Sistema de Alertas Tempranas para identificar situaciones de riesgo académico, utilizando Datos de Caracterización y Aprendizaje de Máquinas: Caso de estudio Universidad Luis Amigó

Camilo García Saldarriaga

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor: Ingrid Durley Torres Pardo
Título de Posgrado

Universidad Católica Luis Amigó Facultad de Ingenierías y Arquitectura
Especialización en Big Data e Inteligencia de Negocios Medellín, Colombia
2024

Dedicatoria

Dedico este trabajo de grado a mi familia, quienes han sido el pilar fundamental de mi vida. Su apoyo incondicional, amor y sacrificio han acompañado cada paso de mi trayectoria académica. Hoy, con orgullo, les agradezco por haberme visto cumplir mis sueños y por inculcarme los valores que me han formado como un profesional apasionado, honesto y trabajador. Además, a la profesora Ingrid Durley Torres Pardo, gracias por acompañarme en cada paso de mi trayectoria académica y por ser una fuente constante de inspiración.

Agradecimientos

A la Universidad Católica Luis Amigó por concederme la beca de Joven Investigador. Su apoyo ha sido crucial para la realización de este trabajo, brindándome la oportunidad de crecer profesional y académicamente.

Al área de bienestar y permanencia académica de la Universidad Católica Luis Amigó por habernos permitido realizar este trabajo de grado, que tiene como finalidad entregar valor a la institución.

A los profesores de la Especialización en Big Data e Inteligencia de Negocios, quienes me

han acompañado en este recorrido brindándonos su apoyo y conocimiento.

Resumen

Actualmente, las instituciones educativas han dado creciente atención a los sistemas de alertas tempranas (SAT) debido a su importancia para identificar oportunamente a estudiantes en riesgo de fracaso académico o deserción. Este artículo subraya la necesidad de diseñar un SAT basado en técnicas de aprendizaje de máquinas, utilizando datos derivados de historiales académicos (como cancelaciones, repeticiones de cursos y créditos aprobados), caracterización sociodemográfica y factores financieros, entre otros, para detectar señales tempranas de alerta que predigan riesgos académicos, facilitando la implementación de intervenciones oportunas. El caso de estudio se desarrolla en la Universidad Católica Luis Amigó, donde la oficina de permanencia académica desempeña un papel crucial tanto en el suministro de datos como en la ejecución de acciones de mitigación. El trabajo resalta la relevancia de utilizar tecnologías avanzadas como el aprendizaje de máquinas para apoyar la toma de decisiones en el ámbito educativo y mejorar la permanencia estudiantil.

Palabras clave: Desempeño académico; Deserción estudiantil; Aprendizaje de Máquinas; Sistema de Alertas Tempranas SAT; Permanencia.

Tabla de Contenido

1. Introducción

En los últimos años, las instituciones educativas han abordado el tema de los Sistemas de Alertas Tempranas (SAT), centrándose en su importancia para la identificación oportuna de estudiantes en riesgo de fracaso académico o deserción. Se destaca en este contexto, la necesidad de desarrollar estrategias efectivas para apoyar a estos estudiantes y garantizar su éxito académico mitigando el riesgo de fracaso académico o la deserción. El objetivo de este trabajo se centra en el desarrollo de un SAT, basado en aprendizaje de máquinas para identificar situaciones de riesgo académico o deserción en estudiantes universitarios, usando datos provenientes de un historial académico de desempeño (cancelaciones, repeticiones de cursos, créditos aprobados), el reporte de la caracterización sociodemográfica, aspectos de índole financiero, entre otros. Reconociendo la importancia de identificar oportunamente a aquellos estudiantes que enfrentan dificultades académicas, para lograr la detección temprana de posibles señales de alerta, que puedan predecir el riesgo de fracaso académico o deserción. El caso de estudio específico del presente trabajo, se concentra en la Universidad Católica Luis Amigó, dónde se destaca la relevancia de las áreas de bienestar y permanencia académica, como fuente de insumo de datos para este estudio, así como la responsable de las acciones de mitigación y atención oportuna a estudiantes en riesgos, específicamente para este escenario el fracaso académico o deserción. En tal escenario, se espera que este trabajo contribuya a mejorar la atención educativa y promuevan el éxito estudiantil, para lo cual, su desarrollo, estará, soportado por la utilización de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para la implementación del SAT basado en aprendizaje de máquinas. Esta metodología proporciona un marco estructurado para abordar proyectos de esta índole, desde la comprensión del negocio y la comprensión de los datos hasta la evaluación y despliegue de los modelos.

2. Planteamiento del Problema

Actualmente, el Ministerio de Educación Nacional (MEN) de Colombia, define “La deserción escolar como un fenómeno complejo multifactorial y multidimensional, que implica el abandono del proceso educativo de niños, niñas, adolescentes y jóvenes, afectando su trayectoria educativa y el desarrollo integral, incidiendo en la formación del capital humano, el desarrollo sostenible, la movilidad social, la superación de la pobreza y la equidad.” Por ello, el gobierno, a través de las instituciones educativas, impulsa la permanencia en el sistema educativo y trabaja para reducir los factores que la amenazan. Esto se logra mediante iniciativas como el reconocimiento de las características y necesidades específicas de los estudiantes, la provisión de alimentación escolar y la creación de entornos de aprendizaje adecuados, entre otras acciones. Conscientes de apoyar la política gubernamental y en consecuencia la institucional educativa, este proyecto propone utilizar datos históricos de caracterización académica, como notas, cancelaciones, repeticiones de cursos y créditos, combinados con técnicas de Aprendizaje de Máquinas, para desarrollar un Sistema de Alertas Tempranas (SAT). Este sistema permitirá identificar patrones y factores de riesgo que podrían llevar al estudiante a enfrentar dificultades académicas. Al anticipar estas situaciones, las instituciones educativas podrán intervenir de manera proactiva para ofrecer apoyo personalizado y estratégico a los estudiantes en riesgo, mejorando así sus posibilidades de éxito académico y reduciendo las tasas de deserción.

