

ARTÍCULO PRODUCTO DE INVESTIGACIÓN

**Diseño de un modelo de estimación de riesgos de deserción
en programas de educación superior**

Design of a model for estimating dropout risks in higher education programs

Oscar Javier Rubiano Rodríguez¹
Efrén Danilo Ariza Ruiz²
Fundación Universitaria del Área Andina

Recibido: 02.02.2026
Aceptado: 30.03.2026

Resumen

El artículo se enfoca en diseñar un modelo de estimación de riesgos de deserción en programas de educación superior en Colombia. Este objetivo se desarrolla a través de la ejecución del modelo de proceso estándar para minería de datos CRISP-DM, en la que se analizan bases de datos numéricas y herramientas con resultados estadísticos para hacer una estimación del impacto de algunas variables relacionadas con la deserción. El alcance de la investigación es correlacional con un enfoque cualitativo en el que se exploran las bases de datos que contienen las características sociodemográficas relevantes que se tendrán en cuenta en la construcción del modelo. La población está conformada por estudiantes de la Fundación Universitaria del Área Andina - Sede Valledupar; el tipo de muestreo es estratificado, usando como base a estudiantes de la sede Valledupar; el tamaño de la muestra es de 3469 estudiantes. El proceso seleccionado es el análisis de datos de matrículas realizado sobre muestras censales y, por otra

¹orodriguez79@estudiantes.areandina.edu.co
<https://orcid.org/0009-0004-8101-0899>

²eariza10@areandina.edu.co
<https://orcid.org/my-orcid?orcid=0000-0001-7762-3562>

parte, resultados de encuestas e investigaciones que indagan sobre características de la población por sectores; con diseño de muestreo probabilístico; el margen de error es de 4% y el nivel de confianza de 92%. El resultado será la construcción de un prototipo de machine learning para identificar los riesgos de deserción en un programa de educación superior.

Palabras clave: CRISP-DM, deserción académica, educación superior, innovación educativa, minería de datos, modelo predictivo

Abstract

The article focuses on designing a model for estimating dropout risks in higher education programs in Colombia. This objective is developed through the execution of the standard process model for CRISP-DM data mining, in which numerical databases and tools with statistical results are analyzed to estimate the impact of some variables related to dropout. The scope of the research is correlational with a qualitative approach in which the databases containing the relevant sociodemographic characteristics that will be considered in the construction of the model are explored. The population is made up of students from the Fundación Universitaria del Área Andina - Valledupar Campus, the type of sampling is stratified using students from the Valledupar campus as the base; and the sample size is 3469 students. The selected process is the data analytics of enrollments made on census-type samples and, on the other hand, results of surveys and research that inquire about characteristics of the population by sectors, with probabilistic sampling design; the margin of error is 4% and the confidence level is 92%. The result will be the construction of a machine learning prototype to identify dropout risks in a higher education program.

Keywords: CRISP-DM, big data, educational dropout, educational innovation, higher education, machine learning

Introducción

La deserción estudiantil es una problemática que se extiende a los estudiantes, instituciones educativas y al desarrollo de los países en general. Según los indicadores, la tasa de cobertura en educación superior de Colombia presenta promedios más bajos que la media de la región. Para la OCDE (2021), la tasa para el año 2020 fue del 51,6%, lo cual representa un 23,5% por debajo del promedio de otros países pertenecientes a la OCDE. Ante este panorama, se debe tener en cuenta que los altos niveles de deserción pueden tener un impacto negativo en la economía y el desarrollo de la región, esto debido a que los estudiantes que abandonan sus estudios pierden oportunidades laborales, además del impacto económico al no producirse investigaciones o innovaciones desde la academia. En síntesis, la deserción puede ser un indicador de problemas más profundos en el sistema educativo, como la ausencia de estrategias efectivas o la implementación inadecuada de las mismas.

Es así como todo aporte hecho para contrarrestar el efecto de la deserción es valioso, más aún desde la interpretación de variables de riesgo, para posteriormente plantear soluciones pertinentes; para el presente artículo se plantea como pregunta base: ¿Cuáles son los elementos constitutivos de un modelo o metodología que permita estimar riesgos de deserción en programas de educación superior? Con el fin de diseñar un modelo de estimación de riesgos de deserción en programas de educación superior, haciendo uso de las tecnologías del siglo XXI y de las capacidades predictivas que ofrecen el big data y el análisis de grandes bases de datos.

Las contribuciones del presente artículo respecto a estudios previos en el tema son, en primera medida, una propuesta detallada y novedosa y con capacidad de brindar información cualitativa en tiempos reducidos, lo que permite una evaluación más precisa y general del panorama; además, pone al alcance de diferentes instituciones educativas el uso de información estadística acorde a su entorno, con un alto nivel de confianza y con una utilidad basada en análisis predictivo. También contribuye a la implementación y uso de nuevas tecnologías en la investigación social, y es que, como se verá más adelante, los estudios

anteriores se han enfocado en su mayoría en el análisis retrospectivo desde técnicas cualitativas como los estudios de caso y otros métodos tradicionales, por lo que para la realización del modelo se dio un uso preferente a las herramientas tecnológicas como el software de análisis estadístico y software para minería de datos.

Para el análisis se llevó a cabo una búsqueda y selección de base de datos para luego plantear un modelo que tenga en cuenta las características de la institución y su entorno y que mida el impacto de cada variable; también, que permita cruzar información con datos de características sociodemográficas disponibles en bases de datos abiertas que permitan complementar la identificación de variables influyentes. El modelo propuesto busca estimar factores de deserción y brindar a las instituciones educativas herramientas para complementar su proceso de evaluación de impacto.

Castaño et al. (2009) destacan que, para identificar adecuadamente las causas de la deserción estudiantil, es crucial analizar unidades más específicas como programas académicos, modalidades y sedes, en lugar de analizar a las instituciones de educación superior como un todo. Esto se debe a que cada institución tiene características particulares según su región, economía y cultura. Por lo tanto, es importante desagregar la información y examinar factores como las trayectorias académicas individuales de los estudiantes, así como características como el acceso a servicios, sexo, edad, composición del hogar y acceso a internet.

Según Cayón et al. (2021), el impacto de las variables en la deserción universitaria cambia con el tiempo. Variables como género, edad y la institución de educación secundaria son significativas. Estudiantes de colegios con buen promedio tienen menor tasa de deserción. González Catalán y Arismendi Vera (2018) señalan que la jornada de clases y el año de finalización de la enseñanza media son factores importantes. Concluyen que la flexibilidad curricular y la ampliación de modalidades educativas mejoran la deserción. En carreras técnicas y tecnológicas, las mujeres tienen mayor propensión a desertar en comparación con las profesionales.

1. Estado del arte y revisión de literatura

Se realizó una revisión sistemática de literatura en bases de datos como Google Scholar, Mendeley, Scopus, con palabras clave: Minería de datos, CRISP-DM, deserción académica, educación superior, innovación educativa y modelo predictivo. Posteriormente, el investigador elaboró fichas biográficas y sistematizó la información en resúmenes analíticos. Los resultados de la revisión se dividen en las siguientes dimensiones: estudios sobre deserción en educación superior, análisis de bases de datos en educación, aplicaciones de modelos predictivos en educación superior y metodologías innovadoras en educación para prevenir problemáticas

Sobre deserción en educación superior

Varios estudios han explorado las causas y efectos de la deserción en programas de educación superior. Estos estudios proporcionan una base teórica para comprender la deserción y definir las variables que se utilizarán en la construcción de modelos predictivo.

La deserción es un fenómeno complejo donde interactúan diferentes factores y dimensiones. Según señalan (González Catalán & Arismendi Vera, 2018; Guzmán Rincón et al., 2022; Himmel, 2002), es relevante investigar sistemáticamente cómo se combinan las diferentes variables, para que las instituciones y organismos estatales puedan actuar sobre aquellos aspectos que son más controlables. Esto incluye medidas institucionales que brinden programas de apoyo financiero, establezcan centros de apoyo familiar y promuevan programas de motivación

Por su parte, (Garcés Prettel et al., 2024; García-Martínez et al., 2024; Gómez-Restrepo et al., 2016; González & Espinoza, 2015) identifican variables externas al sistema educativo (condiciones socioeconómicas del estudiante y su familia, nivel educativo bajo del jefe de hogar, vivir en áreas rurales, estado de pobreza y disfunción familiar), propias del sistema (incremento de la matrícula, la falta de opciones de financiamiento), académicas (formación

académica previa, la falta de apoyo y orientación) y personales (motivación, expectativas, dificultades para adaptarse, la falta de participación en actividades comunitarias, consumo de alcohol y sustancias psicoactivas).

Para (Cayón et al., 2021; González Catalán & Arismendi Vera, 2018; González & Espinoza, 2015; Guerrero, 2018) existen variables como el género y la edad, bajo rendimiento en exámenes como las pruebas colombianas SABER; la jornada a la que se asiste; el tipo de carrera (técnicas, tecnológicas o profesionales); el año y la institución de egreso de la enseñanza media; carreras de menor demanda laboral y la falta de vocación. (González & Espinoza, 2015). Señala que universidades con alta acreditación de calidad tienen un menor número de desertores. Además, las tasas más altas de deserción se dan en los primeros semestres de estudio.

