MODELO DESCRIPTIVO Y PREDICTIVO PARA OPTIMIZAR LA MEJORA EN LAS PRUEBAS SABER PRO: UN ENFOQUE BASADO EN DATOS HISTÓRICOS Y CARACTERIZACIÓN ESTUDIANTIL

Javier Andrés Mendieta López

Maestría en Ingeniería y Analítica de Datos
MIAD

Universidad Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería,
Bogotá D.C., Colombia
2024

MODELO DESCRIPTIVO Y PREDICTIVO PARA OPTIMIZAR LA MEJORA EN LAS PRUEBAS SABER PRO: UN ENFOQUE BASADO EN DATOS HISTÓRICOS Y CARACTERIZACIÓN ESTUDIANTIL

Javier Andrés Mendieta López

Tesis de grado presentada como requisito parcial para optar al título de:
Magister en ingeniería y analítica de datos

Director:
Jorge Aurelio Herrera

Universidad Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería,
Bogotá D.C., Colombia
2024

DEDICATORIA

Dedico con profundo respeto y eterna gratitud este trabajo de grado a Dios y a la Virgen María, cuya guía divina ha sido la luz en cada paso de mi camino hacia la culminación de este trascendental capítulo en mi desarrollo profesional. Este logro representa mucho más que una meta académica; es un testimonio del amor y la fe que constituyen los pilares de mi vida.

Dedico este esfuerzo a mis queridos padres y a mi hermano, quienes han sido la fuente de energía y el apoyo que han sustentado mi espíritu en cada desafío enfrentado. Su presencia inagotable y su apoyo incansable han sido fundamentales para superar los grandes obstáculos y alcanzar las metas personales que me he propuesto.

Dedico este trabajo a mi primo Daniel, quien, aunque ya no está físicamente entre nosotros, continúa siendo una luminosa fuente de inspiración. Desde los días de mi niñez, su ejemplo de fortaleza, determinación y capacidad para superar adversidades ha dejado una huella indeleble en mi alma. Daniel me enseñó que es posible alcanzar grandes logros mediante la integridad, el estudio y la disciplina rigurosa. Su espíritu valiente y su legado perduran en cada página de este trabajo, guiándome hacia la excelencia como la estrella que ilumina mi camino.

Dedico también este logro a toda mi familia, cuyos ánimos y soporte constante me han impulsado a persistir y jamás rendirme. Espero ser una guía inspiradora para mi familia, probando que, con dedicación y amor, los sueños se hacen realidad.

Dedico mi gratitud a mis amigos y compañeros de trabajo por el apoyo recibido y el conocimiento compartido. Cada clase y cada conversación han enriquecido mi experiencia, reforzando la camaradería y el intercambio intelectual que nos caracteriza. Agradezco y dedico este triunfo a todas las personas que, con sus consejos y ejemplos, han forjado en mí a un profesional competente y a una persona íntegra y comprometida. Me siento profundamente orgulloso de los logros alcanzados, que son tanto suyos como míos.

Dedico un sincero reconocimiento a los profesores, cuya dedicación, esfuerzo y sabiduría no solo me formaron académicamente, sino que también me inspiraron a perseverar y completar esta maestría.

Me dedico este trabajo a mí, a mi esfuerzo y mi capacidad resiliente que no importo cual grande fue la meta al final llegue y con orgullo puedo decir lo logre, con muchos aprendizajes en el proceso que me hacen más sabio y mejor ser humano

AGRADECIMIENTOS

Expreso mi más profundo agradecimiento a todas las personas que, tanto de manera directa como indirecta, contribuyeron a la elaboración y culminación de este trabajo de grado. Vuestra presencia en los momentos en que más lo necesité ha sido invaluable. En esos días en los que la inspiración parecía esquiva, vuestras palabras fueron el faro que redirigió mis pensamientos y me encaminó hacia la meta propuesta. Os estoy infinitamente agradecido por estar allí en cada paso crucial de este viaje.

Especial reconocimiento merece mi director de tesis, Jorge Herrera, cuyas constantes palabras de aliento y guía experta han sido fundamentales en este proceso. Gracias, Jorge, por tu paciencia y tu sabiduría, y por impulsarme a alcanzar los objetivos que nos propusimos en este significativo proyecto.

También agradezco sinceramente al profesor Olmer Garcia Bedoya por su apoyo generoso y desinteresado. Su orientación experta fue crucial en la elaboración y finalización de este trabajo, ayudándome a cumplir con los objetivos establecidos y a superar los desafíos con integridad y dedicación.

A todos ustedes, mi eterna gratitud. Su ayuda y apoyo no solo hicieron posible este trabajo, sino que también enriquecieron mi experiencia personal y profesional de manera imborrable.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

La necesidad de medir y conocer la evolución de las instituciones en las propuestas educativas es una constante del Ministerio de Educación en Colombia, con el ánimo de controlar y responder de manera más eficiente a las diferentes circunstancias del entorno educativo, en el proceso de esta búsqueda, el ICFES apoya con la realización de las pruebas saber que buscan medir la evolución de la educación de los estudiantes en diferentes puntos de su ciclo educativo. A través de estas mediciones el ICFES ha venido desarrollando el concepto de valor agregado o también llamado aporte relativo, que busca con dos mediciones una de entrada y otra de salida determinar cuál fue ese aporte o valor que genero la institución educativa en los estudiantes durante su ciclo de formación. [1]

Para el caso puntual de las instituciones de educación superior el punto de partida son las pruebas saber 11 que desarrolla un estudiante durante el proceso de su formación de educación media y que son requisito para su ingreso a la educación superior. Cuando el estudiante cumple con el 75% de su proceso de formación dentro de la institución de educación superior los estudiantes pueden acceder a las pruebas saber pro como un requisito para su grado, el cual cumple como punto de cierre o salida para medir el paso de los estudiantes en una institución universitaria de educación superior.

Con base en lo anterior se cuenta con el proceso de medición de entrada a la educación superior y con el proceso de medición a la salida del proceso formativo en la educación superior que mide, en los dos puntos las competencias básicas establecidas por el MEN. [2, 3][1].

Estas competencias son comparables en el tiempo y permiten identificar como un estudiante ingresó a la educación superior y como en su proceso formativo fue su evolución, lo que el ICFES determina como aporte o valor agregado que dio la institución de educación superior a cada uno de los estudiantes y permite medir las fortalezas y debilidades de cada institución. Lo cual toma mayor relevancia en el decreto 1330 de 2019, donde se establece la necesidad de tener instrumentos de medición internos y externos por parte de las universidades que permitan identificar de manera segmentada y a diferentes niveles la evolución en los procesos formativos y que apoye la generación de acciones para garantizar la calidad de los programas y de la institución.

Para las instituciones universitarias cada día se convierte en un reto para identificar de manera efectiva las acciones más eficaces que se deriven en mejores resultados, esto se ha venido trabajando de una mejor manera cuando las instituciones se apoyan en datos. Es acá cuando estos datos cobran una mayor relevancia, si han estado construidos de una manera adecuada desde años anteriores.

Esta investigación está orientada a entregar una herramienta a la universidad que integra información de los últimos 6 años de las fuentes de ésta y las entregadas por el ICFES

[1] Compilados por el Decreto 1075 de 2015.

que consolidadas y mediante un modelamiento, mostraran las debilidades y necesidades para el proceso que permita ejecutar acciones tempranas, segmentadas y por diferentes niveles en los estudiantes para apoyarlos en el desarrollo de sus competencias básicas y que conlleven a mejorar los resultados en las pruebas saber pro. Esta herramienta también puede ser usada como instrumento de medición al interior de la universidad mostrando la evolución de las diferentes tomas de muestras en el tiempo.

El constante desafío de mejorar la calidad educativa en Colombia ha llevado al Ministerio de Educación a enfocarse en la evaluación y el seguimiento del rendimiento de las instituciones de educación superior. En este contexto, las Pruebas SABER PRO, administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), emergen como un instrumento clave. Estas pruebas no solo miden el desempeño académico de los estudiantes en distintos momentos de su trayectoria educativa, sino que también ofrecen una perspectiva valiosa sobre el valor agregado por las instituciones durante el ciclo formativo de sus estudiantes.

Este estudio se centra en las pruebas SABER 11 y SABER PRO, que respectivamente marcan el ingreso y la culminación del ciclo de educación superior en Colombia. La importancia de estas evaluaciones se ve reforzada por la legislación colombiana, incluyendo la Ley 1324 de 2009 y el Decreto 1075 de 2015, que establecen las competencias básicas a medir en ambos exámenes. Estas competencias comparables en el tiempo permiten rastrear la trayectoria académica de los estudiantes desde su ingreso hasta su graduación, proporcionando así una medida del aporte educativo de las instituciones.

El Decreto 1330 de 2019 subraya aún más la necesidad de herramientas de evaluación internas y externas en las universidades, con el objetivo de monitorear y mejorar continuamente la calidad educativa. En este contexto, la presente investigación propone desarrollar una herramienta analítica basada en datos históricos y la caracterización estudiantil. Este modelo no solo identificará las áreas de fortaleza y debilidad de las instituciones, también facilitará la implementación de estrategias específicas para mejorar el rendimiento en las Pruebas SABER PRO.

Con una metodología que integra información de los últimos seis años, esta herramienta permitirá a las instituciones universitarias realizar un seguimiento detallado y segmentado de sus estudiantes, ofreciendo así un apoyo crucial en el desarrollo de competencias clave. Con este estudio se busca mejorar los resultados de las pruebas estandarizadas y enriquecer la experiencia educativa global de los estudiantes en el sistema de educación superior de Colombia.

2. MARCO TEÓRICO

2.1. ANTECEDENTES

Pruebas PISA: La Organización para la Cooperación y el Desarrollo Económicos (OECD) mide mediante el Programa para la Evaluación Internacional de Alumnos (PISA), 79 países en el mundo, con una participación de 600.000 estudiantes con edad de 15 años, en áreas de matemáticas, lectura y ciencias, cada tres años. Colombia en el año 2012 se ubicó en el puesto 62, en 2015 en el puesto 57 y en 2018 en el puesto 58 ubicándose por debajo de la media en los tres años.

Año	Puesto	Pais	Matemáticas	Lectura	Ciencias
2012	62	Colombia	376	403	399
2015	57	Colombia	390	425	416
2018	58	Colombia	391	412	413

Ilustración 1. *Resultados Pruebas PISA 2012-2018*
Fuente: *Elaboración propia, Basado en* [4]

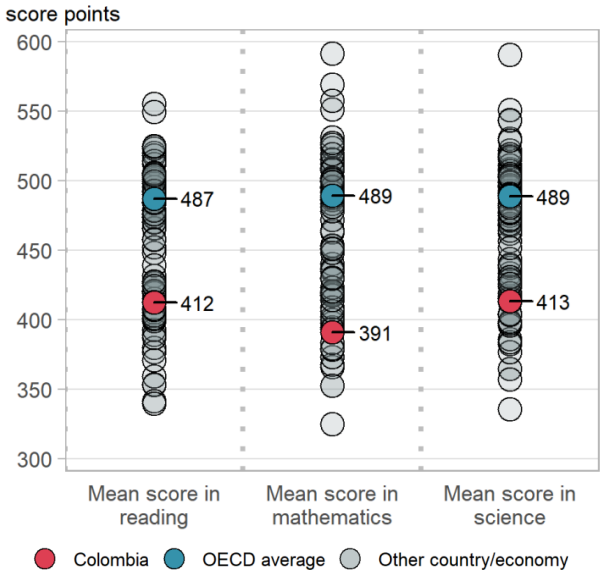


Ilustración 2. *Panorama del rendimiento de lectura, matemáticas y ciencias*
Fuente: Tomado de [5, p. 2].

De acuerdo con el informe de la OECD Colombia presenta una mejoría en los resultados de 2012 a 2015 significativa frente a otros países en todas las áreas; para el 2018 se ve un crecimiento de un punto en matemáticas, pero una desaceleración en lectura y ciencias. **1.** El gobierno nacional ha venido implementando planes que apoyen la mejora de estos puntajes con programas de gobierno como “Ser Pilo Paga” y “Atención a la primera infancia” que apoyan directamente a las poblaciones vulnerables. Sin embargo, no tiene una prueba que se alinee con las mediciones de los estándares internacionales,

pero cuenta con una serie de exámenes que miden y apoyan la toma de muestras en diferentes ciclos del estudiante como lo son las pruebas saber. Bajo esta serie de exámenes encontramos las pruebas saber pro que mide la condición educativa y los aportes que, desde la educación media a la educación superior logra el estudiante.

La Prueba Saber Pro: es una prueba censal que se aplica a los estudiantes que cursan un pregrado y ha completado más del 75% de su malla curricular dentro de una institución de educación superior reconocida por el Ministerio de Educación en Colombia, La prueba saber pro se convierte en herramienta de medición dentro de este ciclo de formación y hace parte de los requisitos de grado de una carrera universitaria exigidos por el Ministerio. Esta prueba desde el año 2016 a 2021 ha tenido el mismo diseño y la misma escala de calificación, lo que nos permite mediante este histórico analizar comportamientos y patrones con los datos, y crear un modelo que nos permita mediante los datos llegar a predecir comportamientos de estudiantes que están próximos a presentar la prueba.

En este proceso se plantean varios retos que se pueden abordar desde las diferentes investigaciones que se encuentran como la de Ortiz y otros [6], donde a partir de un estudio entre dos universidades con características similares ofrecen a sus estudiantes la posibilidad de estudiar bajo una de las dos modalidades que tiene disponibles con sus planes de estudio presencial y virtual; En este estudio se toma un dato importante al no encontrar diferencias significativas entre los estudiantes de una u otra modalidad en las dos universidades, con lo cual se determina manejar el uso de los datos como una sola base sin realizar dos modelos.

Sin embargo, también nos arroja información para el uso de los datos específicos dentro de las competencias genéricas, donde se determina que en la competencia de Ingles se presentan diferencias entre las modalidades, demostrando estadísticamente que la modalidad presencial se ubica en una media y la modalidad virtual en la baja para una de las instituciones universitarias.

2.2. MINERÍA DE DATOS

Con el proceso de minería de los datos se pretende encontrar patrones o asociaciones ocultos que nos permitan generar conocimiento de valor en el comportamiento de los estudiantes, en su mayoría desconocida para la institución universitaria y que aporta significativamente para la comprensión del fenómeno que se pretende predecir con este modelo. [7]

El manejo y explotación de los datos es la parte fundamental de todo modelo analítico, por lo cual el manejo de grandes volúmenes de datos determina el éxito del modelo, para esto es necesario el uso de metodologías ya aplicadas que ayudan a la búsqueda de datos que están por fuera de la estadística descriptiva y que emplean algoritmos determinados para conseguir el mayor provecho de la información.

3. ESTADO DEL ARTE

3.1. ARTICULOS REFERENTES

3.1.1. ANALYZING UNDERGRADUATE STUDENTS' PERFORMANCE USING EDUCATIONAL DATA MINING

Tabla 1. *Analyzing undergraduate students' performance using educational data mining*
Fuente: Elaboración propia, Basado en [8]

Aspecto	Detalles
Objetivos del Artículo	Analizar el rendimiento de estudiantes en un programa de licenciatura de 4 años en Tecnología de la Información.
	Predecir el rendimiento académico al final del programa utilizando solo calificaciones de admisión y de los primeros dos años.
Herramientas y Técnicas	Uso de varios clasificadores para predecir el rendimiento.
	Utilización de árboles de decisión para identificar cursos indicativos del rendimiento del estudiante.
	Análisis de datos de 210 estudiantes utilizando el software RapidMiner.
Limitaciones Vs. Proyecto	El enfoque se limita a calificaciones y no considera factores socioeconómicos o demográficos, que podrían ser relevantes en el contexto.
	La investigación se centra en un contexto específico (universidad pública en Pakistán), lo que podría limitar su aplicabilidad directa al proyecto.
	La precisión de los clasificadores podría no ser suficiente para aplicaciones más complejas o variadas en el proyecto.

3.1.2. MINERÍA DE DATOS EDUCATIVOS: ANÁLISIS DEL DESEMPEÑO DE ESTUDIANTES DE INGENIERÍA EN LAS PRUEBAS SABER-PRO

Tabla 2. *Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO*

Aspecto	Detalles
Objetivos del Artículo	Estudiar los resultados de las pruebas SABER-PRO de estudiantes de ingeniería en Antioquia usando minería de datos.
	Realizar tres modelos analíticos: agrupación de tipos de estudiantes, selección de factores influyentes y predicción de desempeño.
Herramientas y Técnicas	Aplicación de la metodología CRISP-DM para minería de datos educativos.
	Análisis de 49,021 registros con 108 variables académicas, económicas y sociodemográficas.

Aspecto	Detalles
	Uso de algoritmo K-means para clustering, sistema de votación para selección de factores, y algoritmo KNN para predicción.
Limitaciones Vs. Proyecto	El enfoque se centra en estudiantes de ingeniería en una región específica, lo que podría limitar su aplicabilidad a tu contexto más amplio.
	El modelo tiene un mayor desafío con los estudiantes de puntaje medio, lo que podría ser relevante para tu proyecto al considerar a todos los estudiantes.

Fuente: Elaboración propia, Basado en [9]

3.1.3. RESULTADOS EN SABER PRO DE ESTUDIANTES DE MODALIDAD PRESENCIAL Y VIRTUAL EN DOS UNIVERSIDADES COLOMBIANAS

Tabla 3. *Resultados en Saber Pro de estudiantes de modalidad presencial y virtual en dos universidades colombianas*

Aspecto	Detalles
Objetivos del Artículo	Analizar diferencias en los resultados de las pruebas Saber Pro entre estudiantes de modalidades presencial y virtual en administración de empresas de dos instituciones colombianas.
Herramientas y Técnicas	Uso de análisis descriptivos, gráficos de tendencia, coeficientes de correlación Rho de Spearman.
	Pruebas no paramétricas U de Mann-Whitney, Anova y Kruskall-Wallis para comparar modalidades y efectos institucionales.
Limitaciones Vs. Proyecto	El estudio se centra en un campo específico (administración de empresas) y en dos instituciones, lo que puede limitar la aplicabilidad a contextos más amplios.
	- Aunque se evidencian algunas diferencias significativas, no se encuentran diferencias estadísticamente significativas en todas las subpruebas, lo que limita la conclusión definitiva de los hallazgos.

Fuente: Elaboración propia, Basado en [6]

3.1.4. MODELOS PREDICTIVOS DE LA DESERCIÓN ESTUDIANTIL EN UNA UNIVERSIDAD PRIVADA PERUANA

Tabla 4. *Modelos predictivos de la deserción estudiantil en una universidad privada peruana*

Aspecto	Detalles
Objetivos del Artículo	Determinar cómo el uso de modelos predictivos en asignaturas críticas contribuye a identificar a estudiantes en riesgo de deserción.
Herramientas y Técnicas	Uso de la metodología CRISP y técnicas de minería de datos para diseñar siete modelos predictivos aplicados en siete cursos críticos.
	Análisis de patrones de rendimiento académico y aplicación del método de árboles de clasificación y regresión (*Classification and Regression Trees*).
Limitaciones Vs. Proyecto	El enfoque se limita a una universidad privada peruana y a cursos específicos, lo que podría limitar la generalización de los hallazgos a otros contextos.
	Los resultados son específicos para cursos con altos índices de reprobación, lo que puede no ser directamente aplicable a tu proyecto si no se enfrentan problemas similares.