3. Justificación

Las alertas tempranas, permiten identificar a tiempo estudiantes que enfrentan dificultades académicas, lo que facilita a la institución en concreto, la intervención oportuna por parte de los docentes y personal educativo. La implementación de un sistema de alertas tempranas (SAT) es crucial en el entorno educativo, ya que posibilita la detección precoz de problemas que pueden afectar el rendimiento académico de los estudiantes. Los datos de caracterización comprenden información relevante sobre los estudiantes, como su rendimiento académico, participación en clase, comportamiento y asistencia. Estos datos proporcionan una visión holística del alumno, permitiendo identificar posibles señales de alerta, como el bajo rendimiento en ciertas asignaturas o la ausencia reiterada en clase.

Además, el aprendizaje de máquinas juega un papel crucial en la eficacia del SAT al analizar grandes volúmenes de datos de manera automatizada y precisa. Mediante algoritmos avanzados, es posible identificar patrones y correlaciones que podrían pasar por alto al ojo humano. Estos algoritmos no solo pueden detectar estudiantes en riesgo, sino también predecir tendencias y ofrecer recomendaciones específicas para cada caso. Esto permite a las instituciones educativas implementar intervenciones personalizadas y basadas en evidencia, mejorando así las posibilidades de éxito académico de los estudiantes.

4. Marco Referencial

Los Sistemas de Alertas Tempranas (SAT) son herramientas fundamentales en el ámbito educativo, concebidos para anticipar y prevenir situaciones de riesgo académico entre los estudiantes. Estos sistemas, como definidos por (Gordon & Vaughan, 2006), tienen como propósito identificar a tiempo a aquellos estudiantes que podrían enfrentar dificultades en su rendimiento o incluso riesgo de abandono, permitiendo así la implementación de medidas preventivas y de apoyo (Longwell et al., 2012). Los SAT constan de tres componentes esenciales: la recolección de datos, donde se recopila información relevante sobre el desempeño académico, la asistencia y el contexto socioeconómico de los estudiantes (Arnold & Cho, 2015); el análisis de datos, que emplea técnicas estadísticas y algoritmos de aprendizaje automático para identificar patrones y tendencias que sugieren riesgos potenciales (Fan & Steckler, 2012); y la intervención, que implica la implementación de medidas específicas para apoyar a los estudiantes en riesgo, como la tutoría o la consejería (Pascarella & Terenzini, 2005). En este contexto, los Datos de Caracterización juegan un papel crucial. Estos datos, que describen las características individuales o grupales de los estudiantes, incluyen información demográfica, registros académicos y patrones de comportamiento (Snijders et al., 2012). Las fuentes de estos datos pueden variar, desde registros estudiantiles como expedientes académicos y registros de asistencia, hasta encuestas que indagan sobre las actitudes y experiencias de los estudiantes, pasando por datos observacionales que registran el comportamiento en entornos educativos. Utilizar estos datos en los SAT ofrece múltiples beneficios, como proporcionar una comprensión holística de las necesidades y riesgos individuales, facilitar el desarrollo de estrategias de intervención personalizadas y permitir la evaluación de la efectividad de dichas intervenciones.

El Aprendizaje Automático (ML) emerge como una herramienta poderosa en la mejora de los SAT. Este campo de la informática, según (Domingos, 2015), permite a los sistemas aprender de los datos sin necesidad de ser programados explícitamente. En el contexto de los SAT, el ML encuentra aplicaciones en la predicción del riesgo estudiantil, la identificación de factores de riesgo complejos y el desarrollo de intervenciones personalizadas. La utilización del ML en los SAT ofrece ventajas significativas, incluyendo la mejora en la precisión y eficiencia de la identificación de riesgos, la capacidad para identificar factores de riesgo complejos y la posibilidad de desarrollar intervenciones más efectivas y basadas en datos.

Según el Ministerio de Educación Nacional (2022), la deserción escolar es un problema complejo que impacta negativamente la trayectoria educativa, el desarrollo integral de niños, niñas, adolescentes y jóvenes, así como la formación del capital humano, el desarrollo sostenible y la equidad. Este fenómeno, influido por factores individuales, familiares, escolares y contextuales, exige una comprensión amplia para diseñar políticas educativas eficaces. En este sentido, el Gobierno de Colombia, a través del Plan Nacional de Desarrollo 2018-2022 "Pacto por Colombia, pacto por la equidad", ha buscado garantizar una educación de calidad y fomentar la permanencia escolar. Como parte de esta iniciativa, se desarrolló un estudio técnico en colaboración con la Universidad de los Andes, enfocado en analizar la deserción escolar en el país. Este análisis aborda mediciones del fenómeno, explora factores asociados y documenta las estrategias del Ministerio de Educación Nacional para contrarrestarlo. La investigación utiliza un enfoque mixto, combinando análisis cuantitativos y cualitativos basados en datos del Sistema Integrado de Matrícula (Simat) y consultas con actores educativos. Los hallazgos identifican factores protectores y de riesgo, subrayando la necesidad de implementar políticas específicas y estrategias efectivas para enfrentar este desafío.

5. Antecedentes

Se llevó a cabo una búsqueda de estudios en español e inglés, utilizando las bases de datos ScienceDirect, Web of Science, Scopus y Taylor & Francis. La búsqueda se centró en artículos de revistas indexadas y capítulos de libros que reportaban resultados de investigación. Los criterios de búsqueda incluyeron las palabras clave “Sistema de alertas tempranas”, “Universidad”, “Deserción”, “Riesgo académico”, “Aprendizaje de máquinas”, “Registro de notas” y “Datos de caracterización”.