Análisis de datos en educación

Recientemente, avances como el uso de cadenas de bloque “blockchain”, el análisis de datos con lenguajes de programación “big data” y los modelos predictivos están siendo explorados en el ámbito educativo, para generar una prospectiva que permita una mayor calidad del aprendizaje.

Para (Bartolomé Pina et al., 2017), las nuevas tecnologías plantean la posibilidad de aplicar aprendizajes en proyectos de formación que promuevan programas personalizados para cada estudiante. Los autores sugieren abordar estos avances superando viejos paradigmas didácticos.

En relación con el software y lenguajes de programación en el análisis de bases de datos, (Cuevas Acosta et al., 2019; Norma et al., 2016) destacan aspectos como la reproducibilidad computacional para desarrollar modelos predictivos de alertas tempranas utilizando Data Mining (DM), Educational Data Mining (EDM). Con el fin de localizar factores negativos que influyen y proponer la creación de estrategias para tomar acciones oportunas

Aplicaciones de modelos predictivos en educación superior

Los modelos predictivos de machine learning usados para prevenir la deserción en educación superior, según hallazgos, permiten altos porcentajes de confiabilidad en la elaboración de predicciones. Los autores consultados recomiendan tener objetivos estratégicos claros, utilizar datos de alta calidad y evaluar o validar regularmente los modelos. (Alban Taipe, 2019; Camargo, 2020) señalan factores que influyen en la deserción utilizando técnicas de minado de grandes bases de datos. Modelos como el árbol de decisión o el algoritmo Random Forest presentaron mejores precisiones de predicción. Además, combinaron técnicas como redes neuronales y simulaciones para mejorar la precisión de la predicción. Mencionan que entrenar el modelo con más datos podría aumentar la precisión, pero aumenta el tiempo de procesamiento. Por su parte, (Aguilera et al., 2017; Guzmán Rincón et al., 2022; Mella-Norambuena et al., 2022) concluyeron que los modelos predictivos deben implementarse en programas tradicionales teniendo en cuenta las particularidades regionales en Latinoamérica, prometiendo mejorar la calidad y la permanencia.

Metodologías innovadoras en educación para prevenir problemáticas

En su mayoría, los estudios consultados apoyan la hipótesis de que el uso de las tecnologías de la cuarta revolución industrial y otras metodologías innovadoras tiene un impacto positivo en el desarrollo de investigaciones sociales.

En la metodología de la minería de datos, Schonlau y Zou (2020) y Timarán Pereira et al. (2016) proponen fases como la clasificación de datos, segmentación, asociación y patrones. Estas tareas requieren la adopción de algoritmos como el KDD, Decision Tree o el Random Forest, que se han desarrollado en diversos campos para reconocimiento de patrones. (Rodríguez Ojeda (2024) y Villena, 2016) proponen la metodología CRISP-DM para organizar los estudios en fases, subfases y tareas para asegurar la calidad de los resultados.

En relación con instrumentos de investigación mediada por herramientas tecnológicas, (Gros Salvat & Cano García, 2021; Hugo et al., 2019) destacan que el machine learning está ganando importancia debido a su capacidad de facilitar estudios apoyados en tecnologías digitales que pueden contribuir a incentivar la participación de los estudiantes en modalidades virtuales o a distancia.

Según estudios de (Enríquez Vázquez & Navarro Perales, 2024; Vázquez Chicaiza et al., 2024), a medida que las tecnologías digitales se hacen más presentes en los entornos de aprendizaje, la importancia de la autoevaluación y la creación de nuevos esquemas ha aumentado. Estas tecnologías pueden llevar a nuevas prácticas relacionadas que impulsen el pensamiento crítico y la solución a problemas ambientales y del entorno.

2. Materiales y métodos

La investigación se enmarca en un enfoque cuantitativo, que parte del análisis de bases de datos que permitirán estimar el riesgo de deserción en estudiantes de educación superior. La investigación es de tipo correlacional que, a la voz de (Hernández Sampieri, 2014), tiene como objetivo principal describir las relaciones entre variables en un momento determinado.

El método adoptado para el análisis y para crear un modelo de predicción de deserción es el CRISP-DM (Cross-Industry Standard Process for Data Mining), que permitirá identificar los factores clave que contribuyen a la deserción estudiantil y desarrollar un modelo predictivo que pueda predecir con precisión la probabilidad de que un estudiante abandone sus estudios. Un proyecto de minería de datos, según (Villena, 2016), con metodología CRISP-DM, debe ser desarrollado en las siguientes fases:

Recopilación de datos: Se accede a las bases de datos oficiales colombianas de deserción estudiantil, en primer lugar, del Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES), que ofrece datos como sexo, población activa laboralmente al inicio del estudio, núcleo familiar, nivel socioeconómico, vivienda propia, calificación examen de

estado, apoyos económicos oficiales y nivel educativo de los padres. Posteriormente, se compara con base de datos proporcionada por la Fundación del Área Andina – Sede Valledupar, usando 3 cohortes y la totalidad de inscritos en los programas que ofrece.

Como resultado, se obtuvieron tres bases de datos en formato .svc con una totalidad de 19 atributos.

Tabla 1. Descripción de las características de las variables con característica principal, tipo o clasificación según su origen, ejemplo: categórico si son variables basadas en descripciones y numéricas, y comentario.

Número	Característica	Tipo	Ejemplo	Comentario
1	Género	Individual	Categórico	Masculino, femenino, otro. Cualidad que describe la identidad de género del estudiante.
2	Edad	Individual	Numérico	22, 25, 30. Datos sobre la edad del estudiante en años.
3	Trabaja actualmente	Individual	Categórico	Sí. No. Indica si el estudiante trabaja actualmente mientras estudia.
4	Acceso a TIC	Individual	Categórico	Sí. No. Indica si el estudiante tiene acceso a tecnologías de la información y comunicación.
5	Tiene hijos	Individual	Categórico	Sí. No. Indica si el estudiante tiene hijos.
6	Municipio donde vive	Individual	Categórico	Bogotá, Mocoa, Puerto Asís. Nombre del municipio donde reside el estudiante.
7	Zona rural	Individual	Categórico	Sí. No. Indica si el estudiante vive en una zona rural.
8	Estado Civil	Individual	Categórico	Soltero, casado, unión libre, estadocivil del estudiante
9	Resultado prueba saber	Académico	Numérico	50, 100, 200 (puntajes). Puntaje obtenido por el estudiante en la prueba Saber.
10	Promedio semestre anterior	Académico	Numérico	2.5, 3.8, 4.2. Promedio de calificaciones del semestre académico anterior.

11	Promedio estudios secundaria	Académico	Numérico	7.5, 7.9, 9.1. Promedio de calificaciones de estudios de educación secundaria.
12	Jornada	Académico	Categorico	Mañana, tarde, noche. Jornada en la que el estudiante cursa sus estudios.
13	Carácter de institución educativa secundaria	Académico	Categorico	Privada o pública, carácter de la institución donde el estudiante terminó su secundaria
14	Método	Académico	Categorico	Virtual, presencial, adistancia. Modalidad en la que estudia
15	Carácter Ies	Institucional	Categorico	Pública, privada. Tipo de institución de educación superior (IES) en la que estudió el estudiante su secundaria.
16	Apoyo financiero	Institucional	Categorico	Beca, crédito. Tipo de apoyo financiero que recibe el estudiante de la institución de educación superior.
17	Icetex	Institucional	Categorico	Sí. No. Indica si el estudiante ha recibido apoyo financiero del Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior (ICETEX).
18	Ingreso	Núcleo familiar	Numérico	1500000, 2000000, 5000000. Ingresos familiares en pesos.
19	Número de hermanos	Núcleo familiar	Numérico	1, 2, 3. Cantidad de hermanos que tiene el estudiante.
20	Nivel educativo de la madre	Núcleo familiar	Categorico	Primaria, secundaria, superior. Nivel educativo alcanzado por la madre del estudiante.
21	Desertó	Criterio de clase	Categorico	Sí, no. Estudiantes que no tienen matrícula anual

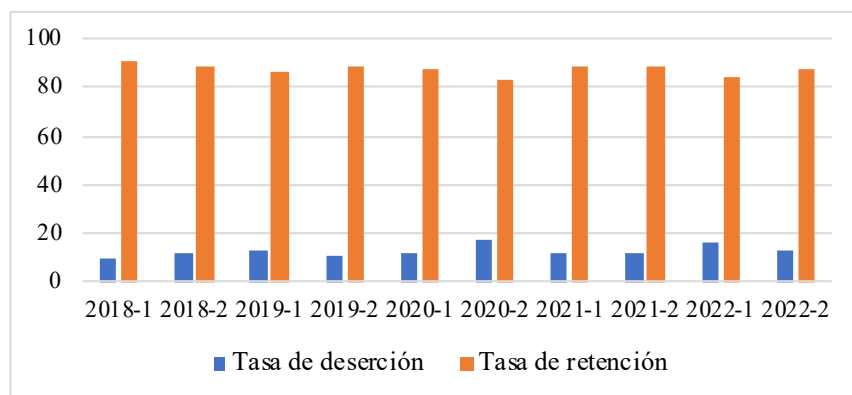
Fuente: propia basada en el modelo CRISP-DM.