Fuente: Elaboración propia, Basado en [10]

3.1.5. LA EDUCACIÓN SUPERIOR EN COLOMBIA: UNA MIRADA A LOS CONCEPTOS DE CALIDAD Y EVALUACIÓN

Tabla 5. *La educación superior en Colombia: Una mirada a los conceptos de calidad y evaluación*

Aspecto	Detalles
Objetivos del Artículo	Explorar el concepto de calidad en la educación superior en Colombia y su relación con las políticas internacionales sobre eficiencia y eficacia educativa.
	Analizar el rol de las pruebas Saber Pro como instrumentos de evaluación en la educación superior colombiana.
Herramientas y Técnicas	Revisión documental sobre el surgimiento y desarrollo del concepto de calidad en la educación superior en Colombia .
	Análisis del marco legal y reglamentario de las pruebas Saber Pro y su impacto en la evaluación de la calidad educativa.
Limitaciones Vs. Proyecto	El enfoque del artículo es más teórico y de revisión, lo que puede limitar su aplicabilidad directa en un contexto de investigación empírica como tu proyecto.
	La concentración en políticas y marcos regulatorios específicos de Colombia puede no reflejar completamente las dinámicas y necesidades de otras regiones o contextos educativos.

Fuente: Elaboración propia, Basado en [11]

3.1.6. APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO

Tabla 6. *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario*

Aspecto	Detalles
Objetivos del Documento	Aplicar cada etapa de la metodología CRISP-DM en datos académicos de la Universidad Carlos III de Madrid.
	Demostrar la eficacia y simplicidad de la metodología CRISP-DM y su utilidad en obtener conclusiones y hacer predicciones a partir de datos.
Herramientas y Técnicas	Enfocado en minería de datos y el uso de CRISP-DM para analizar datos académicos.
	Manejo de datos organizados y limpios para identificar patrones válidos y novedosos mediante diversas tareas y técnicas de minería de datos.
Limitaciones Vs. Proyecto	El proyecto se enfoca en una institución y un conjunto de datos específicos, lo que puede limitar la aplicabilidad de sus hallazgos en contextos más amplios o diferentes.
	El documento se concentra en la teoría y la metodología, sin un enfoque en aplicaciones prácticas específicas que puedan ser directamente relevantes para tu proyecto.

Fuente: Elaboración propia, Basado en [12]

3.1.7. PISA 2015: RESULTADOS CLAVE

Tabla 7. *PISA 2015: Resultados Clave*

Aspecto	Detalles
Objetivos del Documento	Evaluar la calidad, equidad y eficiencia de los sistemas educativos a nivel mundial.
	Identificar las características de los sistemas educativos de alto rendimiento y proporcionar datos para políticas educativas eficaces.
Herramientas y Técnicas	Evaluación de estudiantes de 15 años en 72 países utilizando pruebas en ciencia, lectura y matemáticas, así como resolución colaborativa de problemas.
	Análisis de datos de aproximadamente 540,000 estudiantes para identificar tendencias y competencias.
Limitaciones Vs. Proyecto	PISA se enfoca en una evaluación general de habilidades y conocimientos, no específicamente en el contexto colombiano o en pruebas específicas como las SABER PRO
	Los resultados y análisis de PISA son a nivel macro y pueden no reflejar las dinámicas específicas de las instituciones educativas particulares o regiones específicas.

Fuente: Elaboración propia, Basado en [13]

3.1.8. RESULTADOS DE PISA 2018 EN COLOMBIA

Tabla 8. *Resultados de PISA 2018 en Colombia*

Aspecto	Detalles
Objetivos del Documento	Evaluar el conocimiento y las competencias fundamentales de los estudiantes de 15 años en Colombia en áreas clave como lectura, matemáticas, ciencias y competencia global.
Herramientas y Técnicas	Uso de pruebas computarizadas y evaluaciones adaptativas multifásicas.
	Evaluación de aproximadamente 600,000 estudiantes en 79 países, incluyendo 7,522 estudiantes de Colombia.
Limitaciones Vs. Proyecto	Los resultados se centran en un grupo de edad específico (15 años) y pueden no reflejar directamente el desempeño en pruebas específicas como las SABER PRO.
	Enfoque en una evaluación general de habilidades y conocimientos, no centrado específicamente en el contexto educativo colombiano o en pruebas específicas como las SABER PRO.

Fuente: Elaboración propia, Basado en [14]

3.1.9. UN MODELO ANALÍTICO PARA LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE ESTUDIANTES DE INGENIERÍA

Tabla 9. *Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería*

Aspecto	Detalles
Objetivos del Documento	Desarrollar un modelo predictivo para prevenir la eliminación académica por bajo rendimiento en estudiantes de ingeniería y ciencias de la Universidad de Chile.
	Enfocarse en la predicción temprana de estudiantes en riesgo de reprobar cursos por segunda vez, lo que conduce a la eliminación del programa.
Herramientas y Técnicas	Uso de learning analytics y minería de datos educativos para analizar datos académicos y predecir el rendimiento.
	Aplicación de regresión logística y metodologías de selección de atributos para construir el modelo predictivo.
Limitaciones Vs. Proyecto	El estudio se centra en un contexto específico (estudiantes de ingeniería y ciencias de la Universidad de Chile), lo que puede limitar la aplicabilidad directa de sus hallazgos a tu contexto.
	Se enfoca en prevenir la eliminación académica por bajo rendimiento, que puede ser solo una de las múltiples dimensiones de rendimiento académico consideradas en tu tesis.

Fuente: Elaboración propia, Basado en [15]

3.1.10. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN EL RENDIMIENTO ACADÉMICO Y DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DE CUENCA

Tabla 10. *Aplicación de técnicas de minería de datos en el rendimiento académico y deserción estudiantil en la Universidad de Cuenca*

Aspecto	Detalles
Objetivos del Trabajo	Aplicar técnicas de minería de datos para analizar el rendimiento académico y la deserción estudiantil.
	Desarrollar modelos para predecir la deserción estudiantil, el fracaso por ciclo de estudios y por asignatura.
Herramientas y Técnicas	Uso de la metodología CRISP-DM para el desarrollo de modelos de clasificación y predicción.
	Análisis de datos recolectados por la Universidad de Cuenca para desarrollar modelos predictivos.
Limitaciones Vs. Proyecto	El estudio se enfoca en una sola institución (Universidad de Cuenca), lo que puede limitar la generalización de los hallazgos.
	Se centra en el contexto ecuatoriano, lo que podría requerir adaptaciones para aplicar los hallazgos en el contexto colombiano de tu tesis.

Fuente: Elaboración propia, Basado en [16]

3.1.11. INDUCTION OF DECISION TREES

Tabla 11. *Induction of Decision Trees*

Aspecto	Detalles
Objetivos del Artículo	Resumir un enfoque para sintetizar árboles de decisión utilizados en una variedad de sistemas, incluido el sistema ID3.
Herramientas y Técnicas	Uso de la inducción para formar árboles de decisión a partir de ejemplos, con énfasis en el manejo de información ruidosa o incompleta.
	Exploración de estrategias de aprendizaje subyacentes, representación del conocimiento adquirido y aplicación en diferentes dominios.
Limitaciones Vs. Proyecto	Enfrenta desafíos con datos reales que pueden ser ruidosos o tener valores de atributos desconocidos, lo que puede afectar la precisión y aplicabilidad de los modelos.
	Riesgo de generar árboles de decisión que pueden ser difíciles de interpretar debido a la agrupación de valores de atributos no relacionados y múltiples pruebas en el mismo atributo.

Fuente: Elaboración propia, Basado en *[17]*

3.1.12. Evaluación de Modelos para un Sistema Recomendado para Cursos de Educación Superior Usando Datos del Examen de Admisión Colombiano

Tabla 12. *Evaluación de Modelos para un Sistema Recomendado para Cursos de Educación Superior Usando Datos del Examen de Admisión Colombiano*

Aspecto	Detalles
Objetivos del Artículo	- Desarrollar un sistema de recomendación de cursos de educación superior utilizando datos de resultados de exámenes estatales en Colombia. - Comparar cinco modelos de recomendación diferentes mediante el análisis de precisión, recall, curvas ROC y error de predicción.
Herramientas y Técnicas	- Uso de técnicas de filtrado colaborativo y el modelo de cursos populares. - Evaluación de modelos con el marco "recommenderlab" en R para comparar diferentes métodos de recomendación.
Limitaciones Vs. Proyecto	- El contexto del trabajo es específico de Colombia, pero se espera que los hallazgos puedan aplicarse a otros contextos con exámenes estatales equivalentes. - Se planea implementar el sistema de recomendación como una aplicación móvil para estudiantes

Fuente: Elaboración propia, Basado en *[17]*

3.2. REFERENTES INTERNACIONALES

El apoyo de este proyecto para una retroalimentación de documentos nacionales e internacionales se realiza con fuentes GOOGLE ACADÉMICO y SCOPUS.

El proyecto de Víctor Galán Cortana “Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario”, realiza la aplicación de la metodología CRISP-DM usando cada fase con la información académica de la universidad para la realización de predicciones y conclusiones sobre el conjunto de datos, donde a través del proyecto se puede evidenciar que gracias a esta metodología se tiene la posibilidad de explorar los datos y conocer patrones de comportamiento que se acceden cuando se realiza una buena metodología. [12]

En el estudio realizado en chile “Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería” en año 2015, con base en la información de los estudiantes de ingeniería, se logra predecir con un modelo el rendimiento de los estudiantes, basados en variables de caracterización de los estudiantes apoyados en herramientas sencillas de Learning Analytics, estos modelos predictivos dispuestos para el uso en la toma de decisiones administrativas y de docencia para la intervención de apoyos curriculares, apalancando procesos de aprendizaje más eficientes. [15]

Dentro de las universidades privadas y públicas el manejo de la deserción es clave, generar acciones que contrarresten este fenómeno negativo que perjudica a la universidad, a los estudiantes y a la sociedad. El uso de modelos predictivos que con los datos contribuye a identificar estudiantes con riesgo de deserción, en el 2018 una investigación en una universidad peruana privada genera un modelo apoyado en 10 variables que ellos definen como descriptivas del fenómeno, que al final con la construcción de 7 modelos predictivos ayudaron a generar estrategias de tutorías dirigidas a los estudiantes con una probabilidad alta de perder el curso. El resultado del uso de los modelos generó que entre un 25 y 40 por ciento de estos más de aprobación de los cursos en comparación con años anteriores que no se usó el modelo predictivo. [10]

El proyecto de “Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°” de la Universidad de Nariño en Pasto es un buen referente para este proyecto, por que aborda la metodología CRISP-DM para el manejo de los datos y se complementa con un modelo de clasificación basado en arboles de decisión en donde como resultado del proyecto nos muestra el éxito de la metodología, que con base en estos resultados se determinan factores que en estudiantes de undécimo de bachillerato que presentaron las pruebas saber 11 en los años 2015 y 2016, su estrato socio económico, la jornada de estudio y la edad son variables que ayudan a describir los resultados en las pruebas saber 11 de los estudiantes. [18]

En 2017 se llevó a cabo una investigación para determinar el desempeño académico *“Analyzing undergraduate students' performance using educational data mining”*, con las notas de 4 años de los estudiantes de licenciatura en una universidad búlgara. Se generó un modelo con 10.330 datos distribuidos en 20 atributos de estudiantes que busca apoyar a estudiantes en riesgo durante su ciclo de aprendizaje, generando indicadores que generen alertas tempranas y ayuden a un mejor desempeño en su aprendizaje. Uno de los resultados de esta investigación, muestra que es posible predecir el rendimiento de los estudiantes usando únicamente las notas, sin el uso de datos socioeconómicos. [8]

En el Ecuador en la universidad de Cuenca se realizó un modelo predictivo, con base en notas y en la situación socioeconómica de los estudiantes, donde con la metodología CRISP-DM y el uso de 3 algoritmos se pudo predecir con una precisión del 70% si los estudiantes pueden llegar a perder una materia o llegaran a terminar su carrera. Apoyados en este modelo la universidad puede tomar acciones focalizadas que contrarrestan la deserción de los estudiantes. [16]

La universidad nacional de Colombia aplica un modelo de deserción en donde identifica que el algoritmo de regresión logística no obtiene los mejores resultados para el conjunto de datos que se manejó; adicionalmente también muestra como por sedes entrega información de deserción, lo que permite focalización de esfuerzos por sedes, programas y estudiantes. Dentro de los resultados entregados también se observa que los tres primeros semestres presentan la más alta deserción dentro de los datos analizados. El programa que muestra la mayor deserción con un 44% es la Ingeniería de sistemas, donde concluyen que son las bases de secundaria con respecto al área de matemáticas las que llevan a los estudiantes al abandono de la carrera en el tercer semestre.

4. PLANTEAMIENTO DEL PROBLEMA

El diagnóstico de la situación educativa en los últimos seis años ha revelado una preocupante disminución de los puntajes que obtiene la universidad en Saber Pro, pasando de 147 en 2016 a 136 en 2022. Esta reducción representa una pérdida de 11 puntos en este período. Además, durante estos años, la institución ha mantenido consistentemente resultados por debajo del promedio nacional, estando 3 puntos por debajo en 2016 y aumentando esta brecha a 11 puntos en 2022.

Ilustración 3. *Comparativo de 6 años universidad vs Colombia*
Fuente: Elaboración propia

Esta realidad plantea un problema de gran relevancia para la institución universitaria en términos de acreditación y calidad de programas. Durante las evaluaciones de pares académicos, las estrategias implementadas por la universidad para abordar esta disminución son un tema recurrente de discusión. Los puntajes obtenidos en las pruebas son una medida inmediata sobre la calidad académica de la universidad.

A lo largo de estos años, se han formulado diversas hipótesis para abordar las causas de la disminución de los puntajes. Los informes del ICFES han servido como referencia para el análisis anual de la población y su comportamiento. Sin embargo, aún no se ha implementado un proceso de análisis que permita evaluar y comparar estos seis años con datos estandarizados, principalmente debido al proceso de cambio de sistema que atraviesa la institución. Durante esta transición, la depuración de datos de los estudiantes y otros datos cruciales, como su caracterización, está en curso y se está estabilizando para la población objeto de estudio.

Aunque las estrategias previamente utilizadas han logrado ralentizar la disminución de los puntajes en el último año, la falta de datos confiables y la ausencia de un modelo que utilice la data histórica como base para comprender el comportamiento han impedido un avance significativo y efectivo en este problema.

La presente investigación propone contar con bases de datos validadas y depuradas que contengan variables de caracterización de los estudiantes. Estas bases integrarán datos de los últimos seis años, permitiendo una visión clara y objetiva de cada segmento y

micro segmento de los estudiantes que han presentado las pruebas. Con este aprendizaje, se podrá determinar grupos poblacionales que están próximos a presentar las pruebas y sobre los cuales la universidad debe enfocar mayores esfuerzos, brindando herramientas y apoyo específico para mejorar el puntaje de cada estudiante.

Es crucial destacar que estas bases de datos incorporarán información tanto de la institución universitaria como del ICFES. Al concluir esta investigación, la institución universitaria contará con un modelo predictivo que podrá determinar, en un buen porcentaje la población de estudiantes inscritos para el año en curso. Basándose en los comportamientos de los últimos seis años, se podrán identificar grupos poblacionales que necesitan un apoyo específico para elevar su puntaje en las pruebas Saber Pro y, por ende, mejorar la calidad académica.

5. OBJETIVOS

5.1. OBJETIVO GENERAL

Desarrollar un modelo descriptivo de 2016 a 2022 y predictivo con apoyo de estos datos, que mejore el rendimiento en los resultados en SABER PRO.

5.2. OBJETIVOS ESPECÍFICOS

Validar exhaustivamente la data de los últimos 6 años proveniente de fuentes confiables como ICFES y la institución universitaria. Esto asegurará la calidad y confiabilidad de la información con el que se pueda hacer análisis y modelado.

Consolidar minuciosamente la información recopilada, organizándola por variables relevantes. Se utilizarán herramientas avanzadas de análisis de datos para identificar patrones y relaciones significativas que ayudarán a comprender mejor el comportamiento de los estudiantes en los resultados de SABER PRO.

El desarrollo de este trabajo busca proporcionar a la institución universitaria una herramienta que consolida la información histórica de los últimos 6 años de las pruebas saber pro y generar un modelo que prediga el comportamiento de los nuevos estudiantes que presentaran esta prueba y su resultado.

Desarrollar un modelo predictivo basado en la caracterización de los estudiantes que hayan cursado más del 70% de su malla curricular. El objetivo es proporcionar una base sólida para la implementación de estrategias enfocadas en mejorar el rendimiento de estos estudiantes durante la presentación de las pruebas SABER PRO. Este modelo deberá ser preciso y útil para la toma de decisiones educativas.

6. METODOLOGÍA

Dentro de las metodologías utilizadas para la minería de datos, la selección de un modelo apropiado es crucial para el éxito de la investigación. En la ilustración 4, se identifican las principales metodologías empleadas en minería de datos o ciencia de datos, destacando tres de las más utilizadas y recomendadas para investigaciones exitosas: SEMMA, KDD y CRISP-DM.

La CRISP-DM (Cross Industry Standard Process for Data Mining) es reconocida como una de las metodologías más exitosas, propuesta por IBM para el manejo de proyectos de minería de datos. Su proceso estándar organiza las fases de manera secuencial, permitiendo la revisión continua del desarrollo del proyecto. Esta característica resulta fundamental para corregir desviaciones a tiempo, garantizando que el proyecto se alinee con sus objetivos.

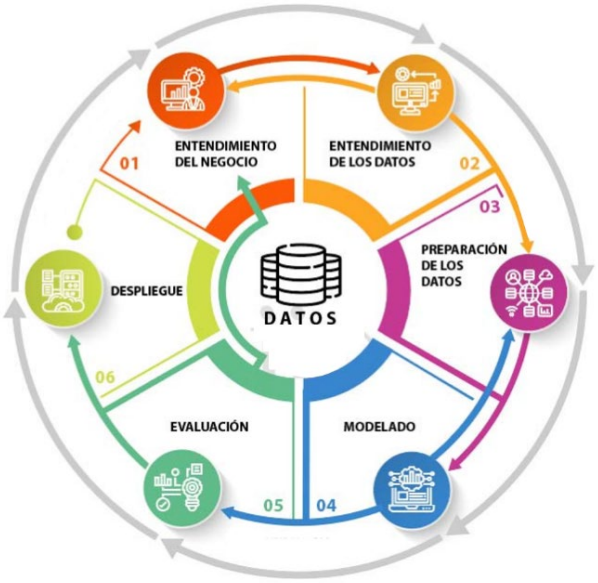


Ilustración 4. *Esquema del ciclo CRISP-DM estándar,*
Fuente: Elaboración propia, Tomado de [19]

Además, la metodología CRISP-DM se desglosa en las siguientes fases, enriquecidas con elementos clave del documento proporcionado:

6.1. ENTENDIMIENTO DEL NEGOCIO

Esta fase se erige como la base del proyecto, derivando objetivos clave del conocimiento del negocio. Un entendimiento profundo del negocio facilita la evolución exitosa del proyecto, minimizando reprocesos y reformulaciones de objetivos finales.

6.2. ENTENDIMIENTO DE LOS DATOS

Tras definir objetivos, se procede a recolectar la materia prima de datos, identificando la información alineada con los objetivos. La calidad de los datos se evalúa minuciosamente, considerando variables que puedan sobredimensionar el modelo o la falta de información necesaria.

6.3. PREPARACIÓN DE LOS DATOS

En esta fase se construye el conjunto de variables y datos para la creación del modelo. Se realiza la limpieza de datos, estandarización de formatos, integración de fuentes y aplicación de técnicas adicionales para garantizar un modelo adecuado.

6.4. MODELADO

El modelado implica la selección de técnicas que se alineen con los objetivos del proyecto, poniendo a prueba la idoneidad de los datos y gestionando tiempos de manera efectiva.

6.5. EVALUACIÓN

Con base en el cálculo de diversas medidas de diagnóstico, el científico de datos valida la idoneidad del modelo y examina la fiabilidad de los resultados, realizando ajustes para mejorar el proceso.

6.6. DESPLIEGUE

Esta fase implica el despliegue del modelo en un entorno de producción o pruebas, demostrando el éxito del proyecto y transformando el conocimiento adquirido en recomendaciones y análisis respaldados por datos.

