development. Previous studies have identified multiple factors and variables associated with this phenomenon, including economic difficulties, low academic performance, lack of institutional support, problems of adaptation to the university environment and problems with the family environment, among others.

That is why the objectives of this monograph are focused on analyzing college dropout through a systematic review of literature on the problem addressed from machine learning and then classifying the techniques and metrics of machine learning techniques most used in these reviews, identifying limitations associated with the implementation of these techniques in the detection of college dropout.

To carry out this systematic review of the literature, a mixed methodology of documentary type is implemented, which is obtained from secondary sources supported by the PRISMA protocol. The methodology is divided into 3 phases, Exploratory, Documentary analysis and Analysis of results.

The results allow us to obtain from the systematic review, the metrics and techniques most implemented and those with the highest performance. These results also provide an analysis of the most recurrent factors and variables within the review and therefore those with the greatest impact on student desertion. Subsequently, a series of factors that limit the implementation of machine learning techniques in predicting the most important variables in the causality of the problem raised are specified.

Finally, the conclusions reflect the use of machine learning as an essential and indispensable tool in the early prediction of the factors that determine dropout in higher

education. It also describes the most accurate model and the metric used by the authors to evaluate these machine learning techniques.

Keywords**:** Machine Learning, College Dropouts, Higher Education, Predictions, Data Science.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La deserción en la educación superior es una problemática de orden global que afecta a todos los estratos sociales, siendo el estrato bajo el más afectado, con unas implicaciones significativas tanto para los estudiantes como para el desarrollo socioeconómico de la nación. La Identificación temprana de los estudiantes en riesgo de abandono estudiantil es esencial para garantizar e implementar estrategias de permanencia.

En el contexto colombiano este fenómeno es preocupante por la alta tasa de deserción estudiantil la cual está dada especialmente por factores socioeconómicos. El Ministerio de Educación, SPADIES y el Informe LEE No. 74 de la Universidad Javeriana, muestran la dimensionalidad de la problemática.

El uso de técnicas de machine learning emergen como una potente herramienta de aporte a la solución que permite predecir los factores que contribuyen a la deserción estudiantil. Sobre esta problemática son diversos los estudios que han explorado la aplicación de las técnicas de machine learning sobre esta problemática. Por ejemplo, Hoyos Osorio y Daza Santacoloma (2023) desarrollaron un sistema de alerta temprana basado en machine learning para identificar a estudiantes de primer semestre con alto riesgo de deserción, logrando una sensibilidad del 61.97% en la predicción de estudiantes "en riesgo".

Asimismo, Kemper et al. (2020) implementaron enfoques de regresión logística y árboles de decisión para predecir la deserción estudiantil en el Karlsruhe Institute of Technology, alcanzando precisiones de hasta el 95% después de tres semestres.

Pues bien, debido a lo anterior es que se formula este interrogante ¿Es relevante y a la vez pertinente utilizar la ciencia de datos desde el machine learning para abordar la problemática planteada?

Esta monografía se propone analizar la deserción en la educación superior, enfocándose en la aplicación de técnicas de machine learning como predictores de los factores que determinan esta problemática. Por medio de una metodología mixta de tipo documental utilizando el protocolo Prisma, se hará una revisión sistemática de estudios previos en los que han implementado algoritmos predictivos. En estos estudios se analizarán el rendimiento de cada uno de los modelos sobre las variables más influyentes en la deserción estudiantil. Por consiguiente, se examinarán los resultados y sus implicaciones en la planificación del diseño de estrategias gubernamentales e institucionales destinadas a disminuir la tasa de deserción estudiantil.

La estructura de este documento consta de diferentes secciones entre las que se destacan el planteamiento del problema, la justificación, los objetivos, un marco de referencia, los fundamentos de la investigación y un análisis de resultados, las recomendaciones y conclusiones de los hallazgos.

Finalmente, lo que se pretende con esta monografía es incentivar a otros a seguir avanzando sobre el estudio de la deserción estudiantil en la educación superior en pro de mejorar los índices de este fenómeno utilizando las técnicas del aprendizaje automático (ML).

Descripción del Problema

De acuerdo con Tinto (1982) y Giovagnoli (2002), “la deserción es la problemática que enfrenta el estudiante al ingresar a una institución de educación superior y no logra graduarse como profesional durante su ciclo educativo, considerando desertor al estudiante que no es activo durante dos semestres consecutivos”.

Planteamiento del Problema

La deserción en la educación superior colombiana es una problemática que afecta significativamente al desarrollo académico y profesional de los estudiantes, así como al progreso socioeconómico del país. Según un estudio del Laboratorio de Economía de la Educación (LEE) de la Universidad Javeriana, cinco de cada diez estudiantes que ingresan a la educación superior no logran graduarse después de 15 semestres. Este fenómeno es más frecuente entre estudiantes de estratos socioeconómicos bajos, hombres y aquellos con una preparación deficiente en la educación media (U. Javeriana, 2022).

Diversos factores contribuyen a esta situación, la pobreza es uno de los principales, debido a que, “limita el acceso y la permanencia de los estudiantes en el sistema educativo. Además, la violencia y la falta de oportunidades laborales también influyen en la decisión de abandonar los estudios” (Garzón, J., 2023).

Las instituciones de educación superior (IES) también presentan diferencias en las tasas de deserción. Las instituciones tecnológicas registran una tasa de deserción del 18,2%, mientras que en las universidades es del 7,7%. Asimismo, las IES oficiales presentan tasas de deserción más altas en comparación con las privadas (U. Javeriana Inf. LEE, 2022).

Este problema no solo afecta a los individuos que abandonan sus estudios, sino que también tiene repercusiones en el mercado laboral y en la economía del país. La falta de

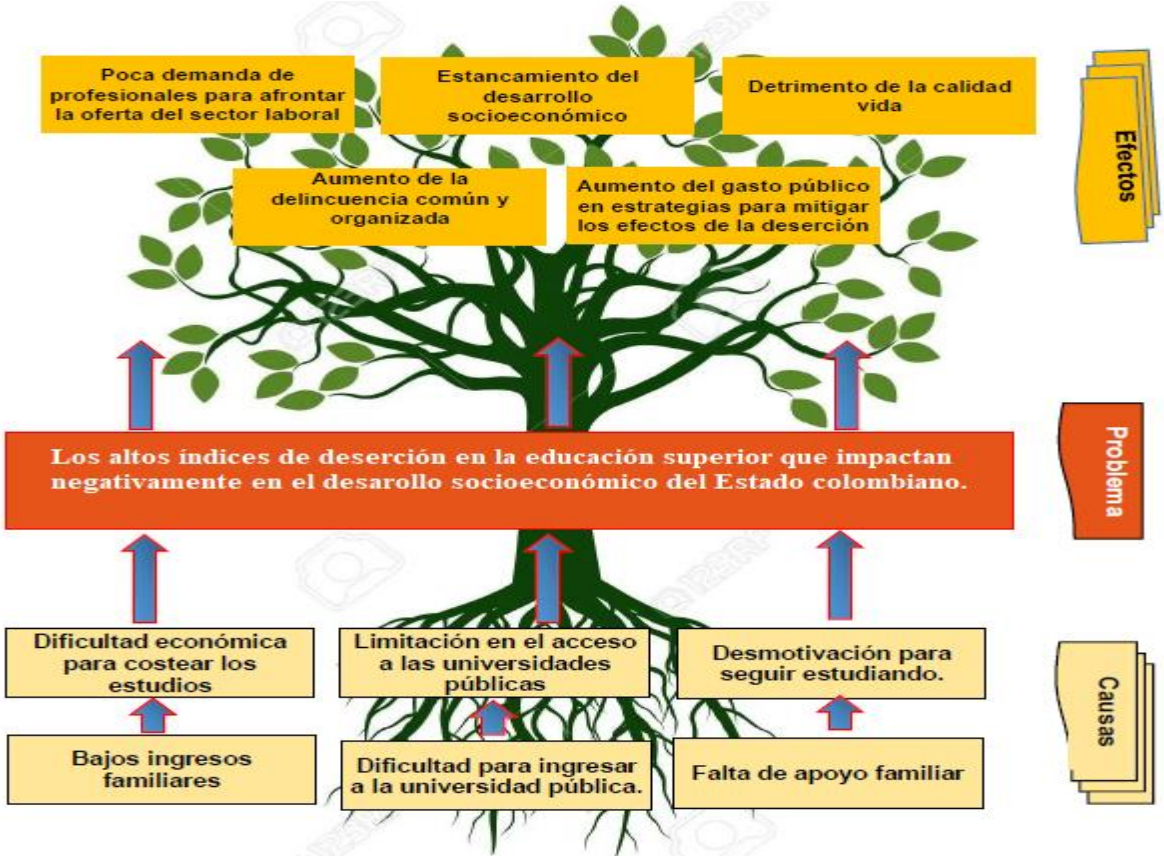
profesionales calificados dificulta la competitividad y perpetúa ciclos de pobreza generacional (Sectorial, 2024).

Según (SPADIES, 2022) entre el año 2000 y año 2021, la problemática se ha mantenido en una serie de tiempo estacionaria con una tasa de deserción entre el 12.3% y el 14%, con una reducción del 1.7% (ver gráfica 7).

En síntesis, hay que alertar de forma imperativa para que el gobierno y las instituciones educativas implementen políticas y programas que aborden las causas de la deserción, fomentando la permanencia y culminación exitosa de los estudios superiores.

Figura 1

Árbol de Problemas - Causa – Efectos



En el árbol de problemas (Ilustración 1) se plasma el contexto de la problemática desde el análisis documental. En él se puede analizar que, los factores económicos, gubernamentales y familiares, son determinantes en la causa de la deserción en la educación superior, lo que consecuentemente desencadena en una serie de efectos que disminuyen la demanda de profesionales, se aumenta la violencia, se estanca la economía, generando un detrimento social y económico en el país.

En síntesis, se justifica esta investigación desde la preocupación que genera la deserción en la educación superior y desde el impacto que causa en el desarrollo social y económico de la nación. En ese orden de ideas es imperioso que se aborde este fenómeno apoyándose en un campo multidisciplinario como lo es la ciencia de datos, lo que permitirá la predicción temprana de los factores determinantes en la deserción de la educación superior.

Pregunta Problema

Luego de una revisión sistemática bibliográfica y de la identificación de investigaciones de diversos orígenes y autores, las cuales tenían como objetivo el análisis de la deserción en la educación superior y la implementación de la ciencia como aporte a la solución por medio de la predicción temprana de los factores determinantes de la creciente problemática, nace la siguiente pregunta.

¿Es posible usar técnicas de machine learning para la predicción de factores y variables más determinantes en la deserción de la educación superior dada la dimensionalidad y la complejidad de la problemática?

Justificación

La educación es un derecho fundamental de la persona y un servicio público, así lo establece nuestra constitución política de 1991 en su Artículo 67. La educación es sin duda uno de los pilares fundamentales para ayudar a cerrar esa brecha de la desigualdad, generando bienestar económico y social, dignificando la vida misma y la de los demás, como lo afirman Sen y Nussbaum, en su libro “El desarrollo como libertad”.

Es por lo que se hace necesario prestar especial atención a esos desafíos que enfrenta la educación entre ellos la conectividad, infraestructura, nuevos métodos de la enseñanza, herramientas digitales, calidad, concentración, aprendizaje colaborativo e inversión (Salazar Q. El País, 2021).

Importancia del Problema

Desde el punto de vista de la relevancia de la problemática planteada, se determina importante analizar la deserción estudiantil desde un contexto nacional y más implícitamente desde el nivel de educación superior debido a la alta tasa de deserción, según datos del Ministerio de Educación Nacional y el Informe LEE No. 74 de la Universidad Javeriana.

Además, es relevante y pertinente, desde el impacto socioeconómico y tecnológico debido a que, el crecimiento de una nación se mide entre otros aspectos, por su PIB y por sus ingresos per cápita, pero para que ese crecimiento sea autosostenible en el tiempo es necesario que la demanda en educación superior esté al nivel de la oferta laboral (Banco de España, 2024), y es relevante desde el nivel tecnológico porque se estaría abordando esta problemática desde una herramienta que promete cambiar la forma de brindar solución a las problemáticas tanto cotidianas como empresariales, tal es el caso del machine learning, lo que le aportaría mucho a la consecución de la reducción significativa de la deserción estudiantil.

Contribución del Estudio a Través de los Resultados Obtenidos

Este trabajo contribuye desde el punto de vista investigativo a partir de una revisión bibliográfica sistemática de referentes que se han interesado por la deserción estudiantil, al fortalecimiento de la implementación desde machine learning en pro de obtener respuestas a través de predicciones tempranas de esos factores que hacen parte de la causa de la problemática planteada, beneficiando a la comunidad estudiantil, a docentes, a las instituciones y la sociedad en general.

Son diversos las investigaciones sobre la problemática abordada desde la ciencia de datos, por ejemplo, Camargo García, A. (2020), implementa la minería de datos con algoritmos de Machine Learning, detectando patrones de comportamiento de variables socioeconómicas causales de la deserción estudiantil, de igual manera la Universidad de Salamanca, (2021) en una investigación destaca la importancia de utilizar la ciencia de datos como herramienta que permite agilizar los procesos para las predicciones de los factores más influyentes en la deserción de la educación superior, aplicando diferentes técnicas entre ellas las de Machine Learning.

Objetivos

Objetivo General

Analizar las aplicaciones de las técnicas de Machine Learning en la detección de la deserción universitaria en Colombia, a través de una revisión sistemática de la literatura.

Objetivos Específicos

Clasificar las técnicas de Machine Learning en los estudios analizados sobre deserción universitaria.

Analizar los factores y variables más relevantes en la incidencia de la deserción universitaria según la literatura revisada.

Identificar las principales limitaciones y desafíos asociados con la implementación de Machine Learning en la detección de la deserción universitaria.

Marco de Referencia

Estado del Arte

En Colombia los primeros estudios que se hacían sobre el tema de deserción en la educación superior se enfocaban principalmente en estudios estáticos sobre programas específicos y no se tenía una conceptualización ni teorización del fenómeno de la deserción, es decir que no se le hacía un seguimiento al problema para observar si las políticas realizadas tenían un efecto positivo en disminuir la deserción (Elcira Solano B. - Mario Barraza N. 2018).

Gordillo y Polanco, (1970) realizaron en la Universidad Nacional de Colombia un estudio cuantitativo de la deserción, la investigación destaca que "más del 50% de los matriculados se gradúan utilizando más tiempo de lo normal". Los porcentajes más relevantes de deserción se dan en Ciencias Puras y en Ciencias Humanas, los más bajos se dan en Ciencias de la Salud; además revela que la mayor tasa de deserción se da en los dos primeros años.

Adicionalmente en un estudio de (Elcira Solano B. y Mario Barraza N., 2018) denominado "Deserción en la educación superior", el cual es un estudio apoyado por el semillero del grupo Economía de la Educación de la Universidad del Atlántico de Barranquilla, donde analizan de forma descriptiva y cuantitativa los diferentes factores, variables e indicadores que promueven la deserción de la educación superior en el Universidad del Atlántico.

Posteriormente en otro estudio de (Aníbal José Camargo, 2020) en un trabajo titulado "Predicción de Deserción de Estudiantes de Pregrado de la Universidad de la Costa (CUC)". El autor, enfoca la problemática abordándola desde la ciencia de datos, aplicando modelos de machine learning para predecir los diferentes factores que promueven la deserción en la

educación superior. Para lo cual desarrolla modelos en la categoría de redes bayesianas y arboles de decisión.

En el trabajo de Jaime Hernando Villamarín, (2017), titulado como “Análisis de la deserción estudiantil en la Universidad Autónoma de Occidente (Cali)”, en este estudio el autor desglosa las diferentes causas y efectos de la deserción en esa universidad. Anidado a la investigación el autor trata de predecir las principales variables de la deserción, por lo cual aplica el estudio del machine learning con el método de mapas auto organizados de Kohone.

Pues bien, se puede decir que hay numerosos estudios tanto a nivel exploratorios como a nivel descriptivo en el contexto local, nacional e internacional sobre el diagnóstico y análisis de la deserción en las universidades. Entonces es pertinente decir que, en lo que respecta a esta monografía, se van a encontrar múltiples coincidencias con las de otros autores, y sobre todo en cuanto a la determinación de querer aportar soluciones al problema planteado. La problemática es compleja, más allá de la implementación del potencial de las ciencias de datos, porque si bien es cierto que la población universitaria es heterogénea, eso no impide que los factores de la problemática pueden varían de un territorio a otro y eso es precisamente lo que hace compleja esta problemática.

Marco Contextual

De forma común el "contexto" se puede definir como "el entorno físico o de situación, ya sea político, histórico, cultural o de cualquier índole, el cual se considera un hecho" (Torres, 2013, p.12).

Martínez, (2006), define de forma mucho más categórica al proceso de contextualización como parte de: un pensamiento analítico del estudiante referente del tema y su contexto implícito, lo que se denomina principio de la investigación; ubicar el objeto de estudio dentro de

su contexto, describir los hechos y realidades que lo circundan, los aspectos, interrogantes y las relaciones que se presentan, definiendo claramente los alcances, el área de estudio, que describen claramente las condiciones contextuales que van a definir el problema del proyecto.

Contexto

Con el fin de dimensionar la deserción estudiantil en la educación superior, se estableció el contexto nacional universitario colombiano como objeto de estudio de esta problemática.

Según el Sistema Nacional de Información de la Educación Superior (SNIES, 2021) con base en datos del ministerio de Educación Nacional, para 2021 la matrícula total en educación superior fue de 2.448.271 estudiantes, lo que representa un aumento del 3,93% respecto a 2020. Tomando como referencia las nuevas proyecciones de población del censo 2018, la tasa de cobertura en educación superior, indicador que da cuenta del acceso de los jóvenes a este nivel de formación, se ubicó en 53,94% para el año 2021, presentando un incremento de 2,36 puntos porcentuales frente a la registrada en 2020.