La variable “desertó” se definió como la variable dependiente del modelo predictivo y se considera cuando el estudiante no cursa estudios durante dos o más cohortes consecutivas. Para el desarrollo del modelo se realizó una preparación exhaustiva de los datos, que incluyó la identificación y tratamiento de valores faltantes, el análisis de la base mediante Tableau, la agrupación de variables de caracterización como edad y nivel de ingresos, y la estandarización de formatos de fecha. Posteriormente, se efectuó una exploración visual de los datos con el fin de identificar tendencias, relaciones entre variables y casos atípicos, generando visualizaciones para cada variable analizada. En la fase de modelado, utilizando el software Knime, se aplicó el algoritmo de árbol de decisiones con una partición de datos 70% para entrenamiento y 30% para prueba. La evaluación se realizó mediante validación cruzada, alcanzando una exactitud del 92% y un margen de error del 4%. Finalmente, el modelo fue implementado en un entorno productivo y presentado a la institución para su integración al sistema de calidad institucional.

3. Resultados

El análisis visual de datos de SPADIES y fuentes internas muestra que, en la Fundación Universitaria del Área Andina – Sede Valledupar, la deserción estudiantil fluctuó en los últimos cinco años. Alcanzó su mínimo en 2019-2, aumentó durante la pandemia en 2020 y mantuvo una tendencia creciente hasta un pico en 2022.

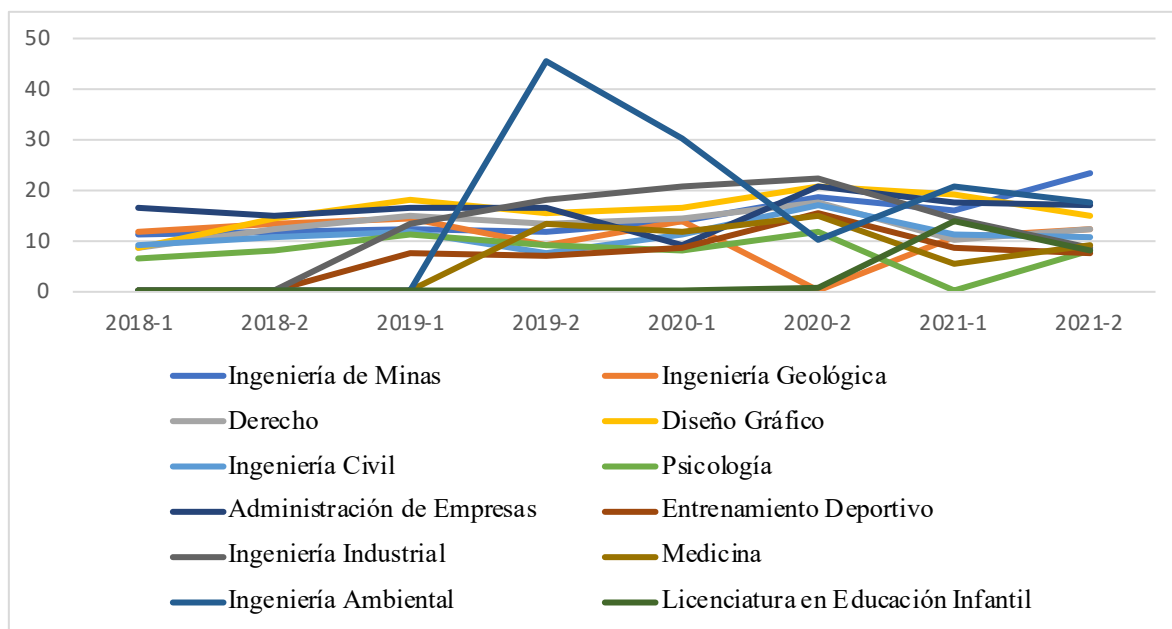
Figura 1. Tabla de matriculados y deserción anual, años 2018 a 2022; comparativo de los últimos 5 años de matriculados y desertores por semestre.



Fuente: Fundación Universitaria del Área Andina, Sede Valledupar, datos reportados por la IES en SPADIES.

Se identificó que la deserción varía significativamente según el programa y el periodo analizado. Ingeniería de Minas, Diseño Gráfico e Ingeniería Ambiental registraron incrementos notables en la deserción durante el segundo semestre de 2020 y el primero de 2021. En contraste, Entrenamiento Deportivo y Medicina presentaron tasas generalmente bajas, aunque Medicina mostró un aumento en el segundo semestre de 2021. De forma atípica, Ingeniería Geológica evidenció un pico abrupto en el segundo semestre de 2019 y un nuevo incremento en el primer semestre de 2021, asociado a la no oferta del programa por ajustes curriculares. Ingeniería Ambiental presentó un pico tras su creación, seguido de una tendencia descendente.

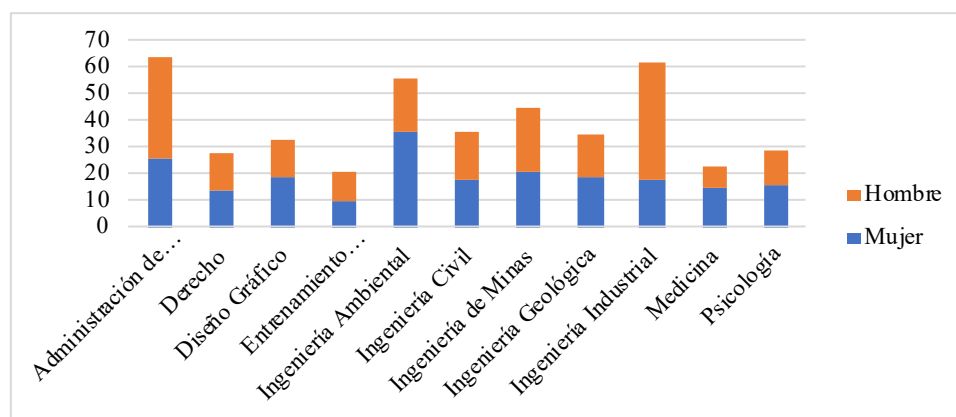
Figura 2. Deserción por programas académicos 2018-2022, Área Andina, Sede Valledupar. Comparativo de los últimos 5 años de deserción por programa académico.



Fuente: Datos institucionales

En cuanto a la deserción acumulada por sexo y programa, los programas con mayor deserción por parte de las mujeres son Medicina con 66.17% de los desertores, Ingeniería Ambiental con un 64.33% y Diseño Gráfico con 56.29%; Ingeniería Industrial con 71.25% y Administración de Empresas con 59.33%. Los programas que presentaron mayor deserción en el 2021 fueron Administración de empresas 30.47%, Ingeniería industrial 20.12% e Ingeniería ambiental 32.89%

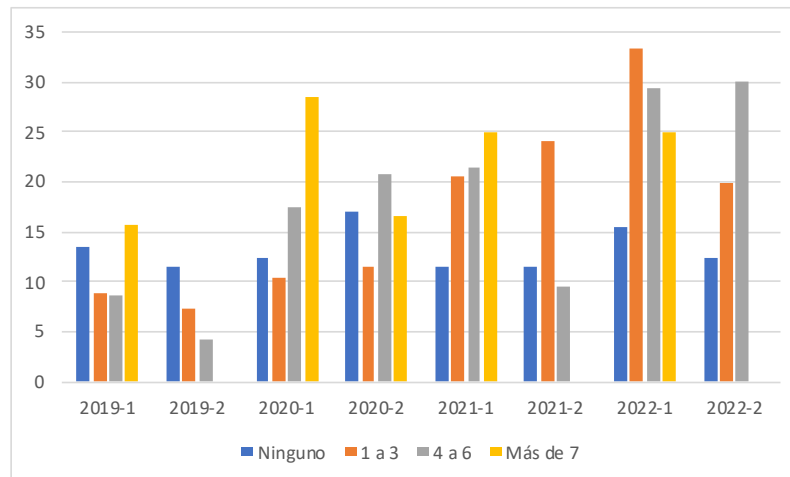
Figura 3. Deserción por sexo y programa académico 2018-2022, Fundación Universitaria del Área Andina, Sede Valledupar. Comparativo de deserción entre sexo y programa académico.