7. DESARROLLO DE LA PROPUESTA

7.1. ENTENDIMIENTO DEL NEGOCIO

Este proyecto tiene lugar en una reconocida institución universitaria con más de cuatro décadas de experiencia en la educación superior. Esta universidad se caracteriza por su enfoque inclusivo, acogiendo a más de 50.000 estudiantes, predominantemente de estratos socioeconómicos 1, 2 y 3. Ofrece dos modalidades educativas: virtual, que abarca el 80% de su población estudiantil, y presencial, que comprende el 20% restante. Su alcance se extiende a más de 900 municipios colombianos y presenta una diversidad de género con una distribución de 58% mujeres y 42% hombres. Actualmente, la institución ofrece 98 programas académicos y opera desde dos sedes principales: Medellín y Bogotá. Además, está en pleno proceso de obtención de una alta acreditación institucional, contando ya con 9 programas acreditados.

Dada la urgente necesidad de mejorar su rendimiento en las pruebas Saber Pro, donde desde 2016 hasta 2022 los resultados han estado consistentemente por debajo del promedio nacional (con una brecha de 12 puntos), la universidad busca proactivamente desarrollar estrategias a corto, mediano y largo plazo. Es crucial abordar las competencias genéricas, ya que la universidad, al ser inclusiva, acoge a estudiantes procedentes de instituciones de educación media que no siempre han consolidado estas habilidades esenciales. Esta situación exige un esfuerzo adicional por parte de la universidad para compensar estas brechas y fortalecer dichas competencias. Por lo tanto, el objetivo principal del modelo es identificar y predecir aquellos estudiantes que, según su perfil, se sitúan en niveles bajos de competencia. Con esta información, la institución podrá diseñar intervenciones específicas y preventivas, garantizando una preparación adecuada antes de que los estudiantes se presenten a las pruebas nacionales.

7.2. COMPRENSIÓN DE LOS DATOS

7.2.1. RECOLECCIÓN DE LOS DATOS

Los datos a los que se tiene acceso provienen de tres fuentes principales, una pública del ICFES, otras dos de acceso privado de la IES que le suministra el ICFES y una de la IES con datos sensibles de los estudiantes.

FUENTE	TIPO	DESCRIPCIÓN	PRIVACIDAD
ICFES SABER PRO (Una a na cada IES)	Base de datos	Esta base de uso publico contiene todas la universidades que participaron con estrudiantes en la presentación de las pruebas SABER PRO del 2016 al 2022.	Publica
ICFES (Uno a uno Estudiantes de la IES)	Base de datos	Esta base es de uno exclusivo de la institución universitario y muesta cada uno de los estudiantes y su desempeño en las pruebas SABER PRO, en las competencias generaicas y especificas.	Privada
IES (Uno a uno Estudiantes de la IES)	Base de datos	Esta base es de uno exclusivo de la institución universitario y muesta cada uno de los estudiantes, con las caracteristicas de cada estudiante, con base en sus procesos de admisión, matricula y desempeño academico.	Privada

Ilustración 5. *Resumen de bases de datos*
Fuente: Elaboración propia

Con base en la información obtenida se realizará la integración de las tres diferentes fuentes para obtener el entendimiento de los datos e ir generando la interacción entre los mismos con el objetivo de dar a conocer los datos e identificar cuales variables son las necesarias para el modelo descriptivo y predictivo.

Ilustración 6. *Flujo de Fuente de Datos*
Fuente: Elaboración propia

7.2.2.1. MODELAMIENTO EN POWER BI

En esta fase se usaron archivos .csv dentro de Excel, para segmentar las bases de acuerdo con las necesidades. Una de las necesidades principales es la construcción de un histórico, que permita realizar comparaciones por años y ver la evolución en el tiempo de los puntajes globales. Para esto se tiene en cuenta que, a partir de 2016 las pruebas Saber Pro presentan un nuevo esquema que no es comparable con los anteriores años. Con base en lo anterior se construye la base de 2016 en adelante.

La primera base de ICFES pública se extrae con los campos de las universidades y el consolidado nacional, que permiten realizar comparaciones de promedio global, carácter académico, sector, IES acreditadas, años y cantidad de evaluados.

Con la base que suministra el ICFES de modo privado a la IES, se valida la recolección de 2016 a 2022 que contiene el uno a uno de los estudiantes que presentan las pruebas, consolidando la historia de los evaluados en el tiempo. De esta base es importante segmentar los diferentes enfoques que presenta la información y estos son: Resultados por competencias y porcentaje de respuestas incorrectas.

Se identifica la necesidad de extraer unas variables adicionales de las fuentes propias de la IES teniendo en cuenta que las bases del ICFES no presentan este nivel de detalle que aporta a la IES, segmentos propios de ella y que permitirán identificar necesidades puntuales con base en estos segmentos, como: facultad, sede, modalidad, nivel de formación, estado civil, etc.

Se identifica que, para la generación de un modelo descriptivo y predictivo, es importante la creación de dos modelos para la interacción dinámica de los datos, por lo que se toma como herramienta de modelado y descripción de los datos Power BI, construyendo así dos modelos internos que permiten generar también un modelo tabular final que será usado en el modelo predictivo.

Se construye una base tabular para modelo descriptivo, y se realiza un comparativo con la demás IES.

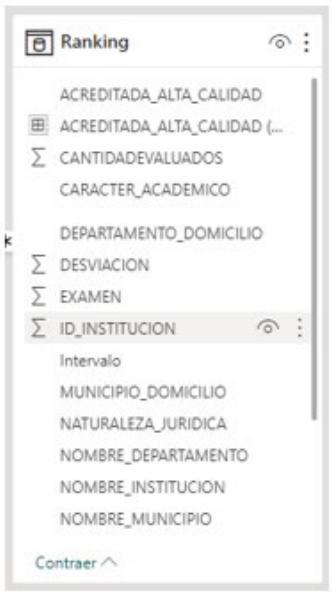


Ilustración 7. *Vista del modelo tabular*
Fuente: Elaboración propia

Luego se genera un modelo estrella para segmentar los resultados con la información obtenida del ICFES y de la IES.

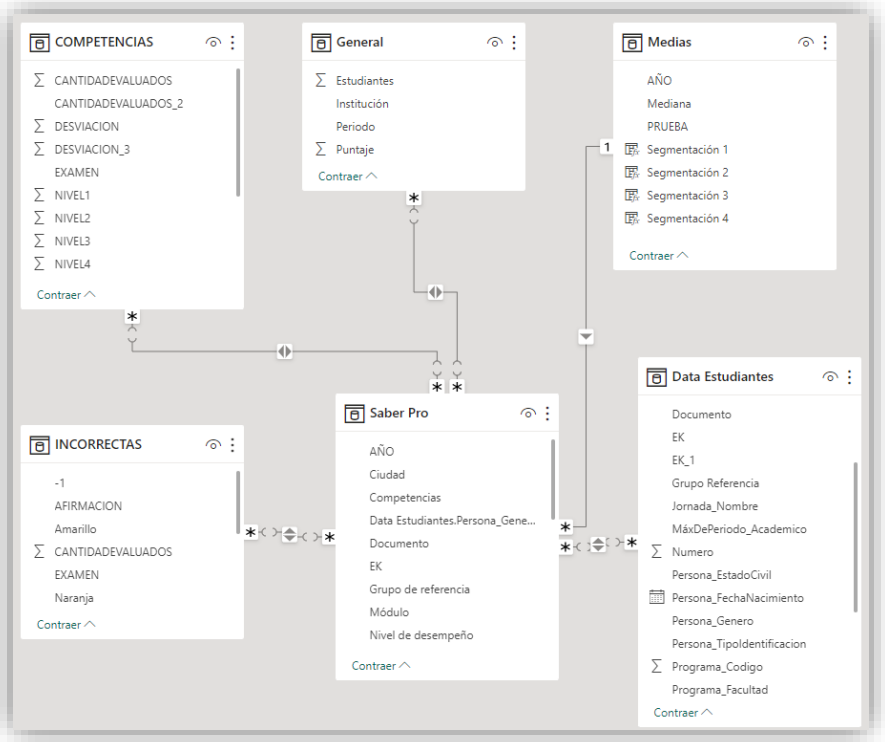


Ilustración 8. *Modelo estrella para modelo descriptivo*
Fuente: Elaboración propia

Ya construidos los dos cubos de información en Power BI realizamos el modelo descriptivo de los datos, para con base en ellos comenzar todo el proceso de conocer las necesidades de la IES, el cual se da en tres segmentos: externo, interno y enfoque, y que proporcionará a la universidad una mirada clara de lo que ha sido su evolución de 2016 a 2022.

7.2.2.2. CREACIÓN DE LA VISUALIZACIÓN DE DATOS

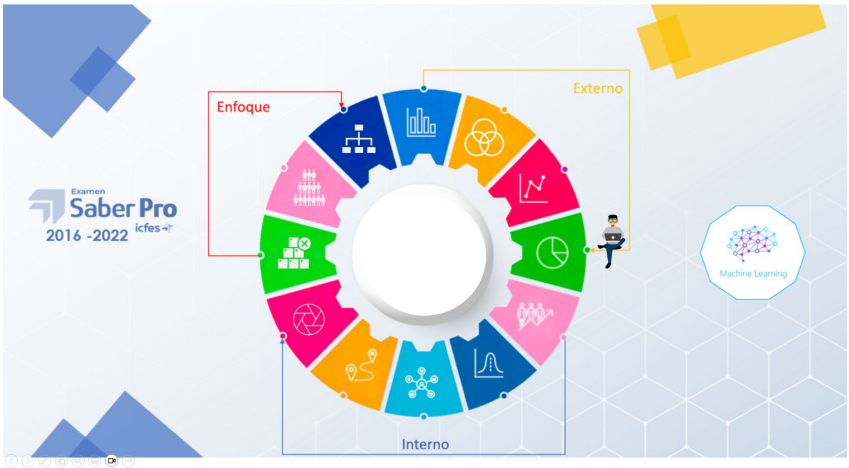


Ilustración 9. *Portada modelo descriptivo*
Fuente: Elaboración propia

En la parte del segmento externo, se muestra cómo está la IES sujeta de esta investigación frente al panorama nacional y al resto de universidades de Colombia.

Ilustración 10. *Modelo descriptivo –Externo*
Fuente: Elaboración propia

En esta parte del modelo descriptivo se muestra el contexto de la IES con respecto a todas la IES para ser comparadas, realizando un raking con base en el puntaje y la cantidad de evaluados de cada IES desde el 2016 al 2022, en otra parte nos permite ver cómo es su evolución en el tiempo con las IES que están más cercanas a ser competencia directa de la IES, de acuerdo con las diferentes modalidades que ofrece la universidad. Luego en un contexto de ver la IES con el promedio nacional y la brecha año tras año como se ha presentado, se genera una previsión partiendo de los datos históricos de Colombia y la IES de 2016 a 2022 lo que nos genera los posibles resultados de 2023 y 2024. Con un intervalo de confianza de 95%.

En resumen, esta parte nos muestra como durante los últimos 7 años la universidad se ha posicionado por debajo del promedio nacional 7 puntos. También se puede ver que la universidad está dentro de las tres IES que tiene mayor participación en la cantidad de los estudiantes que participan de las pruebas Saber Pro con un promedio de 8.932 estudiantes por año. La previsión nos genera una alerta para los dos próximos años donde se ve que la universidad seguirá por debajo del promedio nacional, con una brecha de 12 puntos y una caída en los resultados a 132 en el puntaje global de la IES. Lo que demuestra la necesidad urgente de generar acciones que ayuden a mejorar estos resultados.

En el segmento interno permite con base en las variables de caracterización de los estudiantes identificar los segmentos en los que la universidad debe realizar un mayor esfuerzo para mejorar los resultados.

Ilustración 11. *Modelo descriptivo –Interno*
Fuente: Elaboración propia

En esta parte se describe el comportamiento de los puntajes agrupados por modalidad, facultad, tipo de estudiante y generó, que permiten ver la evolución en los últimos 7 años de cada segmento, con un valor que muestra las variaciones de un año a otro lo que permite encontrar múltiples combinaciones e ir identificando grupos focales sobres los que se tienen fortalezas o debilidades.

En otro análisis se ve la distribución de los estudiantes mostrando la concentración de la distribución normal con el 70% de los datos, lo cual permite ver la evolución positiva o negativa en los máximos y mínimos de los puntajes individuales de cada uno de los estudiantes.

Luego el análisis se concentra en cada una de las competencias genéricas realizando comparativos con el promedio nacional, la distribución normal de cada competencia por estudiantes, clasificándolos por niveles dentro de una escala de percentiles, adicionalmente, se realiza un análisis detallado de cada competencia clasificando 4 características puntuales de los estudiantes (Modalidad, Facultad, Tipo de Estudiantes y Género).

En la parte de enfoque se busca proporcionar a la IES, los grupos focales sobre los cuales se requiere tomar alguna acción puntual.

Ilustración 12. *Modelo descriptivo –Enfoque*
Fuente: Elaboración propia

En una primera parte se muestran los datos segmentados por los programas con mayor participación en las pruebas acumuladas de 2016 a 2022 o año a año, igual que por las variables de modalidad, tipo de estudiante, facultad y género. Las bases del ICFES proporcionan el porcentaje de respuestas incorrectas con base en tres competencias genéricas (Ciudadanas, Lectura Crítica y Razonamiento Cuantitativo), de 2018 a 2022, lo que permite a la IES fortalecer los temas con base en las afirmaciones donde los porcentajes de incorrectas son mayores. En el último tablero creado en esta parte, el

usuario puede revisar diferentes rutas para el conocimiento de puntos críticos o grupos focalizados para identificar donde se puede apoyar con el desarrollo de estrategias focalizadas.

7.2.2.3. ANÁLISIS DE ELEMENTOS INFLUYENTES EN POWER BI

Al finalizar el proceso descriptivo, direccionado a la toma de decisiones académicas, administrativas y estratégicas, se usan las herramientas de *machine learning* que están integradas a Power BI, apoyadas en toda la infraestructura que proporciona Azure. Para esto se usó la visualización de Elementos Influyentes Claves. Donde se dio como objetivo el puntaje de las pruebas Saber Pro, y se proporcionó a la herramienta variables que se consideran pueden ser determinantes para el modelo como, modalidad, convenio, genero, facultad, nivel académico, programa, sede, ciudad de residencia, jornada, materias homologadas, materias perdidas, promedio de notas y edad. Con base en esta caracterización y el resultado de cada estudiante, el modelo fácilmente puede decirnos cuales de ellas son elementos claves al momento de aumentar o disminuir el puntaje.

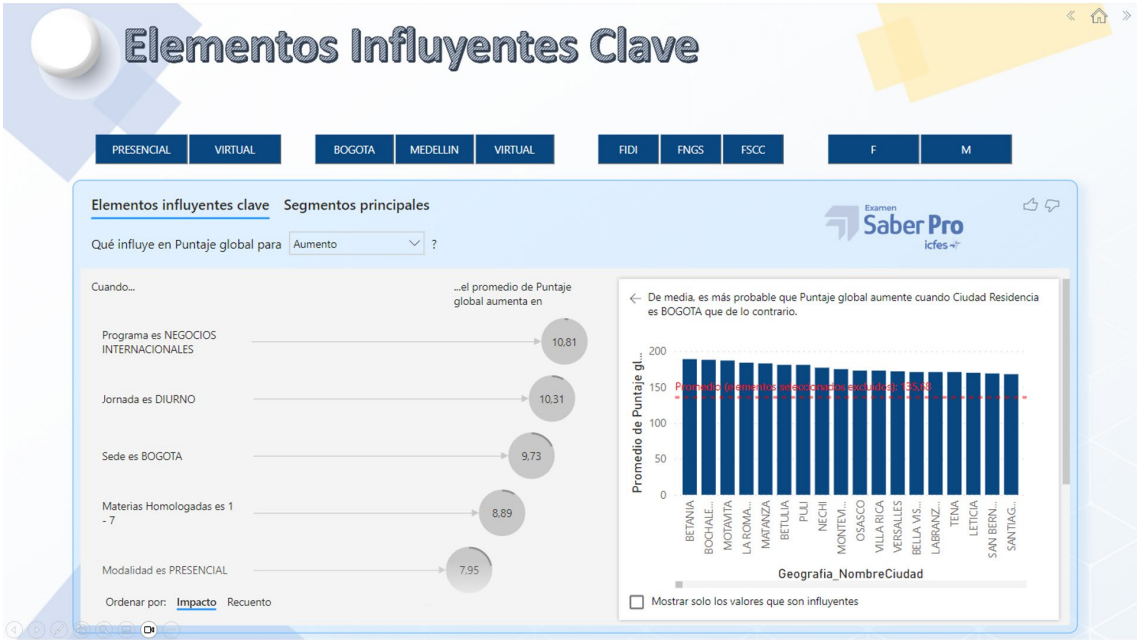


Ilustración 13. *Elementos Influyentes*
Fuente: Elaboración propia

Dentro de este mismo proceso, se tiene la capacidad también de identificar segmentos especiales con características que destacan y que son claves al momento de identificar donde se tienen las fortalezas o debilidades.

Ilustración 14. *Elementos Influyentes Clave 1*
Fuente: Elaboración propia

Estos segmentos nos muestran las características especiales y que son predominantes en una población importante de estudiantes y con base en la información histórica, determinar cómo solo ese segmento hubiera mostrado un mejor o menor comportamiento en los resultados finales de la universidad.

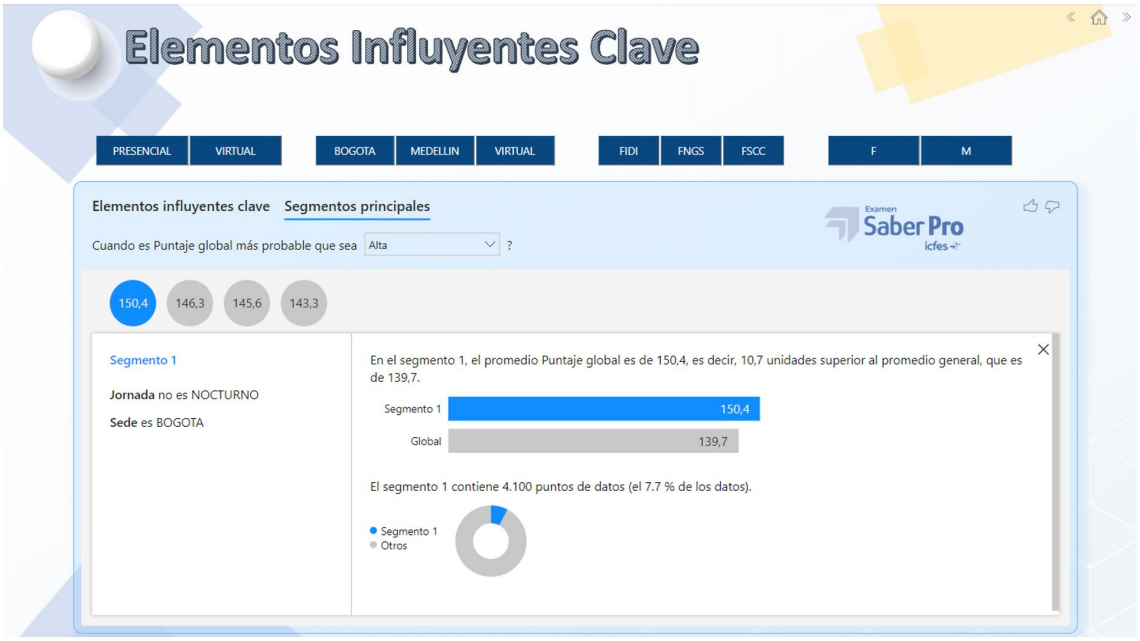


Ilustración 15. *Elementos Influyentes Clave 2*
Fuente: Elaboración propia

Gracias a la interacción entre los datos externos del ICFES y las bases propias de la IES que se realizó en Power BI, este proceso permite conocer muy bien y a profundidad las problemáticas de la IES y su evolución en los últimos 7 años; con el apoyo de la herramienta de *Machine Learning* se logra identificar variables que son influyentes y que pueden apoyar la creación del modelo predictivo, así que con este modelamiento de datos en Power BI se genera un modelo tabular con las variables que caracterizan a cada estudiante, lo que se convierte en un punto de partida interesante, porque ya se tienen varias fuentes agrupadas y evaluadas por un modelo adaptado para este fin.