Desde la época de la Colonia, la educación ya se le hacían transformaciones sustanciales, los cuales se pueden apreciar en dos fases o periodos: el periodo de la Formación, enmarcado entre el año 1580 al año 1736; y el periodo Ilustrado enmarcado entre el año 1736 al año 1826 (International Institute for Higher Education in Latin America and the Caribbean and United Nations Educational, Scientific and Cultural Organization [IESALC y UNESCO], 2002). Es allí en esas fases en la que se inicia el concepto de la universidad como institución de educación superior, con la creación de la primera universidad en Santafé de Bogotá en el año de 1580, la cual tuvo el nombre de Universidad de Santo Tomás, creada por la comunidad católica de la orden de los predicadores Dominicos de Santo Domingo de España.

Constitucionalmente se han venido incluyendo normativas que empoderen la educación superior. Luego de la entrada en vigor de la Ley 30 de 1992, se dictan los lineamientos que deben seguir las instituciones de educación superior; con lo que se fortalecen las bases de la educación universitaria privada y pública en Colombia.

Marco Conceptual y Teórico

El abandono estudiantil en la educación superior es un fenómeno no solo de las universidades colombianas sino a nivel latinoamericano, esta problemática está enmarcada por indicadores que varían de acuerdo, al modo tiempo y lugar, estas variaciones están cohesionadas por políticas educativas de las IES, como también por factores personales, demográficos, psicológicos, socioeconómicos, sociológicos y familiares (Elcira Solano B. y Mario Barraza N, 218).

Las primeras bases conceptuales de la deserción se tomaron de la teoría del suicidio de Durkheim (1897), y de los análisis costo-beneficio de la educación desde una perspectiva económica (Elcira B. y Mario B., 218).

Según Samaria V. Muñoz-Camacho, (2018) en la actualidad, el abandono de la educación superior constituye un fenómeno expandido que tiene alcance en todo el contexto Latinoamericano. Dentro de las que han realizado los investigadores sobre la deserción universitaria, se visualiza entorno diverso y complejo, donde se estudian diferentes factores con sus respectivas variables desde la mirada de distintas disciplinas y modelos de análisis.

En Vicent Tinto, (1989) define la deserción estudiantil como la frustración para terminar un curso o lograr un fin determinado. Se plantea que la deserción depende de los deseos individuales, sociales e intelectuales, por las cuales las personas elaboran metas deseadas en cierta universidad.

Es por lo anterior que el Instituto Internacional de la UNESCO para la Educación Superior en América Latina y el Caribe (IESALC) plantea la necesidad de realizar estudios que ayuden a detectar los factores causales, el costo y la búsqueda de solución al problema.

La Ciencia de Datos

La ciencia de datos es un campo de la inteligencia artificial (AI), que combina técnicas de estadística, matemáticas e informática para extraer conocimiento y valor a partir de datos estructurados y no estructurados (IBM). Su propósito es identificar patrones, hacer predicciones y respaldar la toma de decisiones en diversos ámbitos (Provost y Fawcett, 2013).

Machine Learning

El machine learning, es un área en el desarrollo de los algoritmos computacionales diseñados para imitar la inteligencia del ser humano, aprendiendo del entorno. Se considera el caballo de batalla del big-data. Las técnicas basadas en machine learning se han aplicado en varios campos desde el reconocimiento de patrones, la visión por ordenador, la ingeniería de naves espaciales, las finanzas, el entretenimiento y la biología computacional hasta las aplicaciones biomédicas y médicas. La capacidad de estos modelos de machine learning para aprender del contexto actual y generalizar en tareas desconocidas permitirían mejorar tanto la seguridad como la eficacia de la práctica de la radioterapia, lo que redundaría en mejores resultados (Issam El Naga, et al., 2015).

Según un estudio de (Kuleto, V. et al., 2021), Las universidades más importantes, han comprendido que la IA y el ML representan el presente y el futuro tanto en la educación como en el desarrollo progresivo del mundo. Estas tecnologías proporcionan una experiencia educativa interactiva y avanzada a sus estudiantes. Los resultados son impresionantes: el 65% de las universidades de los Estados Unidos, apoyan el aprendizaje asistido por IA y ML. Además, estos

sistemas proporcionan una valiosa ayuda a los profesores de las mejores universidades, facilitando y mejorando el aprendizaje de diversas maneras. Por ejemplo, las estimaciones indican que la IA en la educación en Estados Unidos aumentó un 47,5% entre 2017 y 2021.

Minería de Datos

De manera análoga, la minería de datos debería haber sido más apropiadamente llamada "minería de conocimiento a partir de datos", lo cual desafortunadamente es algo largo. Sin embargo, en el corto plazo, la minería del conocimiento puede no reflejar el énfasis en minería a partir de grandes cantidades de datos. Por lo tanto, un nombre tan inapropiado que incluía tanto "datos" como "minería" se convirtió en un término muy popular (ver ilustración 2).

¿Qué tipos de datos se pueden extraer? La minería de datos se puede aplicar a cualquier tipo de datos siempre que estos sean significativos para una aplicación de destino. Las aplicaciones se pueden dar en las bases de datos (database data), datos de almacén de datos (data warehouse), y datos transaccionales (transactional data).

La minería de datos también se puede aplicar a otras formas de datos (por ejemplo, flujos de datos, datos ordenados/de secuencia, datos gráficos o en red, datos espaciales, datos de texto, datos multimedia y la web). El tratamiento en profundidad se considera un tema avanzado.

La minería de datos ciertamente continuará adoptando nuevos tipos de datos a medida que surjan más información.

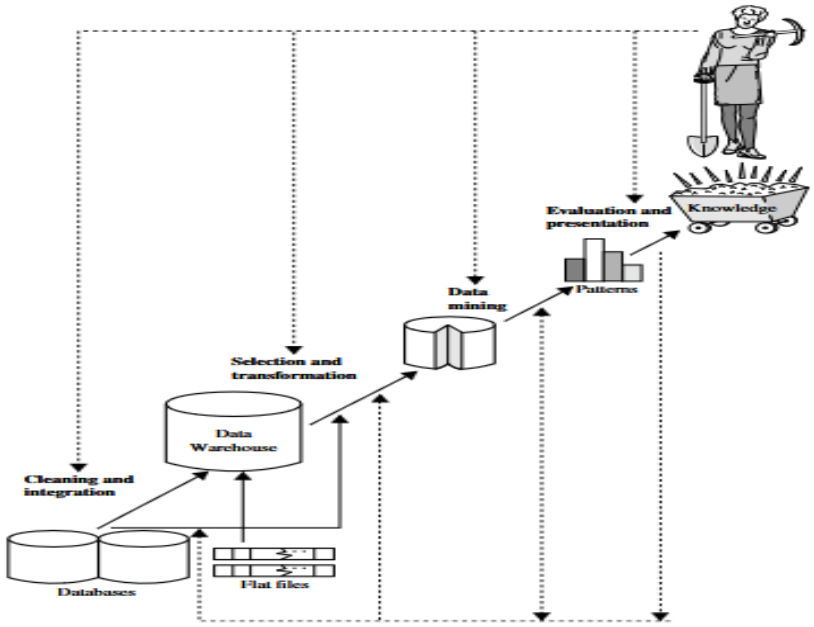
Pasos del Proceso de la Minería de Datos

- Limpieza de datos (para eliminar ruido y datos inconsistentes).
- Integración de datos (se pueden combinar múltiples fuentes de datos).
- Selección de datos (los datos relevantes para la tarea de análisis se recuperan del base de datos).

- Transformación de datos (los datos se transforman y consolidan en formularios apropiado para la minería mediante la realización de operaciones de resumen o agregación).
- Minería de datos (un proceso esencial donde se aplican métodos inteligentes para extraer patrones de datos).
- Evaluación de patrones (para identificar los patrones verdaderamente interesantes que representan el conocimiento)
- Presentación del conocimiento (las tecnologías de visualización y representación del conocimiento y las técnicas se utilizan para presentar el conocimiento extraído a los usuarios).

Figura 2

El Proceso de Descubrimiento del Conocimiento en Base a los Datos



Nota. La figura2, muestra los pasos del proceso que desarrolla la minería de datos. Tomado de: (Han, J., Pei, J., & Kamber, M., 2011).

Marco Normativo

Para el desarrollo del presente proyecto se tienen en cuenta referencias legales y normativas, esto con el fin de adaptarlo y trabajar dentro de un marco legal que evite inconvenientes a mediano o largo plazo, además el estudio de este marco apoya el fundamento teórico a la investigación del proyecto.

Normas Generales de la Educación Superior

La Ley 115 de 1994 (Ley General de Educación), de conformidad con el artículo 67 de la Constitución Política, define y desarrolla la organización y la prestación de la educación formal en sus niveles de preescolar, básica y media, no formal e informal.

La educación superior, por su parte, es reglamentada por la Ley 30 de 1992 que define el carácter y autonomía de las Instituciones de Educación Superior -IES-, el objeto de los programas académicos y los procedimientos de fomento, inspección y vigilancia de la enseñanza.

Estas dos leyes indican los principios constitucionales sobre el derecho a la educación que tiene toda persona, por su parte, las condiciones de calidad que debe tener la educación se establecen mediante el Decreto 2566 de 2003 y la Ley 1188 de 2008.

El Decreto 2566 de 2003 reglamentó las condiciones de calidad y demás requisitos para el ofrecimiento y desarrollo de programas académicos de educación superior, norma que fue derogada con la Ley 1188 de 2008 que estableció de forma obligatoria las condiciones de calidad para obtener el registro calificado de un programa académico, para lo cual las Instituciones de Educación Superior, además de demostrar el cumplimiento de condiciones de calidad de los programas, deben demostrar ciertas condiciones de calidad de carácter institucional.

Esta normatividad se complementa con la Ley 749 de 2002 que organiza el servicio público de la educación superior en las modalidades de formación técnica profesional y

tecnológica, amplía la definición de las instituciones técnicas y tecnológicas, hace énfasis en lo que respecta a los ciclos propedéuticos de formación, establece la posibilidad de transferencia de los estudiantes y de articulación con la media técnica.

Constitución Política de Colombia 1991

Nuestra carta magna establece los lineamientos sobre los derechos a la educción mediante los siguientes artículos constitucionales.

Artículo 27. El Estado garantiza las libertades de enseñanza, aprendizaje, investigación y cátedra. Artículo 67. La educación es un derecho de la persona y un servicio público que tiene una función social; con ella se busca el acceso al conocimiento, de la ciencia, la técnica, y los demás bienes y valores de la cultura. La educación formará al colombiano en el respeto a los derechos humanos, a la paz y a la democracia; y en la práctica del trabajo y la recreación, para el mejoramiento cultural, científico, tecnológico y para la protección del ambiente. El Estado, la sociedad y la familia son responsables de la educación, que será obligatoria entre los cinco y los quince años y que comprenderá como mínimo, un año de preescolar y nueve de educación básica. La educación será gratuita en las instituciones del Estado, sin perjuicio del cobro de derechos académicos a quienes puedan sufragarlos. Corresponde al Estado regular y ejercer la suprema inspección y vigilancia de la educación con el fin de velar por su calidad, por el cumplimiento de sus fines y por la mejor formación moral, intelectual y física de los educandos; garantizar el adecuado cubrimiento del servicio y asegurar a los menores las condiciones necesarias para su acceso y permanencia en el sistema educativo. La Nación y las entidades territoriales participarán en la dirección, financiación y administración de los servicios educativos estatales, en los términos que señalen la Constitución y la ley.

Artículo 69. Se garantiza la autonomía universitaria. Las universidades podrán darse sus directivas y regirse por sus propios estatutos, de acuerdo con la ley. La ley establecerá un régimen especial para las universidades del Estado. El Estado fortalecerá la investigación científica en las universidades oficiales y privadas y ofrecerá las condiciones especiales para su desarrollo. El Estado facilitará mecanismos financieros que hagan posible el acceso de todas las personas aptas a la educación superior.

Artículo 70. El Estado tiene el deber de promover y fomentar el acceso a la cultura de todos los colombianos en igualdad de oportunidades, por medio de la educación permanente y la enseñanza científica, técnica, artística y profesional en todas las etapas del proceso de creación de la identidad nacional. La cultura en sus diversas manifestaciones es fundamento de la nacionalidad.

Artículo 71. La búsqueda del conocimiento y la expresión artística son libres. Los planes de desarrollo económico y social incluirán el fomento a las ciencias y, en general, a la cultura. El Estado creará incentivos para personas e instituciones que desarrollen y fomenten la ciencia y la tecnología y las demás manifestaciones culturales y ofrecerá estímulos especiales a personas e instituciones que ejerzan estas actividades.

Ley Estatutaria 1581 del 2012

La Gobernanza de datos es un conjunto de procesos, roles, políticas, estándares y métricas que aseguran el uso efectivo y controlado de la información dentro de una organización. En Colombia, este concepto ha tomado relevancia debido al crecimiento exponencial de datos y la necesidad de cumplir con normativas legales, como la Ley de Protección de Datos Personales (Ley 1581 de 2012).

Esta Ley también es conocida como ley de protección de datos y la cual está definida por el Congreso de la Republica en el Titulo 1 *Objeto, Ámbito de aplicación y Definiciones* y en su Artículo 1.

La presente ley tiene por objeto desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, y los demás derechos, libertades y garantías constitucionales a que se refiere el artículo 15 de la Constitución Política; así como el derecho a la información. consagrado en el artículo 20 de la misma. (Congreso de la República de Colombia, 2012, p.1)

La importancia de esta normatividad en el proyecto radica en que se tratan datos en las muestras que en ocasiones pueden ser datos personales.

Datos para la Fundamentación de la Investigación

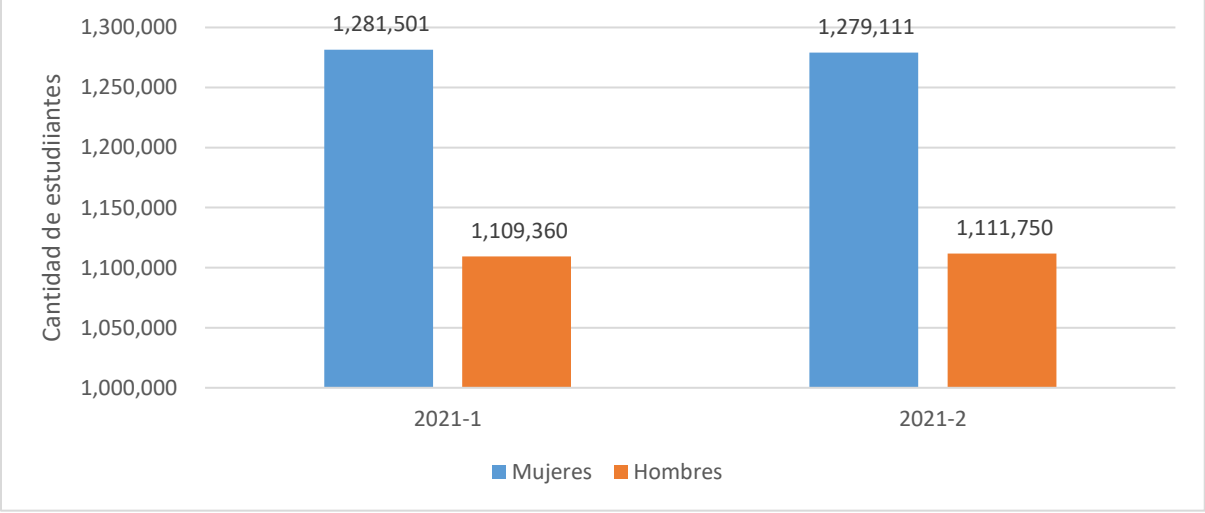
Para fundamentar esta investigación se incluyeron datos sobre la educación superior, suministrados por del Ministerio de Educación Nacional de Colombia (MEN), de los cuales se destacan la población estudiantil registrada en el sistema de educación nacional, la tasa deserción en sus diferentes aspectos y niveles académicos.

Población Estudiantil Registrada en el Sistema Nacional de Educación Superior

De acuerdo con Sistema Nacional de Información de la Educación Superior (SNIES), en primer curso entre 2021-1 y 2021-2 se matricularon 2.390.851 estudiantes incluyendo técnicos, tecnólogos, pregrados y postgrados, de los cuales el 53.6% (1.281.583) fueron mujeres y el 46.4% (1.109.360) hombres (ver figura 3).

Figura 3

Número de Estudiantes Matriculados en Primer Curso de 2021-1 a 2021-2



Nota. La gráfica de la figura 3, muestra las cifras de los estudiantes matriculados en el sistema educativo en la educación superior, formado por mujeres y hombres. Se considera todos los niveles de la educación superior: técnico, tecnológico y universitario. Tomado de (SPADIES, 2022).

Ahora bien, la población matriculada equivale a 2.390.861 en 2021-2 y se distribuye en tres niveles de educación (técnico, tecnólogo y superior (IES)), siendo las Instituciones de Educación Superior (IES) de mayor población estudiantil con un 52.6%, seguida de las Instituciones Tecnológicas con un 23.8%. La Tabla 1, muestra que 5 de cada 10 estudiantes estaban matriculados en universidades.