Fuente: Datos institucionales

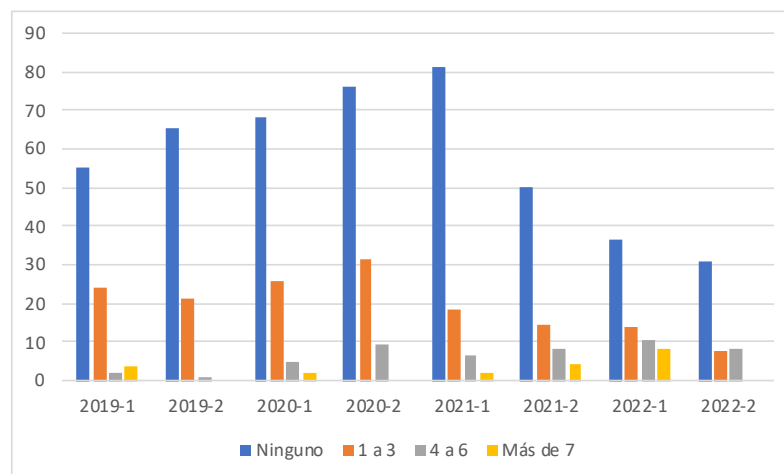
La deserción de estudiantes con líneas de crédito del ICETEX se mantiene relativamente estable a lo largo de los periodos analizados, con porcentajes que oscilan entre 11,42% y 16,84%. En contraste, los estudiantes sin líneas de crédito presentan una variabilidad moderada en sus tasas de deserción. No obstante, en el grupo que cuenta con entre una y tres líneas de crédito se evidencia un incremento progresivo y significativo de la deserción, que pasa del 10,53% en el primer semestre hasta alcanzar el 33% en el último año. Este grupo registra la mayor variabilidad, lo que sugiere que la gestión simultánea de múltiples créditos puede generar dificultades adicionales que influyen en la decisión de abandonar los estudios.

Figura 4. Deserción en programas de apoyo Icetex 2019-2022, Fundación Universitaria del Área Andina, Sede Valledupar. Variación de deserción en los últimos 4 años en estudiantes con líneas de crédito Icetex.



Fuente: Datos institucionales y del Icetex

Figura 5. Deserción en programas de apoyo financiero 2019-2022, Fundación Universitaria del Área Andina, Sede Valledupar. Variación de deserción en los últimos 4 semestres en estudiantes con líneas de crédito Icetex y estudiantes con apoyo financiero.

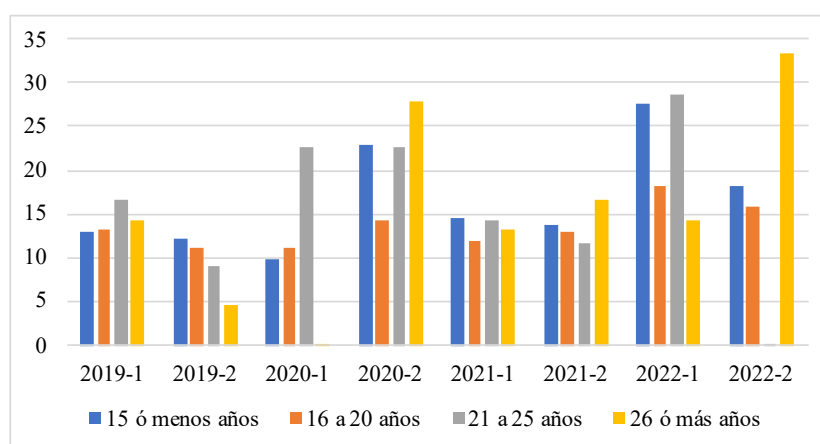


Fuente: Datos institucionales

En contraste con lo anterior, los estudiantes que tienen algún tipo de apoyo económico, como programas de gratuidad, descuentos o responsabilidad social con convenios con entidades como Instituto Colombiano de Bienestar Familiar (ICBF), Proyecto Sueños, Cerrejón, entre otros. Presentaron resultados distintos: la deserción es notablemente alta en estudiantes que no contaron con ningún tipo de apoyo, alcanzando su punto máximo en el segundo semestre de 2022, en el que de la totalidad de los estudiantes que desertaron, el 74,03% no tenían dichos apoyos, sugiriendo entonces que un nivel moderado de apoyo financiero condonable puede contribuir a una mayor retención estudiantil.

Ahora bien, el rango de edad en el que más se matriculan los estudiantes es de 16 a 20 años, lo que equivale al 76.2% de los estudiantes nuevos. En cuanto a la deserción, los estudiantes entre 21 y 25 años muestran una tendencia descendente en general; este patrón sugiere que este rango de edad puede estar experimentando una mayor retención en comparación con otros grupos. Los estudiantes de 26 años o más muestran mayores incrementos en la deserción en varios semestres, lo que permitiría inferir que fueron los más afectados por los desafíos que se presentaron durante la pandemia.

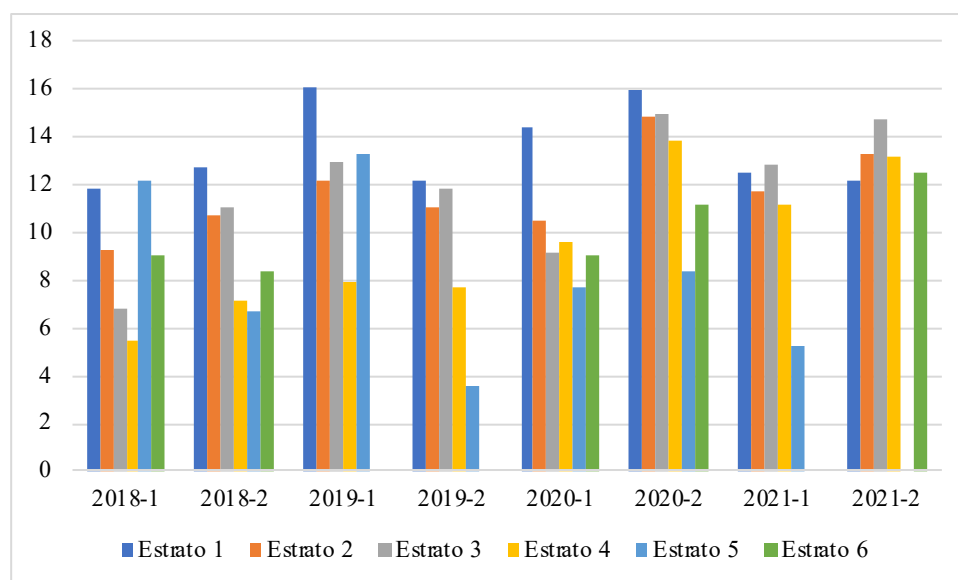
Figura 6. Deserción por edad 2019-2022, Fundación Universitaria del Área Andina, Sede Valledupar. Variación de deserción por rangos de edad en los últimos 8 semestres.



Fuente: Datos institucionales

En cuanto al estrato socioeconómico, los estudiantes que viven en estratos 1 y 3 presentan mayor deserción, con picos en algunos semestres, especialmente el 2020 y 2022. Aunque la deserción es relativamente alta en ciertos períodos, se mantiene en un rango moderado. Para el estrato 4, 5 y 6 las tasas son relativamente bajas, lo que sugiere una mayor retención en este grupo.

Figura 7. Deserción por estrato socioeconómico 2018-2021, Fundación Universitaria del Área Andina, Sede Valledupar. Variación de deserción por estrato socioeconómico en los últimos 8 semestres.

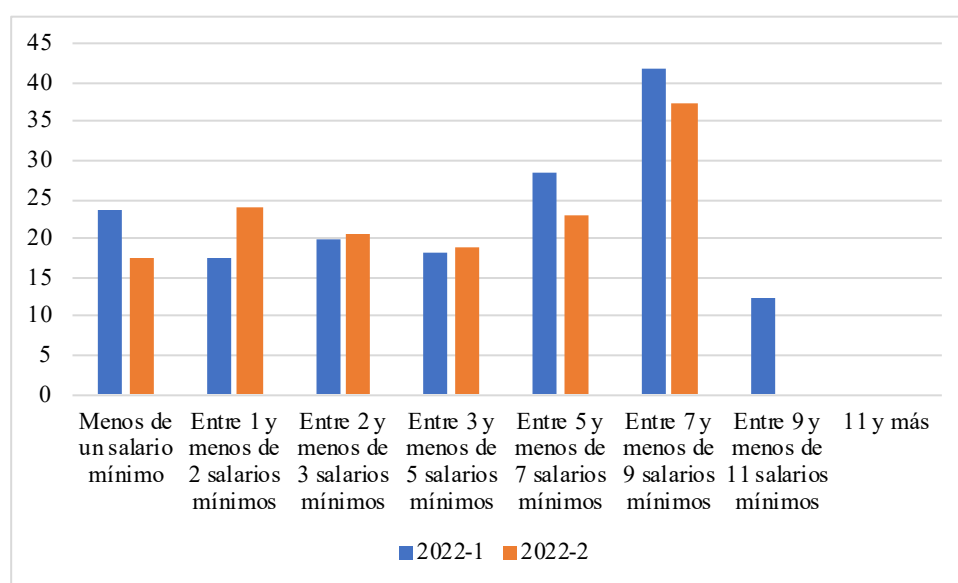


Fuente: Datos institucionales

El análisis por nivel de ingresos familiares muestra que los estudiantes con ingresos entre 2 y 3 salarios mínimos constituyen el 47,2% de la muestra, mientras que aquellos con ingresos inferiores a un salario mínimo representan el 20,93%. En ambos grupos se evidencian tasas de deserción más altas a lo largo del tiempo, aunque relativamente estables. No obstante, se registra un incremento significativo durante el segundo semestre de 2020 y el primero de 2021, lo que sugiere dificultades asociadas al contexto de la pandemia. En 2022, la deserción de estudiantes con ingresos entre 5 y 9 salarios mínimos presentó un aumento notorio, tras mantenerse estable en años previos. Este comportamiento podría estar relacionado con la

implementación de la política pública de gratuidad, orientada a beneficiar a los grupos A, B y C del SISBÉN y a población vulnerable, lo que fortaleció la permanencia de estudiantes con apoyo financiero. Cabe señalar que la población con ingresos superiores a 5 salarios mínimos es reducida en la institución, por lo que pequeñas variaciones generan picos más pronunciados en el análisis.