Esta base se constituye con datos de 2016 a 2022 y con los evaluados en este periodo de tiempo, se tiene un total de 59.585 registro con tenidos en 29 variables.

7.2.2. DESCRIPCIÓN DE LOS DATOS

La descripción de los datos que se usaran en el modelo predictivo, luego de la consolidación y depuración de los datos obtenidos de las tres fuentes principales.

Tabla 12. *Descripción de variables para el modelo predictivo*

Variable	Descripción	Tipo de Datos	Longitud Max
PIDM	Identificador único de estudiante asignado por la universidad.	Numérico	10
Puntaje global	Calificación total en las pruebas SABER PRO.	Numérico	3
Percentil nacional global	Posición percentil del estudiante en los resultados nacionales del SABER PRO.	Numérico	2
Materias	Total, de asignaturas inscritas por el estudiante.	Numérico	3
Promedio	Media aritmética de calificaciones de la carrera universitaria.	Numérico	2
Materias Homologadas	Cantidad de créditos convalidados al ingresar a la universidad.	Numérico	3
Materias Perdidas	Total, de asignaturas reprobadas durante el programa académico.	Numérico	3
EK	Identificador asignado por el ICFES para las pruebas estudiantiles.	Alfanumérico	15
AÑO	Año en que se presentaron las pruebas.	Numérico	4
Programa_Codigo	Código oficial del programa asignado por el MEN.	Numérico	10
Programa_Facultad	Siglas representativas de cada facultad dentro de la universidad.	Alfabético	5

Variable	Descripción	Tipo de Datos	Longitud Max
Programa_Nombre	Denominación específica del programa académico.	Alfabético	50
Programa_NivelAcademico	Nivel académico del programa, asignado exclusivamente a pregrado.	Alfabético	10
Programa_Modalidad	Modalidad de impartición del programa: presencial o virtual.	Alfabético	10
Convenio_Homolog	Tipo de convenio de homologación: Sena o Regular.	Alfabético	10
Convenio_Tipo	Diferentes tipos de convenios educativos que maneja la universidad.	Alfabético	25
Jornada_Nombre	Denominación de la jornada educativa y horario correspondiente.	Alfabético	10
Sede_Nombre	Nombres de las sedes que ofrecen el programa en modalidad presencial.	Alfabético	10
CSU_Nombre	Nombre del centro de servicios universitarios asociado.	Alfanumérico	25
Geografia_NombreCiudad	Ciudad de residencia del estudiante.	Alfabético	10
Persona_TipoIdentificacion	Tipo de documento de identidad del estudiante.	Alfabético	3
Persona_EstadoCivil	Estado civil actual del estudiante.	Alfabético	10
Persona_Genero	Sexo biológico del estudiante.	Alfabético	8
Persona_FechaNacimiento	Fecha de nacimiento del estudiante.	Fecha	10
CSU_Region	Región educativa a la que pertenece el estudiante según su residencia.	Alfabético	25
Edad	Edad del estudiante en el momento de la investigación.	Numérico	2
Edad_2	Edad del estudiante en el momento de la presentación de la prueba.	Numérico	2

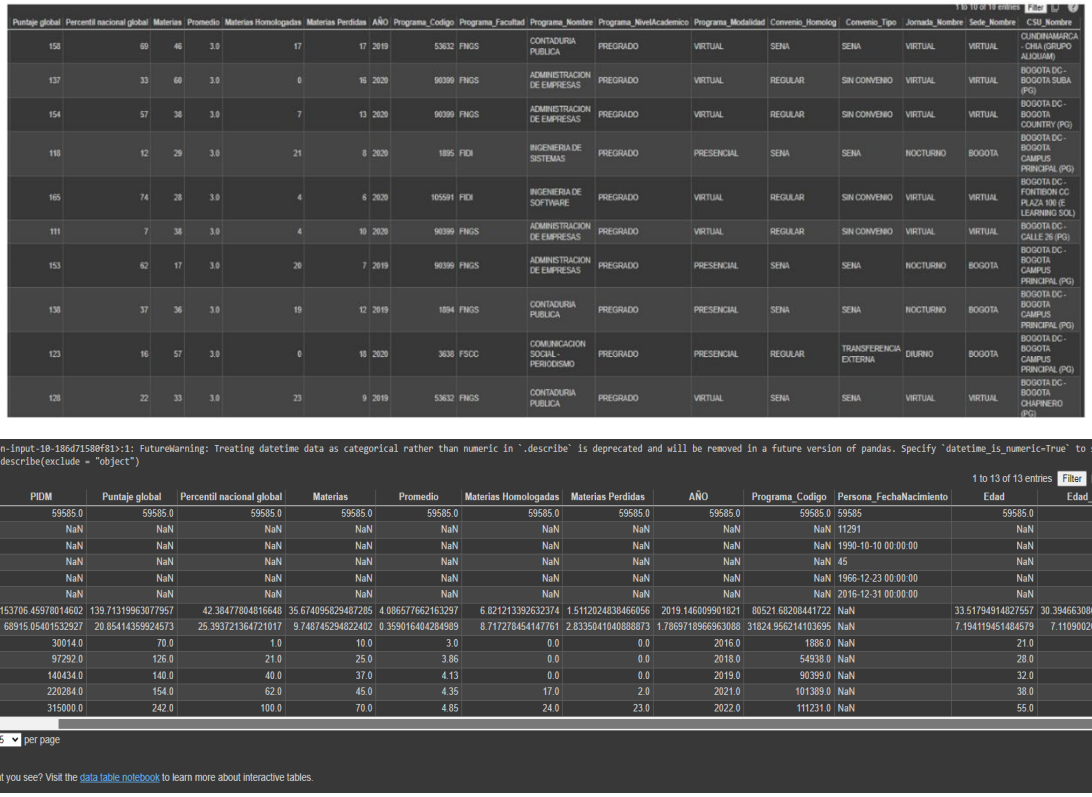
Fuente: Elaboración propia

7.2.3. EXPLORACIÓN DE LOS DATOS

Para el modelo predictivo se cargan los datos a la herramienta de Python a través de *Google Colab* que nos facilita el uso de múltiples librerías con un acceso más sencillo al momento de instalarlas, con ayuda de varias librerías de *Python*, se permite evaluar varios modelos y explorar diferentes simulaciones necesarias para llegar a encontrar el resultado esperado. Es acá donde se comienza a realizar la preparación de estos datos para ser incluidos al modelo predictivo.

Se cargan las librerías necesarias para este proceso en *Python* y, que permiten la transformación, validación, descripción, prueba y medición del resultado para el modelo.

Se importan los datos al notebook.



Puntaje global	Percentil nacional global	Materias	Promedio	Materias Homologadas	Materias Perdidas	AÑO	Programa_Codigo	Programa_Facultad	Programa_Nombre	Programa_NivelAcademico	Programa_Modalidad	Convenio_Homolog	Convenio_Tipo	Jornada_Nombre	Sede_Nombre	CSU_Nombre
158	69	46	3.0	17	17	2019	53632	FNGS	CONTADURIA PUBLICA	PREGRADO	VIRTUAL	SENA	SENA	VIRTUAL	VIRTUAL	CUNDINAMARCA - CHIA (GRUPO ALIQUAM)
137	33	60	3.0	0	16	2020	90399	FNGS	ADMINISTRACION DE EMPRESAS	PREGRADO	VIRTUAL	REGULAR	SIN CONVENIO	VIRTUAL	VIRTUAL	BOGOTA DC - BOGOTA SUBA (PG)
154	57	38	3.0	7	13	2020	90399	FNGS	ADMINISTRACION DE EMPRESAS	PREGRADO	VIRTUAL	REGULAR	SIN CONVENIO	VIRTUAL	VIRTUAL	BOGOTA DC - BOGOTA COUNTRY (PG)
118	12	29	3.0	21	8	2020	1895	FIDI	INGENIERIA DE SISTEMAS	PREGRADO	PRESENCIAL	SENA	SENA	NOCTURNO	BOGOTA	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)
165	74	28	3.0	4	6	2020	105591	FIDI	INGENIERIA DE SOFTWARE	PREGRADO	VIRTUAL	REGULAR	SIN CONVENIO	VIRTUAL	VIRTUAL	BOGOTA DC - FONTIBON CC PLAZA 100 (E LEARNING SOL)
111	7	38	3.0	4	10	2020	90399	FNGS	ADMINISTRACION DE EMPRESAS	PREGRADO	VIRTUAL	REGULAR	SIN CONVENIO	VIRTUAL	VIRTUAL	BOGOTA DC - CALLE 26 (PG)
153	62	17	3.0	20	7	2019	90399	FNGS	ADMINISTRACION DE EMPRESAS	PREGRADO	PRESENCIAL	SENA	SENA	NOCTURNO	BOGOTA	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)
138	37	36	3.0	19	12	2019	1894	FNGS	CONTADURIA PUBLICA	PREGRADO	PRESENCIAL	SENA	SENA	NOCTURNO	BOGOTA	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)
123	16	57	3.0	0	18	2020	3638	FSCC	COMUNICACION SOCIAL - PERIODISMO	PREGRADO	PRESENCIAL	REGULAR	TRANSFERENCIA EXTERNA	DIURNO	BOGOTA	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)
128	22	33	3.0	23	9	2019	53632	FNGS	CONTADURIA PUBLICA	PREGRADO	VIRTUAL	SENA	SENA	VIRTUAL	VIRTUAL	BOGOTA DC - BOGOTA CHAPINERO (PG)

```
ipython-input-10-186d71580f81>:1: FutureWarning: Treating datetime data as categorical rather than numeric in `.describe` is deprecated and will be removed in a future version of pandas. Specify `datetime_is_numeric=True` to silence
  data.describe(exclude = "object")
```

1 to 13 of 13 entries Filter

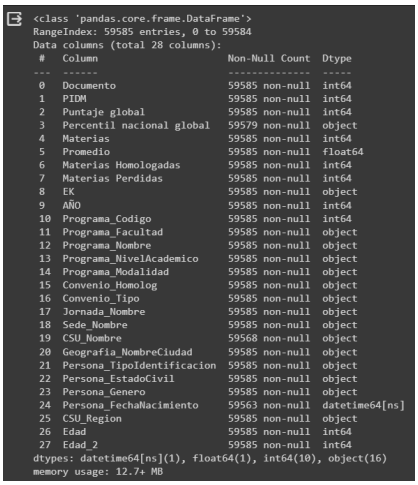
to	PIDM	Puntaje global	Percentil nacional global	Materias	Promedio	Materias Homologadas	Materias Perdidas	AÑO	Programa_Codigo	Persona_FechaNacimiento	Edad	Edad_2
9585.0	59585.0	59585.0	59585.0	59585.0	59585.0	59585.0	59585.0	59585.0	59585.0	59585	59585.0	59585.0
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	11291	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1990-10-10 00:00:00	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	45	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1966-12-23 00:00:00	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2016-12-31 00:00:00	NaN	NaN
34731	153706.45978014602	139.71319963077957	42.38477804816648	35.674095829487285	4.086577662163297	6.821213392632374	1.5112024838466056	2019.146009901821	80521.68208441722	NaN	33.51794914827557	30.3946630863472
79545	68915.05401532927	20.85414359924573	25.393721364721017	9.748745294822402	0.359016404284989	8.717278454147761	2.8335041040888873	1.7869718966963088	31824.956214103695	NaN	7.194119451484579	7.110900203172
2312.0	30014.0	70.0	1.0	10.0	3.0	0.0	0.0	2016.0	1886.0	NaN	21.0	15.0
9097.0	97292.0	126.0	21.0	25.0	3.86	0.0	0.0	2018.0	54938.0	NaN	28.0	25.0
9408.0	140434.0	140.0	40.0	37.0	4.13	0.0	0.0	2019.0	90399.0	NaN	32.0	29.0
8402.0	220284.0	154.0	62.0	45.0	4.35	17.0	2.0	2021.0	101389.0	NaN	38.0	35.0
1334.0	315000.0	242.0	100.0	70.0	4.85	24.0	23.0	2022.0	111231.0	NaN	55.0	54.0

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Ilustración 16. *Descripción de los datos*
Fuente: Elaboración propia

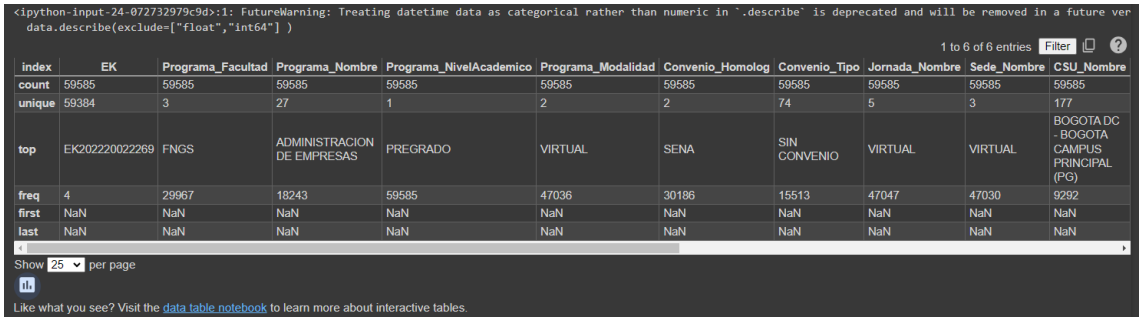
Se verifica la información cargada.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59585 entries, 0 to 59584
Data columns (total 28 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Documento                   59585 non-null  int64
 1   PIDM                        59585 non-null  int64
 2   Puntaje global              59585 non-null  int64
 3   Percentil nacional global   59579 non-null  object
 4   Materias                    59585 non-null  int64
 5   Promedio                    59585 non-null  float64
 6   Materias Homologadas        59585 non-null  int64
 7   Materias Perdidas           59585 non-null  int64
 8   EK                          59585 non-null  object
 9   AÑO                         59585 non-null  int64
 10  Programa_Codigo             59585 non-null  int64
 11  Programa_Facultad           59585 non-null  object
 12  Programa_Nombre             59585 non-null  object
 13  Programa_NivelAcademico     59585 non-null  object
 14  Programa_Modalidad          59585 non-null  object
 15  Convenio_Homolog            59585 non-null  object
 16  Convenio_Tipo               59585 non-null  object
 17  Jornada_Nombre              59585 non-null  object
 18  Sede_Nombre                 59585 non-null  object
 19  CSU_Nombre                  59568 non-null  object
 20  Geografia_NombreCiudad      59585 non-null  object
 21  Persona_TipoIdentificacion  59585 non-null  object
 22  Persona_EstadoCivil         59585 non-null  object
 23  Persona_Genero              59585 non-null  object
 24  Persona_FechaNacimiento     59563 non-null  datetime64[ns]
 25  CSU_Region                  59585 non-null  object
 26  Edad                        59585 non-null  int64
 27  Edad_2                      59585 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(10), object(16)
memory usage: 12.7+ MB
```

Ilustración 17. *Conjunto de columnas y número de registros*
Fuente: Elaboración propia

En nuestro proceso exploratorio se encuentran tres tipos de datos, enteros, flotantes y categóricos.