Tabla 1

Estudiantes Matriculados en el 2021-2 Según el Tipo de Institución

	2021-1	2021-2
Total, estudiantes matriculados	2.359.899	2.390.861
Universidad	55%	52.6%
Institución universitaria/Escuela Tecnológica	25.2%	23.8%
Institución tecnológica	17.7%	21.3%
Institución técnica profesional	2.1%	2.3%

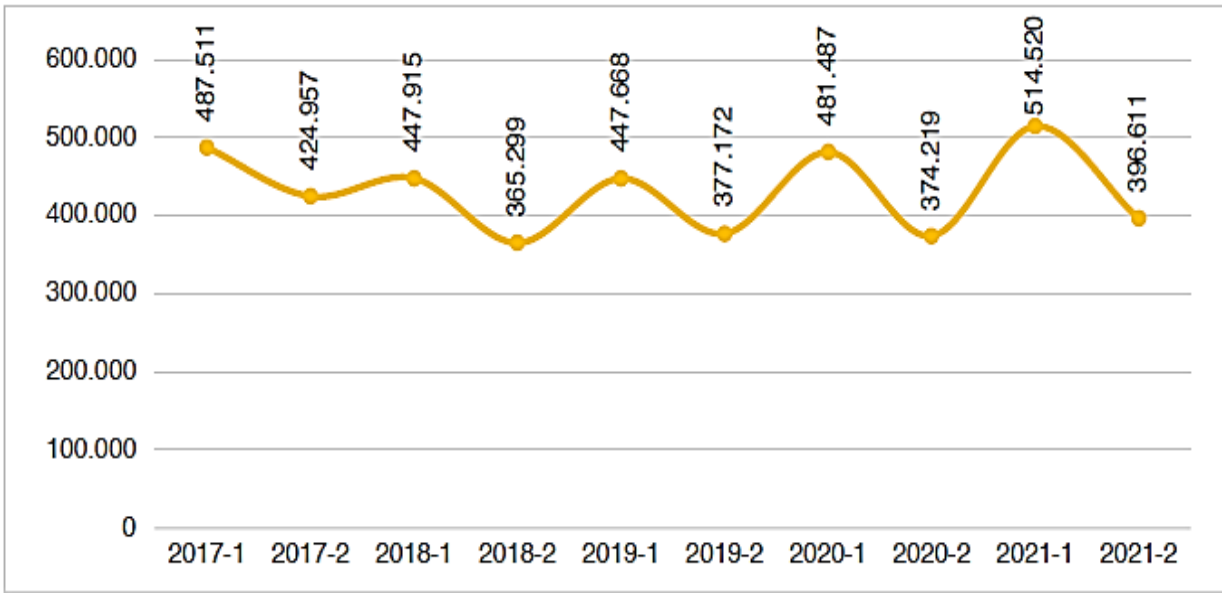
Nota. La tabla 1, muestra la distribución de la población matriculada en sus diferentes niveles académicos. *T*omado de los datos del (MEN, 2022)

La Deserción Estudiantil en la Educación Superior

Sistema de Información Especializado para el Análisis de la Permanencia en la Educación Superior Colombiana (SPADIES) y el Informe Análisis Estadístico LEE de la Universidad Javeriana, revela que, entre el año 2017 y 2021 se matricularon en primer semestre de un programa de educación superior en promedio 432 mil personas, pero entre el primer y el segundo semestre de cada año se nota una reducción considerable, teniendo en cuenta que en el primer semestre hubo un promedio de 476 mil estudiantes y para el segundo semestre solo se matricularon 388 mil estudiantes, como se observa en la gráfica 2.

Figura 4

Número de Estudiantes Matriculados en Primer Curso en I ES. 2017-2021



Nota. La gráfica de la figura 4, muestra la cantidad de estudiantes que dejan de matricularse entre el semestre 1 y 2 del mismo año en un rango de diez años. Tomado de: (Informe Lee U. Javeriana, 2022).

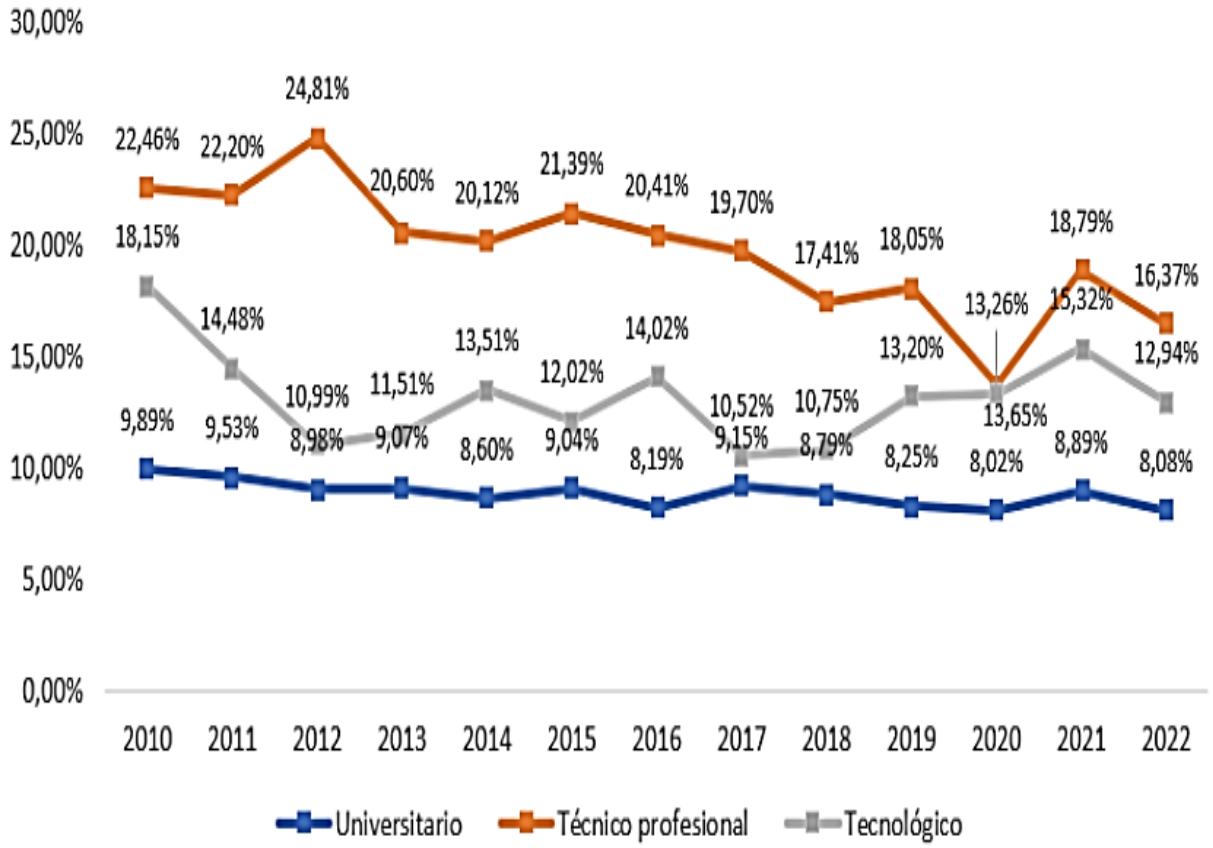
No obstantes, según (SPADIES 2021), entre el año 2000 y el 2021, en promedio de la tasa de deserción anual fue de 11,0%. es decir que, en cada semestre, 1 de cada 10 estudiantes que estaban matriculados en la educación superior no continúan con sus estudios y salieron del sistema.

Deserción Según el Nivel de Formación Académica

La deserción impacta a todos los niveles de la educación superior desde el nivel técnico profesional, tecnólogos hasta el universitario. Según Spadies, para el año 2021 la deserción en el nivel técnico se ubicó en un 18.79%, seguida de los tecnólogos con un 15.32%, y no menos preocupante están las universidades con una tasa del 8.08%.

Figura *5*

Tasa de Deserción Anual Según Nivel de Formación Académica



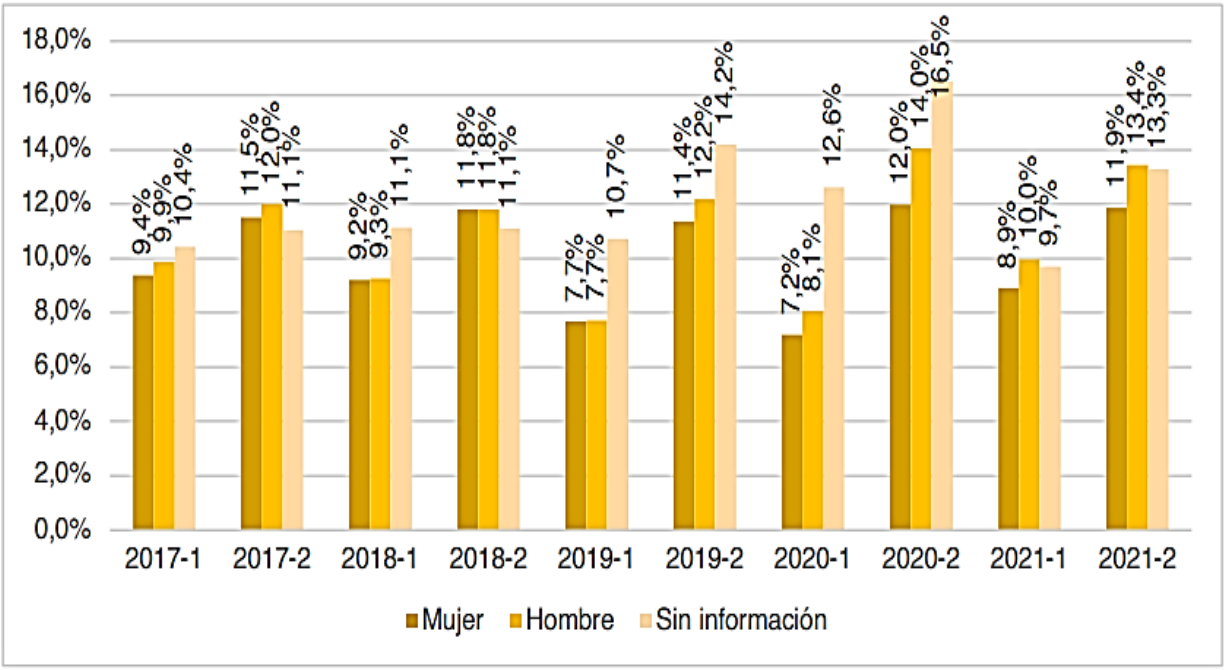
Nota. La gráfica de la figura 5, muestra el porcentaje de la deserción en los tres niveles de educación superior, siendo el nivel técnico profesional el del porcentaje más alto. Tomado de (SPADIES, 2022)

Deserción Anual Según el Sexo

La problemática de la deserción no distingue entre sexos, es proporcional, es decir que golpea a mujeres y a hombres por igual, puede que en un periodo A la deserción de la mujer sea mayor, pero en un periodo B puede ser lo contrario, como se observa en la gráfica 4.

Figura 6

Tasa de Deserción Anual del Sistema, Según Sexo. 2017-2021



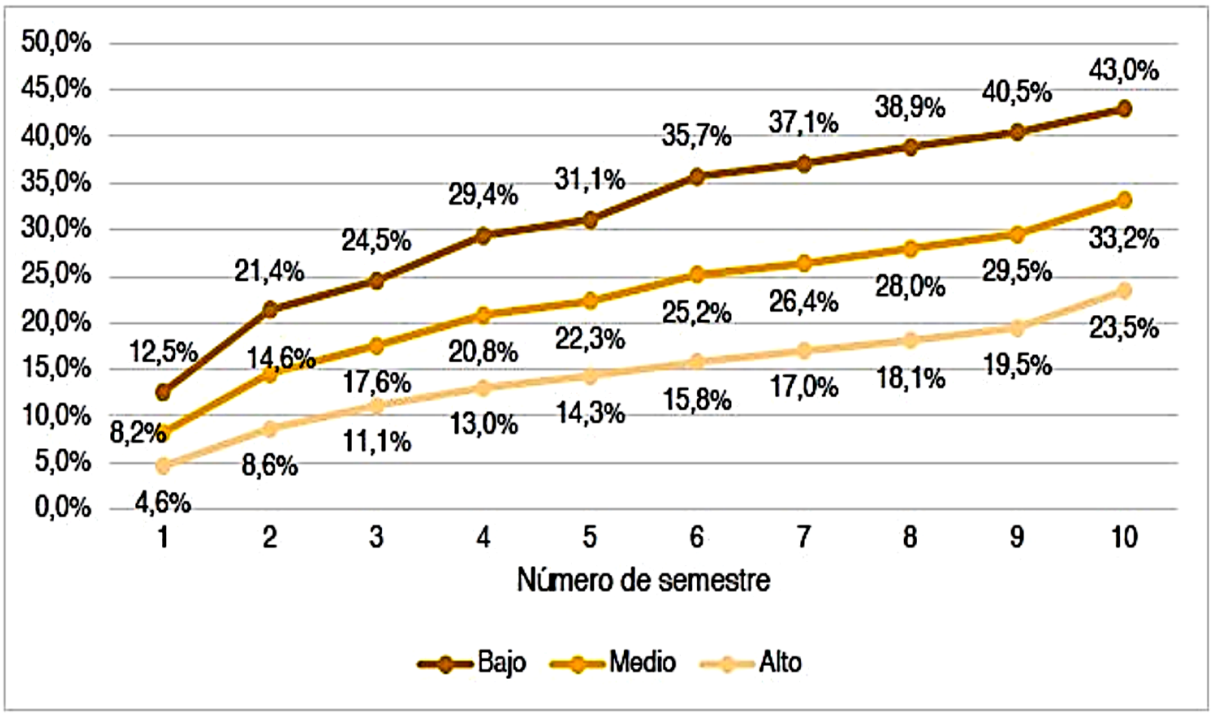
Nota. La gráfica de la figura 6, muestra los porcentajes de deserción tanto en hombres como en mujeres. La tasa de deserción del sistema mide "estudiantes que no se matriculan en ningún programa académico en dos o más períodos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios. Tomado de (SPADIES, 2022)

La Deserción Según el Puntaje Obtenido en las Pruebas Saber 11

Una buena preparación previa al ingreso de la educación superior es fundamental, debido a que puede garantizar en un alto porcentaje la permanencia del estudiante. Una de las métricas o indicadores que pueden medir esa preparación, es precisamente la prueba saber 11. La Gráfica de la figura 7, evidencia la tasa de deserción estudiantil según el puntaje obtenido en las pruebas saber 11.

Figura 7

Rendimiento Académico, Según Pruebas Saber 11 del 2016



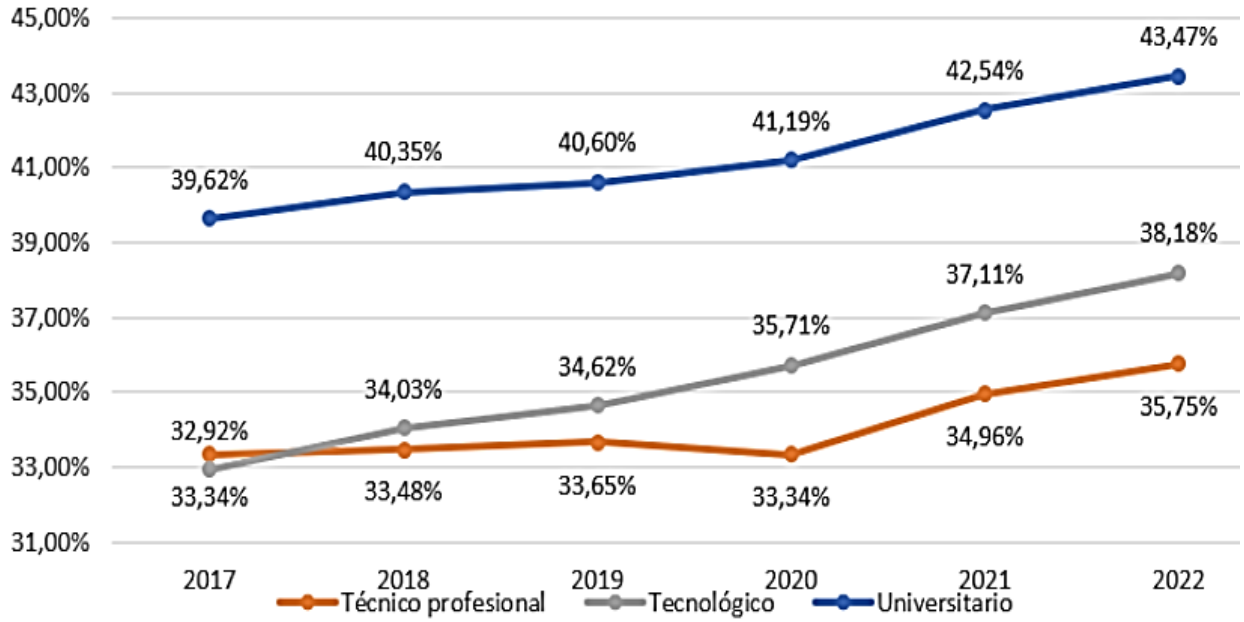
Nota. La gráfica de la figura 7, deja ver como los estudiantes que tuvieron puntajes bajos en las pruebas Saber 11, tuvieron una mayor tasa de deserción, por lo que al 5to semestre el 31.1% de esos estudiantes con pruebas bajas ya abrían desertado. Tomado de (U. Javeriana, Informe Lee, 74, 2022).

Tasa de Graduación Acumulada

La tasa de estudiantes que no logran graduarse en el tiempo real es bastante preocupante. La gráfica 6 muestra una tasa acumulada semestre a semestre desde el año 2017 al año 2022 de aquellos estudiantes que debieron haberse graduado en el año real, es decir en el año que culmina su carrera.