Figura 8. Deserción por ingresos semestre 2022, Fundación Universitaria del Área Andina, Sede Valledupar. Deserción por rangos de ingresos para los semestres 2022-1 y 2022-2.

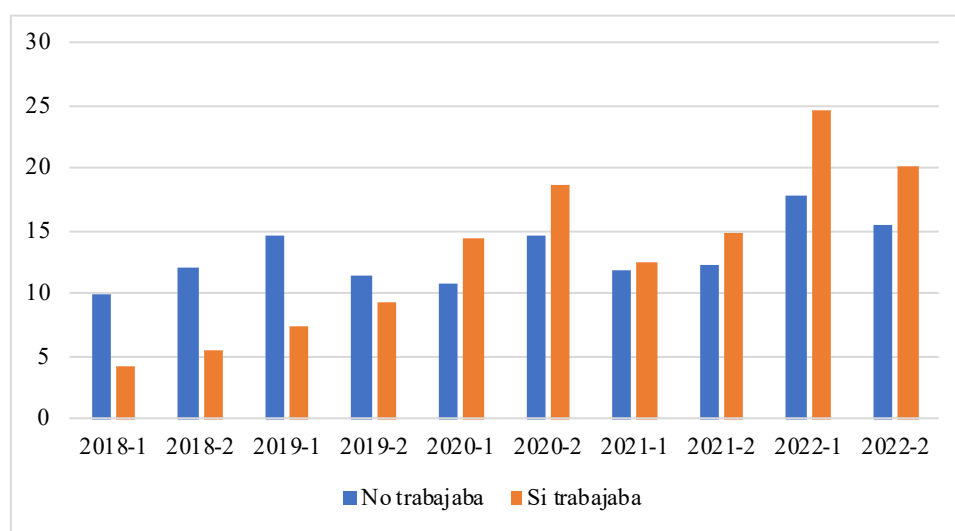


Fuente: Datos institucionales

El análisis evidencia que la probabilidad de deserción es mayor entre estudiantes provenientes de instituciones de educación media oficiales, comportamiento que coincide con la tendencia observada en los niveles de ingresos familiares, donde los estudiantes con menores ingresos presentan mayor riesgo de abandonar sus estudios. En relación con la variable “trabajaba al inicio de los estudios”, los estudiantes que no laboraban al comenzar su carrera registran tasas de deserción moderadas y relativamente estables, que oscilan entre el 10,01% y el 17,72%. En contraste, quienes sí trabajaban al inicio presentan, en general, menores tasas de deserción, aunque con mayor variabilidad, ubicándose entre el 4,26% y el 24,59%. No obstante, en este

grupo se identifica una tendencia creciente en la deserción a lo largo del tiempo, con un pico en el segundo semestre de 2022, lo que sugiere que las exigencias derivadas de combinar trabajo y estudio pudieron intensificarse en el periodo pospandemia.

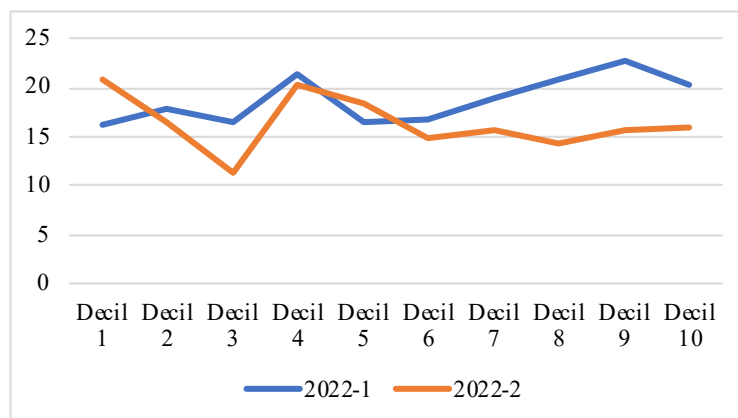
Figura 9. Deserción según situación laboral al inicio de estudios 2018-2022, Área Andina, Sede Valledupar. Deserción por variable de trabajo al inicio de estudios.



Fuente: Datos institucionales

Ahora bien, la variable de resultados de Pruebas Saber 11 fue relativamente poco influyente; sin embargo, se encontró que existe una relación entre el promedio académico acumulado del cuarto semestre con el promedio de las pruebas. Además, existe una tendencia a que los estudiantes con puntajes más bajos en Pruebas Saber 11 presenten tasas de deserción ligeramente más altas, fluctuando entre el 10.24% y el 20.69%, en comparación con los estudiantes que obtuvieron puntajes intermedios con fluctuaciones entre el 9.58% y el 18.29%. La tendencia vuelve a subir en los deciles más altos, entre el 8.27% y el 20.18%.

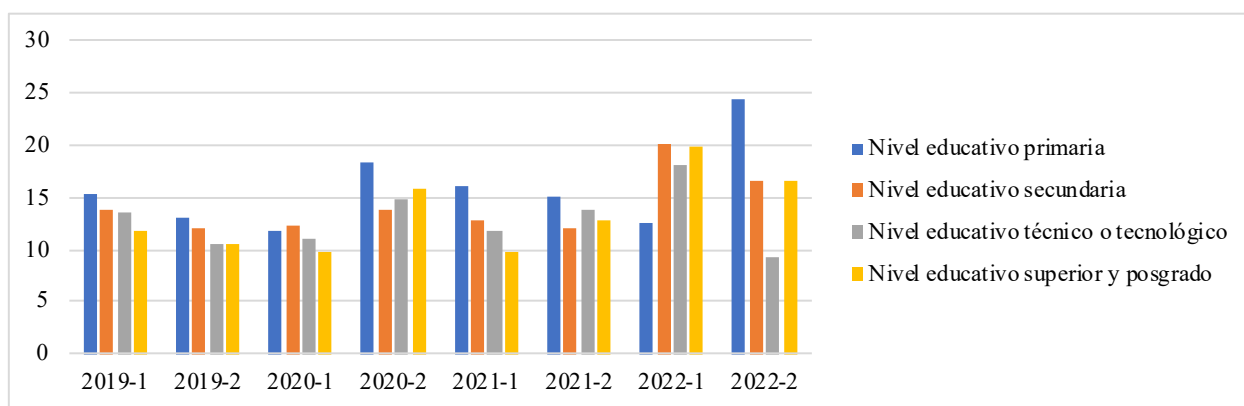
Figura 10. Deserción por resultados Prueba Saber 11, año 2022, Fundación Universitaria del Área Andina, Sede Valledupar. Deserción por deciles de puntaje en Pruebas Saber 11, año 2022.



Fuente: Datos institucionales e ICFES.

El análisis del promedio académico de los últimos semestres muestra que los estudiantes que continúan presentan un promedio ligeramente superior al de quienes desertan; sin embargo, aunque existe una diferencia en el último semestre cursado, esta no resulta estadísticamente significativa. Un comportamiento similar se observa en variables como el número de créditos matriculados por semestre y los créditos aprobados, las cuales, de manera individual, no constituyen predictores sólidos de la deserción en esta institución de educación superior. En contraste, la variable “último periodo cursado” evidencia que los estudiantes que desertan suelen hacerlo en periodos más recientes de su trayectoria académica, lo que sugiere una mayor propensión a la deserción en las etapas iniciales de la deserción: a mayor nivel educativo materno, menores tasas de abandono, pasando de rangos entre 11,89% y 24,44% en primaria a entre 9,74% y 19,96% en educación superior y posgrado.

Figura 11. Deserción por nivel educativo de la madre 2019-2022, Fundación Universitaria del Área Andina, Sede Valledupar. Deserción por variable nivel educativo de la madre en los últimos 8 semestres.