```
<ipython-input-24-072732979c9d>:1: FutureWarning: Treating datetime data as categorical rather than numeric in `.describe` is deprecated and will be removed in a future ver
  data.describe(exclude=["float","int64"] )
```

1 to 6 of 6 entries Filter

index	EK	Programa_Facultad	Programa_Nombre	Programa_NivelAcademico	Programa_Modalidad	Convenio_Homolog	Convenio_Tipo	Jornada_Nombre	Sede_Nombre	CSU_Nombre
count	59585	59585	59585	59585	59585	59585	59585	59585	59585	59585
unique	59384	3	27	1	2	2	74	5	3	177
top	EK202220022269	FNGS	ADMINISTRACION DE EMPRESAS	PREGRADO	VIRTUAL	SENA	SIN CONVENIO	VIRTUAL	VIRTUAL	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)
freq	4	29967	18243	59585	47036	30186	15513	47047	47030	9292
first	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Ilustración 18. *Descripción de los datos 2*
Fuente: Elaboración propia

Se realiza el análisis descriptivo de cada variable, donde se evidencia que contamos con 59.585 registros con la siguiente información:

- El Puntaje global muestra un mínimo de 70 y un máximo de 242 con una media de 139 puntos, con una desviación de 20.85 lo que muestra unos datos cercanos a la media.
- Las materias tienen un promedio de 35 materias con un mínimo de 10 y un máximo de 70, con una desviación de 9.74 lo que indica que existe variabilidad importante en los datos.
- El promedio tiene un mínimo de 3.0 y un máximo de 4.85 con una media de 4.08, la desviación estándar de 0.35 evidencia una variabilidad en los datos, pero, casi todos los datos están cercanos a la media.
- Las Materias homologadas cuentan con un mínimo de 0 y un máximo de 24 materias, y una media de 6, con una desviación de 8.71 que muestra una variabilidad a tener en cuenta.
- Las materias perdidas muestran una mediana de 1.5 con un máximo de 23 y un mínimo de 0, la desviación estándar de 2.83 nos muestra una variabilidad moderada.
- La variable Año presenta una desviación moderada con un mínimo de 2016 y un máximo de 2022. Se debe validar si se convierte en una variable categórica para el modelo.
- Programa código presenta una desviación amplia dada la cantidad de programas diferentes que presenta la universidad.
- Edad_2 presenta una desviación de 7.1 lo que indica una dispersión en los datos moderados, pero centrándose en la mediana de 30 años, con un mínimo de 15 y un máximo 54.
- Variables como Documento, PIDM y Código programa, son variables que no aportan información al modelo, y se deben eliminar.

Ahora se analizarán las variables categóricas:

- Para la variable Programa_Facultad se encuentran 3 clases de datos.
- Programa_Nombre muestra 27 programas únicos.

- Nivel Académico tiene una única clase.
- Existen dos Modalidad en el conjunto de datos.
- Convenio_Homolg existen dos tipos.
- Convenio_Tipo muestra 74 datos únicos.
- Jornada_Nombre presenta 5 jornadas de estudio.
- Sede Nombre existe con 3 sedes para sus estudiantes.
- CSU_Nombre presenta 177 puntos diferentes.
- Geografia_NombreCiudad muestra presencia en 793.
- Persona_Tipoidentificacion tiene 5 clases.
- Persona_EstadoCivil muestra 7 tipos de estado.
- Persona_Genero tiene dos tipos.
- Persona_FechaNacimiento presenta variedad en la información.
- CSU_Region muestra 4 tipos.

Luego de verificar la estructura de los datos, se valida que no contenga información faltante o nula en cada una de su columna.

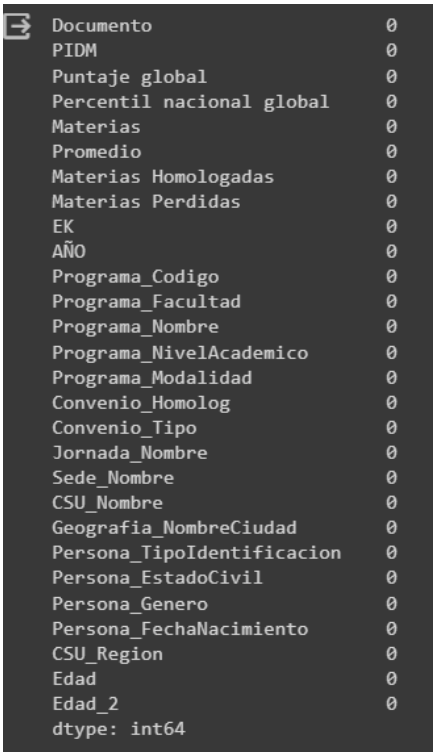


Ilustración 19. *Datos sin registros faltantes*
Fuente: Elaboración propia

De acuerdo a la ilustración encontramos una base sin datos nulos o faltantes.

Se procede al análisis univariado exploratorio con el apoyo de los gráficos, para detectar datos atípicos y entender mejor el comportamiento de los mismos, sus distribuciones y sus posibles correlaciones entre variables.

7.2.3.1. ANÁLISIS UNIVARIADO

A continuación, cada una de las variables se graficará para comprender mejor su comportamiento y así ir conociendo la importancia para nuestro modelo.

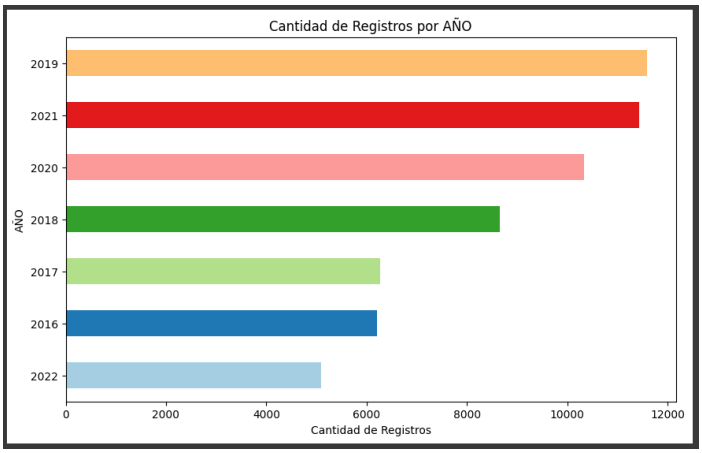


Ilustración 20. *Cantidad de registros por año*
Fuente: Elaboración propia

En la ilustración anterior se observa que existen 7 años de información con una gran concentración de los datos en los años de 2019, 2020 y 2021, así mismo se evidencia que los años 2016, 2017 y 2022, presentan una variación en los datos en su cantidad con respecto a los otros años, algo que hay que tener en cuenta para la siguiente fase.

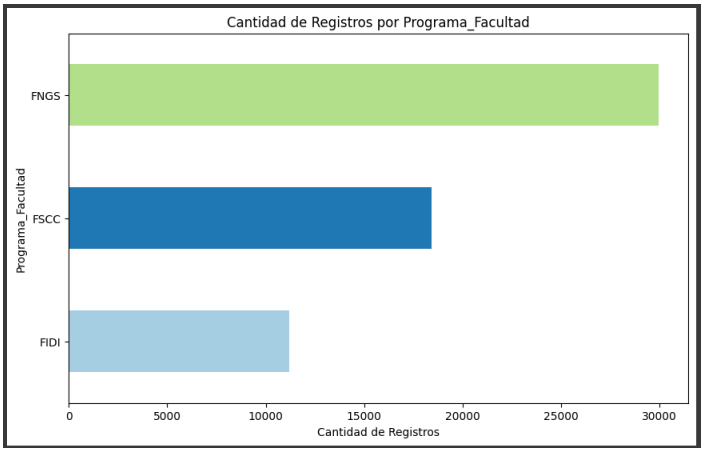


Ilustración 21. *Cantidad de registros por Facultad*
Fuente: Elaboración propia

En la ilustración se evidencian 3 facultades que agrupan los datos con un liderazgo de estudiantes que presentaron las pruebas saber pro en la facultad de FNGS, y con una facultad con más pocos registros FIDI.

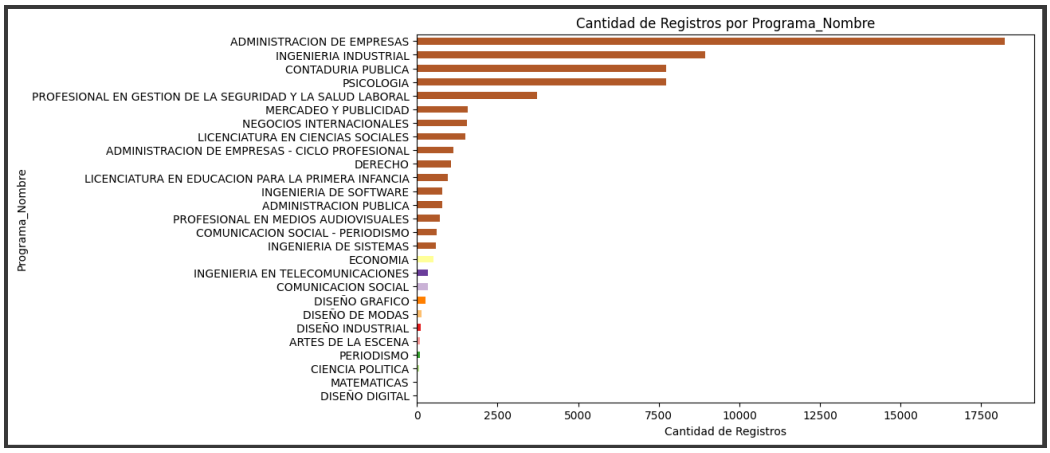


Ilustración 22. *Cantidad de registros por Nombre*
Fuente: Elaboración propia

En la ilustración se evidencia gran diversidad de programas, pero con un protagonismo de 5 programas principales, Administración de Empresas, Ingeniería Industrial, Contaduría Pública, Psicología y Profesional en gestión de la seguridad y la salud laboral. Información relevante para el diseño del modelo.

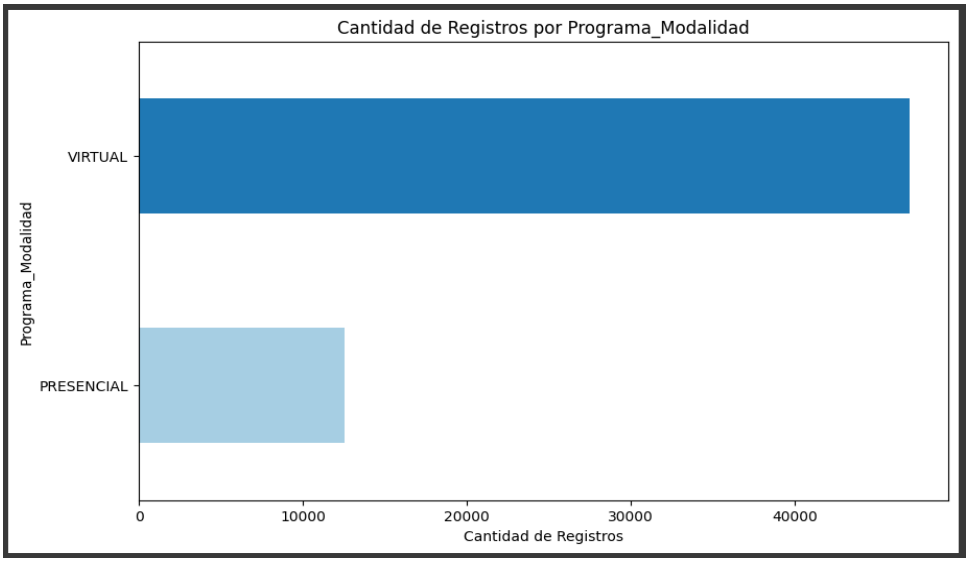


Ilustración 23. *Cantidad de registros por Modalidad*
Fuente: Elaboración propia

En la ilustración anterior se observar como la modalidad virtual representa casi 4 veces los estudiantes que se encuentran en la modalidad presencial.

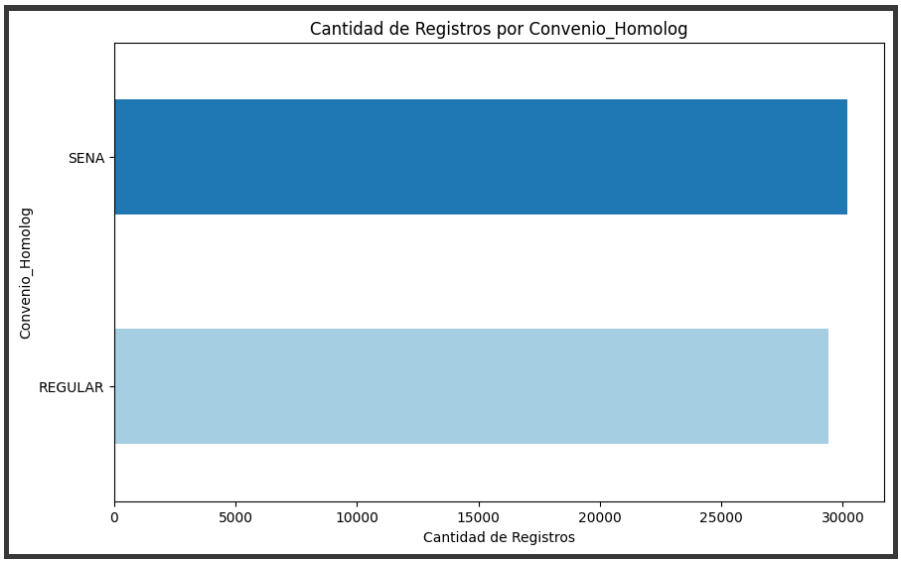


Ilustración 24. *Cantidad de registros por convenio*
Fuente: Elaboración propia

En la ilustración se evidencia que, convenio homologación tiene dos clases SENA y regular con una participación de casi un 50% en cada una.

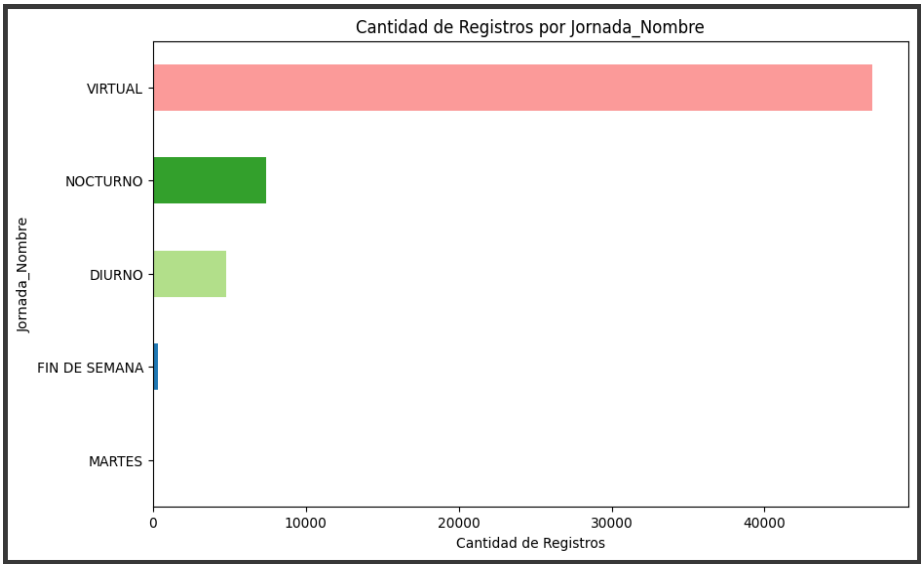


Ilustración 25. *Cantidad de registros por jornada*
Fuente: Elaboración propia

En la ilustración se muestra que la variable jornada presenta 5 jornadas, pero que la jornada denominada virtual es la que concentra la gran cantidad de datos, mientras que las demás como nocturno, diurno, fin de semana y martes tienen una participación más baja.

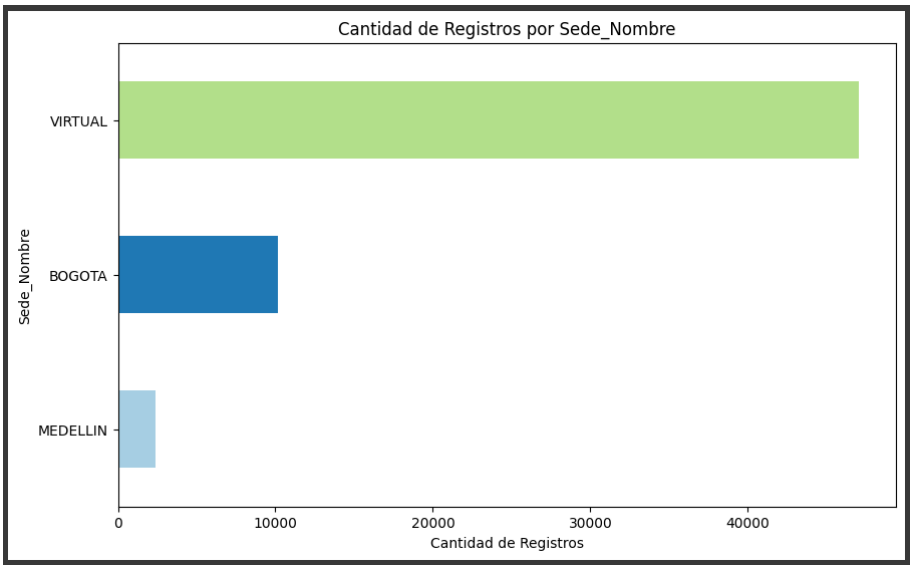


Ilustración 26. *Cantidad de registros por sede*
Fuente: Elaboración propia

En la ilustración se ve como dentro de las sedes existe la virtualidad que también ocupa una parte muy importante en esta variable, ya que compone más de 70% de los datos.

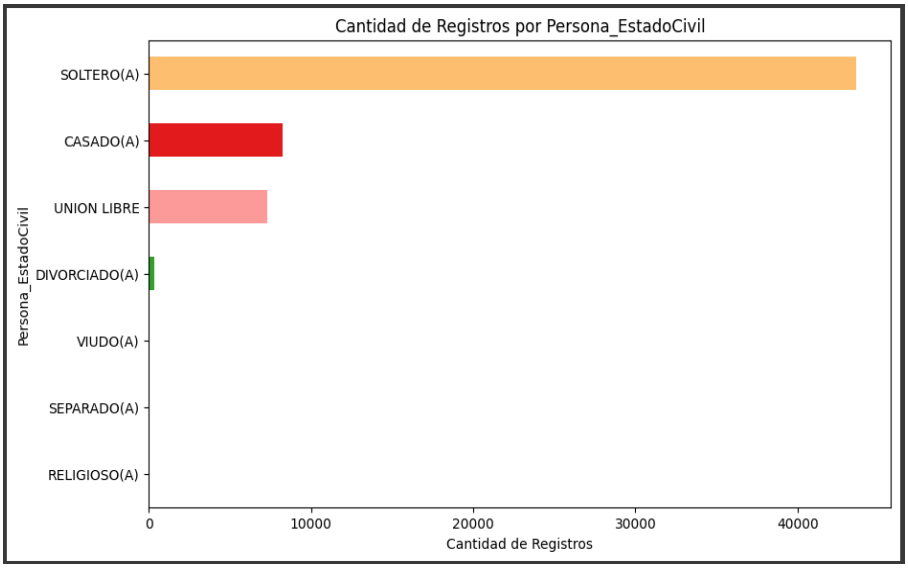


Ilustración 27. *Cantidad de registros por estado civil*
Fuente: Elaboración propia

La ilustración muestra el estado civil de los estudiantes con 7 tipos de estado y una concentración principal en SOLTERO(A), tipos como RELIGIOSO(A), SEPARADO(A) y VIUDO(A), tiene muy baja representación en los datos.

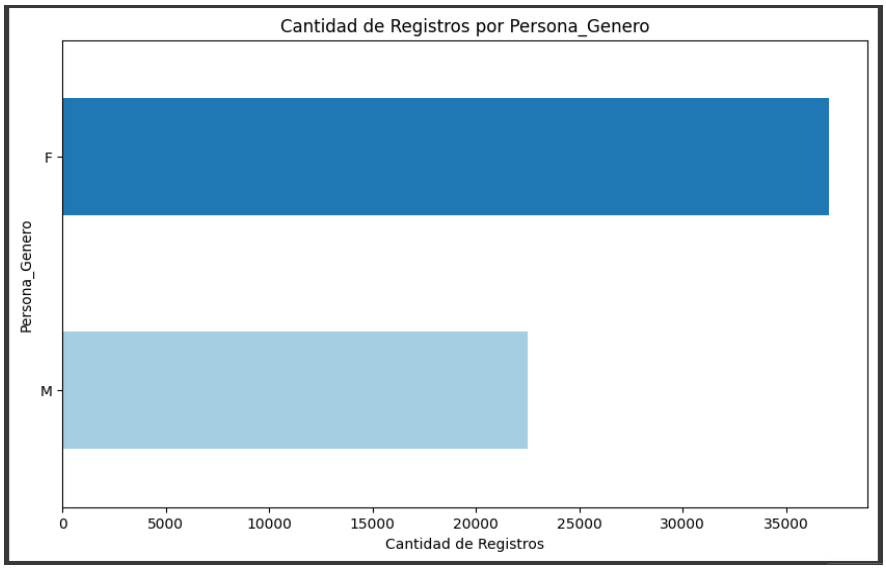


Ilustración 28. *Cantidad de registros por género*
Fuente: Elaboración propia

La ilustración se muestra el género de los estudiantes que han participado en las pruebas Saber Pro, donde se evidencia una mayor participación del género femenino con un 30% más por encima del género masculino.

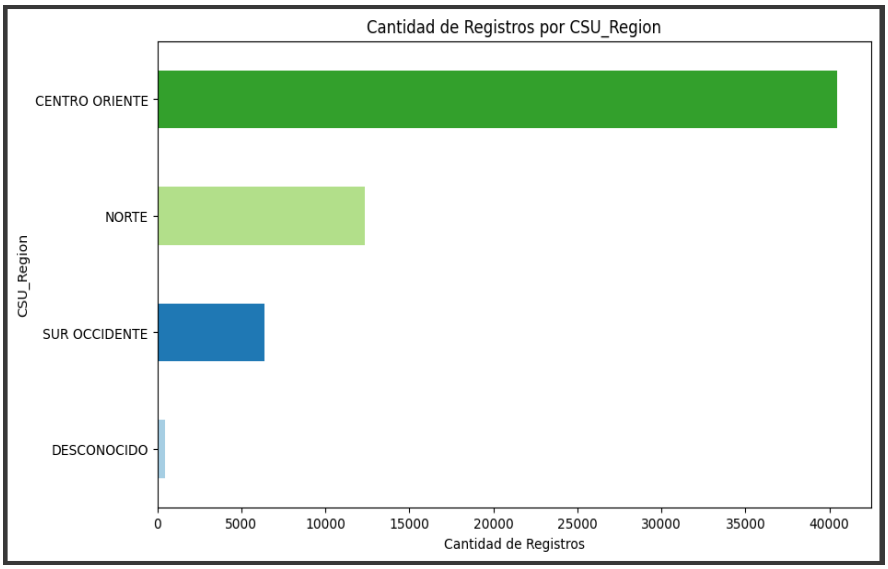


Ilustración 29. *Cantidad de registros por región*
Fuente: Elaboración propia

En la ilustración se evidencian 4 tipos de CSU_Region, una de ellas desconocido que se debe tener en cuenta al momento de trabajar los datos para el modelo.

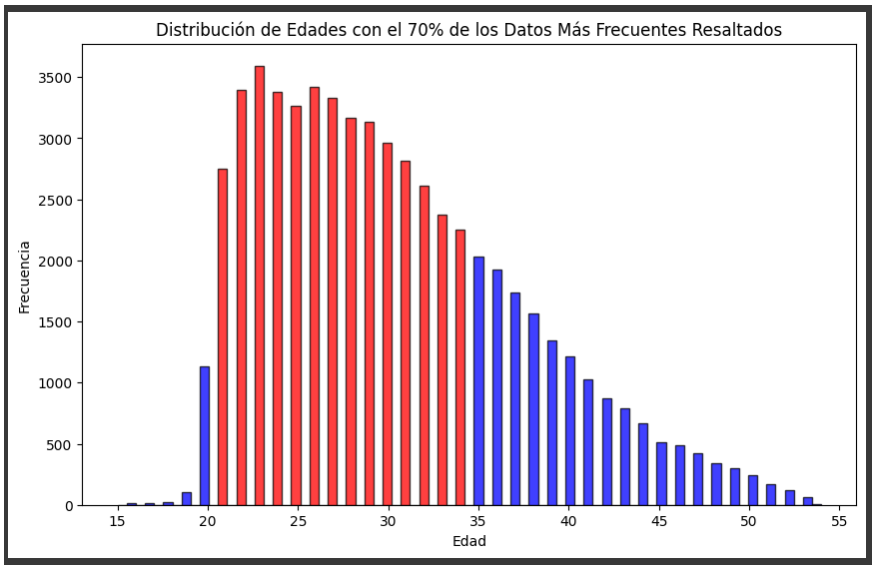


Ilustración 30. *Distribución por edades*
Fuente: Elaboración propia

En la ilustración se ve la composición de las edades de 15 a 55 años para la presentación de las pruebas, pero se evidencia una concentración de los datos del 70% en las edades de 21 a 34 años aproximadamente. También se evidencia que los mayores de 45 años son una muestra baja, al igual que los menores de 20 años.

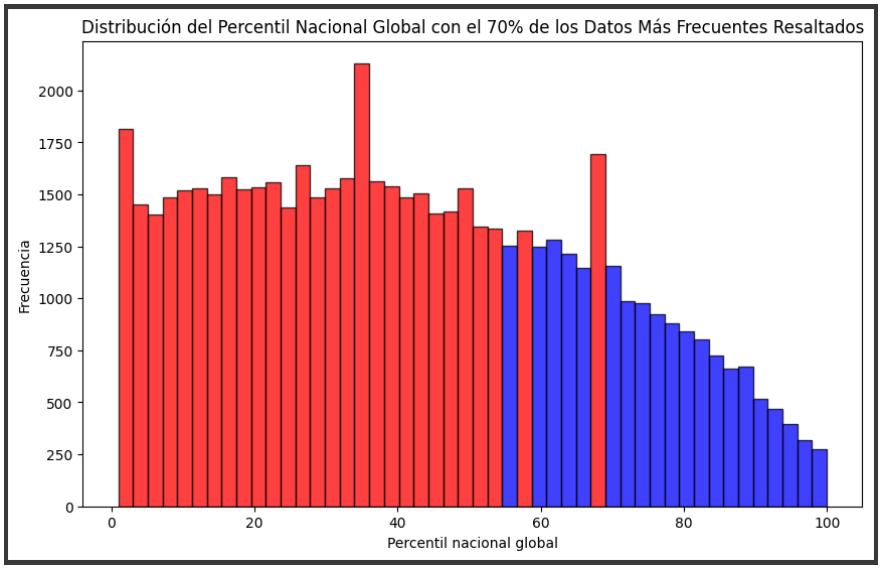


Ilustración 31. *Distribución del percentil nacional Global*
Fuente: Elaboración propia

En la ilustración la variable percentil nacional global muestra una concentración principal en los datos de 1 a 58 aproximadamente, y una frecuencia baja en los percentiles de 70 en adelante.

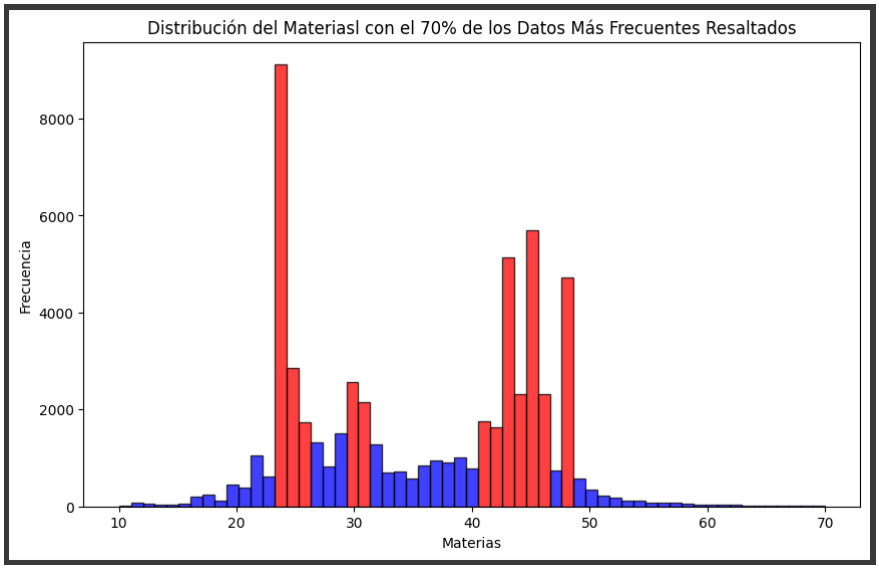


Ilustración 32. *Distribución de materias*
Fuente: Elaboración propia

En la ilustración se ve una tendencia de 24 materias seguido por 45, 43 y 48 materias. También se evidencia una muy baja frecuencia en los datos para materias superiores a 50 y menores de 20.

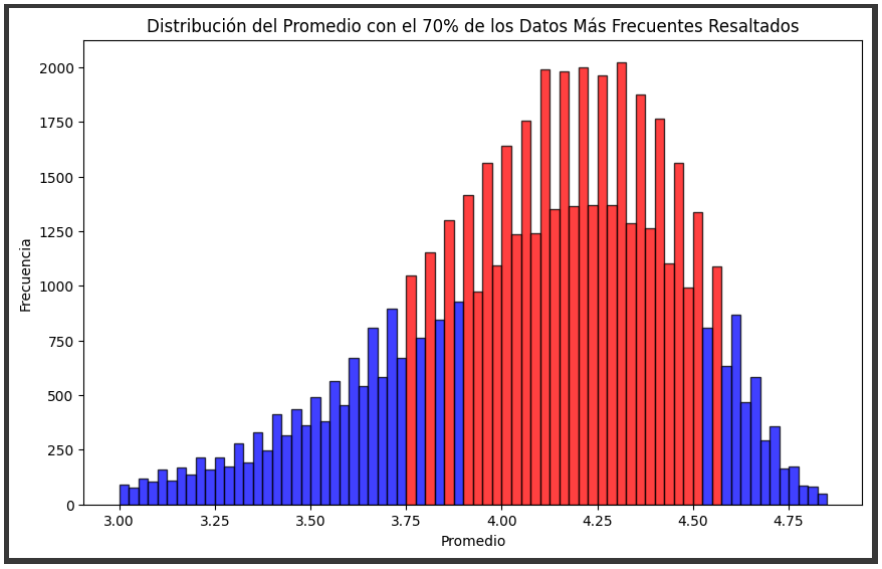


Ilustración 33. *Distribución del promedio*
Fuente: Elaboración propia