La búsqueda se realizó siguiendo la metodología PRISMA (Preferred Reported Items for Systematic Review and Meta-Analyses). La Figura 1 muestra un resumen gráfico de la selección de documentos. Inicialmente, se obtuvieron 3469 documentos, de los cuales 59 cumplieron con los criterios de elegibilidad; sin embargo, tras la revisión del texto completo, se excluyeron aquellos que no desarrollaban un sistema de alertas tempranas. Finalmente, se incluyeron 21 documentos en el análisis de esta revisión.

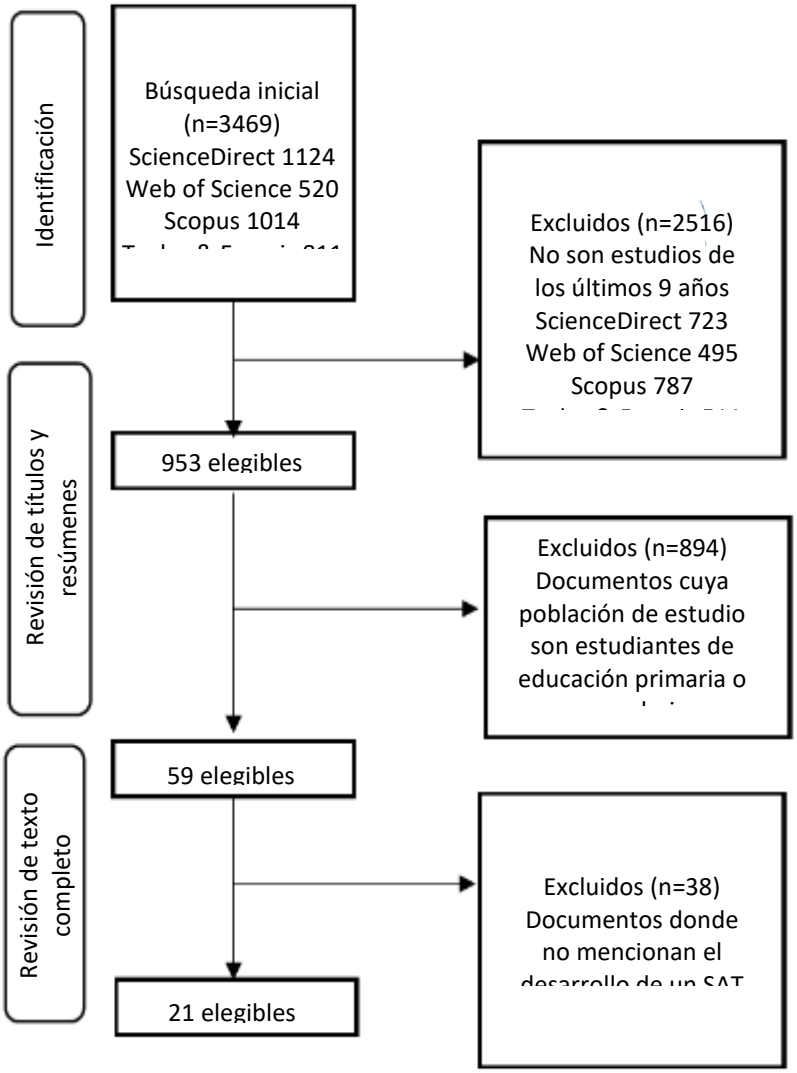

Figura 1. Metodología Prisma

Los Sistemas de Alertas Tempranas (SAT) han cobrado relevancia en el ámbito educativo como herramientas fundamentales para la identificación oportuna de estudiantes en riesgo de fracaso académico. Estos sistemas permiten analizar una amplia gama de datos, incluyendo el rendimiento académico, la asistencia, el comportamiento y el contexto socioeconómico, con el objetivo de detectar patrones que podrían indicar un potencial riesgo de abandono o bajo rendimiento.

En su tesis de maestría, Beltrán Assia (2016) describe el desarrollo e implementación de

un SAT para el programa de Ingeniería de Sistemas en la Universidad de Cartagena. El sistema utiliza tecnología web 2.0 para recopilar y analizar datos de rendimiento académico, asistencia y participación en actividades extracurriculares. Los resultados del estudio muestran que el SAT es una herramienta efectiva para identificar estudiantes en riesgo de bajo rendimiento y deserción (Beltrán Assia, 2016).

Los autores presentan una revisión de la literatura sobre los SAT en la educación superior. El estudio identifica diferentes tipos de SAT, así como los beneficios y desafíos de su implementación. Los autores concluyen que los SAT pueden ser una herramienta valiosa para mejorar la retención y el éxito de los estudiantes en la educación superior (López-Alcalde et al., 2017).

Este estudio analiza la efectividad de un SAT en una escuela de farmacia. Los resultados del estudio muestran que el SAT es una herramienta efectiva para identificar estudiantes en riesgo de bajo rendimiento académico. El estudio también destaca la importancia de la intervención temprana para prevenir el fracaso académico (Stratton et al., 2021).

El Instituto Departamental de Bellas Artes presenta una guía para la implementación de un SAT en instituciones educativas. La guía describe los pasos para desarrollar e implementar un SAT, así como las herramientas y recursos disponibles (Instituto Departamental de Bellas Artes, 2024).

El diseño e implementación del sistema de alerta temprana (SAT) Centinela en la Universidad Católica de la Santísima Concepción marca un avance significativo en la gestión académica universitaria. Este sistema, creado como una herramienta de seguimiento y análisis del proceso docente, se distancia de las concepciones tradicionales de los SAT al proponer un enfoque más dinámico y relacional del aprendizaje. Surgido en un contexto de crecimiento y masificación de la educación superior en Chile, donde la

diversidad socioeconómica de los estudiantes es cada vez más marcada, el SAT Centinela desarrollado en el trabajo busca enfrentar el desafío de la deserción estudiantil. La preocupación por el abandono estudiantil, evidenciada por numerosos estudios y revisiones sistemáticas, ha impulsado la implementación de intervenciones como los SAT, que buscan detectar y prevenir el riesgo de abandono a través del manejo inteligente de la información estudiantil. En este sentido, el SAT Centinela se erige como una herramienta valiosa, aprovechando tecnologías de gestión de la información y visualización de datos para identificar oportunamente situaciones de riesgo académico. Sin embargo, su diseño y funcionamiento se sustentan en una perspectiva que va más allá de la mera identificación de estudiantes en riesgo, promoviendo una mejora institucional y un enfoque proactivo en la atención de las necesidades académicas y socioemocionales de los estudiantes. En este proceso, se destaca la importancia de una reflexión teórica y metodológica previa, así como una visión crítica de los datos y su interpretación. (Casanova Daniel, 2021).