Fuente: Datos institucionales

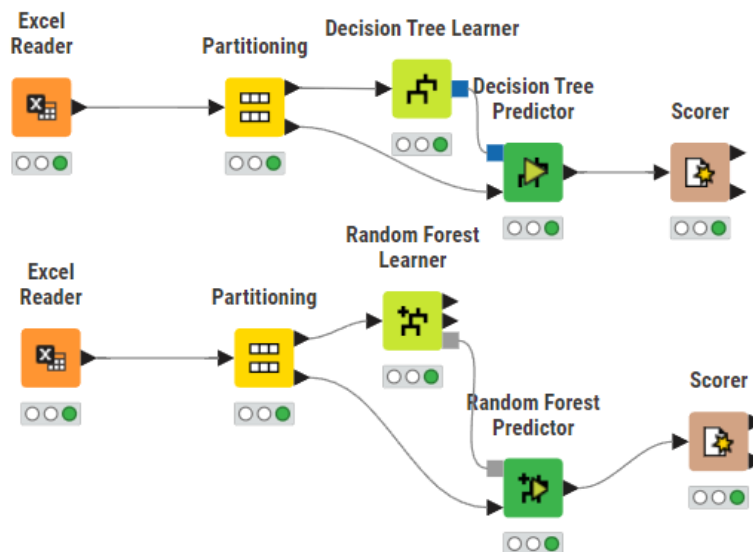
El análisis de las variables asociadas a salud mental y apoyos psicosociales evidenció que los estudiantes que viven solos presentan mayores niveles de deserción, alcanzando un 33%, en comparación con un promedio cercano al 16% en quienes conviven con una o más personas. Asimismo, se identificó que los estudiantes que cuentan con servicio médico o han hecho uso de los servicios de acompañamiento psicológico ofrecidos por la universidad registran una tasa de deserción aproximadamente 45% menor que aquellos que no acceden a este tipo de apoyos, lo que resalta la importancia del acompañamiento institucional en la permanencia estudiantil.

4. Modelado del prototipo

Tras la identificación de las variables más influyentes, se desarrolló el modelado del prototipo predictivo. Para ello, se importó el conjunto de datos de las cohortes, aplicando procesos de limpieza, conversión de variables, eliminación de valores atípicos e imputación de datos faltantes mediante promedios. Posteriormente, el dataset se dividió en conjuntos de entrenamiento y prueba. Se evaluaron dos algoritmos de aprendizaje automático con

diferentes cortes de partición: árbol de decisiones y Random Forest. Ambos presentaron niveles de exactitud similares, 95,76% y 95,72% respectivamente; no obstante, se seleccionó el árbol de decisiones por su mayor simplicidad, eficiencia y facilidad de interpretación, características especialmente relevantes en investigaciones educativas y sociales.

Figura 12. Flujo de trabajo del prototipo de predicción de deserción desarrollado utilizando el software KNIME. Se observan los diferentes módulos y nodos utilizados en el proceso, desde la importación de datos hasta la generación de predicciones.



Fuente: Elaboración propia basada en el modelo desarrollado en Knime.

Las pruebas resultaron fundamentales para validar la eficacia del modelo predictivo. Se evaluaron distintos algoritmos y se ajustaron de forma iterativa los parámetros de precisión, sensibilidad y especificidad, hasta alcanzar niveles óptimos de desempeño superiores al 95%. Para la validación del prototipo se construyó una base de datos con 30 patrones de estudiantes, de los cuales 10 habían desertado en los últimos semestres. El sistema identificó correctamente como de alta probabilidad los 10 casos de deserción y solo señaló dos falsos positivos entre los estudiantes sin abandono.

Posteriormente, el prototipo fue presentado a la institución de educación superior, permitiendo cargar datos de nuevas cohortes y generar alertas tempranas sobre estudiantes con mayor riesgo de deserción, junto con su caracterización según las variables analizadas. Esta información facilita la implementación de estrategias de acompañamiento y ajustes institucionales orientados a la permanencia estudiantil. El objetivo del estudio se centra en la prevención de la deserción, por lo que no se contempla una validación longitudinal a largo plazo del modelo.

Figura 13. Ejemplo de dataset previo a la partición de datos: Importación de los datos en formato .svc con 19 variables

Preview File Content

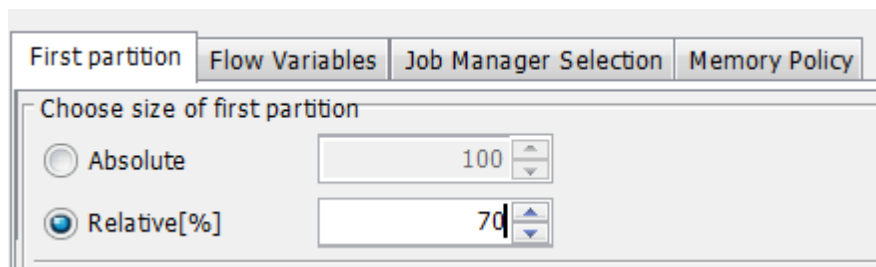
Preview with current settings

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	I 1. ID o a...	S 2. Sexo	I 3. Edad	S 4. Programa ...	S 5. Moda...	S 6. Estad...	S 7. Tipo ...	S 8. Parta...
Row0	1	Masculino	20	Ingeniería de Siste...	Presencial	Sojero	Contributivo	Sí
Row1	2	Femenino	22	Psicología	Virtual	Sojero	Subsidado	No
Row2	3	Masculino	24	Administración de E...	Presencial	Sojero	Contributivo	No
Row3	4	Femenino	21	Medicina	Presencial	Sojero	Subsidado	Sí
Row4	5	Masculino	23	Derecho	Virtual	Sojero	Contributivo	No
Row5	6	Femenino	25	Ingeniería Civil	Presencial	Sojero	Contributivo	No
Row6	7	Masculino	19	Contaduría Pública	Virtual	Sojero	Subsidado	Sí
Row7	8	Femenino	20	Ingeniería Industrial	Presencial	Sojero	Contributivo	Sí
Row8	9	Masculino	22	Enfermería	Presencial	Sojero	Contributivo	Sí
Row9	10	Femenino	21	Comunicación Social	Virtual	Sojero	Subsidado	No
Row10	11	Femenino	20	Ingeniería de Siste...	Presencial	Sojero	Contributivo	No
Row11	12	Masculino	25	Administración de E...	Virtual	Casado	Subsidado	Sí
Row12	13	Femenino	22	Psicología	Presencial	Sojero	Contributivo	No
Row13	14	Masculino	23	Medicina	Presencial	Sojero	Contributivo	Sí
Row14	15	Femenino	24	Ingeniería Civil	Presencial	Casado	Contributivo	No
Row15	16	Masculino	21	Derecho	Virtual	Sojero	Subsidado	Sí
Row16	17	Femenino	22	Ingeniería de Siste...	Presencial	Casado	Contributivo	No

Fuente: Captura de pantalla del proceso de partición en KNIME, elaboración propia

Figura 14. Proceso de partición de datos en el software KNIME. Se muestra cómo se divide el conjunto de datos en grupos de entrenamiento y prueba, un paso crucial para el desarrollo y validación del modelo predictivo.

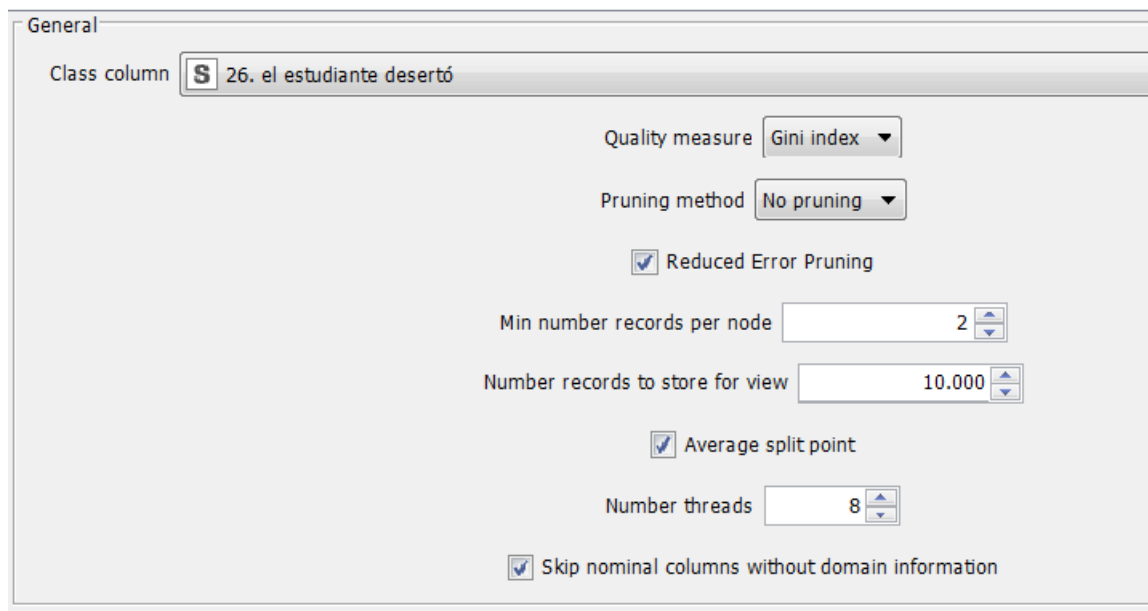


Fuente: Captura de pantalla del proceso de partición en KNIME, elaboración propia

. Se probó el algoritmo de árbol de decisiones utilizando el nodo "Decision Tree Learner" de Knime; también se probó el algoritmo Random Forest utilizando el nodo "Random Forest Learner". Se llevaron a cabo varias pruebas variando la proporción de datos de entrenamiento y prueba.