Figura 8

Tasa de Graduación Acumulada Semestre a Semestre



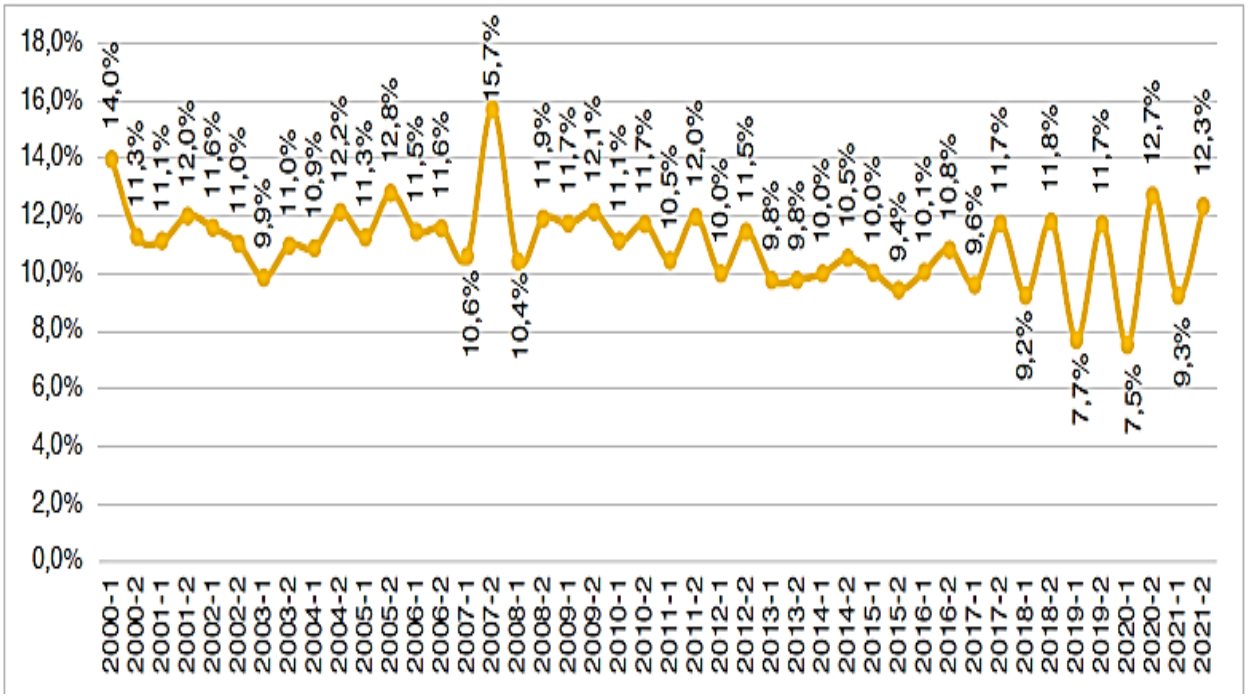
Nota. La gráfica figura de la 8, muestra el año 2022 con un aumento de 0,93 para el nivel universitario, 1,07 en el nivel tecnológico y 0,79 en el nivel técnico profesional. Es decir que solo 4 de cada 10 estudiantes logra graduarse. Tomada del: (U. Javeriana, Informe Lee 74, 2022).

Tasa de Deserción Anual del 2000 al 2021

Finalmente se muestra el impacto de la de la deserción en la educación superior. La gráfica muestra que la pandemia por Covid-19, en el segundo semestre de 2020 la tasa de deserción del sistema fue de 12,7%, un porcentaje superior al promedio de las dos décadas anteriores, mientras que para el segundo semestre de 2021 fue de 12,3%. Sin embargo, en el periodo estudiado, la mayor tasa de deserción del sistema se registró en el segundo semestre de 2007 (15,7%) (ver figura 9).

Figura 9

Tasa de Deserción Anual del Sistema. 2000 - 2021



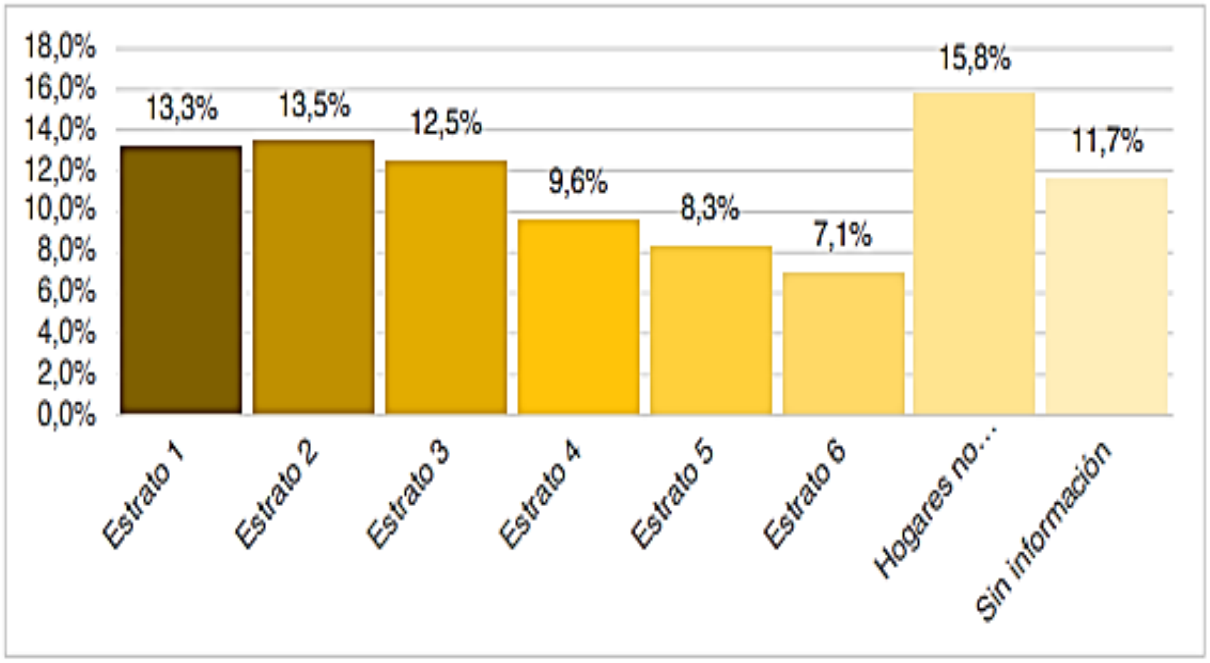
Nota. La gráfica de la figura 9, deja ver la tasa de deserción del sistema mide los estudiantes que no se matriculan en ningún programa académico de ninguna Institución de Educación Superior durante dos o más períodos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios. Se consideran todos los niveles de la educación superior: técnico, tecnológico, universitario y posgrado. Tomado del (U. Javeriana, Informe Lee 74, 2022).

Deserción Según el Estrato Socioeconómico

El nivel económico es fundamental, de este depende en gran manera la permanencia del estudiante en el sistema educativo, la tabla 8 muestra la diferencia abismal de la deserción que hay entre los estratos altos y los estratos bajos, con una tasa del 13.3% y un 7.1% respectivamente.

Figura 10

Tasa de Deserción Anual del Sistema por Estrato 2021-2



Nota. La gráfica de la figura 10, muestra los porcentajes de deserción según el estrato socioeconómico. Tomado de (U. Javeriana, Informe Lee 74, 2022).

Finalmente, La gráfica 7 deja ver la real dimensión de la problemática de la deserción en Colombia. La gráfica muestra que en el año 2000 la tasa de deserción se ubicaba en 14% mientras que para el año 2021 se encontraba en un 12.3%. es decir que en más de dos décadas la problemática solo disminuyó en 1.7%, manteniéndose prácticamente en una estacionalidad.

La Deserción en Contexto Colombiano Según la Comunidad Internacional

Son varias las organizaciones que se han manifestado sobre el fenómeno de la deserción, Según la ONU, (2020) en su informe de políticas sobre la educación en el 2020, presentó cifras preocupantes debido a la afección de la pandemia, afección que ha repercutido en todos los

niveles de formación académica. Según ese informe, fueron más de 1.600 millones de estudiantes de más de 195 países los que se vieron afectados por la emergencia sanitaria. Esta situación ha contribuido a agudizar la problemática de la deserción estudiantil.

El Banco Interamericano de Desarrollo (BID), y específicamente en el caso colombiano, ha demostrado gran interés y preocupación resaltando el impacto negativo en el desarrollo económico y social que el fenómeno de la deserción causa a la nación, razón por la cual este organismo internacional, aprobó y desembolsó en 2022 un préstamo de US$81,5 millones con el objetivo de aumentar la cobertura estudiantil. El BID también ha señalado que la discontinuidad en la educación tiene efectos no solo educativos y sociales, sino que también impacta en la economía en su conjunto. La pérdida de aprendizajes se traduce en una disminución del Producto Interno Bruto (PIB) de los países de la región (BID, 2022, https://www.iadb.org)

De igual forma el Banco Mundial, se ha manifestado sobre este fenómeno y ha señalado que solo el 50% de los estudiantes que inician sus estudios superiores en América Latina logran graduarse. Esta alta tasa de deserción se atribuye a factores como la falta de programas educativos de calidad que mantengan el interés de los estudiantes y los preparen adecuadamente para el mercado laboral actual. (Bancomundial.org, 2017).

En conclusión, se puede decir que muy a pesar de los esfuerzos realizados de manera mancomunada entre los gobiernos y las instituciones de educación superior para mitigar este fenómeno la deserción en la educación superior en Colombia sigue siendo un desafío complejo que afecta no solo a los estudiantes y sus familias, sino también al desarrollo social y económico de todo un país.

Se puede ver que en más de dos décadas (2000 – 2021) la problemática no ha tenido una mejora relevante, por lo que, si observamos el gráfico 7, se muestra que solo se ha mejorado en

1.7%, esto debido a que en el año 2000 la tasa de deserción era del 14% y en 2021, es decir veintiún años después la tasa es del 12.3%.

Los factores económicos, las limitaciones en el acceso a instituciones públicas, la falta de apoyo integral para los estudiantes y los efectos recientes de la pandemia han agravado esta problemática. Es necesario implementar muevas estrategias innovadoras y sostenibles que permitan no solo aumentar la permanencia en los programas educativos, sino también garantizar el acceso equitativo y la calidad en la formación académica, especialmente para los sectores más vulnerables de la población estudiantil.

Investigaciones Previas Acerca de la Deserción Abordada desde la Ciencia de Datos

En el contexto de la problemática de la deserción en la educación superior, se han llevado a cabo numerosas investigaciones desde la perspectiva de la minería de datos y el machine learning. Desde este proyecto de monografía se analizarán estas investigaciones con el fin de identificar las distintas técnicas de Machine Learning utilizadas para el estudio de la problemática planteada.

La investigación de Thomas, Emily & Galambos, Nora (2004), analiza, cómo las características y experiencias de los estudiantes afectan la satisfacción, para lo cual utilizaron la regresión y análisis de árboles de decisión con el algoritmo CHAID. Los datos fueron obtenidos de una encuetas de opinión de los estudiantes de una universidad pública. En esta encuesta se tuvieron en cuenta distintas variables tales como características, experiencias y planes; su satisfacción con el entorno del campus, clima, servicios e instalaciones; sus percepciones del crecimiento; y las razones para su elección universitaria.

Tabla 2

Satisfacción con la Calidad de la Educación

Estudiantes que reportaron un crecimiento intelectual muy grande (n = 324) El 91% calificó la calidad de la educación como buena o excelente.		
Insatisfecha (n = 62) 71% buena / excelente	Calidad de la instrucción Satisfecho (n = 143) 94% bueno/excelente	Muy satisfecha (n = 119) 97% buena / excelente
	Estimulado intelectualmente	
	No siempre (n = 46) 93% bueno/excelente	Siempre (n = 73) 100% bueno/excelente

Nota. Tomado de (Thomas, Emily & Galambos, Nora 2004)

Gutiérrez F. & Pineda (2021) analiza las causas de la deserción estudiantil en una institución de educación superior en Bogotá. Para la identificación el origen del fenómeno, emplearon las técnicas de aprendizaje no supervisado tales como el análisis de componentes principales (PCA) y la descomposición en valores singulares (SVD) para la reducción de la dimensionalidad a 24 componentes con el 71% de la variabilidad explicada y el algoritmo de k-means posibilitó el agrupamiento en tres grupos con información recolectada de 207 estudiantes que desertaron durante el 2020. Los estudios arrojaron que las variables de tipo económico fueron la principal causa de deserción en hombres (67.6%) de los programas de ingeniería mecánica y ambiental (tabla 3).

Tabla 3

Análisis Descriptivo de las Variables Numéricas

	Edad	Materias Inscritas	Materias Reprobadas	Promedio	Numero hijos	Personas cargo
Count	207	207	207	207	207	207
mean	25,497,585	5,014,493	1.410628	3.3756.4	0.323671	0.908213
std	4,968,712	2,343,092	1,824,984	0.705398	0.643542	1.073186
min	18	1	0	0.390000	0	0
25%	22	3	0	2,995	0	0
50%	25	5	1	3,460	0	1
75%	28	7	2	3,855	0	2
max	46	11	7	4,730	3	5

Nota. Tomado de: (Gutiérrez, Fonseca & Pineda, 2021)

Figura 11

Matriz de Correlación de Variables Numéricas

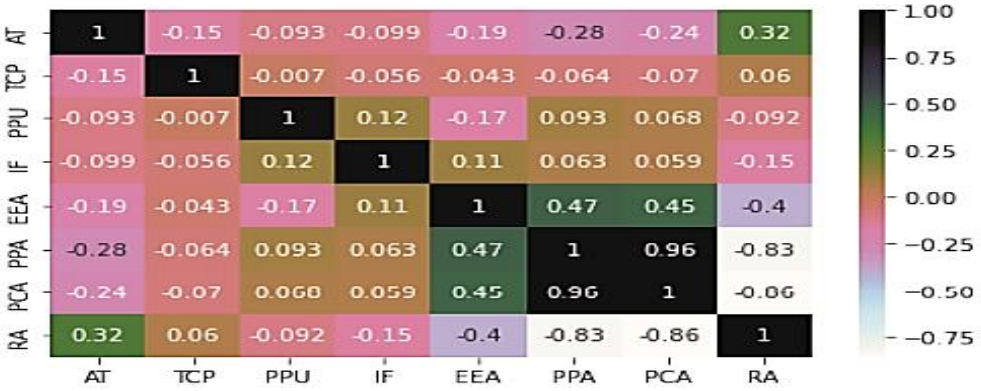
Nota. La matriz de la figura 11, muestra la correlación que hay entre las variables. Tomado de (Fonseca Gutiérrez, & Pineda, 2021).

Dos aspectos que se pueden apreciar en la matriz de correlación (Ilustración 3) de las variables numéricas, es que entre la edad y el número de hijos hay mediana correlación del 49 %, mientras que entre el promedio y el número de materias reprobadas se muestra una correlación negativa y medianamente alta (-58.4 %). Según los autores, el algoritmo K-means fue el que mejor agrupó los datos. Este algoritmo agrupa los datos tratando de separar muestras en n grupos de igual varianza y minimizando la inercia o suma de cuadrados dentro del grupo (Suárez, 2015).

Valero Cajahuanca, Julio Elvis, (2022), realizaron una investigación descriptiva, que tiene como propósito determinar el algoritmo de Machine Learning con mejor desempeño para detectar la deserción universitaria. Este estudio se hizo sobre los datos de la deserción universitaria de Perú entre 2018 y 2021. La población está compuesta de 652 estudiantes. Mediante los resultados del análisis determinaron que, el algoritmo K-Nearest-Neighbor fue la mejor precisión con 90.6% (ver tabla 4), para pronosticar la deserción universitaria con las variables académicas y socioeconómicas de los estudiantes.

Figura 12

Matriz de Correlación de las Variables de Estudio



Nota. La matriz de la figura 12, muestra la correlación que hay entre las variables. Se aprecia una fuerte correlación entre las variables de la dimensión académica del estudio, promedio ponderado acumulado (PPA) y proporción de créditos aprobados (PCA). Tomado de (Valero Cajahuanca, Julio Elvis, 2022).

Tabla 4

Precisión de los Algoritmos por Tipo de Variables

Algoritmos	Variables		
	Total	Académico	Socioeconómico
SVM	0,875	0,812	0,6562
Logistics Regression	0,812	0,812	0,6875
Decision Tree	0,812	0,843	0,5937
KNN	0,906	0,875	0,5937

Nota. La tabla 4, muestra al algoritmo KNN, con mejor precisión con un 0.875. Tomado de: *(Valero Cajahuanca, Julio Elvis, 2022).*

Anibal C, (2020), plantea el desarrollo de una aplicación o prototipo que permita predecir la deserción estudiantil a nivel de pregrado en la Universidad de la Costa - CUC, a partir del análisis de información socioeconómica de los estudiantes de pregrado, mediante la implementación de un modelo funcional basado en técnicas de Aprendizaje Automático. En este estudio el autor utilizó diferentes técnicas de clasificación con distintas métricas con datos sin balacear y obtuvo los resultados de la tabla 5.

Tabla 5

Resultados de Clasificación Usando el Conjunto de Datos sin Balancear

Técnica de clasificación	TP Rate	FP Rate	Precision	Accuracy	F-Measure	ROC Area
BayesNet	76.50%	56.80%	73.60%	76.50%	74.40%	69.20%
Naive Bayes	73.80%	53.10%	72.60%	73.80%	73.20%	67%
Naive Bayes Simple	73.50%	52.80%	72.50%	73.50%	73%	66.80%
Naive BayesUpdateable	73.80%	53.10%	72.60%	73.80%	73.20%	67%
SMO	77.80%	72.50%	74.10%	77.80%	70.20%	52%
ADTree	78.10%	64.70%	74.30%	78.10%	73.40%	71.50%
Random Forest	75.70%	70.40%	68.90%	75.70%	70%	60.10%
J48	77.50%	62.90%	73.60%	77.50%	73.60%	61.10%

Nota. Tomado de (Anibal C, 2020).