En el contexto de la preocupación por la deserción estudiantil en la educación superior chilena, se ha llevado a cabo una investigación orientada a comprender los factores asociados a este fenómeno y desarrollar estrategias para prevenirlo. Este estudio se enfoca en identificar variables predictoras de deficiencias en habilidades de lenguaje y matemáticas entre los estudiantes de primer año de la UMCE, con el objetivo de establecer un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción. Utilizando regresión logística y análisis de curva ROC, los autores identificaron ciertas variables, como los puntajes de la prueba de selección universitaria, el promedio de notas de enseñanza media, género y edad de ingreso, que permiten pronosticar estas deficiencias. (Henríquez Natalia, 2016)

Otro ejemplo es el Sistema de Alertas Tempranas (SAT) implementado por la Universidad Santo Tomás (USTA) que busca abordar el problema de la deserción estudiantil, que afecta

al 53,49% de los estudiantes en Colombia. Coordinado por la Unidad de Desarrollo Integral Estudiantil (UDIES) y liderado por profesionales como Helvy Sierra y Oscar Hernández, el SAT integra sistemas de información SAC y Moodle de la USTA. Este sistema detecta factores de riesgo académico, socioeconómico, institucional y personal, permitiendo intervenciones tempranas para fortalecer la permanencia estudiantil. Financiado a través del convenio 626 con el Ministerio de Educación Nacional, el SAT se adapta a las necesidades cambiantes de la universidad y se ha integrado con otros sistemas para mejorar la gestión de la información y brindar una respuesta más efectiva a las necesidades de los estudiantes en situación de vulnerabilidad. (H Sierra, 2014)

Adicionalmente, se tiene un modelo predictivo realizado en una universidad pública chilena, donde las reformas curriculares desde 2009, orientadas por el Ministerio de Educación, han generado desafíos para los docentes, especialmente en cuanto a las habilidades básicas de los estudiantes. La UMCE ha identificado déficits en lenguaje y matemáticas en sus estudiantes de primer año entre 2013 y 2015, señalando la necesidad de ajustar las estrategias de enseñanza. A nivel regional, América Latina enfrenta altas tasas de deserción universitaria, incluyendo a Chile, donde al menos el 50% de los estudiantes abandonan sus estudios antes de graduarse. Estudios han identificado factores como rendimiento académico previo, nivel socioeconómico y puntajes en las pruebas de selección como predictores de la deserción. En la UMCE, se ha implementado un sistema de alerta temprana basado en modelos logísticos, adaptados a cada facultad, para predecir el rendimiento académico y mitigar la deserción (N Henríquez, 2022).

Se evidencia el trabajo realizado en la Universidad Nacional de Catamarca sobre la deserción estudiantil en ingeniería ha sido ampliamente estudiada a nivel global, en donde se reconoce esta problemática, especialmente en los primeros años de estudio. Los análisis descriptivos tradicionales a menudo no logran identificar a tiempo las situaciones

de riesgo. La minería de datos educativos (EDM) ha emergido como una herramienta clave para abordar este desafío. y donde se han desarrollado métodos para extraer información valiosa de los datos educativos, permitiendo una mejor toma de decisiones en políticas educativas. (HC Ahumada, 2015)

Por último, Juan Sebastián Parra Sánchez, docente de la Universidad Católica Luis Amigó, en su estudio "Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial", identifica los principales factores que contribuyen a la deserción universitaria y cómo son abordados desde la inteligencia artificial (IA). A través de una revisión de 31 documentos seleccionados de un total de 2745 reportados, el docente agrupa los factores explicativos en categorías académicas, motivacionales, institucionales, y socioeconómicas. El estudio revela que el 92% de las investigaciones se enfocan en predecir la deserción utilizando modelos supervisados, siendo los árboles de decisión el método más común (84%), seguido por métodos probabilísticos (36%) y la regresión logística (28%). Sin embargo, solo el 6% de los estudios evaluó las intervenciones sobre la deserción. Además, se destaca la necesidad de una definición clara del concepto de deserción y de un análisis más amplio del rendimiento académico, no solo basado en notas promedio. El estudio concluye que, aunque la aplicación de IA en este campo está en auge, aún falta evaluar intervenciones concretas para mitigar la deserción y considerar aspectos motivacionales (Parra Sánchez, 2024).

Los estudios revisados demuestran que los SAT son una herramienta efectiva para identificar estudiantes en riesgo de fracaso académico. Los sistemas pueden analizar una amplia gama de datos para detectar patrones que podrían indicar un potencial riesgo de abandono o bajo rendimiento. La intervención temprana puede ser crucial para prevenir el fracaso académico y mejorar el éxito de los estudiantes.