En la ilustración observamos que el promedio este concentrado entre 3.75 y 4.50 sobre el total de notas obtenidas en la carrera. Se ve claramente una frecuencia baja en los promedios superiores a 4.50 e inferiores de 3.50.

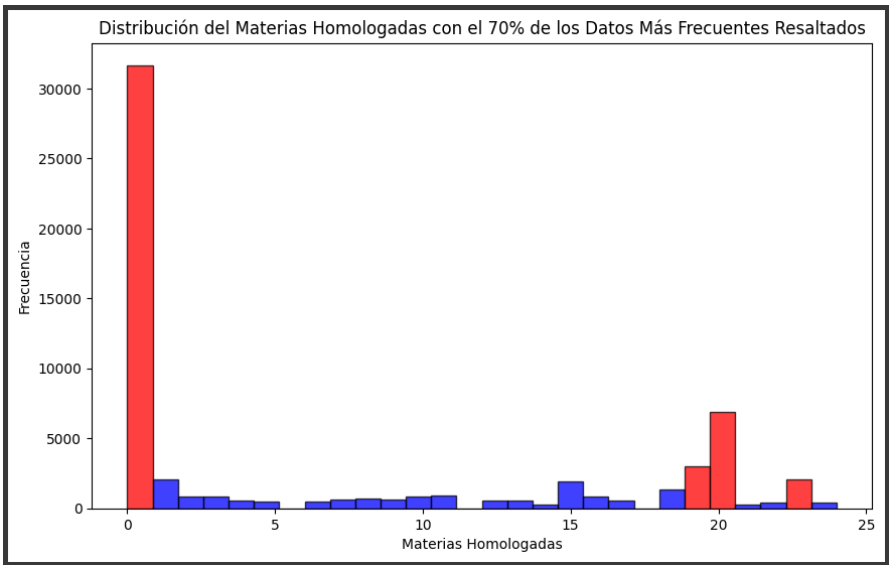


Ilustración 34. *Distribución de materias homologadas*
Fuente: Elaboración propia

En la ilustración se ve como las 0 materias homologadas es la frecuencia más importante, luego se ve una ligera concentración en homologaciones entre a 19 y 20 materias, también se ve una gran dispersión de los datos de los que no están con mayor a 0 materias homologadas.

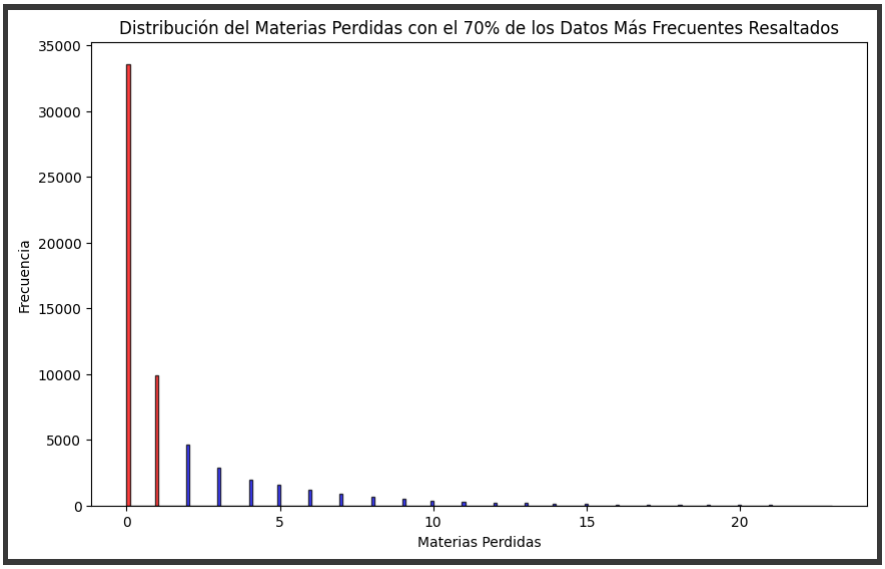


Ilustración 35. *Distribución de materias perdidas*
Fuente: Elaboración propia

En la ilustración se evidencia una concentración principalmente en 0 y 1 materia perdida y los mayores de 1 con una frecuencia muy baja con respecto al resto de los datos.

Teniendo en cuenta que la variable objetivo es el puntaje global obtenido por cada uno de los estudiantes, se deben clasificar en dos clases, para facilitar el aprendizaje del

modelo. Por lo que con base en la previsión para 2023 y el resultado de los 2 últimos años a nivel nacional, se toma como bajo todo lo que se encuentra por debajo o igual a 145 puntos y lo que está por encima como medio, creando una nueva variable llamada CLASE_1.

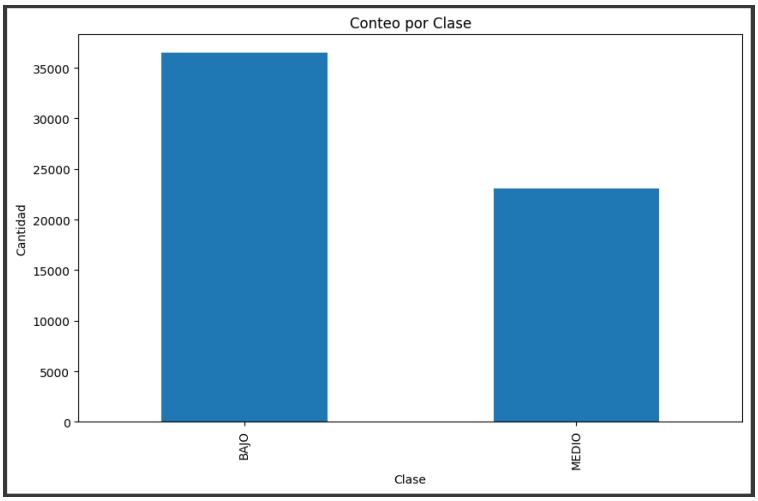


Ilustración 36. *Conteo por clase*
Fuente: Elaboración propia

La ilustración muestra cómo es la distribución de la nueva clase creada, con base en el puntaje que se convierte en la variable objetivo para nuestro modelo de clasificación.

7.1.3.2. ANÁLISIS COMPUESTO POR LA VARIABLE OBJETIVO

Se realiza a continuación el análisis de las variables principales con respecto a la variable objetivo CLASE_1.

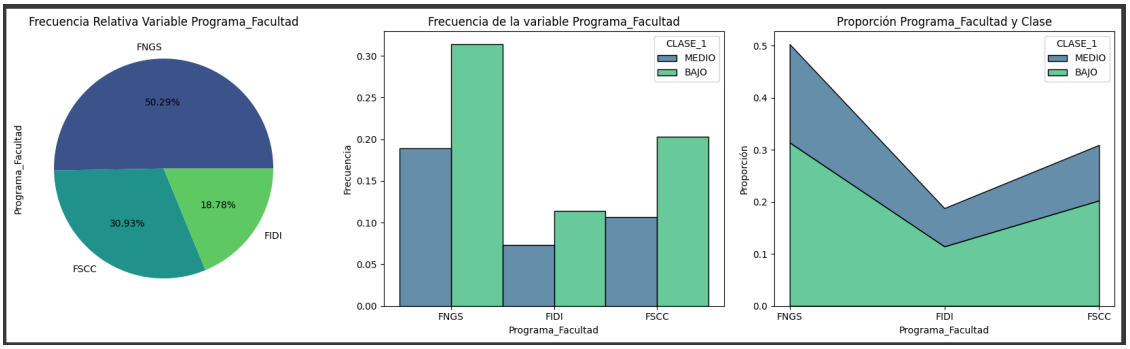


Ilustración 37. *Variable objetivo vs Facultades*
Fuente: Elaboración propia

En la ilustración se ve como la facultad FNGS tiene una concentración mayor de los datos y como existe una frecuencia importante de la clase bajo. Lo que muestra, como a través del tiempo el resultado que más predomina en las diferentes facultades es el bajo.

La facultad FIDI presenta un comportamiento un poco diferente a las otras dos facultades ya que presenta un mejor equilibrio entre la clase media y la baja.

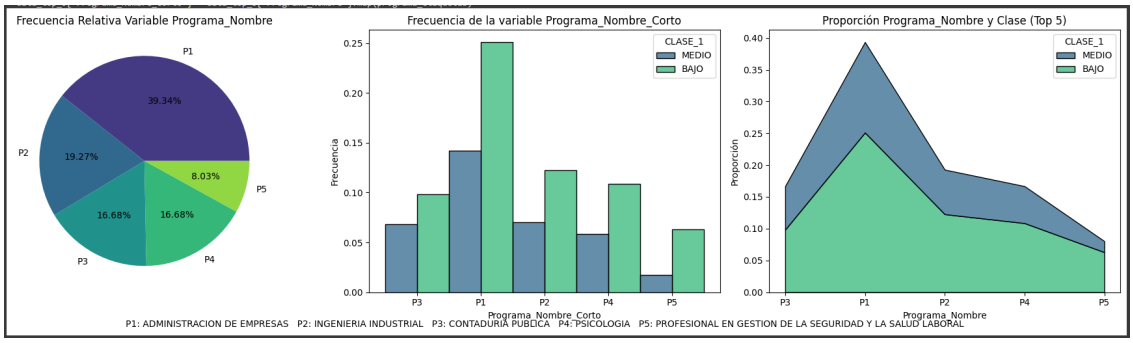


Ilustración 38. *Variable objetivo vs TOP 5 Programas*
Fuente: Elaboración propia

La ilustración de programa nombre al tener 27 nombres, se centra en los 5 principales identificados anteriormente, donde se observa que el programa de Administración de empresa tiene una participación muy representativa de la clase bajo.

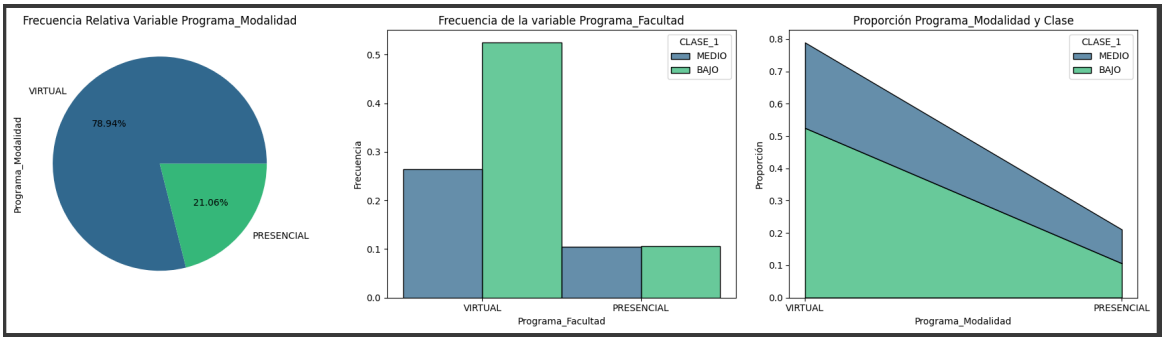


Ilustración 39. *Variable objetivo vs Modalidad*
Fuente: Elaboración propia

La anterior ilustración evidencia la participación en los resultados de las modalidades con una concentración en los datos en la virtualidad, lo que muestra que esta modalidad virtual cuenta con una mayor participación en la clase bajo, mientras la modalidad presencial muestra una participación casi igual en las clases bajo y medio.

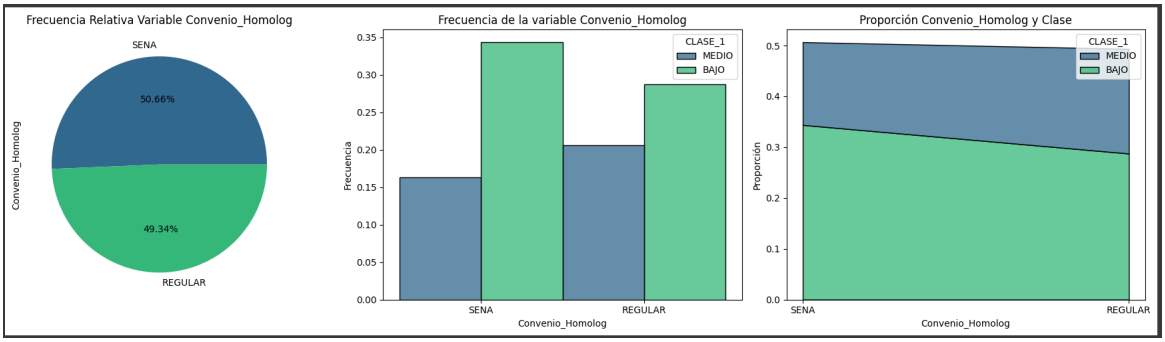


Ilustración 40. *Variable objetivo vs Convenio*
Fuente: Elaboración propia

La ilustración muestra como la variable convenio homologación tiene dos tipos Sena y regular, lo que muestra que los convenios Sena presentan una mayor participación con la clase bajo.

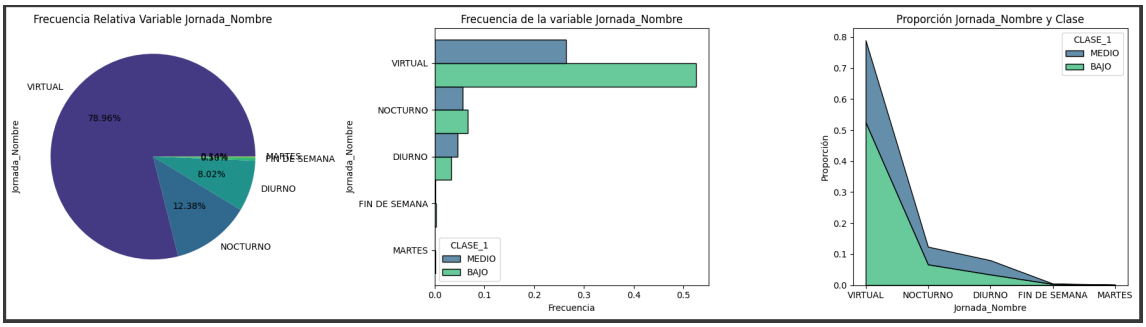


Ilustración 41. *Variable objetivo vs Jornada*
Fuente: Elaboración propia

La ilustración muestra la variable jornada con sus 5 tipos de jornadas la cual se centra principalmente en la VIRTUAL, con una participación muy importante en la jornada virtual con la clase bajo. Es importante destacar como la jornada DIURNO tiene una mayor participación en los datos con la clase medio.

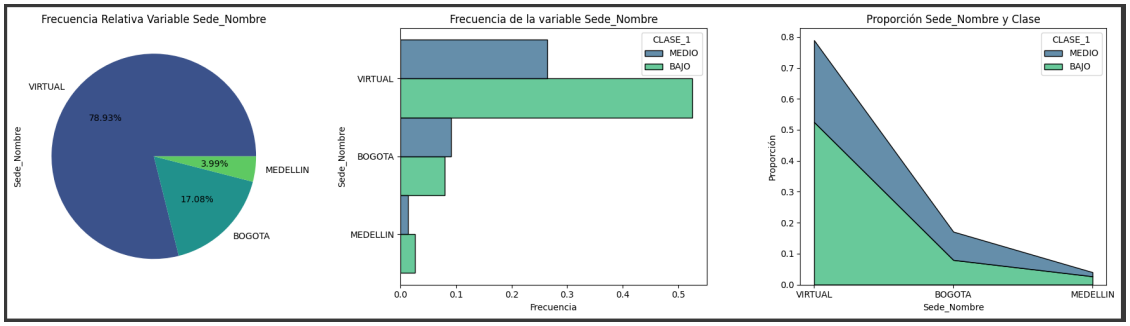


Ilustración 42. *Variable objetivo vs Sede*
Fuente: Elaboración propia

La ilustración muestra la variable sede, con tres sedes principales con una participación alta de la VIRTUAL con la clase bajo. Destacable la sede BOGOTÁ donde se ve una participación mayor en la clase medio.

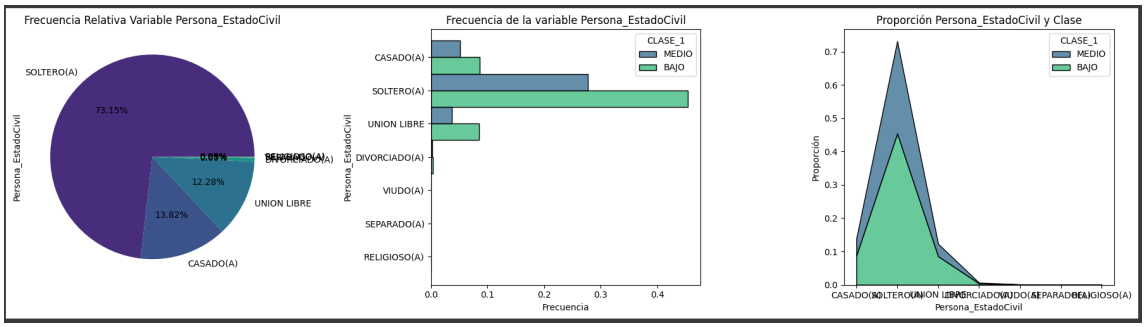


Ilustración 43. *Variable objetivo vs Estado Civil*
Fuente: Elaboración propia

La ilustración muestra como la variable estado civil tiene tres categorías principales, pero con una participación mayor de la SOLTERO(A) que en su mayor cantidad de información tiene clase bajo.

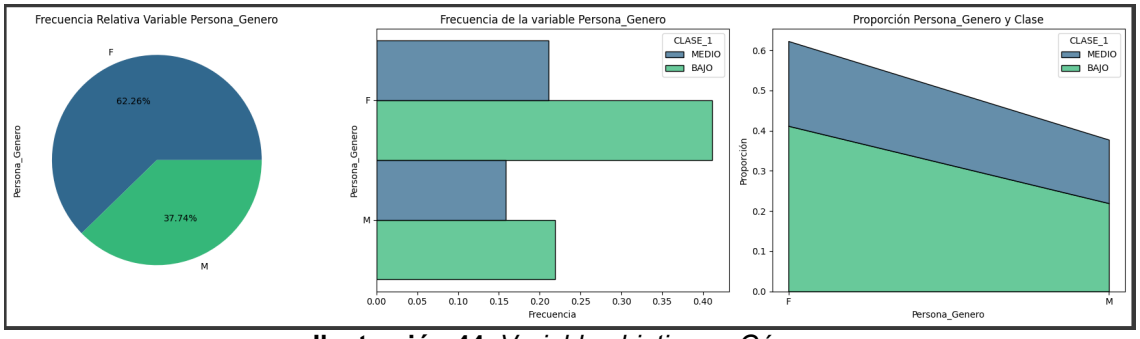


Ilustración 44. *Variable objetivo vs Género*
Fuente: Elaboración propia

La ilustración muestra la variable genero con sus dos categorías, donde la femenina tiene mayor participación con la clase bajo, casi el doble de la clase bajo de los masculinos.

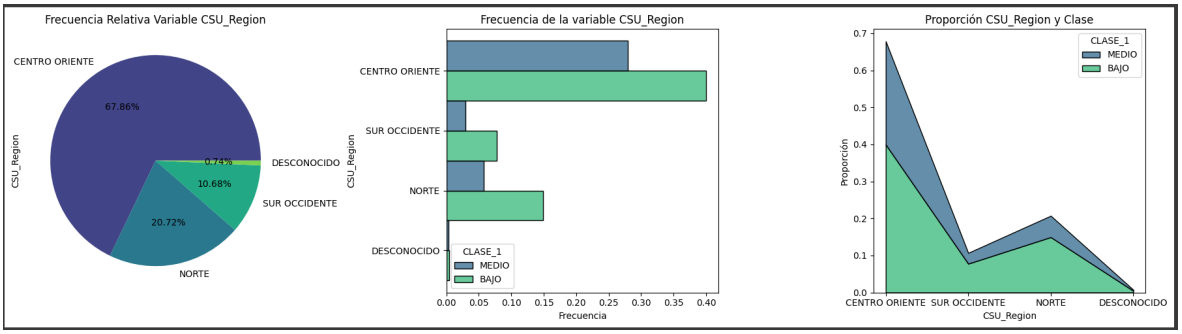


Ilustración 45. *Variable objetivo vs CSU*
Fuente: Elaboración propia

La ilustración muestra los CSU región con tres categorías, principalmente donde CENTRO ORIENTE muestra mayor participación de la clase bajo. También se observa una categoría DESCONOCIDO con una muy baja participación, pero que se puede tener en cuenta en la próxima fase.

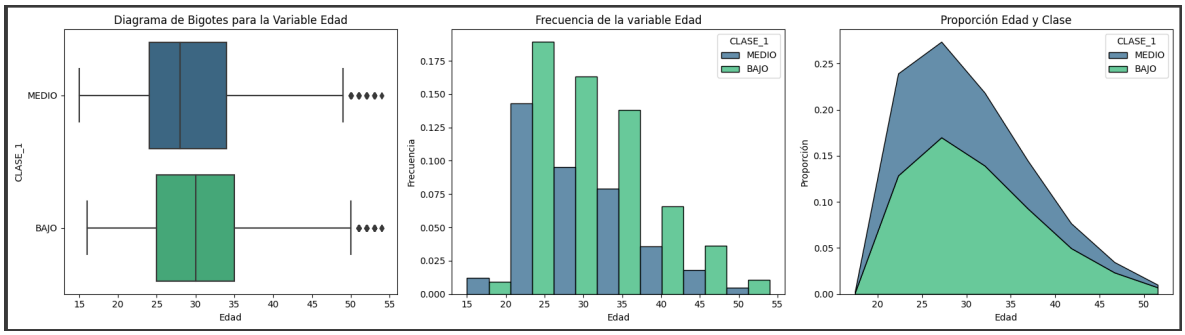


Ilustración 46. *Variable objetivo vs Edad*
Fuente: Elaboración propia

La ilustración muestra la variable Edad con una concentración en edades mayores de 20 y menores de 35 años. Se observa también, que la edad de 20 a 25 tiene una buena participación en las clase medio y bajo.

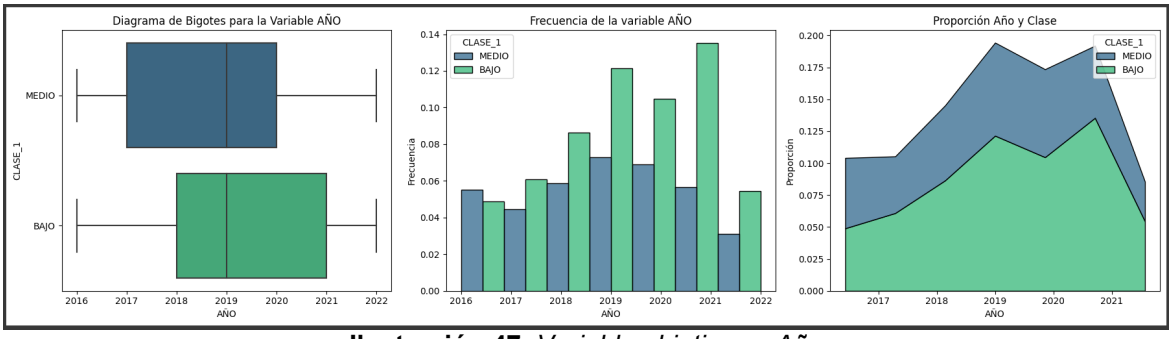


Ilustración 47. *Variable objetivo vs Año*
Fuente: Elaboración propia

En la ilustración se ve como los datos se concentran en 2018 a 2021 y la participación de la clase bajo es muy representativa en los años 2021 y 2019.

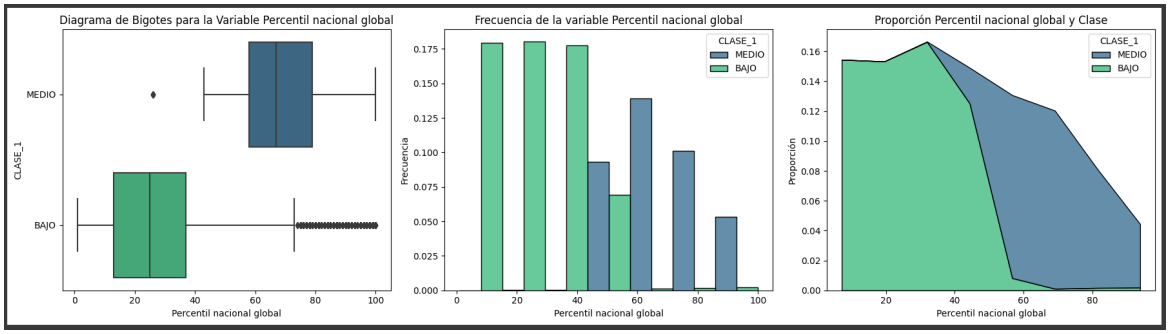


Ilustración 48. *Variable objetivo vs Percentiles*
Fuente: Elaboración propia

La ilustración muestra la variable percentil nacional global y su comportamiento con la clase, que evidencia que pueden tener una correlación y debe ser trabajada en la próxima fase.

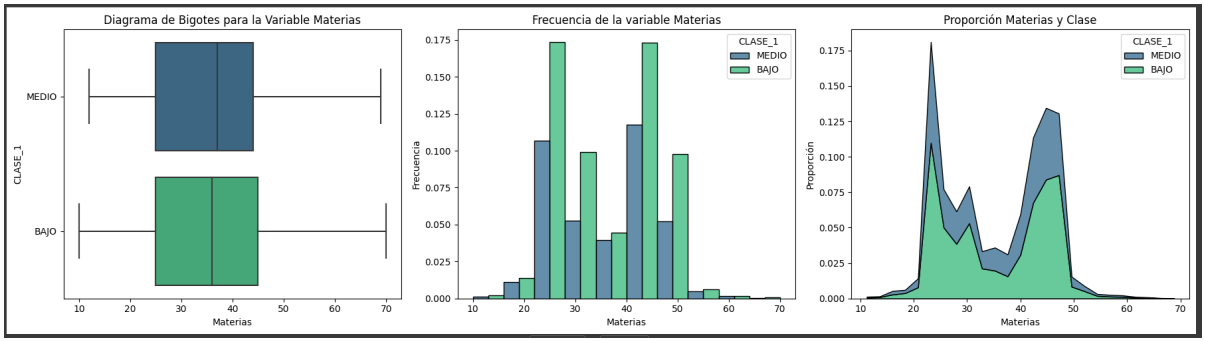


Ilustración 49. *Variable objetivo vs No de Materias*
Fuente: Elaboración propia

La ilustración muestra la cantidad de materias de la malla académica que cursó un estudiante, donde se evidencia una participación importante en 24 materias y mayores de 40. La clase predominante en esta variable es la baja.

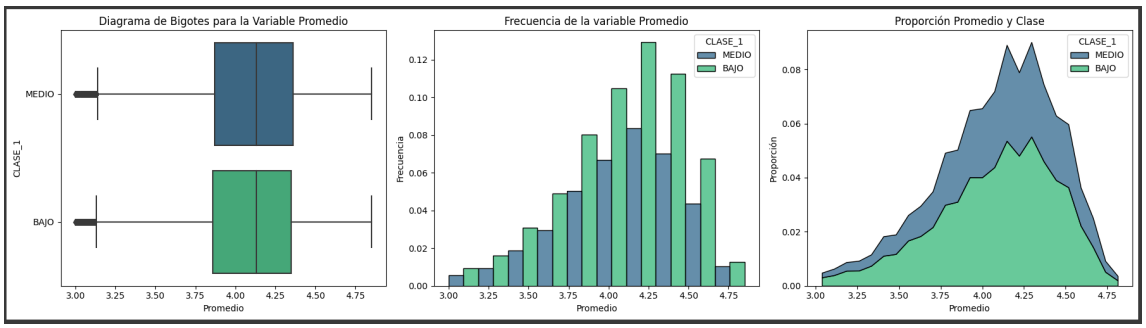


Ilustración 50. *Variable objetivo vs Promedio*
Fuente: Elaboración propia

En la ilustración se evidencia que el promedio este concentrado entre 3.75 y 4.50 sobre el total de notas obtenidas en la carrera con una destacable participación de la clase bajo.

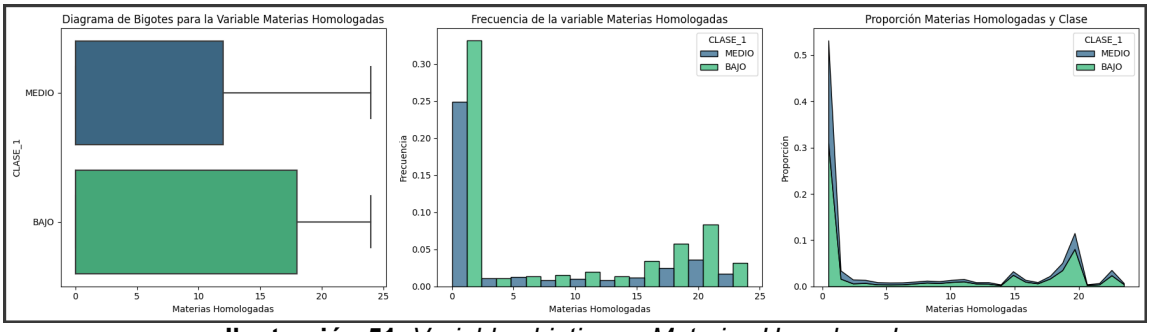


Ilustración 51. *Variable objetivo vs Materias Homologadas*
Fuente: Elaboración propia

La ilustración muestra que, la mayor cantidad de registros están en 0 materias homologadas con una participación mayor de la clase bajo.

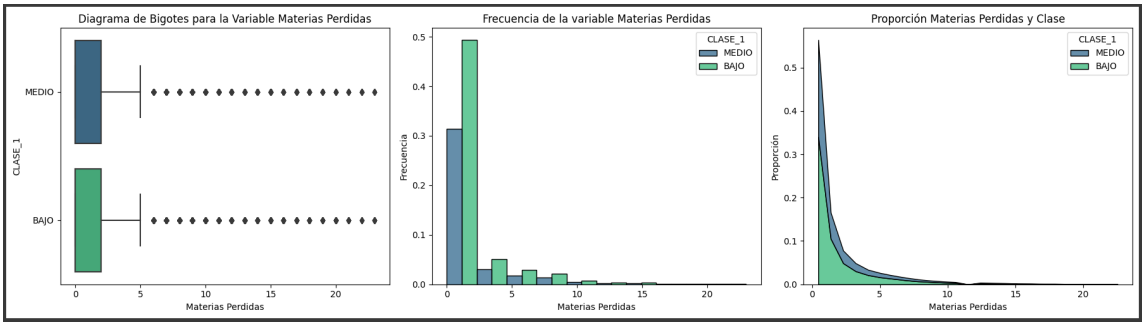
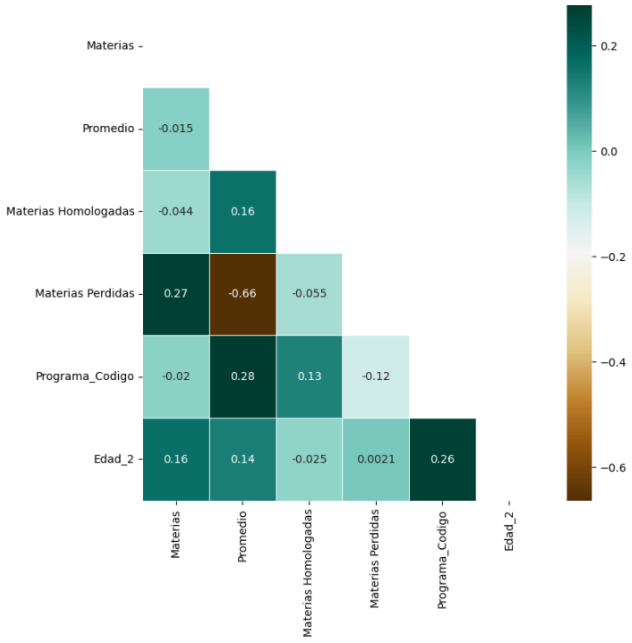


Ilustración 52. *Variable objetivo vs Materias Perdidas*
Fuente: Elaboración propia

En la ilustración se muestra el comportamiento de las materias perdidas, con una clara participación de materias perdidas igual a cero centrada en la clase bajo.

7.2.3.2. ANÁLISIS CORRELACIÓN VARIABLES NUMÉRICAS



	Materias	Promedio	Materias Homologadas	Materias Perdidas	Programa_Codigo	Edad_2
Materias	1.00	-0.02	-0.04	0.27	-0.02	0.16
Promedio	-0.02	1.00	0.16	-0.66	0.28	0.14
Materias Homologadas	-0.04	0.16	1.00	-0.06	0.13	-0.02
Materias Perdidas	0.27	-0.66	-0.06	1.00	-0.12	0.00
Programa_Codigo	-0.02	0.28	0.13	-0.12	1.00	0.26
Edad_2	0.16	0.14	-0.02	0.00	0.26	1.00

Ilustración 53. *Análisis de Correlación*
Fuente: Elaboración propia

En la ilustración se ve lo siguiente:

- **Materias y Materias perdidas (0,27):** Existe una correlación moderadamente positiva. Esto puede indicar que a medida que aumenta el número de materias, también aumenta el número de materias perdidas. Sin embargo, la correlación no es lo suficientemente fuerte.

- **Promedio y Programa Código (0,28):** Ligera correlación positiva. Esto puede sugerir que algunas carreras están asociadas con promedios más altos, aunque la correlación no es significativa.

- **Promedio y Materias Perdidas (-0,66):** Esta es el más fuerte y el que tiene la correlación más negativa de la matriz. Esto muestra que a medida que aumenta el