La tabla 5 muestra al algoritmo ADTree, basado en árboles de decisión con una exactitud (accuracy) del 78.1%. En la siguiente fase, el autor realiza el mismo proceso utilizando el conjunto de datos balanceados, con suavizado (SMOTE) y con la técnica de prueba (test) validación cruzada (cross-validation) con 10 pliegues, en lo que el RandomForest fue superior con una precisión del 84.1%. como se muestra en la tabla 6.

Tabla 6

Clasificación Conjunto de Datos Balanceados con Cross-Validation 10 Pliegue

Técnica de clasificación	TP Rate	FP Rate	Precision	Accuracy	F-Measure	ROC Area
BayesNet	84%	16.10%	85.30%	84.00%	83.80%	86.90%
Naive Bayes	74.90%	25.10%	75.00%	74.90%	74.80%	84.30%
Naive Bayes Simple	74.90%	25.00%	75.10%	74.90%	74.90%	84.40%
Naive BayesUpdateable	74.90%	25.10%	75.00%	74.90%	74.80%	84.30%
SMO	80.40%	19.60%	80.40%	80.40%	80.40%	80.40%
ADTree	77.20%	22.90%	77.50%	77.20%	77.10%	84.20%
Random Forest	84.30%	15.70%	85.70%	84.10%	84.20%	88.90%
J48	79.60%	20.40%	79.80%	79.60%	79.50%	81.90%

Nota. Tomado de: (Anibal C, 2020)

Entre tanto Miranda & Guzmán, (2017), realizan un trabajo para determinar la importancia de las variables que conllevan al abandono universitario, en la carrera de Ingeniería de la Universidad Católica del Norte de Santiago de Chile. Para dicho estudio, los autores construyen tres clasificadores que permiten categorizar a los estudiantes entre las clases abandono y no abandono. Para ello utilizamos tres algoritmos: redes bayesianas, redes neuronales y árbol de decisión. La red bayesiana mostró una mejor precisión con un 76%, seguida de la red neuronal con una precisión de 73% y el árbol de decisión con una precisión del 72% como se observa en la tabla7.

También, de los resultados obtenidos en los experimentos de los anteriores autores, se puede decir que los tres modelos pueden ser útiles para analizar el problema teniendo en consideración sus ventajas y desventajas.

Tabla 7

Comparación de los Parámetros de Evaluación de los Clasificadores

	Precisión	Recall	TPR	TNR	FPR	FNR	ROC Curve	F-Measure	Clasificados correcta	Clasificados incorrectos
Red neuronal	73%	65%	65%	88%	12%	35%	83%	69%	80%	20%
Árbol de decisión	72%	64%	64%	87%	12%	36%	74%	68%	82%	18%
Red Bayesiana	76%	76%	76%	70%	30%	24%	76%	76%	76%	24%

Nota. Tomado de: (Miranda & Guzmán, 2017).

S. Kotsiantis et al., (2004), en su investigación sobre la Universidad Abierta Helénica en Grecia, en la cual identificaron estudiantes con bajo rendimiento académico. Para predecir el desempeño de los estudiantes, los autores utilizaron seis de las técnicas de aprendizaje automático más comunes, a saber, árboles de decisión (Murthy, 1998), redes neuronales (Mitchell 1997), algoritmo Bayes ingenuo (Domingos y Pazzani (1997), algoritmos de aprendizaje basado en instancias (Aha 1997), regresión logística (Long 1997) y máquinas de vectores de soporte (Burges 1998), los cuales se entrenaron utilizando conjuntos de datos con 500 observaciones del curso de computación. Se analizaron variables socioeconómicas tales como la ocupación, él entorno familiar, el conocimiento, el trabajo en equipo. La evaluación de los modelos el Naive Bayes resultó ser el mejor como se observa. En la Tabla 8 se presenta la precisión de predicción promedio de cada algoritmo para todos los pasos de prueba del segundo experimento (tabla 8).

Tabla 8

Precisión (Accuracy) de los Algoritmos

	Naive Bayes	C4.5	BP	SMO	3-NN	Logistic
DEMOGR	62.95%	61.65%	61.85%	64.47%	58.84%	61.38%
FTOF-1	62.72%	61.56%	61.14%	64.47%	59.12%	61.56%
FTOF-2	69.78%	62.93%	68.18%	68.58%	62.41%	69.40%
WRI-2	75.36%	74,16%	75.55%	75.99%	68.45%	75.88%
WRI-3	78.58%	79.22%	80.11%	77.71%	72.62%	79.20%
FTOF-4	79.20%	74.84%	78.02%	78.37%	75.14%	80.14%
WRI-4	82.14%	77.80%	82.14%	80.68%	76.77%	82.01%

Nota. Tomado de: (S. Kotsiantis et al., 2004)

Zhang & Oussena, (2010), argumentan cómo la minería de datos puede ayudar a detectar a los estudiantes "en riesgo". Para lo cual llevan a cabo un experimento para predecir el abandono escolar de los estudiantes en función del perfil del estudiante. Los datos se dividen en grupo de entrenamiento y grupo de evaluación en una proporción de 2:1. eligen tres algoritmos: Naive Bayes (Harry, 2004), Support Vector Machine (Cristianini et al., 2000) y Decision Tree (Quinlan, 1986). prueban diferentes configuraciones para cada algoritmo para encontrar el resultado óptimo. Finalmente, y como se muestra en la Tabla 9, Naive Bayes con 85.9%, 93.1% y 89.5%, logró la mayor precisión de predicción, mientras que el árbol de decisiones obtuvo la menor.

Tabla 9

Resultado de la Predicción

	Naïve Bayes	Support Vector Machine	Decision Tree
Negative Accuracy	85.90%	78.70%	71.20%
Positive Accuracy	93.10%	88.30%	91.40%
Average Accuracy	89.50%	83.50%	81.30%

Nota: tomada de: (Zhang & Oussena, 2010)

Kalles & Pierrakeas, (2006), tomaron como datos, la información de los estudiantes que se inscriben en el programa de pregrado de informática, en la Universidad Abierta Helénica, los cuales demuestran una dificultad significativa para avanzar más allá del curso introductorio. Para ello dividieron el ejercicio en 2 fases. En la primera fase implementaron 5 clasificadores con validación cruzada de 10 pliegues. En la tabla 10 se muestra que, el clasificador Naive Bayes fue el mejor para la predicción éxito/fracaso predictor con un 72.48%.

Tabla 10

Resultados de Precisión de la Experimentación Inicial

Classifier	Overall accuracy
CA.5	69.99
BP	72.26
Naïve Bayes	72.48
3-NN	66.93
Logistic regression	72.32
SMO	72.17

Nota. Tomado de: (Kalles & Pierrakeas, 2006).

Luego (Kalles & Pierrakeas) realizaron una segunda fase en la que implementaron seis algoritmos (J48, MLP, Logistic regression, Naïve Bayes, 3-NN, GATREE) por lo que al aplicar estos modelos y compararlos, obtuvieron evidencia suficiente para determinar que, el algoritmo Naive Bayes con un 84.148% era el más apropiado para ser utilizado en la predicción del desempeño de los estudiantes, dada también la facilidad de implementación, seguido del modelo Logistic regression que tuvo una precisión en las dos fases de 72.32 % y 73.73% respectivamente (ver tabla 11).

Tabla 11

Resultados Resumidos del Conjunto de Datos

Classifier	Average accuracy	Accuracy on training set
J48	(65.41) 81.73	(82.27) 89.96
J48 (with REP)	(69.48) 83.73	(75.00) 82.73
J48 (with bagging)	(69.48) 82.93	(86.92) 90.76
MLP	(66.86) 82.13	(93.02) 84.14
Logistic regression	(67.44) 83.73	(75.58) 83.53
Naïve Bayes	(63.08) 84.14	(65.12) 84.54
3-NN	(63.66) 81.93	(75.58) 88.15
GATREE 250/250	(...) 83.27	(83.14) 84.34
GATREE 1000/1000	(78.82) 83.27	(84. 30) 84.34

Nota. Las tablas 10 y 11 muestran el comportamiento de los modelos aplicados en sus dos fases experimentales. Tomado de: (Kalles & Pierrakeas, 2006).

Cheewaprakobkit, P. (2013), analiza los factores que afectan el rendimiento académico y que contribuyen a la predicción de estos, en los estudiantes. El conjunto de datos estaba compuesto por 1.600 estudiantes, 22 atributos de estudiantes matriculados entre año 2001 y 2011 en una universidad de Tailandia. El investigador aplicó el conjunto de datos para diferenciar clasificadores (Decision Tree, Neural Network) Se utilizó una cross-validation con 10 pliegues para evaluar la precisión de la predicción. Como resultado el investigador muestra que el clasificador de árbol de decisión logra una alta precisión de 85,188%, que es más alto que el del clasificador de redes neuronales. en un 1,313%.

Tabla 12

Comparación de Rendimiento del Árbol de Decisión y Redes Neuronales

Performance Measures	J48 (C4.5)	MLP
Correctly Classified Instances (%)	85.188	83.875
Incorrectly Classified Instances (%)	14.812	16.125
Precision	0.852	0.838
Recall	0.852	0.838

Nota. Tomado de: (Cheewaprakobkit, P. 2013).

Moseley & Mead, (2008) utilizaron arboles de decisiones para predecir la deserción de estudiantes de enfermería de una Universidad Británica. en este artículo los autores hacen uso del paquete Answer Tree de SPSS para ese propósito. El conjunto de datos principal consistió en 3978 registros de 528 estudiantes de enfermería, divididos en un conjunto de entrenamiento y un conjunto de prueba. La fuente fueron los registros estándar de estudiantes universitarios. Cómo resultados: El método obtuvo una sensibilidad del 84 %, una especificidad del 70 % y una precisión del 94 % en casos no vistos (ver tabla 13).

Tabla 13

Éxito con Casos Invisibles Utilizando Tres Medidas

	% Success
Sensitivity	84
Specificity	70
Accuracy	94

Nota. Tomada de: (Moseley & Mead, 2008).

T. Mishra, D. Kumar and S. Gupta, (2014), aplicaron modelos de clasificación para la predicción del rendimiento en la integración socioacademica del estudiante. Los autores implementaron dos modelos el J48 (Implementación de C4.5) y Random Fores, esto se

implementaron sobre los datos de estudiantes de MCA de las universidades afiliadas a Guru Gobind Singh Indraprastha University, en Delhi India, con el objetivo de predecir el rendimiento del tercer semestre. Random Fore, resulto ser el más preciso con un porcentaje del 94.4%, mientras que J48 precisó el 88.3%, como se aprecia en la tabla 14.

Tabla 14

Comparación de Rendimiento de J48 y Random Forest

	J48			Random Forest		
	TP Rate	Precision	Recall	TP Rate	Precision	Recall
ABVG	0.792	0.924	0.792	0.948	0.924	0.948
EXCL	1.000	0.857	1.000	0.952	0.952	0.952
	J48			Random Forest		
AVG	0.900	0.851	0.900	0.914	0.970	0.914
BAVG	0.923	0.923	0.923	1.000	0.929	1.000
Weighted Average	0.884	0.887	0.884	0.944	0.945	0.944
Correctly Classified Instances		88.3721%			94.4186%	
Incorrectly Classified Instances		11.627%			5.5814%	

Nota. Tomado de: (T. Mishra, D. Kumar and S. Gupta, 2014).

Ara, N-B., Halland, R., Igel, C., & Alstrup, S. (2015), estos autores mediante un análisis para la predicción de alumnos que no terminarían 11° de educación media, con una data pública del sistema de administración del estudio MaCom Lectio de Dinamarca, utilizaron un dataset con 36.299 observaciones para lo cual lo dividieron, 70% para entrenamiento y 30% para la prueba. Como resultado, el clasificador que obtuvo mayor precisión fue Random Forrest, con una precisión del 93.5% mientras que la curva ROC de 0.965, como se observa en la tabla 15.

Tabla 15

Precisión de Predicción y Área Bajo la Curva (AUC) en los Datos de Prueba

	Random Forest	CART	SVM	Naïve Bayes
Accuracy (in %)	93.5	89.8	90.4	85.6
AUC (-100)	96.5	86.9	94.8	93.1

Nota. Tomado de:(Ara, N-B., Halland, R., Igel, C., & Alstrup, S. 2015).

M. Solis, et al., (2018), analizaron sobre una data compuesta por información de estudiantes que se matricularon en un programa de pregrado en el Instituto Tecnológico de Costa Rica (ITCR) entre 2011 y 2016. Con un total de 90.067 observaciones. Para lo cual implementaron 4 modelos dentro de los cuales Random Forest fue el algoritmo que obtuvo una predicción del 91% y una sensibilidad del 87%, como lo podemos ver en la tabla 16.

Tabla 16

Comparación de la Evaluación del Modelo Randon Forest

Indicators	Random Forest	SVM	NNET	LOGIT
Kappa	0.80	0.80	0.76	0.74
Sensitivity	0.87	0.85	0.83	0.83
Specificity	0.92	0.94	0.93	0.91
Positive	0.91	0.91	0.90	0.88
Negative	0.90	0.89	0.87	0.87

Nota. Tomado de: (M. Solis, et al., 2018).

Los autores concluyeron que el método requiere grandes cantidades de datos de alta calidad. Cuando se dispone de esos datos, la inducción de reglas ofrece una manera de reducir la deserción. Sería conveniente comparar sus resultados con los de las predicciones realizadas por tutores que utilizan métodos convencionales más informales.

Variables de Estudio Analizadas sobre la Problemática de la Deserción

Luego de analizar las bibliografías en las que diferentes autores analizaron la deserción estudiantil desde el enfoque de la ciencia de datos, se observó que, para llevar a cabo esos análisis utilizaron variables determinantes en la problemática planteada. Esas variables en su común denominador fueron las siguientes:

Según Hernández-Jiménez, et al, (2019), las variables sociodemográficas, económicas, y de entorno familiar permiten identificar patrones de deserción y diseñar estrategias para mitigarla, como apoyo financiero, tutorías, programas de adaptación y políticas inclusivas. Considerar estos factores es clave para mejorar la retención estudiantil y garantizar una educación superior más equitativa.

Variable Sociodemográfica

Estas variables revelan aspectos que no pueden ser modificadas, por ejemplo:

- Edad: Los estudiantes mayores pueden tener más responsabilidades (trabajo, familia), lo que afecta su permanencia.
- Estado civil: Estudiantes casados o con hijos pueden enfrentar dificultades para equilibrar el estudio con la vida personal.
- Nivel educativo de los padres: Un mayor nivel educativo suele estar asociado con mayor apoyo académico y motivación.

Variable Socioeconómica

Estas variables hacen referencia a problemas exteriorizados, dicho de otra manera, se relacionan con la capacidad económica tomando como base el ingreso, lo cual impacta en el estilo de vida afectando factores como la alimentación, el entorno en el que vive, etc.; así como

los conceptos de pobreza y de estilo de vida del estudiante. Es decir, reflejan la posición económica y social de la persona o grupo dentro de una sociedad. Por ejemplo:

- Nivel de ingresos
- Nivel educativo
- Condiciones de vivienda
- Tipo de empleo y estabilidad laboral
- Acceso a servicios básicos (salud, educación, transporte, etc.)

Variable del Entorno Familiar

En esta variable se relacionan aspectos basados en el apoyo y el afecto al interior del núcleo familiar del estudiante. Por ejemplo:

- Apoyo familiar: Un entorno que no valore la educación puede desmotivar al estudiante.
- Conflictos familiares: Problemas como violencia, divorcio o crisis familiares pueden afectar la concentración y el desempeño académico.
- Presión familiar: Exigir carreras específicas o imponer expectativas poco realistas
- Nivel educativo de los padres: Si los padres tienen bajo nivel educativo, es menos probable que puedan brindar apoyo académico y motivacional. Puede generar ansiedad y abandono.

Metodología

Método

Para el desarrollo de esta monografía se implementó una metodología mixta entre cualitativa y cuantitativa, de datos que se obtienen de la revisión y análisis de fuentes secundarias. La metodología sigue un diseño documental, basado en la recopilación, comparación y análisis de estudios previos. Este enfoque metodológico se dividió en 3 fases: (Exploratoria, Análisis documental y Análisis de resultados). Además, para apoyar el proceso de la revisión sistemática, en cuanto a la recolección de las fuentes bibliográficas se implementa junto a esta metodología el protocolo PRISMA 2020, como se observa en el flujograma de la ilustración 5.

Fase 1 – Exploratoria

- Búsqueda y recopilación de la información genérica relacionada con la problemática de deserción desde un ámbito meramente estadístico y académico.
- Búsqueda y recopilación de la información relacionada con la problemática en donde se haya abordado la ciencia de datos para la detección temprana de los factores que determinan la deserción en la educación superior.
- Se selecciona solo la documentación relevante para abordar la monografía.

Fuentes de Información Recopilada

Para la recolección de la información se utilizaron artículos científicos, libros, tesis y documentos académicos relevantes sobre la aplicación de la ciencia de datos en la deserción universitaria. Se utilizaron bases de datos tales como:

- Scopus
- Web of Science
- SciELO

- IEEE Xplore
- Google Academic
- Repositorios universitarios: UNAD, Universidad del Norte, Universidad Politécnico Gran Colombiano.

Repositorios de Datos Abiertos

- Datos abiertos del gobierno
- Banco Mundial
- Banco Interamericano de Desarrollo (BID)
- Organización de las Naciones Unidas (ONU)
- Congreso Nacional de Colombia.