6. Objetivos

6.1 Objetivo General

Desarrollar un Sistema de Alertas Tempranas (SAT) basado en aprendizaje de máquinas para identificar situaciones de riesgo académico en estudiantes universitarios, integrando datos de caracterización como: notas, cancelaciones, repeticiones de cursos, créditos, entre otros

6.2 Objetivos Específicos

- Recopilar y procesar datos de caracterización académica de los estudiantes, incluyendo notas, historial de cancelaciones, repeticiones de cursos y créditos.
- Implementar un modelo usando algoritmos de aprendizaje de máquinas para analizar y clasificar patrones en los datos académicos recopilados.
- Construir el modelo del SAT que identifique de manera temprana posibles situaciones de riesgo académico como: bajo rendimiento, deserción o reprobación, basándose en los resultados del análisis de datos.
- Evaluar la efectividad del Sistema de Alertas Tempranas mediante prueba piloto y el análisis de la capacidad predictiva y precisión en la identificación de estudiantes en riesgo

7. Viabilidad

A continuación, se presentan algunos conceptos clave para facilitar la comprensión del desarrollo de los objetivos de este trabajo de grado. En primer lugar, se encuentra **CRISP-DM** (Cross-Industry Standard Process for Data Mining), una metodología ampliamente utilizada en la minería de datos. Esta propone un enfoque estructurado y gradual, basado en un ciclo de vida, para llevar a cabo proyectos de análisis de datos. Esta metodología permite optimizar la planificación de los objetivos planteados, ya que se ejecuta en fases definidas, cada una de las cuales detalla las actividades a realizar y la manera en que se desarrollarán para alcanzar los objetivos establecidos.

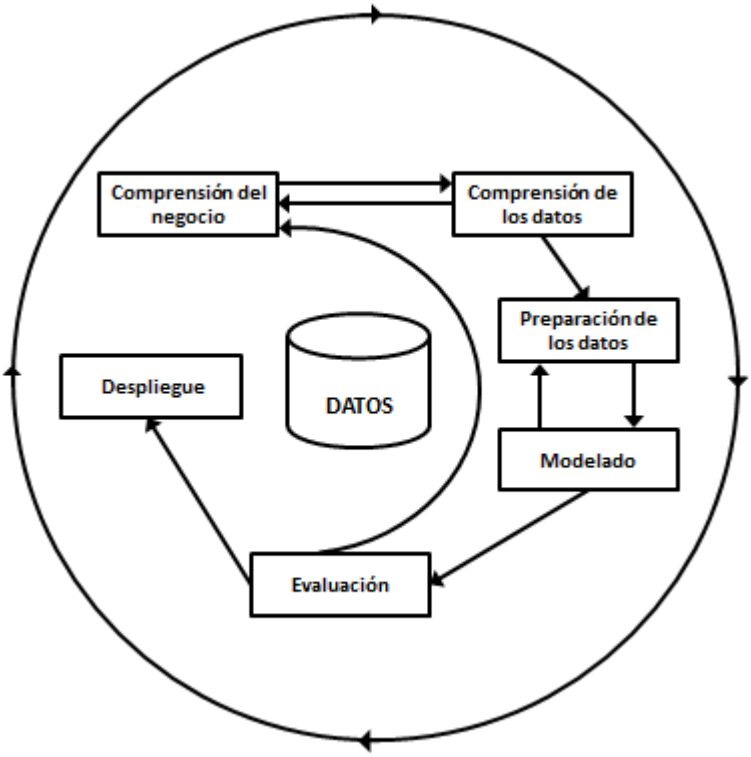

Figura 2. Metodología CRISP-DM

Big Data, según Vargas y Peñaloza (2019), se refiere a una evolución digital en la forma de procesar y gestionar grandes volúmenes de datos, transformándolos en conocimiento. Este concepto se caracteriza por la “automatización y gestión del conocimiento en todas

las dimensiones de la sociedad actual", con el propósito de que dicho conocimiento pueda ser "potenciado, difundido e intercambiado", contribuyendo al bienestar económico y social. En el contexto de la especialización, la minería de datos resulta ideal para tareas de clasificación y predicción.

Python, un lenguaje de programación creado en 1989, se utiliza ampliamente en áreas como la administración de sistemas, ciencia de datos, computación científica, inteligencia artificial e Internet de las cosas (García, 2017, p. 151). Las siguientes librerías de Python se aplicaron en este proyecto:

NumPy: Librería para cálculos numéricos que permite trabajar de manera eficiente con matrices y vectores (Numpy.org, s.f.).

Pandas: Herramienta para ciencia de datos y aprendizaje automático, capaz de manejar tareas como lectura, filtrado, limpieza, normalización, análisis estadístico y visualización de datos (pandas.pydata.org, s.f.).

Matplotlib: Librería enfocada en la creación de gráficos bidimensionales como diagramas de barras e histogramas (Matplotlib.org, s.f.).

Seaborn: Librería especializada en gráficos estadísticos atractivos e informativos (seaborn.pydata.org, s.f.).

8. Metodología

El desarrollo del proyecto se basó en la metodología **CRISP-DM** (Cross Industry Standard Process for Data Mining). Este método permitió organizar el proceso en etapas claras, asegurando la calidad en cada paso y la alineación con los objetivos del proyecto.

8.1 Comprensión del negocio

En la fase inicial, se realizó un análisis exhaustivo de los objetivos institucionales y de las necesidades específicas de bienestar institucional y permanencia académica en relación con la identificación temprana de riesgos académicos en los estudiantes pertenecientes a pregrado de la Universidad Católica Luis Amigó. Este análisis permitió comprender el contexto educativo y definir los requisitos clave para el diseño e implementación del Sistema de Alertas Tempranas (SAT), asegurando que estuviera alineado con las expectativas institucionales.

8.2 Comprensión de los datos

En esta etapa, se identificaron y recopilaron datos históricos a partir del primer semestre del año 2022. Esta decisión se tomó para evitar incluir datos anómalos derivados de la pandemia de COVID-19 vivida entre los años 2020 y 2021. La información recolectada abarca aspectos del rendimiento académico de los estudiantes, como notas, cancelaciones y repeticiones de cursos, además de datos sociodemográficos provenientes de encuestas de caracterización, reportes de ausencia intersemestral y seguimientos realizados por los docentes durante las clases. También se incluyeron datos relevantes de las áreas de servicios médicos, enfermería y psicología de la institución, lo que permitió obtener una perspectiva integral del estudiante y sus circunstancias tanto dentro como fuera de la universidad.

Posteriormente, se evaluó la calidad de estos datos, detectando problemas de integridad, como valores faltantes e inconsistencias. Este análisis permitió establecer una base confiable para las fases posteriores del proyecto.