número de materias perdidas, la nota promedio tiende a disminuir. Estas relaciones son obvias y reflejan patrones esperados de rendimiento académico.

- **Edad_2 y varias variables:** La edad generalmente tiene correlaciones bajas con otras variables, lo que indica que la edad no tiene un efecto significativo sobre estas variables. La correlación más alta es con Program_Code (0,26), lo que puede indicar que existe un ligero sesgo hacia estudiantes de cierta edad en ciertos cursos, aunque la correlación sigue siendo débil.
- **Programa_Codigo y otras variables:** Las correlaciones entre Programa Código y otras variables académicas son generalmente bajas, lo que indica que la selección de cursos no tiene una fuerte correlación con el número de materias, el promedio, la cantidad de materias aprobadas o faltantes.

Las correlaciones que se evidencian no necesariamente indican causalidad, por que varias de ellas pueden estar influenciadas por otros factores, que más adelante se evidenciaran o que no están incluidas en el análisis.

7.1.4. CALIDAD DE DATOS

Los anteriores análisis muestran una gran cantidad de información necesaria para continuar con la siguiente fase de nuestra metodología CRISP-DM, orientando el modelo a depurar variables que no son útiles y a ajustar otras, dado los hallazgos encontrados durante el proceso exploratorio.

7.3. FASE DE PREPARACIÓN DE LOS DATOS

7.3.1. SELECCIÓN DE LOS DATOS

La base de datos que se generó luego desde la construcción del modelo descriptivo en Power BI entrega una data limpia y con gran potencial para el modelo predictivo, sin embargo, después del análisis exploratorio realizado en la fase anterior se identificaron algunas variables que no se deben considerar para la construcción del modelo predictivo. Así mismo, como algunos datos atípicos que deben ser tratados; es importante también resaltar que algunas variables deben ser transformadas antes de ser incluidas en el modelo.

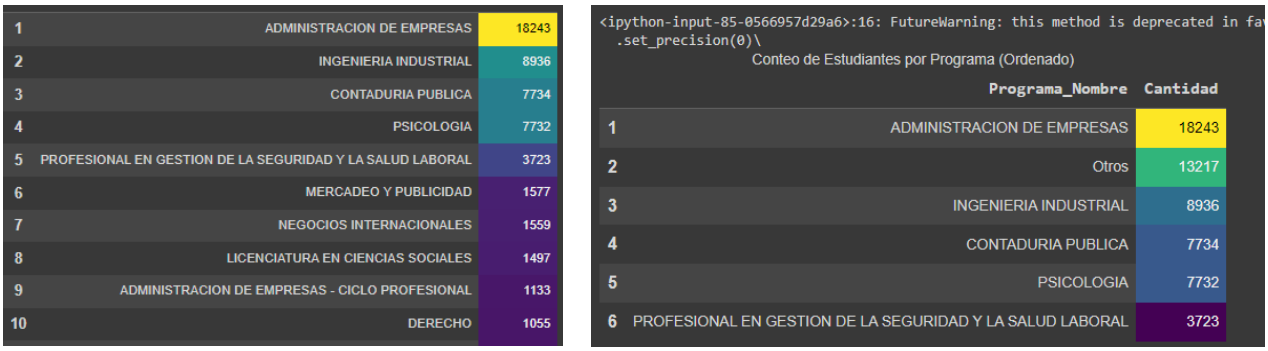
En el análisis de la fase anterior se evidencian algunas oportunidades, en cuanto agrupamiento de algunas categorías dentro de las variables para ser más consistentes y para proporcionar al modelo únicamente la información más relevante y que debe considerar para el modelo de manera más limpia.

Luego de este proceso, se realizará uno nuevo de análisis sobre las variables tratadas.

7.3.2. LIMPIEZA DE LOS DATOS

Una de las principales transformaciones, es el manejo de los lumbrales en las variables para dejar las principales categorías.

La primera variable para transformar, es la variable de Programa nombre, donde se encuentran 27 categorías con una gran dispersión en los datos, para esto se dejará únicamente los programas que tienen más de 2000 registros y se creará otra categoría como otros, donde se agruparan todos los programas con frecuencia inferior a 2000.



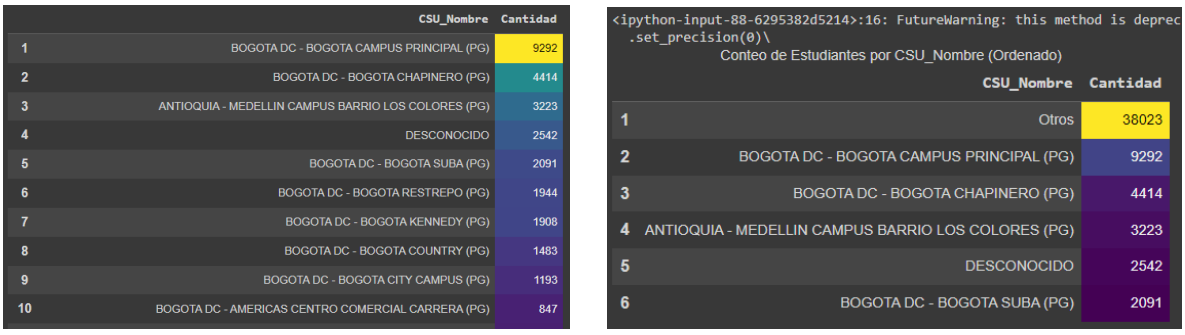
1	ADMINISTRACION DE EMPRESAS	18243
2	INGENIERIA INDUSTRIAL	8936
3	CONTADURIA PUBLICA	7734
4	PSICOLOGIA	7732
5	PROFESIONAL EN GESTION DE LA SEGURIDAD Y LA SALUD LABORAL	3723
6	MERCADEO Y PUBLICIDAD	1577
7	NEGOCIOS INTERNACIONALES	1559
8	LICENCIATURA EN CIENCIAS SOCIALES	1497
9	ADMINISTRACION DE EMPRESAS - CICLO PROFESIONAL	1133
10	DERECHO	1055

	Programa_Nombre	Cantidad
1	ADMINISTRACION DE EMPRESAS	18243
2	Otros	13217
3	INGENIERIA INDUSTRIAL	8936
4	CONTADURIA PUBLICA	7734
5	PSICOLOGIA	7732
6	PROFESIONAL EN GESTION DE LA SEGURIDAD Y LA SALUD LABORAL	3723

Ilustración 54. *Programas Principales por número de registros*
Fuente: Elaboración propia

En la ilustración se ven los 10 primeros programas, donde se puede evidencia la dispersión de los datos con los 27 programas existentes en la data. En la ilustración se muestra cómo se deja la variable con 6 categorías que tienen más de 2000 registros.

Una de las variables con mayor número de categorías es CSU_Nombre, con 177 categorías lo que mostraba una dispersión de los datos muy alta para el modelo



	CSU_Nombre	Cantidad
1	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)	9292
2	BOGOTA DC - BOGOTA CHAPINERO (PG)	4414
3	ANTIOQUIA - MEDELLIN CAMPUS BARRIO LOS COLORES (PG)	3223
4	DESCONOCIDO	2542
5	BOGOTA DC - BOGOTA SUBA (PG)	2091
6	BOGOTA DC - BOGOTA RESTREPO (PG)	1944
7	BOGOTA DC - BOGOTA KENNEDY (PG)	1908
8	BOGOTA DC - BOGOTA COUNTRY (PG)	1483
9	BOGOTA DC - BOGOTA CITY CAMPUS (PG)	1193
10	BOGOTA DC - AMERICAS CENTRO COMERCIAL CARRERA (PG)	847

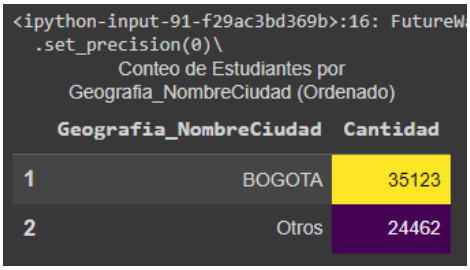
	CSU_Nombre	Cantidad
1	Otros	38023
2	BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG)	9292
3	BOGOTA DC - BOGOTA CHAPINERO (PG)	4414
4	ANTIOQUIA - MEDELLIN CAMPUS BARRIO LOS COLORES (PG)	3223
5	DESCONOCIDO	2542
6	BOGOTA DC - BOGOTA SUBA (PG)	2091

Ilustración 55. *CSU principales por número de registros*
Fuente: Elaboración propia

En la ilustración se muestra la necesidad de agrupar las categorías por las más representativas, así que se procede a deja únicamente las categorías con más de 2000 registros y se crea una categoría de Otros, que consolida todas las categorías con registros inferiores a 2000.

Otra variable con más de 700 categoría es la Geografia_Nombreciudad, por lo que debe ser también agrupada con las principales categorías y con mayor cantidad de registros.

	Geografia_NombreCiudad	Cantidad
1	BOGOTA	35123
2	MEDELLIN	3226
3	DESCONOCIDO	2732
4	SOACHA	1077
5	CALI	705
6	BELLO	567
7	ZIPAQUIRA	500
8	MOSQUERA	435
9	CHIA	402
10	ITAGUI	346



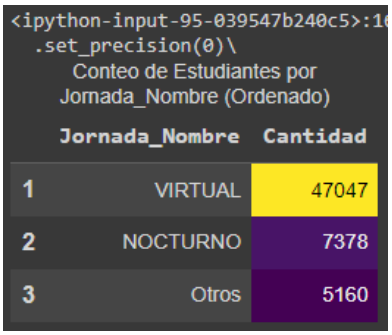
	Geografia_NombreCiudad	Cantidad
1	BOGOTA	35123
2	Otros	24462

Ilustración 56. *Geografia_Nombreciudad por número de registros*
Fuente: Elaboración propia

En la ilustración se ve lo disperso de los datos entre BOGOTA y las demás, por esto solo se dejan 2 categorías BOGOTA y Otras para centralizar los registros que se tienen en esta variable.

La variable jornada muestra una gran dispersión en sus diferentes categorías, por lo que también es necesario la agrupación de los datos.

	Jornada_Nombre	Cantidad
1	VIRTUAL	47047
2	NOCTURNO	7378
3	DIURNO	4778
4	FIN DE SEMANA	299
5	MARTES	83



	Jornada_Nombre	Cantidad
1	VIRTUAL	47047
2	NOCTURNO	7378
3	Otros	5160

Ilustración 57. *Jornada por número de registros*
Fuente: Elaboración propia

La ilustración muestra que existen 5 categorías, pero con una frecuencia muy baja lo que plantea agrupar las categorías con registros menores a 5000 registros en una sola como otros.

La variable estado civil tiene 7 categorías con una dispersión muy alta, por lo que se necesita agrupar estos datos por debajo de 500 en una categoría como otros.

	Persona_EstadoCivil	Cantidad
1	SOLTERO(A)	43584
2	CASADO(A)	8233
3	UNION LIBRE	7317
4	DIVORCIADO(A)	359
5	VIUDO(A)	56
6	SEPARADO(A)	32
7	RELIGIOSO(A)	4

<ipython-input-98-325928c86c35>:16: Futu
.set_precision(0)\
Conteo de Estudiantes por
Persona_EstadoCivil (Ordenado)

	Persona_EstadoCivil	Cantidad
1	SOLTERO(A)	43584
2	CASADO(A)	8233
3	UNION LIBRE	7317
4	Otros	451

Ilustración 58. *Estado Civil por número de registros*
Fuente: Elaboración propia

En la ilustración muestra cómo se crea la nueva categoría de se mantienen 4 categorías que representa una dispersión menor con los datos ya agrupados.

La variable convenio tipo muestra múltiples categorías donde se evidencia una dispersión en los datos alta, lo que hace necesario dejar los que tiene una frecuencia superior de 2000.

	Convenio_Tipo	Cantidad
1	SIN CONVENIO	15513
2	ARTICULACION	12633
3	SENA ARTICULACION	9528
4	SENA	8296
5	REGULAR	5983
6	PRIMERA VEZ	1584
7	NORMALISTAS	1028
8	DESCONOCIDO	832
9	PONAL	787
10	COMERCIAL	583

<ipython-input-101-efa65faab560>:16:
.set_precision(0)\
Conteo de Estudiantes por
Convenio_Tipo(Ordenado)

	Convenio_Tipo	Cantidad
1	SENA	17824
2	SIN CONVENIO	15513
3	ARTICULACION	12633
4	Otros	7632
5	REGULAR	5983

Ilustración 59. *Convenio por número de registros*
Fuente: Elaboración propia

En la ilustración se ve la dispersión en los datos con las tres primeras categorías de la variable y la necesidad de unificar las dos que contiene datos de Sena, por lo anterior se agrupan Sena articulación y Sena después del entendimiento del negocio, y se crea una categoría como Otros para unificar los valores inferiores a 2000.

Luego de este manejo a los datos para agrúpalos y manejar los más dispersos en cada variable, se identifican las variables que no están generando valor para el análisis de la información y posteriormente para la construcción del modelo.

La variable materias perdidas presenta una gran dispersión en los datos, por lo que se convierte en categórica y se deja solo en dos categorías SI perdió materias y NO perdió materias.