Fase 2 – Análisis Documental

En esta fase, luego de la búsqueda y recopilación de la información filtrada, se hacen los siguientes análisis:

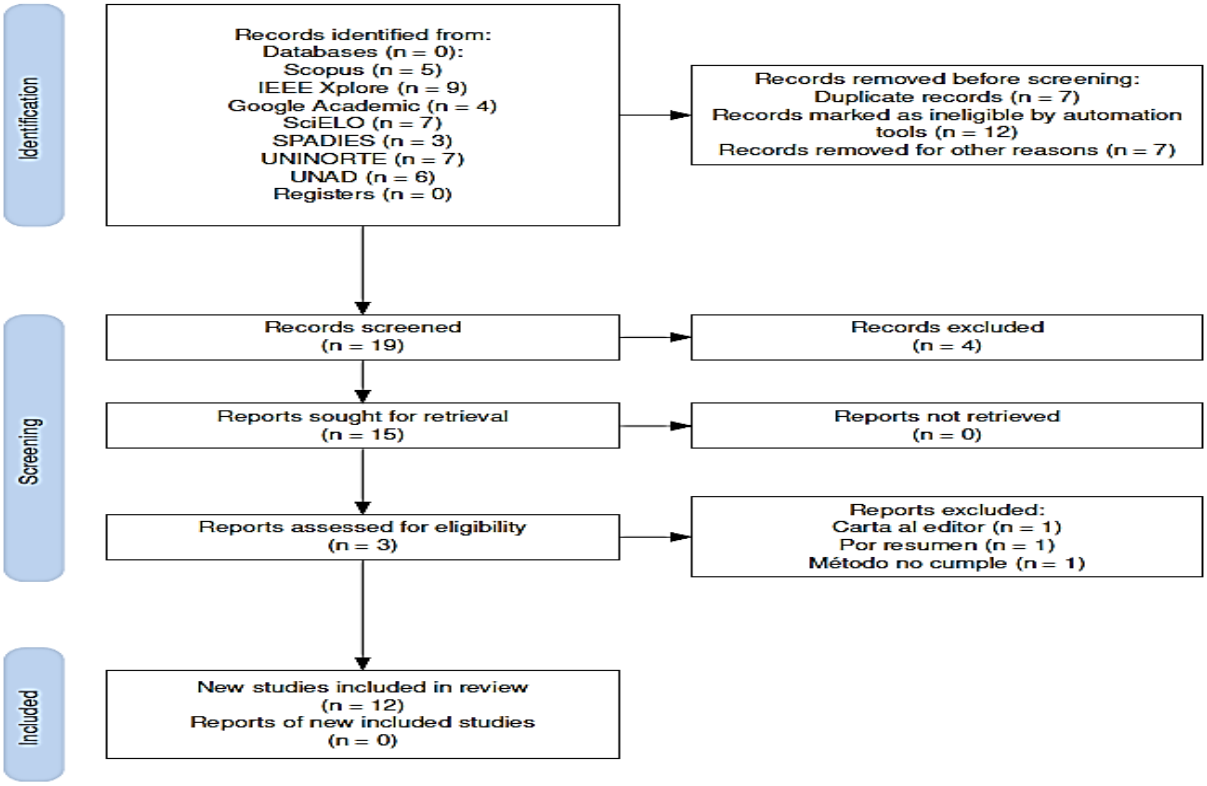
- Se analizan datos estadísticos (tablas y gráfico) relacionados con la deserción estudiantil, en las que se identifiquen los principales factores que causan la deserción en la educación superior.
- Se analizan las diversas formas de la implementación de la ciencia de datos sobre la problemática en mención. En este análisis se identifican los modelos, las métricas y las variables tenidas en cuenta para el análisis de la predicción temprana de los factores determinantes de la deserción.
- Se analizan series de tiempo identificando tendencias y estacionalidad de la problemática en diferentes periodos.

Fase 3 - Análisis de Resultados

En esta fase se analizan los resultados obtenidos de la aplicación y la evaluación de los modelos y métricas usados por loa autores de la bibliografía revisada. Estos resultados se analizan por medio de tablas, gráficas e ilustraciones.

Figura 13

Flujograma del Protocolo PRISMA



Nota. El flujograma de la figura 13, muestra el proceso de selección de la bibliografía que se utilizó en esta monografía. Tomada de: (Haddaway, N. R et al, 2020).

El flujograma de la ilustración 5. Nos permite realizar la clasificación de la documentación con elegibilidad deseada. Esto nos permite lo siguiente:

1. Documentar el proceso de selección de estudios: Permite mostrar cuántos artículos fueron identificados, eliminados, evaluados y finalmente incluidos en la revisión.

2. Garantizar la transparencia: Facilita que otros investigadores comprendan cómo se filtró la información y evalúen la validez de la revisión.

3. Evitar sesgos de selección: Ayuda a justificar la inclusión o exclusión de ciertos estudios, evitando interpretaciones arbitrarias.

4. Estandarizar la metodología: PRISMA es una referencia internacional en revisiones sistemáticas, lo que aumenta la credibilidad del trabajo.

Análisis de Resultados

Desarrollo de los Objetivos

Clasificar las Técnicas de Machine Learning en los Estudios Analizados sobre Deserción Universitaria

En el desarrollo de las investigaciones de bibliografías sobre deserción estudiantil se identificaron el uso de distintos clasificadores y métricas, resaltando de entre los clasificadores el uso de árboles de decisión y en cuanto a las métricas que permiten valorar la calidad del modelo, la exactitud (Accuracy). En esta exhaustiva revisión literatura científica, en este ámbito de conocimiento, ha permitido identificar los clasificadores más usados en las investigaciones relacionadas con la problemática, en la tabla 17 se muestran los clasificadores.

Tabla 17

Los Clasificadores más Utilizados en las Investigaciones de la Problemática

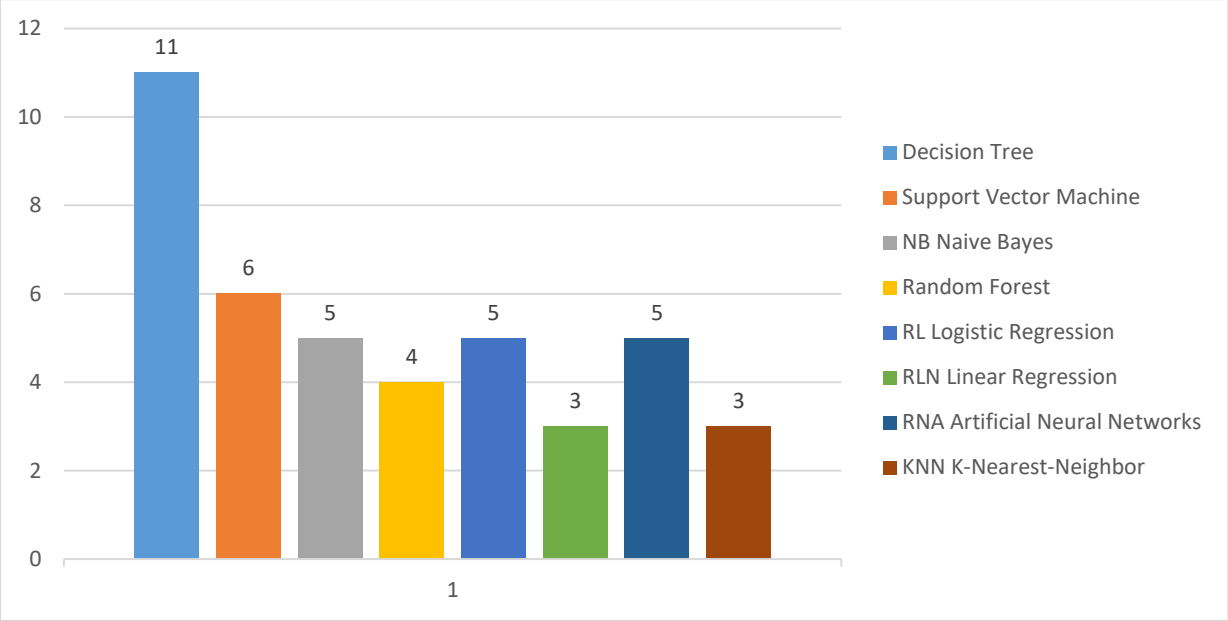
Referencias	Decision Tree	Support Vector Machine	NB Naive Bayes	Random Forest	RL Logistic Regression	RLN Linear Regression	RNA Artificial Neural Networks	KNN K-Nearest-Neighbor
(Thomas, Emily & Galambos, Nora 2004)	**X**					**X**		
(Juber Gutiérrez, Lida Fonseca & Wilmer Pineda, 2021),	**X**							
(Valero C. Julio Elvis, 2022),	**X**	**X**			**X**			**X**
(Aníbal C, 2020)	**X**		**X**	**X**				
(Hernández González et al., 2016)	**X**				**X**		**X**	

Referencias	Decision Tree	Support Vector Machine	NB Naive Bayes	Random Forest	RL Logistic Regression	RLN Linear Regression	RNA Artificial Neural Networks	KNN K-Nearest-Neighbor
(Miranda & Guzmán, 2017)	**X**		**X**				**X**	
(S. Kotsiantis et al., 2004)	**X**	**X**				**X**	**X**	
(Zhang & Oussena, 2010)	**X**	**X**	**X**			**X**		**X**
(Aníbal Alviz M. y Erika J. quintero, 2022),					**X**			
(Kalles & Pierrakeas, 2006),	**X**	**X**	**X**		**X**		**X**	**X**
Cheewaprakobkit, P. (2013)	**X**						**X**	
(Moseley & Mead, 2008)	**X**							
(T. Mishra et al., 2014)				**X**				
(ara et al., 2015)		**X**	**X**	**X**				
(Solis et al., 2018)		**X**		**X**	**X**			
TOTALES	**11**	**6**	**5**	**4**	**5**	**3**	**5**	**3**

Nota. La tabla 17 relaciona los autores con los modelos que utilizaron en su respectiva investigación, es así como la X en cada celda significa que el autor utilizó ese modelo. Al final en la celda Totales se muestra la cantidad de veces que fue usado el modelo dentro de la revisión bibliográfica, siendo el modelo Decision Tree, seguido del modelo Support Vector Machine con 11 y 6 veces respectivamente, los más usados en las diferentes fases experimentales.

Figura 14

Algoritmos más Usados en las Investigaciones



Nota. La gráfica de la figura 14, representa los valores obtenidos de la tabla 17.

Tabla 18

Métricas más Aplicadas a Técnicas más Usadas en la Literatura Analizada

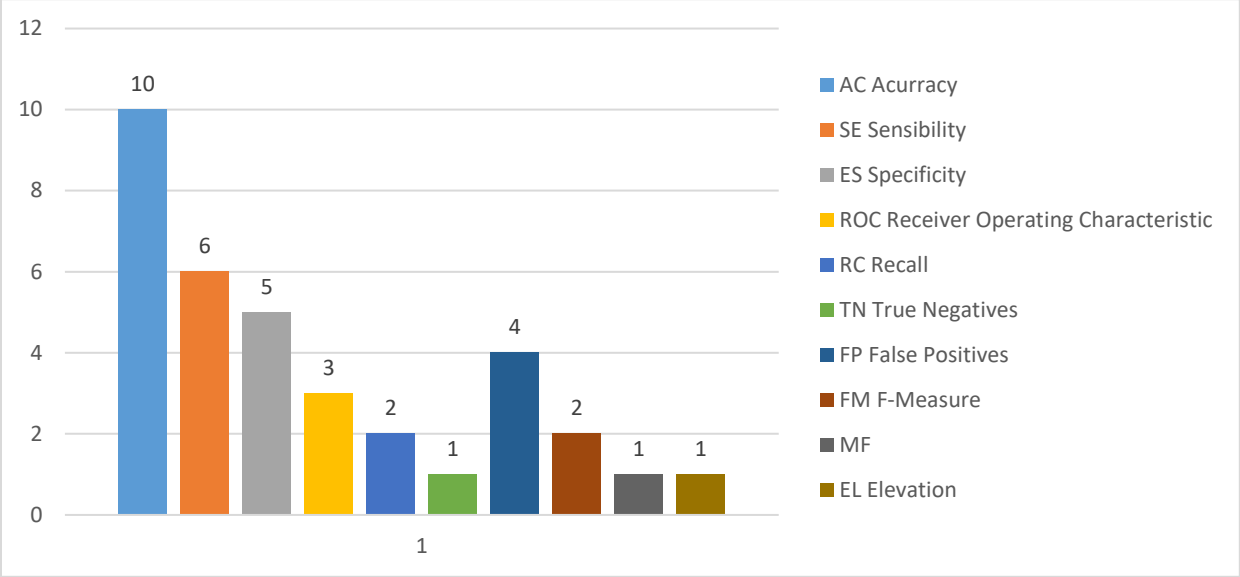
Referencia	AC Acurracy	SE Sensibility	ES Specificity	ROC Receiver Operating Characteristic	RC Recall	TN True Negative	FP False Positive	FM F-Measure	MF	EL Elevation
(Thomas, Emily & Galambos, Nora 2004)	**X**	**X**	**X**							
(Valero Cajahuanca, Julio Elvis, 2022)		**X**	**X**							

Referencia	AC Acurracy	SE Sensibility	ES Specificity	ROC Receiver Operating Characteristic	RC Recall	TN True Negative	FP False Positive	FM F-Measure	MF	EL Elevation
(Aníbal C, 2020)	**X**			**X**			**X**	**X**		
(Hernández González et al., 2016)	**X**	**X**							**X**	**X**
(Miranda & Guzmán, 2017)	**X**	**X**	**X**	**X**	**X**		**X**	**X**		
(Zhang & Oussena, 2010)	**X**									
(Kalles & Pierrakeas, 2006),	**X**									
Cheewaprakobkit, P. (2013)	**X**									
(Moseley & Mead, 2008)	**X**	**X**	**X**							
(T. Mishra et al., 2014)	**X**				**X**		**X**			
(ara et al., 2015)	**X**			**X**						
(Solis et al., 2018)		**X**	**X**			**X**	**X**			
TOTALES	**10**	**6**	**5**	**3**	**2**	**1**	**4**	**2**	**1**	**1**

Nota. En el Apéndice A, se puede tener el concepto de cada una de las métricas

Figura 15

Las métricas más Usados en las Investigaciones



Nota. La gráfica de la figura 15 muestra las métricas más aplicadas a las técnicas de machine learning, dentro del revisión bibliográfica.

Tabla 19

Contrastación de Modelos y sus Respectivas Métricas de Evaluación

Técnica	Descripción	Métricas Comunes de Evaluación
Regresión Lineal	Algoritmo supervisado utilizado para predecir valores continuos.	- **Error Cuadrático Medio (MSE):** Medida de la diferencia entre las predicciones y los valores reales. - **R^2 (Coeficiente de Determinación):** Mide qué tan bien las predicciones se ajustan a los datos observados.
Regresión Logística	Algoritmo supervisado para clasificación binaria, donde las salidas son probabilidades.	- **Precisión (Precision)**: Proporción de verdaderos positivos entre todos los casos clasificados como positivos. - **Recall (Sensibilidad)**: Proporción de verdaderos positivos entre todos los casos que realmente son positivos. - **F1-Score**: Promedio armónico entre precisión y recall. - **AUC-ROC**: Área bajo la curva ROC

Técnica	Descripción	Métricas Comunes de Evaluación
Máquinas de Vectores de Soporte (SVM)	Utiliza un hiperplano para clasificar datos en clases.	- **Precisión (Precision)**: Mide qué tan exactos son los positivos predichos. - **Recall (Sensibilidad)**: Proporción de casos positivos correctamente identificados. - **F1-Score**: Equilibrio entre precisión y recall. - **AUC-ROC**: Mide la capacidad de clasificación del modelo.
Árboles de Decisión	Modelo de clasificación o regresión basado en una estructura de árbol de decisiones.	- **Precisión (Accuracy)**: Porcentaje de predicciones correctas. - **Precisión (Precision)**: Relación de casos positivos correctos sobre los predichos. - **Recall (Sensibilidad)**: Identificación de positivos verdaderos. - **F1-Score**: Balance entre precisión y recall.
Bosques Aleatorios (Random Forest)	Conjunto de árboles de decisión para clasificación o regresión.	- **Precisión (Accuracy)**: Proporción de predicciones correctas. - **F1-Score**: Evaluación global del rendimiento. - **AUC-ROC**: Evalúa la capacidad de clasificación del modelo.
Redes Neuronales Artificiales	Modelos inspirados en el cerebro humano para aprender patrones complejos.	- **Precisión (Accuracy)**: Proporción de aciertos. - **Precisión (Precision)** y **Recall (Sensibilidad)**: Miden la calidad de clasificación, especialmente en clases desbalanceadas. - **F1-Score**: Compensa precisión y recall.
K-Vecinos más Cercanos (KNN)	Clasificación basada en la similitud entre instancias cercanas.	- **Precisión (Precision)**: Casos positivos correctamente predichos. - **Recall (Sensibilidad)**: Identificación de verdaderos positivos. - **F1-Score**: Balance entre precisión y recall.
Naive Bayes	Clasificador probabilístico basado en el teorema de Bayes.	- **Precisión (Precision)**: Proporción de verdaderos positivos sobre los predichos. - **Recall (Sensibilidad)**: Proporción de positivos verdaderos que el modelo detectó. - **F1-Score**: Promedio armónico entre precisión y recall. - **AUC-ROC**: Evalúa la capacidad de clasificación en problemas binarios.
Algoritmos de Clustering (como K-means)	Técnica no supervisada para agrupar datos en clústeres.	- **Silhouette Score**: Mide la calidad del clustering. Un valor cercano a 1 indica un buen agrupamiento. - **Índice de Dunn**: Mide la separación entre los clústeres. - **Índice de Calinski-Harabasz**: Evalúa la dispersión entre los clústeres.

Técnica	Descripción	Métricas Comunes de Evaluación
Reducción de Dimensionalidad (PCA, t-SNE)	Técnicas utilizadas para reducir el número de características de los datos.	- **Varianza Explicada (para PCA)**: Mide cuánta varianza de los datos es capturada por las componentes principales. - **Kullback-Leibler Divergence (para t-SNE)**: Mide la similitud entre la distribución de los datos originales y la reducción dimensional.