8.3 Preparación de los datos

La preparación de los datos implicó una serie de tareas orientadas a garantizar su calidad y adecuación para el análisis. Gracias a la cooperación de la Universidad Católica Luis Amigó, en particular de las áreas de Bienestar Universitario y Permanencia Académica, se obtuvieron ocho bases de datos distintas provenientes de diversas fuentes:

1. Encuestas de caracterización.
2. Ausencia intersemestral.
3. Descuentos.
4. Reportes médicos y de enfermería.
5. Reportes psicológicos.
6. Reporte y seguimiento docente.
7. SNIES.
8. Registro académico.

Cada una de estas bases fue sometida a actividades de limpieza específicas, como la eliminación de duplicados y la imputación de valores faltantes. Además, las variables fueron normalizadas para garantizar consistencia en los formatos y escalas, y se derivaron nuevas variables que ofrecieron una perspectiva más profunda del riesgo académico.

Finalmente, se consolidó una base de datos unificada por estudiante, que integra información tanto actualizada como datos históricos, proporcionando una visión completa para el análisis. Todas estas tareas se realizaron utilizando Python, con el apoyo de librerías especializadas como Pandas y NumPy, las cuales facilitaron el procesamiento

eficiente de grandes volúmenes de datos.

8.4 Modelado y evaluación

Con los datos preparados, se emplearon algoritmos de aprendizaje automático, específicamente el modelo de agrupación K-means, para identificar patrones que indican riesgo académico. Este enfoque se basó en el modelado no supervisado, ya que los datos iniciales no estaban etiquetados con categorías predefinidas de riesgo. La técnica permitió agrupar a los estudiantes en clusters basados en similitudes detectadas en variables como rendimiento académico, asistencia y características sociodemográficas.

El uso de K-means fue clave para identificar grupos de estudiantes que compartían características comunes, destacando aquellos con un perfil asociado a un mayor riesgo de deserción o bajo desempeño. Este tipo de análisis no supervisado es particularmente útil en contextos donde no existe una clasificación previa y se requiere explorar los datos para descubrir patrones ocultos.

Para determinar el número óptimo de grupos en el análisis de agrupamiento, se aplicaron dos métodos ampliamente conocidos: el método del codo y el método de la silueta. El método del codo se utilizó para evaluar la variación explicada por los diferentes números de clusters, permitiendo identificar un punto de inflexión donde agregar más clusters no aporta beneficios significativos. Por otro lado, el método de la silueta permitió medir la cohesión y separación de los grupos formados, ayudando a garantizar que los clusters fueran internamente homogéneos y externamente distintos.

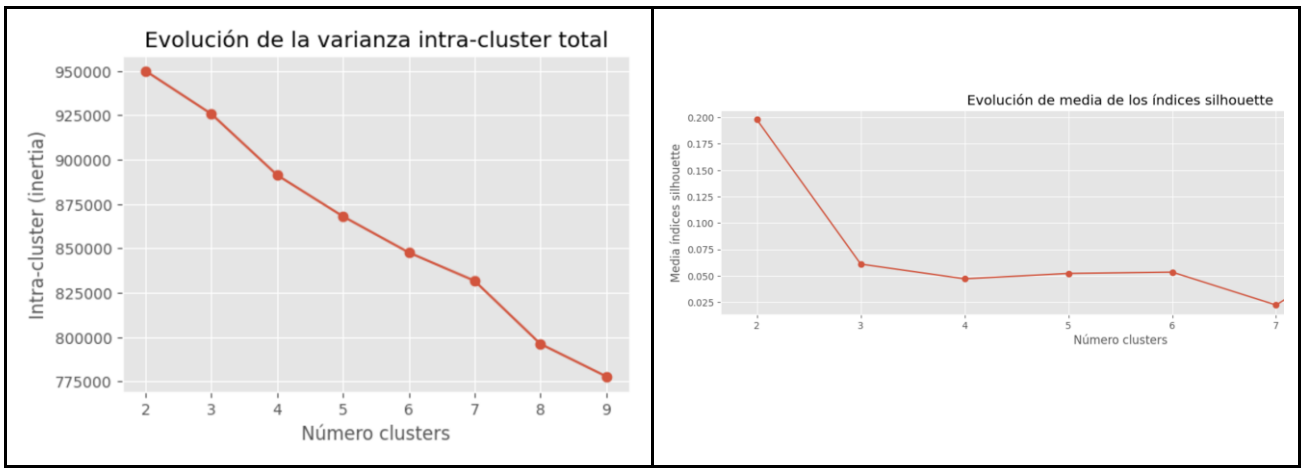

Figura 3. Resultados métodos del codo y la silueta

Aunque los resultados técnicos sugirieron varias configuraciones posibles para el número de clusters, se optó por consolidar tres grupos, en respuesta a las necesidades y recomendaciones expresadas por los expertos del área educativa y de bienestar universitario. Esta decisión no solo respalda la interpretabilidad del modelo, sino que también facilita la implementación de estrategias de intervención específicas. Los tres grupos se definen de la siguiente forma:

- Grupo Verde: Estudiantes que no requieren intervención.
- Grupo Amarillo: Estudiantes que requieren una intervención preventiva.
- Grupo Rojo: Estudiantes que requieren una intervención inmediata.

Los tres grupos definidos reflejan patrones consistentes en el comportamiento y las características de los estudiantes, teniendo en cuenta factores como el rendimiento académico, asistencia, historial de cancelaciones, y variables sociodemográficas. Estos grupos serán utilizados para desarrollar estrategias diferenciadas de intervención que permitan atender de manera efectiva las diversas situaciones de riesgo detectadas.

Una vez ejecutado el algoritmo de modelado no supervisado, que clasificó a todos los estudiantes en los tres grupos definidos previamente, se procedió a realizar un análisis

complementario mediante el algoritmo de Análisis de Componentes Principales (PCA, por sus siglas en inglés). Este método permitió reducir la dimensionalidad de los datos originales, conservando la mayor cantidad de variabilidad posible y facilitando su representación en un espacio tridimensional.