. Configuración de parámetros: Se ajustaron los parámetros de precisión, sensibilidad y especificidad para optimizar el rendimiento del modelo. Se ajustó la profundidad máxima del árbol. Se configuraron los criterios para la división de nodos (ganancia de información, índice Gini, etc.). Se establecieron los parámetros de poda para evitar el sobreajuste.

Figura 15. Configuración del algoritmo de árbol de decisión aplicado en el modelo. Se pueden ver los diferentes parámetros y ajustes utilizados para optimizar el rendimiento del algoritmo en la predicción de la deserción estudiantil.



Fuente: Captura de pantalla de la configuración del algoritmo en KNIME, elaboración propia.

Pruebas:

. Validación cruzada (cross-validation): Se utilizó la técnica de validación cruzada para evaluar el rendimiento del modelo. Este proceso implica dividir repetidamente los datos en conjuntos de entrenamiento y prueba, y calcular el rendimiento promedio en todos los casos. Se utilizó el nodo "X-Partitioner" de Knime para realizar validación cruzada. En esta etapa se dividió el dataset en K particiones o "folds"; se entrenó el modelo con K-1 particiones y se evaluó con la partición restante.

Este proceso se repitió varias veces, utilizando cada vez una partición diferente para la evaluación. Se calcularon las métricas promedio de rendimiento (precisión, sensibilidad, especificidad, etc.) en todos los folds.

Figura 16. Pruebas estadísticas utilizadas para evaluar los promedios en el modelo. Se muestran los resultados de las métricas de rendimiento que permiten validar la precisión y confiabilidad del modelo predictivo.

Rows: 25 | Columns: 14

Table

Statistics

Name	Type	# Missing val...	# Unique valu...	Minimum	Maximum	25% Quantile	50% Quantile ...	75% Quantile
10. Puntaje pr...	Number (inte...	0	41	260	400	291	320	351
11. promedio ...	Number (inte...	0	48	1,077,755	4,450,335	1,611,159.25	2,600,899	3,619,587.25

Fuente: Resultados de las pruebas estadísticas en KNIME, elaboración propia.

. Ajuste de parámetros: Se llevaron a cabo ajustes adicionales en los parámetros del modelo para alcanzar niveles óptimos de precisión superiores al 95% y un margen de error del 4%; se modificaron la profundidad máxima, los criterios de división de nodos y los parámetros de poda.

. Evaluación del prototipo: Se evaluó el rendimiento del prototipo utilizando la base de datos de prueba. Se utilizó el modelo entrenado y ajustado para realizar predicciones en la base de datos de prueba.

Figura 17. Resultados de la evaluación del prototipo. En la última columna se puede observar el resultado de aciertos entre los estudiantes que desertaron, lo que permite evaluar la eficacia del modelo en la predicción de la deserción.

16. Pobra... String	17. en m... String	18. Prese... String	19. Pade... String	21. ¿Ha p... String	22. ¿Cuá... Number (inte...)	23. ¿Quié... String	24. ¿Uste... String	25. Estrato Number (inte...)	26. el est... String	Predictio... String	Predictio... Number (dou...)
No	Amigo	No	No	Si	20	Padres y trab...	No	1	Si	Si	1
No	Familiar de ot...	No	Depresión, an...	No	50	Padres y trab...	No	2	Si	Si	1
No	Psicólogo	No	No	Si	45	Padres y cré...	No	3	Si	Si	1
No	Familiar con ...	No	No	Si	60	Trabajo y cré...	No	2	Si	Si	1
No	Amigo	No	No	Si	45	Padres	Si	1	Si	Si	1
No	Familiar con ...	No	No	No	60	Estudiante	No	4	Si	Si	1
No	Amigo	No	Pensamiento...	Si	30	Padres	No	1	Si	Si	1
No	Familiar de ot...	Si	No	No	45	Padres	No	3	Si	Si	1

Fuente: Resultados de la evaluación del prototipo en KNIME, elaboración propia

5. Presentación del prototipo a la IES

La presentación del prototipo a los responsables de Bienestar Universitario y Retención Estudiantil permitió validar su pertinencia y aceptación institucional. Los participantes resaltaron su potencial para identificar de manera temprana a estudiantes con mayor riesgo de deserción y optimizar la focalización de acciones y recursos. Se evidenció consenso sobre la necesidad de articular el prototipo con los sistemas de información y procesos existentes, así como de definir protocolos claros de actuación frente a las alertas generadas. Asimismo, se destacó la importancia de capacitar a los equipos encargados en el uso e interpretación de la herramienta. Aunque se reconocieron desafíos asociados a su implementación tecnológica, presupuestal y a la gestión del cambio, la percepción general fue positiva, considerando que el prototipo puede fortalecer un abordaje integral de los factores de riesgo y promover la innovación en las estrategias de acompañamiento estudiantil. Las reflexiones de los actores universitarios confirman la pertinencia práctica de los modelos basados en minería de datos y machine learning para abordar la deserción estudiantil.

Tabla 2. Análisis cualitativo del focus group realizado con actores clave de la institución educativa. Se muestran los conceptos más recurrentes y las ideas principales expresadas por los participantes en relación al prototipo de predicción de deserción.

Concepto	Idea
Abordaje integral y diferenciado	La alta precisión del modelo genera confianza en su utilidad. Las variables prioritarias (salud mental, factores socioeconómicos, rendimiento académico) requieren un abordaje integral y diferenciado. El prototipo se percibe como un habilitador de un abordaje más integral de los diversos factores de riesgo de deserción. Se propone desarrollar estrategias diferenciadas haciendo énfasis en los factores económicos, académicos y de salud mental. Es importante

	mantener un abordaje integral sin descuidar ningún aspecto.
Capacitación del equipo	Es necesaria la capacitación de todo el equipo en el uso e interpretación del modelo. Se requerirá un esfuerzo importante de capacitación a todo el equipo de bienestar para apropiar esta innovación.
Desafíos de implementación	Uno de los principales desafíos identificados es la integración tecnológica del modelo con los sistemas existentes. Otro desafío es la gestión presupuestal para garantizar los recursos necesarios para las intervenciones sugeridas. El manejo de posibles resistencias al cambio también se identifica como un desafío. Abordar estos desafíos requerirá un trabajo articulado con otras áreas, una comunicación estratégica y un claro liderazgo institucional. Se brindaron consideraciones valiosas para tener en cuenta en la implementación y aprovechamiento del prototipo.
Estrategias de acompañamiento	El modelo predictivo se percibe como una herramienta valiosa para identificar tempranamente estudiantes en riesgo de deserción. El diseño de protocolos claros de acción ante alertas tempranas será fundamental. El modelo abre oportunidades para innovar en estrategias de acompañamiento estudiantil. El focus group evidenció una percepción altamente positiva sobre la utilidad potencial del prototipo. Se validó la pertinencia y aceptación del prototipo por parte de actores clave de la universidad. El modelo debería generar alertas para activar rutas de

	acompañamiento oportunas. Se sugiere crear un programa integral de acompañamiento desde los programas, especialmente para estudiantes en alto riesgo. Se recomienda una ruta de apoyo psicosocial fortalecida para aquellos con mayores riesgos en salud mental. Se sugiere una ruta de acompañamiento académico intensivo para los estudiantes de bajo rendimiento. Se resalta la importancia de poder anticipar las razones de retiro de los estudiantes para diseñar estrategias preventivas.
Integración del modelo con sistemas existentes	Se requiere una integración efectiva del modelo con los sistemas de información y procesos actuales de bienestar. Se identificó la necesidad de una estrecha integración de esta herramienta con los sistemas de información y procesos existentes en la universidad. El modelo podría incorporarse desde el proceso de inducción y caracterización inicial de los estudiantes. Sería ideal que el modelo se nutra de los datos de sistemas como Brújula, las encuestas de facultad y bienestar. Se cuenta con recursos tecnológicos en la sede que permiten trabajar con herramientas como el modelo predictivo.

Fuente: Elaboración propia basada en el análisis de las transcripciones del focus group realizado en la Fundación Universitaria del Área Andina, Sede Valledupar.

Discusión y conclusiones

El presente estudio confirma que la deserción estudiantil en la educación superior es un fenómeno complejo, dinámico y altamente dependiente del contexto institucional y sociodemográfico. Los resultados obtenidos mediante la aplicación del modelo predictivo

basado en la metodología CRISP-DM y algoritmos de Machine Learning no solo son coherentes con los análisis descriptivos y visuales realizados, sino que además refuerzan hallazgos ampliamente documentados en la literatura, especialmente aquellos relacionados con factores socioeconómicos, académicos y psicosociales.