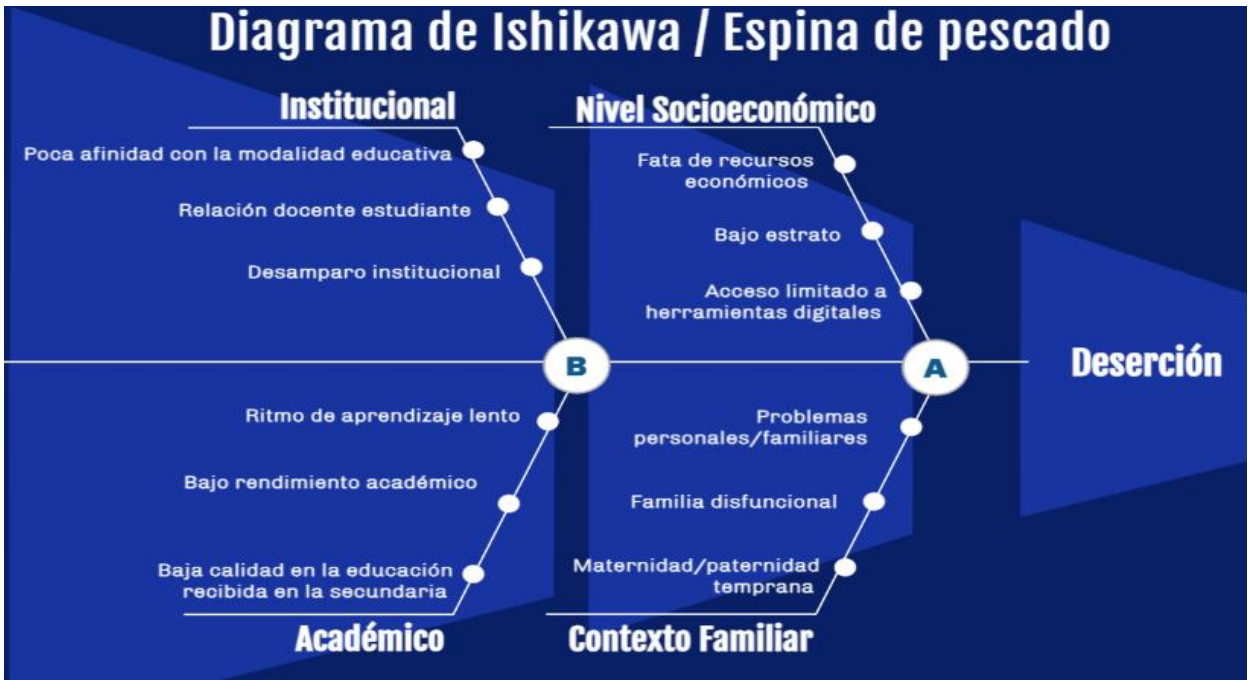
Nota. La tabla 19, muestra un concepto de las técnicas y sus respectivas métricas de evaluación modelos. Tomado de: (James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013).

Análisis de las Variables y Factores Más Comunes que Repercuten en la Deserción Universitaria

Estos factores están asociados a variables Institucionales, socioeconómicas, académicas y al contexto familiar como se aprecia en la siguiente ilustración de la figura 6.

Figura 16

Diagrama de Ishikawa de Factores Asociados a la Deserción



Nota. Tomado de: EDIT.org sobre los datos de (Mario Luis Ávila, 2021).

Factores Económicos

El análisis de las variables del factor socioeconómica que impactan la deserción en la educación superior en Colombia debe considerar diversos aspectos relacionados con el entorno del lugar de residencia del estudiante, como son, el nivel de ingreso familiar, el acceso a recursos educativos, el empleo, la disponibilidad de becas o apoyos financieros.

El aspecto económico es esencial en la permanencia; es por ello por lo que estudiantes provenientes de hogares con niveles de ingresos bajos enfrentan mayores obstáculos para continuar sus estudios debido a las dificultades que afrontan y que no les permiten cubrir los costos pecuniarios, materiales educativos, transporte, alimentación y alojamiento. En la gráfica 8, se muestra la gran diferencia entre la deserción de estratos bajos relacionados con los estratos más altos.

Factores Familiares

La falta de apoyo emocional y económico de las familias afecta la continuidad en los estudios. La educación superior puede ser vista como una carga para algunas familias, especialmente cuando los estudiantes provienen de entornos donde la educación no es vista como una prioridad.

El entorno familiar influye significativamente en la permanencia del estudiante en el sistema educativo. Aspectos como el nivel educativo de los padres y la estabilidad familiar son determinantes. Además, eventos familiares como enfermedades o situaciones de crisis pueden afectar la continuidad académica.

Spady (1970) basado en el modelo suicida de Durkheim (1951), aduce que el medio familiar es una de las muchas fuentes que expone a los estudiantes a influencias, expectativas y demandas, las que a su vez afectan su nivel de integración social en la universidad.

Factores Académicos

El rendimiento académico es uno de los factores que puede incidir en la decisión de abandonar los estudios (Olave & Cisneros, 2013). Este rendimiento académico está determinado por otros factores tales como la preparación previa al ingreso a la educación superior la cual es crucial, parte de esa preparación se refleja en una buena presentación de las pruebas saber 11; esta preparación previa puede predecir en un alto porcentaje sobre el futuro del estudiante en cuanto a la permanencia en la universidad.

Según el Informe LEE No. 74, la tasa de deserción es mayor entre los estudiantes con bajo desempeño en el examen Saber 11. La gráfica 5, muestra cómo, aquellos estudiantes que ingresaron a la universidad en el primer semestre de 2016, el 31,1% de los estudiantes con bajo desempeño en el examen Saber 11 desertaron al quinto semestre, el doble del porcentaje entre los estudiantes con alto desempeño en estas pruebas (14,3%).

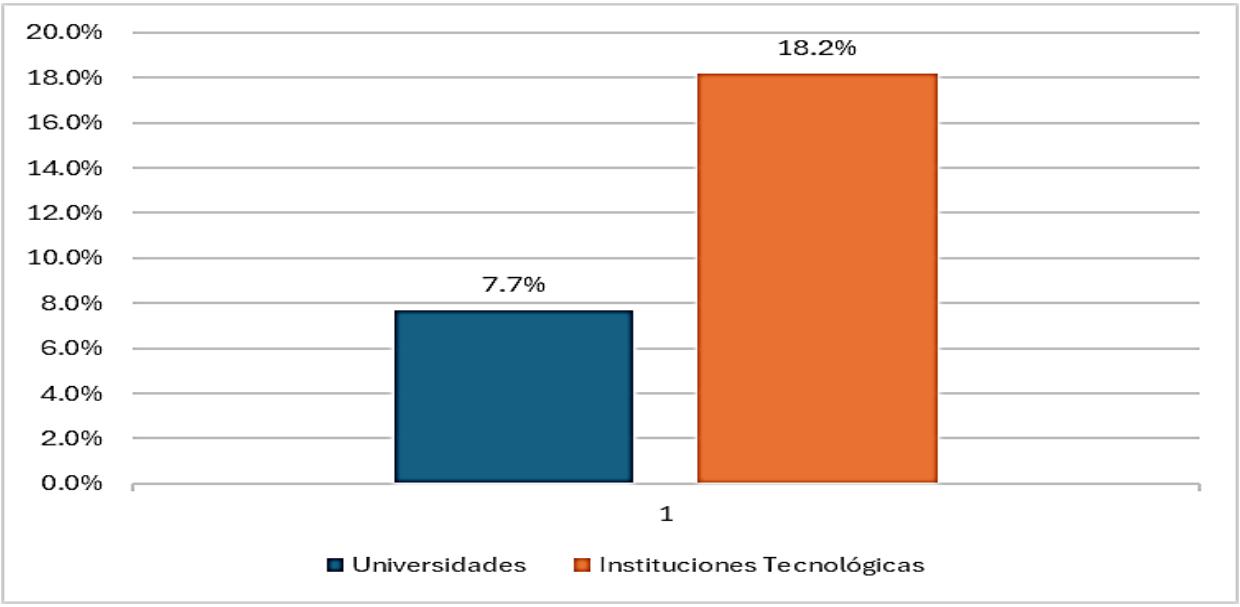
Transcurridos diez semestres desde el ingreso, la tasa de deserción alcanza el 43,0% para los estudiantes con bajo desempeño y 23,5% para los estudiantes con alto desempeño.

Factores Institucionales

La infraestructura de las instituciones de educación superior (IES) también influyen en la deserción. El ambiente, las metodologías, los procesos, los modelos pedagógicos y los diseños curriculares no llenan las expectativas de los estudiantes que llegan con una percepción y se encuentran con una realidad totalmente diferente. Las tasas de deserción varían según el tipo de institución y el programa académico. Además, la calidad de los programas, la disponibilidad de recursos académicos y el apoyo institucional son factores determinantes en la decisión de los estudiantes de continuar o abandonar sus estudios. Por ejemplo, la gráfica 10 muestra la tasa de deserción por tipo institución en el segundo semestre de 2021.

Figura 17

Tasa de Deserción Según el Tipo de Institución



Identificar las Limitaciones y Desafíos en la Aplicación de Machine Learning para la Detección de la Deserción Universitaria

Al momento de entrenar un modelo para tratar de obtener resultados precisos, en cuanto a la problemática de la deserción en la educación superior, se van a encontrar muchos desafíos y limitaciones, entre ellos tenemos los siguientes**:**

Calidad y Disponibilidad de Datos

Datos faltantes**:** si la datasetet tiene datos nulos, faltantes o incompletos, hace que se dificulte la identificación de las variables de estudio y por ende dificulta el entrenamiento de los modelos impidiendo obtener la precisión o rendimiento esperado (Moseley & Mead, 2008; Sangodiah et al., 2015; Thomas & Galambos, 2004).

Conjunto de Datos Desbalanceados

Esto ocurre cuando una clase tiene muchas más muestras que otra. Según (Delen, 2010; Moseley & Mead, 2008; Sarker et al., 2014). Cuando hay un porcentaje alto de estudiantes que permanecen y pocos datos de estudiantes que han desertado.

Sesgos en los Datos

Los conjuntos de datos pueden contener sesgos relacionados con factores como género, etnia o nivel socioeconómico, lo que afecta la equidad y la generalización del modelo.

Data Desactualizada

Los patrones de deserción en la educación superior pueden cambiar con el tiempo, por lo que el uso de datos antiguos puede reducir la eficacia en la predicción del modelo.

Protección de los Datos

Las instituciones educativas deben cumplir con normativas de privacidad de los datos, respetando la Ley de protección de los mismo (Ley 1582 de 2012).

Complejidad de los Factores Asociados a la Deserción

La deserción universitaria depende de múltiples factores (académicos, emocionales, económicos, sociales), lo que hace difícil capturar todas las variables relevantes. Algunas variables como la motivación personal o el estrés no siempre están reflejadas en los datos disponibles.

Interpretación del Modelo

Los modelos de ML, especialmente los de tipo "caja negra" (p. ej., redes neuronales profundas), pueden ser difíciles de interpretar, lo que dificulta la justificación de sus predicciones ante responsables académicos o estudiantes.

Recursos y Capacidad Técnica

La implementación y el mantenimiento de modelos de Machine Learning, requiere recursos computacionales muy rigurosos, que no todas las instituciones pueden costear.

Estrategias para Mitigar estos Desafíos

Estos desafíos destacan la importancia de una planificación cuidadosa y un enfoque ético al aplicar ML en el ámbito educativo (Tufféry, S. 2011).

Sobre Entrenamiento (Overfitting)

Este se causa cuando el clasificador identifica que una clase se relaciona demasiado con el conjunto de datos de entrenamiento y se generaliza mal las nuevas observaciones (Alkhasawneh & Hobson, 2011; Barnes et al., 2009; Beaulac & Rosenthal, 2019; Lee & Chung, 2019; Lykourentzou et al., 2009).

Conclusiones

El desarrollo de esta monografía deja en evidencia el uso de la ciencia de datos como herramienta esencial para la predicción temprana de los factores que determinan la deserción en la educación superior, para lo cual se apoya en algoritmo de aprendizaje automático supervisado y no supervisado (machine learning) y Aprendizaje profundo (Deep Learning)

Dentro de la revisión sistemática de los artículos analizados, se evidencia que el modelo más utilizado en la literatura fue el modelo Árbol de decisión AD (Decision Tree) (ver tablas 17), esto se debe a que el algoritmo ADTree es una técnica con simplicidad, interpretabilidad y capacidad para manejar datos heterogéneos (para tener un mayor concepto sobre ADTree diríjase al Apéndice_B).

De igual forma también se evidenció que la métrica fue la que más se utilizó en los análisis realizados por los autores, esto se debe a que es una de las métricas más utilizadas en clasificación supervisada debido a su fácil interpretación, es ideal para clases balanceadas, sus cálculos son rápidos, es de amplia adopción (ver Apéndice_C).

El análisis muestra que el modelo Random Forest resultó ser el de mejor rendimiento para la predicción de las variables objetivo de la deserción, obteniendo altos porcentajes de precisión como se observa en la tabla 20.

Tabla 20

Autores y Precisión del Modelo Random Forest

Random Forest	
Autores	**Precisión**
(Aníbal C, 2020)	84.1%
(T. Mishra et al., 2014)	94.4%
(ara et al., 2015)	93.5%
(Solis et al., 2018)	87.0%

Nota. La tabla 20, muestra los autores y el rendimiento del modelo Random Forest.

Es pertinente tener en cuenta que, para obtener mejores resultados, es decir una mejor precisión en los modelos implementados, es de vital importancia aplicar un buen preprocesamiento de datos (limpieza de datos, transformación de variables, selección de atributos y rebalanceo de los datos etc.), además de aplicar las métricas de mejor desempeño.

De acuerdo con el análisis realizado sobre la literatura referente a la problemática, se observó que las variables que más utilizaron los autores para predecir patrones de comportamiento fueron las de los factores socioeconómico y de entorno familiar, Siendo estos factores los que consecuentemente determinan en mayor proporción la causa de la deserción en la educación superior según SPADIES en su informe del año 2022-2.

Finalmente, se puede decir que el fenómeno de la deserción es un problema intrínseco en un contexto ambiguo en el que de nada servirá la implementación de todo el potencial de la ciencia de datos como herramienta imperativa para la predicción temprana de los patrones de comportamiento de la variables que determinan la problemática, si, las instituciones no diseñan nuevas estrategias donde el constructor social sea objetivo principal de esas estrategias y si el gobierno no transforma las necesidades del clamor estudiantil en políticas públicas que en el mediano plazo permitan afrontar esos desafíos que exige la educación superior de calidad.

¿Por qué Random Forest tuvo mejor rendimiento, al predecir los factores que influyen en la problemática de la deserción estudiantil?

Random Forest es un algoritmo de aprendizaje automático supervisado que se usa para solucionar problemas de clasificación y regresión a partir de la construcción de árboles de decisión individuales para cada muestra.

Una de las características más importantes del Random Forest es que puede manejar un conjunto de datos que contenga variables continuas, como en el caso de la regresión, y variables

categóricas, como en el caso de la clasificación. Por eso ofrece mejores resultados para problemas de clasificación.

El análisis de los factores que determinan de la deserción implica múltiples variables: lo que hace a Random Forest esencial para la deserción, debido a que este algoritmo puede manejar este tipo de datos mixtos (numéricos y categóricos) sin necesidad de preprocesamiento extenso.

Lo que se Pretende con esta Monografía

Por medio de esta monografía se pretende minimizar el impacto de la problemática de la deserción en la educación superior, apoyando y aportando desde esta revisión documental a las investigaciones realizadas por otros autores; además de incentivar a las instituciones educativas para que desde los primeros niveles de la educación comiencen a utilizar la Ciencia de datos como herramienta esencial para dar solución a diferentes problemáticas tanto en lo cotidiano como en lo empresarial implementando técnicas de machine learning para la predicción temprana de los factores y variables que influyen en la problemática de estudio.

Recomendaciones

Desde lo Institucional: es necesario que se hagan seguimientos desde el departamento de psicología, en el que se brinde acompañamiento al estudiante involucrando a la familia. Esto facilitará la detección de patrones de comportamientos no deseados en los estudiantes

Es importante que las instituciones de educación superior cuenten con un sistema de predicción temprana de los factores que determinan la de deserción.

Es necesario que se hagan seguimientos desde el departamento de psicología, en el que se brinde acompañamiento al estudiante involucrando a la familia. Esto facilitará la detección de patrones de comportamientos no deseados en los estudiantes

En el desarrollo de la investigación se pudo ver que se utiliza mucho el software WEKA. Weka, es un software muy interactivo e intuitivo que contiene una colección integral de algoritmos de modelado y herramientas de análisis de datos (ver Apéndice_D para ampliar detalles sobre este software). Es por eso por lo que, desde este trabajo encaminado a analizar la problemática de la deserción universitaria, se recomienda este software de uso libre desarrollado en Java por la Universidad de Waikato en Nueva Zelanda (Birjali et al., 2018).

Desde lo Gubernamental (MEN**):** es urgente que, desde el gobierno, el Congreso de la República y desde las Altas Cortes, se asuma esta problemática con toda la responsabilidad social del caso, decretando políticas públicas que le permitan al sector de la educación transferirle una mayor inversión.

Referencias

Alexánder B. (2023) "Accuracy vs. Precision vs. Recall in Machine Learning". Recuperado de: https://encord.com/blog/classification-metrics-accuracy-precision-recall/

Alonso, P., & García, M. (2019) *Introducción al aprendizaje automático: Fundamentos, algoritmos y aplicaciones*. Editorial McGraw-Hill.

Aníbal Alviz M. y Erika J. quintero, (Universidad Nacional, 2022), *Proyecto Aplicado en Ciencia de Datos*: Visualización y Análisis de Datos del Ministerio de Educación de Colombia 2022, https://www.researchgate.net/publication/362781514.