La visualización en 3D obtenida a través de PCA ofreció una representación clara y comprensible de la distribución de los estudiantes dentro de los tres grupos. Esto permitió observar patrones clave, como la distancia y la cohesión entre los clusters. Estos resultados proporcionan una base sólida para interpretar y validar la agrupación realizada.

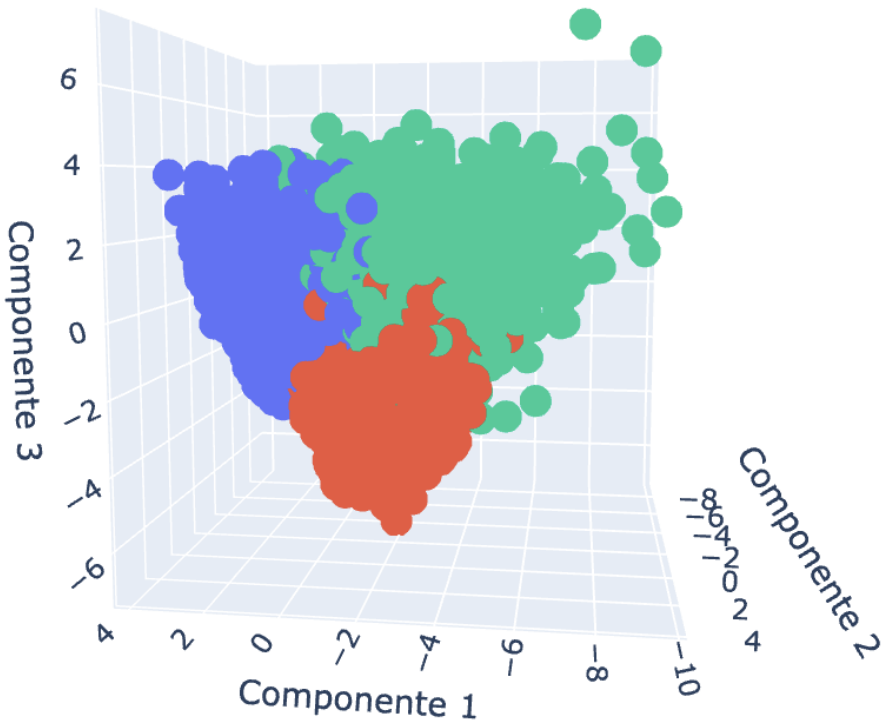

Figura 4. Resultados Modelamiento No Supervisado

Con los resultados del modelado, se procedió a un análisis detallado para identificar y caracterizar los grupos obtenidos. Este proceso se llevó a cabo en tres etapas principales:

- Reunión con expertos: Se presentó la visualización de los resultados a expertos

institucionales en bienestar y permanencia académica. Al contrastar estos resultados con la información previa que manejaban, se logró obtener una interpretación preliminar basada en su experiencia y conocimiento del contexto estudiantil.

- Cálculo de un score de riesgo acumulativo: Se asignaron puntajes a variables críticas relacionadas con el rendimiento y las condiciones académicas, como notas bajas, cancelaciones o repetición de cursos. Este análisis cuantitativo permitió generar una nueva interpretación preliminar que resaltaba a los estudiantes con mayor riesgo.
- Uso de una IA generativa: Los resultados del modelo fueron enviados a una inteligencia artificial generativa, que analizó y sugirió interpretaciones adicionales sobre los patrones y características de los grupos.

Con estas tres interpretaciones preliminares, se consolidó un análisis final que permitió identificar los grupos y asociar las características buscadas, asegurando que reflejaran tanto los objetivos del proyecto como las necesidades de la institución.

9. Resultados

Como resultado del proyecto, se consolidó un data warehouse que integra la información limpia y transformada necesaria para la implementación del Sistema de Alertas Tempranas (SAT). Este almacén de datos centralizó variables clave, como notas, cancelaciones y repeticiones de cursos, así como datos sociodemográficos provenientes de las encuestas de caracterización, reportes de ausencia intersemestral y seguimiento académico. Esta integración garantiza una base sólida, actualizada y confiable para el análisis predictivo.

Además, a partir del modelamiento no supervisado, se lograron resultados significativos mediante el uso del algoritmo K-Means. Este permitió clasificar a los estudiantes en tres grupos principales, definidos en función de patrones ocultos en los datos. Para determinar el número óptimo de grupos, se aplicaron los métodos del codo y la silueta, asegurando una segmentación adecuada de la población estudiantil.

Tras la ejecución del modelo, los resultados iniciales fueron analizados mediante un proceso estructurado que incluyó validación por los expertos, score de riesgo acumulativo y análisis con IA generativa, la combinación de estas tres aproximaciones permitió afinar la caracterización de los grupos. Posteriormente, se aplicó un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y visualizar los resultados en un espacio tridimensional, lo que facilitó la comprensión y comunicación de los patrones encontrados.

Estos resultados no solo proporcionan una herramienta predictiva confiable, sino que también permiten a la institución diseñar estrategias de intervención más personalizadas y efectivas para cada grupo, contribuyendo así a la mejora de la permanencia estudiantil y la reducción de la deserción académica.

Como resultado adicional del proyecto, se desarrolló una plataforma web denominada Apolo-SAT, diseñada para facilitar el uso y la implementación del Sistema de Alertas Tempranas (SAT). Esta herramienta permite a las áreas de Bienestar Académico y Permanencia Académica cargar los datos de los estudiantes y obtener, de manera automática, el nivel de alerta asignado por el modelo. La plataforma también incluye un dashboard interactivo que permite a los usuarios visualizar, a través de gráficos, los registros y análisis correspondientes al semestre actual. Este panel proporciona estadísticas clave, como la distribución de los estudiantes en los diferentes niveles de alerta (verde, amarillo y rojo), y permite realizar seguimientos más detallados de los grupos en riesgo.