Desde una perspectiva comparativa, el modelo demuestra que variables como el apoyo financiero, la salud mental, la situación laboral, el nivel educativo del núcleo familiar y el momento de la trayectoria académica en el que se encuentra el estudiante son determinantes en la probabilidad de deserción. No obstante, la relevancia específica de cada variable no debe interpretarse como universal, sino como el resultado de interacciones propias del entorno institucional analizado. Este hallazgo pone en evidencia una de las principales limitaciones de los sistemas estandarizados de seguimiento, como SPADIES, y al mismo tiempo justifica la necesidad de modelos predictivos adaptables a cada institución. Es necesario brindar herramientas más expeditas a las universidades para estimar estos riesgos en menos tiempo.

En este sentido, uno de los principales aportes del estudio radica en demostrar que el modelo propuesto es replicable y escalable, siempre que se alimente con datos contextualizados y de calidad. La estructura metodológica empleada —basada en CRISP-DM— permite que otras instituciones de educación superior, tanto públicas como privadas, ajusten el modelo a sus propias realidades territoriales, modalidades académicas y perfiles estudiantiles, sin necesidad de replicar exactamente las mismas variables o ponderaciones.

Las herramientas de big data permiten estimar los riesgos de deserción con un margen de error altamente confiable y, junto con un sistema de permanencia, ayudan a tomar decisiones oportunas sobre los estudiantes en riesgo, incluyendo la prevención de problemas de salud mental. Aunque estas herramientas se están popularizando y ofrecen resultados más rápidos y efectivos a un costo menor, es necesaria la capacitación en su uso para aprovecharlas al máximo. Asimismo, el uso de algoritmos interpretables, como el árbol de decisiones, facilita la apropiación del modelo por parte de equipos no especializados en ciencia de datos, lo que incrementa su viabilidad en contextos educativos diversos

El ejercicio de acercamiento inicial y posterior presentación realizada con responsables de áreas clave de bienestar universitario y retención estudiantil de la IES permitió validar la pertinencia y aceptación del prototipo de estimación de riesgo de deserción desarrollado. La validación con actores universitarios ratificó el valor del prototipo para la detección temprana y diseño de estrategias preventivas focalizadas; se evidenció un consenso sobre la necesidad de integrar estrechamente este prototipo con los sistemas de información y procesos existentes en la universidad.

Desde el punto de vista práctico, el prototipo no es una herramienta de predicción determinista, sino un sistema de alerta temprana que habilita la toma de decisiones preventivas. Su valor reside menos en anticipar con certeza la deserción futura y más en permitir intervenciones oportunas, focalizadas y basadas en evidencia, como acompañamiento psicosocial, apoyos financieros diferenciados, ajustes curriculares o estrategias pedagógicas flexibles.

La implementación de este prototipo en la institución de educación superior representa una oportunidad para mejorar los procesos de retención estudiantil y tomar acciones preventivas oportunas. Con información precisa sobre los riesgos de deserción, las autoridades educativas podrán diseñar programas de acompañamiento personalizado, ajustes curriculares, servicios de apoyo financiero y psicosocial, entre otras medidas focalizadas. Este estudio abre la puerta a nuevas líneas de investigación orientadas a la validación longitudinal del modelo, su integración con sistemas nacionales de información y su adaptación a otros niveles educativos o modalidades formativas.

Referencias

Aguilera, J., Campos, J., & Rodríguez, M. (2017). MODELO PREDICTIVO PARA LA PERMANENCIA EN LA EDUCACIÓN SUPERIOR Línea Temática: Factores asociados. Tipos y perfiles de abandono. Montserrat. En VII CLABES.

Alban Taipe, M. S. (2019). Contribuciones a la predicción de la deserción universitaria a través de minería de datos. Universidad Nacional Mayor de San Marcos.

Bartolomé Pina, A. R., Bellver Torlà, C., Castañeda Quintero, L., & Adell Segura, J. (2017). Blockchain en educación: introducción y crítica al estado de la cuestión. *Edutec. Revista Electrónica de Tecnología Educativa*, 61. <https://doi.org/10.21556/edutec.2017.61.915>

Camargo, A. J. (2020). Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos. En Universidad de la Costa - CUC.

Cayón, E., Correa, J. S., & Scarpetta, L. E. (2021). Caracterización de los programas académicos y determinación de las variables que afectan el rendimiento de los estudiantes en el Colegio de Estudios Superiores de Administración, CESA. *Formación Universitaria*, 14(5). <https://doi.org/10.4067/S0718-50062021000500175>

Cuevas Acosta, J. H., Solís, M. C., & Silva Contreras, I. (2019). Programación computacional y análisis de datos en educación estadística. *Areté: Revista Digital del Doctorado en Educación de la Universidad Central de Venezuela*, ISSN-e 2443-4566, Vol. 5, No. 9, 2019, págs. 11-27, 5(9).

Enríquez Vázquez, L., & Navarro Perales, J. (2024). Explorar los matices: aprendizaje personalizado y adaptativo en la educación digital. *Revista Digital Universitaria*, 25(1). <https://doi.org/10.22201/cuaieed.16076079e.2024.25.1.10>

Garcés Prettel, M., De la Ossa Robinson, S., Arellano Cartagena, W., Alvis Arrieta, J., & Figueroa Castillo, L. C. (2024). ¿Volver o no volver a las clases presenciales? Motivaciones y temores que influyen en la deserción universitaria en Colombia en tiempos de postpandemia. *Salud Uninorte*, 40(1). <https://doi.org/10.14482/sun.40.01.159.456>

García-Martínez, R., Poblano-Ojinaga, E. R., & García-Gámez, L. (2024). Factores determinantes en la elección de una carrera universitaria. *Investigación Administrativa*, 53–1. <https://doi.org/10.35426/iav53n133.02>

Gómez-Restrepo, C., Padilla Muñoz, A., & Rincón, C. J. (2016). Deserción escolar de adolescentes a partir de un estudio de corte transversal: Encuesta Nacional de Salud Mental

Colombia 2015. Revista Colombiana de Psiquiatría, 45.
<https://doi.org/10.1016/j.rcp.2016.09.003>

González Catalán, F., & Arismendi Vera, K. J. (2018). Deserción estudiantil en la educación superior técnico-profesional: Explorando los factores que inciden en alumnos de primer año. Revista de la Educación Superior, 47(188), 109-137.
<https://doi.org/10.36857/resu.2018.188.510>

González, L. E., & Espinoza, O. (2015). Deserción en educación superior en América Latina y el Caribe. Paideia, 45 (julio-diciembre).

Guerrero, S. C. (2018). Caracterización de la deserción en la Universidad Pedagógica y Tecnológica de Colombia durante el período 2008-2015. Revista Lasallista de Investigación, 15(1). <https://doi.org/10.22507/rli.v15n1a2>

Guzmán Rincón, A., Barragán Moreno, S., Cala-Vitery, F., & Segovia-García, N. (2022). Deserción en la educación superior rural: Análisis de causas desde el pensamiento sistémico. Qualitative Research in Education, 11(2). <https://doi.org/10.17583/qre.10048>

Hernández Sampieri 2014. (2014). Metodología de la investigación, 6. McGraw-Hill / Interamericana Editores, S.A. de C.V. ISBN: 978-1-4562-2396-0

Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior. Calidad en la educación, 17. <https://doi.org/10.31619/caledu.n17.409>

Hugo, V., Garcia, M., Victoria, A., Motta, A., Daniela, M., & Ariza, A. (2019). Predictive models of knowledge management in virtual education Modelo predictivo de gestión del conocimiento en educación virtual. Industry, Innovation, And Infrastructure for Sustainable Cities and Communities.

Mella-Norambuena, J., Badilla, M. G., & López Angulo, Y. (2022). Modelos predictivos basados en el uso de analíticas de aprendizaje en educación superior. Texto libre, 15. <https://doi.org/10.35699/1983-3652.2022.36310>

Norma, P., Barron-Adame, R. A., & Miguel, J. (2016). Revista de Aplicación Científica y Técnica: Propuesta del diseño de un modelo predictivo de alerta temprana en indicadores educativos de nivel superior aplicando minería de datos. Junio, 2(4).

OCDE. (2021). Health at a Glance 2021. En Health at a Glance 2021.

Rodríguez Ojeda, M. J. (2024). Determinación de perfiles profesionales mediante técnicas de minería de datos. *Ciencia Latina Revista Científica Multidisciplinar*, 8(1). https://doi.org/10.37811/cl_rcm.v8i1.9783

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, 20(1). <https://doi.org/10.1177/1536867X20909688>

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, 8(26).

Vanneschi, L., & Silva, S. (2023). Decision Tree Learning. En *Natural Computing Series*. https://doi.org/10.1007/978-3-031-17922-8_6

Vásquez Chicaiza, F. P., Vega Cocha, D. P., Defaz Lasso, M. L., Vazco Silva, C. D., & López Velasco, J. E. (2024). Estrategias educativas por medio de herramientas digitales basadas en inteligencia artificial, revisión bibliográfica. *Ciencia Latina Revista Científica Multidisciplinar*, volumen 7 (número 6). 5691-5708. https://doi.org/10.37811/cl_rcm.v7i6.9110

Villena, J. (2016). CRISP-DM: La metodología para poner orden en los proyectos de data science. *Singular*.