Ara, N-B., et al, S. (2015). *High-school dropout prediction using machine learning: a Danish large-scale study*. In M. Verleysen (Ed.), Proceedings. ESANN 2015: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (pp. 319-324). i6doc.com https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-86.pdf

Banco Mundial. (2017). *Graduating: Only half of Latin American students manage to do so*. https://www.bancomundial.org/es/news/feature/2017/05/17/graduating-only-half-of-latin-american-students-manage-to-do-so

Camargo García A. (2020), *Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos*. Corporación Universidad de la Costa. Disponible en: https://hdl.handle.net/11323/7077

Cavelier,Andres, (2022) *Apoyo a Colombia para aumentar la cantidad de estudiantes graduados de educación superior*, https://www.iadb.org/es/noticias/bid-apoya-colombia-para-aumentar-la-cantidad-de-estudiantes-graduados-de-educacion?utm_source=chatgpt.com

Constitución Política de Colombia 1991, https://minciencias.gov.co/sites/default/files/upload/reglamentacion/ConstitucionPoliticaColombia-1991.pdf

Dekker, G.et al (2009). *Predicting Students Drop Out: A Case Study. Educational Data Mining*. https://eric.ed.gov/

Domínguez L. (2023), *El fenómeno va en aumento año tras año y las carreras técnicas y tecnológicas son las más afectadas*. El Tiempo. https://www.eltiempo.com/vida/educacion/tristes-estadisticas-de-desercion-universitaria-mitad-de-estudiantes-no-se-gradua-789914

Doricela Gutiérrez C. et al, (2017). *Taller con Weka*, http://ri.uaemex.mx/bitstream/handle/20.500.11799/69982/secme-30553_1.pdf?sequence=1

Evidently AI Team (2025), *Accuracy vs. precision vs. recall in machine learning*. https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

FasterCapital (2020), *Ventajas y desventajas de los árboles de decisión*, https://fastercapital.com/es/tema/ventajas-y-desventajas-de-los-%C3%A1rboles-de-decisi%C3%B3n.html

Fredy romero Chacon (2021), *La evolución de la educación universitaria en Colombia*, https://repository.unimilitar.edu.co/server/api/core/bitstreams/36a761a9-29a9-4630-8a0e-e6ad23e83777/content.

Gabriel Quirós R., (2024). *Más allá del PIB: cómo medir el bienestar económico*, https://www.bde.es/wbe/es/noticias-eventos/blog/mas-alla-del-pib-como-medir-el-bienestar-economico.html

Garzón, J. (2023). *Deserción universitaria en Colombia: una problemática que afecta a millones de jóvenes. Medium,* https://medium.com/%40juanjose.garzon/deserci%C3%B3n-universitaria-en-colombia-una-problem%C3%A1tica-que-afecta-a-millones-de-j%C3%B3venes-fc6ac69a3ab9

González, R., & Uribe, D. (2018). *Variables Sociodemográficas y Académicas Explicativas de la Deserción Universitaria en una Universidad Pública Chilena*. Formación Universitaria, https://www.scielo.cl/scielo.php?pid=S0718-50062018000200003&script=sci_arttext

Google Developers. *Clasificación: Exactitud, recuperación, precisión y métricas relacionadas*. Curso intensivo de aprendizaje automático. recuperado de: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419

Gordillo, E. & Polanco, J. (1970). *Deserción Estudiantil: Análisis Cuantitativo*. Oficina de Planeación, División de Programación Económica. Bogotá: Universidad Nacional de Colombia.

Haddaway, N. R. et al (2022). PRISMA2020: *An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams,* with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230

Han, J., et al (2012), *Data Mining – Concepts and Techniques*, 3rd ed.

Hernandez Gonzalez, et al (2016), *Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico.*

Hernández Jiménez, et al (2019). *Estudio descriptivo de variables sociodemográficas y motivacionales asociadas a la deserción: la perspectiva de personas universitarias de primer ingreso*. Revista Educación, 44(1). Tomado de: https://www.redalyc.org/journal/440/44060092023/44060092023.pdf

Hoyos Osorio, et al (2023). *Predictive Model to Identify College Students with High Dropout Rates*. Revista electrónica de investigación educativa, 25, e13. Epub 26 de junio de 2023.https://doi.org/10.24320/redie.2023.25.e13.5398

IBM (s.f.), *¿Qué es un árbol de decisión*?, https://www.ibm.com/es-es/topics/decision-trees

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*: With Applications in R.

Javier García E. *Principales Retos de la Educación Media y Universitaria en Colombia* (2023), https://cider.uniandes.edu.co/es/retos-educacion-media-universitaria-Colombia-09-23

Juber Gutiérrez et al (2021), *Estimación de las principales causas de la deserción universitaria mediante el uso de técnicas de Machine Learning*, AGLALA ISSN 2215-7360 2021; julio-diciembre. Vol. 12, N°2. PP. 293-311.

Kabathova, Janka, and Martin Drlik. 2021. "*Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques*" Applied Sciences 11, no. 7: 3130. https://doi.org/10.3390/app11073130

Kalles, D., & Pierrakeas, C. (2006). *Analyzing Student Performance In Distance Learning With Genetic Algorithms And Decision Trees*. Applied Artificial Intelligence, 20(8), 655–674. https://doi.org/10.1080/08839510600844946

Kemper, Lorenz, et al (2020). *Predicting student dropout: A machine learning approach*. European Journal of Higher Education. 10. 1-20. 10.1080/21568235.2020.1718520.

Kuleto, V., et al (2021). *Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher* Education Institutions. Sustainability, 13(18), 10424. https://doi.org/10.3390/su131810424Education Institutions. Sustainability. 2021; 13(18):10424. https://doi.org/10.3390/su131810424.

Kuz, A., & Morales, R. (2023). *Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México*. Education in the Knowledge Society (EKS), 24, e30080. https://doi.org/10.14201/eks.30080.

M. Solis, et al 2018, *Perspectives to Predict Dropout in University Students with Machine Learning*, pp. 1-6, doi: 10.1109/IWOBI.2018.8464191.

McKenzie, et al D. (2021). *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Ministerio de Educación Nacional (2023), A*umento de la Deserción Escolar en los Últimos Años en Colombia.* https://www.mineducacion.gov.co/portal/salaprensa/Comunicados/415819:Ante-el-aumento-de-la-desercion-escolar-en-los-ultimos-anos-en-Colombia-el-Ministerio-de-Educacion-ha-implementado-estrategias-para-prevenir-que-los-estudiantes-abandonen-las-aulas#:~:text=Adicionalmente%2C%20los%20estudiantes%20que%20abandonan,de%20desarrollo%20personal%20y%20profesional.

Ministerio de Educación Nacional (2023), *Estadísticas de Deserción y Permanencia en Educación Superior Spadies 3.0 - Indicadores 2022,* https://www.mineducacion.gov.co/sistemasinfo/spadies/secciones/Estadisticas-de-desercion/

Miranda, Mauricio A, & Guzmán, Jheser. (2017). *Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos*. Formación universitaria, 10(3), 61-68. https://dx.doi.org/10.4067/S0718-50062017000300007

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & *The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses*: The PRISMA Statement. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moseley, Laurence & Mead, Donna. (2008). *Predicting who will drop out of nursing courses: A machine learning exercise. Nurse education today*. 28. 469-75. 10.1016/j.nedt.2007.07.012.

Muñoz-Camacho, et al. (2018). *Probabilidad de Deserción Estudiantil en Cursos de Matemáticas Básicas en Programas Profesionales de la Universidad de Los Andes-Venezuela.* Formación universitaria, 11(4), 33-42. https://dx.doi.org/10.4067/S0718-50062018000400033

Nagesh Singh Chauhan (2023), "*Métricas De Evaluación De Modelos En El Aprendizaje Automático*", https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Noemí Guillamón Cano, (2003), *Variables socioeconómicas y problemas interiorizados y exteriorizados en niños y adolescentes,* https://www.tesisenred.net/bitstream/handle/10803/5432/ngc1de1.pdf?sequence=1&isAllowed=y

ONU, (2020), Organización de las Naciones Unidas, *Informe de políticas: La educación durante la COVID- y después de ella: 2020*, https://www.un.org/...id-19_and_beyond_spanish.pdf

Provost, F., & Fawcett, T. (2013). *Data Science for Business*

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Quintero Salazar (2021), *Diez retos para la educación en Colombia,* https://www.elpais.com.co/educacion/diez-retos-de-la-en-colombia-para-el-2021.html

Reyes Rodríguez, I. (2022). *Diseño de un modelo de caracterización de la deserción de un estudiante de pregrado del departamento de Ingeniería Eléctrica y Electrónica de la Universidad de Los Andes por medio de técnicas de machine learning. Universidad de los Andes*. Disponible en: http://hdl.handle.net/1992/58923

Sectorial. (2024). *Con una deserción promedio anual del 11% en educación superior, en Colombia se puede perpetrar una pobreza generacion*al. https://sectorial.co/articulos-especiales/con-una-desercion-promedio-anual-del-11-en-educacion-superior-en-colombia-se-puede-perpetrar-una-pobreza-generacional/?utm_source=chatgpt.com

SNIES (2018), *El Ministerio de Educación Nacional pone a disposición la información estadística de educación superior 2021,* https://snies.mineducacion.gov.co/portal/401926:#:~:text=Para%202021%20la%20matr%C3%ADcula%20total,%2C93%25%20respecto%20a%202020.

Solano Benavides, E. S. B., & Barraza Niebles, M. (2018, September 29). *Deserción en la educación superior*. Editorial Universidad del Atlántico. https://investigaciones.uniatlantico.edu.co/omp/index.php/catalog/catalog/book/34

T. Mishra, D. Kumar and S. Gupta, "*Mining Students' Data for Prediction Performance*," 2014 Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 2014, pp. 255-262, doi: 10.1109/ACCT.2014.105.

Team Asana (2023). "*El árbol de decisiones: un análisis de 5 pasos para tomar mejores decisiones*", https://asana.com/es/resources/decision-tree-analysis

Thomas, Emily & Galambos, Nora 2004, WHAT SATISFIES STUDENTS?, *Mining Student-Opinion Data with Regression and Decision Tree Analysis*, Research in Higher Education, Vol. 45, 2004), https://www.researchgate.net/publication/226068922_What_Satisfies_Students_Mining_Student-Opinion_Data_with_Regression_and_Decision_Tree_Analysis

Tinto, V. (1982). *Límites de la teoría y la práctica en el desgaste estudiantil*. Revista de Educación Superior, 53(6), 687-700.

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making*. Wiley.

UED (2020), *Árbol de Problemas*, https://www.youtube.com/watch?v=BvcAhSv-BRE&t=87s

UNESCO (s.f.). https://en.unesco.org/higher-education/iesalc.

Universidad de Navarra (2020), *Revisiones sistemáticas: PRISMA 2020: guías oficiales para informar (redactar) una revisión sistemática*, *Prisma 2020*: Guías oficiales para informar (redactar) una revisión sistemática.

Universidad Javeriana, (2023), *Informe LEE No. 74.* https://lee.javeriana.edu.co/-/lee-informe-74, Deserción en la educación superior en Colombia

Valero Cajahuanca, Julio Elvis, 2022, Revista de Ciencias Sociales, *Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción*. (Ve) ISSN: 1315-9518 rcs_luz@yahoo.com Universidad del Zulia República Bolivariana de Venezuela

Yeny Álvarez C. (2014), *¿Cómo hacer un árbol de objetivos*?, https://www.youtube.com/watch?v=d9mPMLGy3D4

Zhang, Ying, et al (2010). *Use Data Mining to Improve Student Retention in Higher Education - A Case Study*. 190-197.

Apéndices

Apéndice A

Concepto de Métricas más Utilizadas por los Autores en la Literatura de Estudio

Métrica	Concepto	Formula
Accuracy – AC	Es una métrica utilizada en estadística, aprendizaje automático y clasificación para evaluar el desempeño de un modelo o sistema de predicción. Representa la proporción de predicciones correctas en relación con el total de predicciones realizadas.	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity, SE	También conocida como recall o tasa de verdaderos positivos, mide la capacidad de un modelo para identificar correctamente los casos positivos dentro de un conjunto de datos. Es especialmente útil en problemas donde los falsos negativos (FN) son costosos o peligrosos, como en el diagnóstico médico.	$Sensitivity = \frac{TP}{TP + FN}$
Specificity, ES	Mide la capacidad de un modelo para identificar correctamente los casos negativos dentro de un conjunto de datos. Es decir, evalúa qué tan bien el modelo evita clasificar erróneamente como positivos aquellos casos que en realidad son negativos.	$Specificity = \frac{TN}{TN + FP}$
GM – Geometric Mean	es una medida estadística que se utiliza para calcular el promedio de un conjunto de valores positivos, teniendo en cuenta su crecimiento proporcional en lugar de su suma directa. Se emplea comúnmente en análisis de tasas de crecimiento, modelos de clasificación en aprendizaje automático y economía.	$GM = \sqrt[n]{x_1 \times x_2 \times ... \times x_n}$
Recall (RC)	mide la capacidad de un modelo de clasificación para identificar correctamente los casos positivos dentro de un conjunto de datos.	$Recall = \frac{TP}{TP + FN}$
F-Measure	Esta métrica se utiliza en modelos de clasificación para encontrar un equilibrio entre precisión (Precision) y recall (Sensibilidad). Se utiliza especialmente cuando los datos están desbalanceados.	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
TN – True Negatives	Son los casos en los que un modelo de clasificación predice correctamente que una instancia pertenece a la clase negativa.	
FP (False Positives),	son los casos en los que un modelo de clasificación predice incorrectamente que una instancia pertenece a la clase positiva, cuando en realidad debería ser clasificada como negativa. Es	

	decir, el modelo "marca" erróneamente un ejemplo negativo como positivo.	
Métrica	**Concepto**	**Formula**
Elevación (EL)	Se utilizada en achine learning y análisis de modelos de clasificación, especialmente en problemas de clasificación. La elevación mide cuán eficaz es un modelo al predecir los casos positivos en comparación con un modelo aleatorio o un modelo base.	$Elevation = \frac{\text{Tasa de Verdaderos Positivos del Modelo}}{\text{Tasa de Verdaderos Positivos del Modelo Aleatorio}}$
Curva ROC (Receiver Operating Characteristic)	Es una herramienta gráfica utilizada para evaluar el rendimiento de un modelo de clasificación binaria. La ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación.	

Nota: Tomado de:(Alonso, P., & García, M. (2019).

Apéndice B

Ventajas del Árbol de Decisión (Decision Tree) sobre Otros Clasificadores

Concepto	Característica
Fácil de interpretar	Puede representarse gráficamente, lo que facilita su comprensión incluso para personas sin experiencia en machine learning.
Manejo de datos mixtos	Funciona bien con variables numéricas y categóricas sin necesidad de preprocesamiento extenso.
Poca necesidad de normalización o estandarización	A diferencia de modelos como SVM o regresión logística, no requiere escalar los datos.
Capacidad de manejar valores faltantes	Puede dividirse en función de los datos disponibles sin necesidad de eliminación de valores faltantes.
Rápido en entrenamiento y predicción	En comparación con modelos como redes neuronales o SVM, los árboles de decisión entrenan rápidamente.
Capacidad para capturar interacciones entre variables	Detecta relaciones no lineales y combinaciones de factores que influyen en la clasificación.
No necesita supuestos estadísticos	A diferencia de la regresión logística, que asume relaciones lineales entre variables.

Nota. Tomado de: IBM, FasterCapital, Asana y QuestionPro

Apéndice C

Importancia y Ventajas de Acurracy sobre Otras Métricas

Importancia:

- Fácil de interpretar: Representa el porcentaje de predicciones correctas, lo que facilita su comprensión incluso para no expertos en machine learning.
- Adecuada para clases balanceadas: Cuando las clases están bien distribuidas en el dataset, la Accuracy refleja correctamente el rendimiento del modelo.
- Rápida de calcular: Es una métrica computacionalmente eficiente y fácil de obtener.
- Amplia adopción: Es la métrica base en muchas aplicaciones y permite comparaciones rápidas entre modelos.

Ventajas de la Accuracy Sobre Otras Métricas:

Métrica	Ventajas
Precisión (Precision)	Accuracy evalúa el rendimiento global, mientras que la precisión solo mide la exactitud en las predicciones positivas.
Sensibilidad (Recall)	Accuracy no se enfoca solo en los verdaderos positivos, sino en todas las predicciones correctas.
F1-Score	Accuracy es más sencilla de calcular y entender, mientras que F1-Score requiere un balance entre precisión y recall.
Matriz de Confusión	Accuracy da un resumen global del desempeño del modelo sin necesidad de analizar múltiples valores.

Nota. Tomado de Google Developers (2023), Evidently AI, Nagesh Singh C., Alexander B. (2023)

Apéndice D

Software para Implementar Machine Leraning y Minería de Datos Weka

Weka (Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL.

Contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. La versión original de Weka fue un front-end en TCL/TK para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para pre-procesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura, pero la versión más reciente basada en Java (WEKA 3), que empezó a desarrollarse en 1997, se utiliza en muchas y muy diferentes áreas, en particular con finalidades docentes y de investigación.

Dentro de los principales aspectos relevantes de Weka, destacan: licencia pública general de GNU, portabilidad; está completamente implementado en Java y puede correr en casi cualquier plataforma, contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado, es fácil de utilizar gracias a su interfaz gráfica de usuario. Weka soporta varias tareas estándar de minería de datos, especialmente, pre-procesamiento de datos, clustering, clasificación, regresión, visualización, y selección. Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un archivo plano (flat file) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (normalmente

numéricos o nominales, aunque también se soportan otros tipos), también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos (Doricela Gutiérrez C. et al, 2017).