Gracias a Apolo-SAT, la institución puede realizar análisis de riesgo de manera semestral, optimizando la identificación de patrones y el diseño de estrategias de intervención temprana. Esto representa un avance significativo en la gestión y monitoreo del bienestar académico, ofreciendo una solución integral que combina análisis predictivo con una interfaz accesible y orientada al usuario.

Figura 5. Plataforma Web Apolo-SAT

10. Recomendaciones

Se recomienda normalizar las ocho bases de datos existentes para garantizar que toda la información utilice un formato uniforme y consistente. En su estado actual, las bases presentan discrepancias en su estructura y formato según los distintos períodos, lo que dificulta su integración y análisis. Este proceso permitirá que la información sea más accesible y funcional para las necesidades del proyecto.

Posteriormente, se sugiere unificar todas las bases de datos en una única estructura consolidada. Esta integración facilitará el acceso y manejo eficiente de los datos, permitiendo que la información sirva como un insumo centralizado no solo para este proyecto, sino también para futuros desarrollos. Una base de datos consolidada proporcionará una plataforma sólida para análisis integrales y estrategias basadas en datos.

11. Referencias

Stratton, T. P., Janetski, B. K., Ray, M. E., Higginbotham, M. C., Lebovitz, L., & Martin, B. A. (2021). Detection of Academic Early Warning Signs and Effective Intervention “Takes a Village”. American Journal of Pharmaceutical Education 2022; 86 (7) Article 8743. https://doi.org/10.5688/ajpe8743

López-Alcalde, M., García-García, M.J., & García-Peñalosa, C. (2017). Sistemas de Alerta Temprana para estudiantes en riesgo de abandono de la Educación Superior. https://www.scielo.br/j/ensaio/a/zxq7jgfN9gQSyVYYxgPc9HM/

Instituto Departamental de Bellas Artes. (2024). Sistema de Alertas Tempranas: Bienestar Institucional. https://bellasartes.edu.co/images/informacionciudadano/sistema_alertas_tempranas.pdf

Beltrán Assia, J. D. (2016). Sistema de alertas tempranas para la identificación de bajo rendimiento, pérdida de la calidad y seguimiento estudiantil, a través de tecnología web 2.0, en el programa de ingeniería de sistemas presencial de la Universidad de Cartagena. (Tesis de maestría, Universidad de Cartagena). https://repositorio.unicartagena.edu.co/

Arnold, K. M., & Cho, S. Y. (2015). Using early warning systems to identify and support struggling students. https://www.edweek.org/leadership/most-schools-have-early-warning-systems-how-well-do-they-work/2024/02

Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

Fan, W., & Steckler, M. (2012). Prediction and intervention for early warning systems in education. https://www.apa.org/pubs/journals/edu

Gordon, D., & Vaughan, J. S. (2006). Early warning systems: A framework for decision-

making. Emergency Management Journal, 31(1), 21-40.

Longwell, J. A., Reynolds, J. R., Durlak, J. A., Clark, L. H., & Meglina, J. M. (2012). A meta-analysis of early warning systems for student success. Journal of Educational Psychology, 104(4), 795.

Pascarella, E. T., & Terenzini, P. T. (2005). How college affects students: A third decade of research. Jossey-Bass.

Snijders, T. A., van de Bunt, G., & Bosker, R. (2012). Network analysis in social research. Cambridge University Press.

Ministerio de Educación Nacional. (2022). Deserción escolar en Colombia: Análisis, determinantes y política de acogida, bienestar y permanencia: Nota técnica. Bogotá D.C.

Casanova Cruz, D., Miranda Díaz, C., & Yáñez Corvalán, A. M. (2021). Sistema de alerta temprana: Centinela, una experiencia para la retención estudiantil en la Universidad Católica de la Santísima Concepción. Calidad en la Educación, (55), 156-174. https://www.scielo.cl/scielo.php?pid=S0718-45652021000200156&script=sci_arttext

Henríquez, N., & Escobar, D. (2016). Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la Universidad Metropolitana de Ciencias de la Educación. Revista mexicana de investigación educativa, 21(71), 1221-1248. https://www.scielo.org.mx/scielo.php?pid=S1405-66662016000401221&script=sci_arttext

Sierra, H., & Hernández, O. (2014). Sistema de alertas tempranas como herramienta de innovación tecnológica en la Universidad Santo Tomás para el fortalecimiento de la permanencia estudiantil y graduación oportuna. In Congresos CLABES. https://revistas.utp.ac.pa/index.php/clabes/article/view/1056

Henríquez Cabezas, N., & Vargas Escobar, D. (2022). Modelos predictivos de

rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena. Revista de estudios y experiencias en educación, 21(45), 299-316. https://www.scielo.cl/scielo.php?pid=S0718-51622022000100299&script=sci_arttext&tlng=pt

Ahumada, H. C., Dip, H., Herrera, C. G., & Leguizamón Almendra, J. C. (2015, May). Minería de datos para un sistema de alerta temprana de deserción en carreras de ingeniería. In XVII Workshop de Investigadores en Ciencias de la Computación (Salta, 2015). https://sedici.unlp.edu.ar/handle/10915/45515

Matplotlib.org (s.f). Matplotlib: visualización con Python[en línea] (Consulta realizada el 20 de octubre de 2022). https://matplotlib.org/

Numpy. Org (s.f). NumPy documentation. [en línea] (Consulta realizada el 20 de octubre de 2022). https://numpy.org/doc/stable/

pandas.pydata.org (s.f). 10 minutos para pandas [en línea] (Consulta realizada el 20 de octubre de 2022). https://pandas.pydata.org/docs/user_guide/10min.html

Parra Sánchez, J. S., Torres Pardo, I. D., & Martínez de Merino, C. Y. (2023). Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial. Revista electrónica de investigación educativa, 25. https://doi.org/10.24320/redie.2023.25.e18.4455