```
[108] # Convertimos la columna 'Materias Perdidas' a una categoría 'SI'/'NO'
      data['Materias Perdidas'] = data['Materias Perdidas'].apply(lambda x: 'NO' if x == 0 else 'SI')

      print(data["Materias Perdidas"].value_counts())

      NO    33587
      SI    25998
      Name: Materias Perdidas, dtype: int64
```

Ilustración 60. *Materias perdidas convertidas en categórica*
Fuente: Elaboración propia

La variable materias homologadas y promedio medio presenta una interesante correlación que permite la creación de una nueva variable, luego de la multiplicación entre sí de estas. Así que, creamos una nueva variable llamada Promedio_x_MateriasHomologadas.

```
data['Promedio_x_MateriasHomologadas'] = data['Promedio'] * data['Materias Homologadas']
```

Ilustración 61. *Creación de nueva variable*
Fuente: Elaboración propia

La variable Materias homologadas también se convierte en categórica y se deja en dos Categorías, materias homologadas SI y materias homologadas NO.

```
[110] # Convertimos la columna 'Materias Homologadas' a una categoría 'SI'/'NO'
      data['Materias Homologadas'] = data['Materias Homologadas'].apply(lambda x: 'NO' if x == 0 else 'SI')

      print(data["Materias Homologadas"].value_counts())

      NO    31675
      SI    27910
      Name: Materias Homologadas, dtype: int64
```

Ilustración 62. *Materias Homologadas convertidas en categórica*
Fuente: Elaboración propia

La variable AÑO es una variable numérica, que por su naturaleza se convertirá en categórica. Luego de la validación de los datos y el comportamiento con los datos de materias homologadas y materias de 2019 a 2022. Se determinó con la universidad un cambio en políticas de homologación y de mallas curriculares, lo que lleva a tomar como referencia los datos de 2019 a 2022, para la creación del modelo predictivo para una mejor efectividad. La variable año luego de filtrar la información es eliminada, ya que esta variable no se tiene con los estudiantes que van a presentar la prueba saber pro.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59585 entries, 0 to 59584
Data columns (total 30 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Documento                       59585 non-null  int64
 1   PIDM                            59585 non-null  int64
 2   Puntaje global                  59585 non-null  int64
 3   Percentil nacional global       59585 non-null  int64
 4   Materias                        59585 non-null  int64
 5   Promedio                        59585 non-null  float64
 6   Materias Homologadas            59585 non-null  object
 7   Materias Perdidas               59585 non-null  object
 8   EK                              59585 non-null  object
 9   AÑO                             59585 non-null  int64
 10  Programa_Codigo                 59585 non-null  int64
 11  Programa_Facultad               59585 non-null  object
 12  Programa_Nombre                 59585 non-null  object
 13  Programa_NivelAcademico         59585 non-null  object
 14  Programa_Modalidad              59585 non-null  object
 15  Convenio_Homolog                59585 non-null  object
 16  Convenio_Tipo                   59585 non-null  object
 17  Jornada_Nombre                  59585 non-null  object
 18  Sede_Nombre                     59585 non-null  object
 19  CSU_Nombre                      59585 non-null  object
 20  Geografia_NombreCiudad          59585 non-null  object
 21  Persona_TipoIdentificacion      59585 non-null  object
 22  Persona_EstadoCivil             59585 non-null  object
 23  Persona_Genero                  59585 non-null  object
 24  Persona_FechaNacimiento         59585 non-null  datetime64[ns]
 25  CSU_Region                      59585 non-null  object
 26  Edad                            59585 non-null  int64
 27  Edad_2                          59585 non-null  int64
 28  CLASE_1                         59585 non-null  object
 29  Promedio_x_MateriasHomologadas  59585 non-null  float64
dtypes: datetime64[ns](1), float64(2), int64(9), object(18)
memory usage: 13.6+ MB
```

Ilustración 63. *Tipos de Variables dentro del set de datos*
Fuente: Elaboración propia

La ilustración muestra las variables existentes y se identifica luego de los análisis realizados que las variables Documento, PIDM, EK y Persona tipo de Identificación, son variables que ayudan con la identificación del estudiante dentro de la institución pero que no deben ser tenidas en cuenta para el modelo por su naturaleza.

Programa nivel académico por ser una variable con una categoría única que es pregrado, no aporta información relevante para el análisis y debe ser descartada también.

Las variables Puntaje Global y Percentil nacional global, son variables que son obtenidas luego de la presentación de las pruebas y no se tendrán por cada estudiante antes de ello, por esto esta variable es la variable objetivo del modelo mediante la clasificación de puntaje global, se obtuvo la variable objetivo CLASE_1 que muestra dos clases bajo y medio. Esta variable clasifica el puntaje global de lo que está por debajo del promedio nacional, en el año 2022 (<145), sea baja y lo que está por encima (>145) del promedio nacional sea media.

La variable fecha de nacimiento y edad, estas dos ya están representadas en el modelo como EDAD_2, por lo que al volver a usarlas puede perjudicar el buen funcionamiento del modelo.

Con base en el análisis anterior se dejarán las siguientes variables, que se observan en la ilustración.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59585 entries, 0 to 59584
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Materias                        59585 non-null  int64
 1   Promedio                        59585 non-null  float64
 2   Materias Homologadas            59585 non-null  object
 3   Materias Perdidas               59585 non-null  object
 4   AÑO                             59585 non-null  int64
 5   Programa_Codigo                 59585 non-null  int64
 6   Programa_Facultad               59585 non-null  object
 7   Programa_Nombre                 59585 non-null  object
 8   Programa_Modalidad              59585 non-null  object
 9   Convenio_Homolog                59585 non-null  object
 10  Convenio_Tipo                   59585 non-null  object
 11  Jornada_Nombre                  59585 non-null  object
 12  Sede_Nombre                     59585 non-null  object
 13  CSU_Nombre                      59585 non-null  object
 14  Geografia_NombreCiudad          59585 non-null  object
 15  Persona_EstadoCivil             59585 non-null  object
 16  Persona_Genero                  59585 non-null  object
 17  CSU_Region                      59585 non-null  object
 18  Edad_2                          59585 non-null  int64
 19  CLASE_1                         59585 non-null  object
 20  Promedio_x_MateriasHomologadas  59585 non-null  float64
dtypes: float64(2), int64(4), object(15)
memory usage: 9.5+ MB
```

Ilustración 64. *Variables del set de datos*
Fuente: Elaboración propia

7.3.2.1. ANÁLISIS UNIVARIADO Y COMPUESTO DESPUÉS DEL PREPROCESAMIENTO

Luego del tratamiento a cada una de las variables que se consideran para el modelo, se volverán a realizar los gráficos para su análisis.

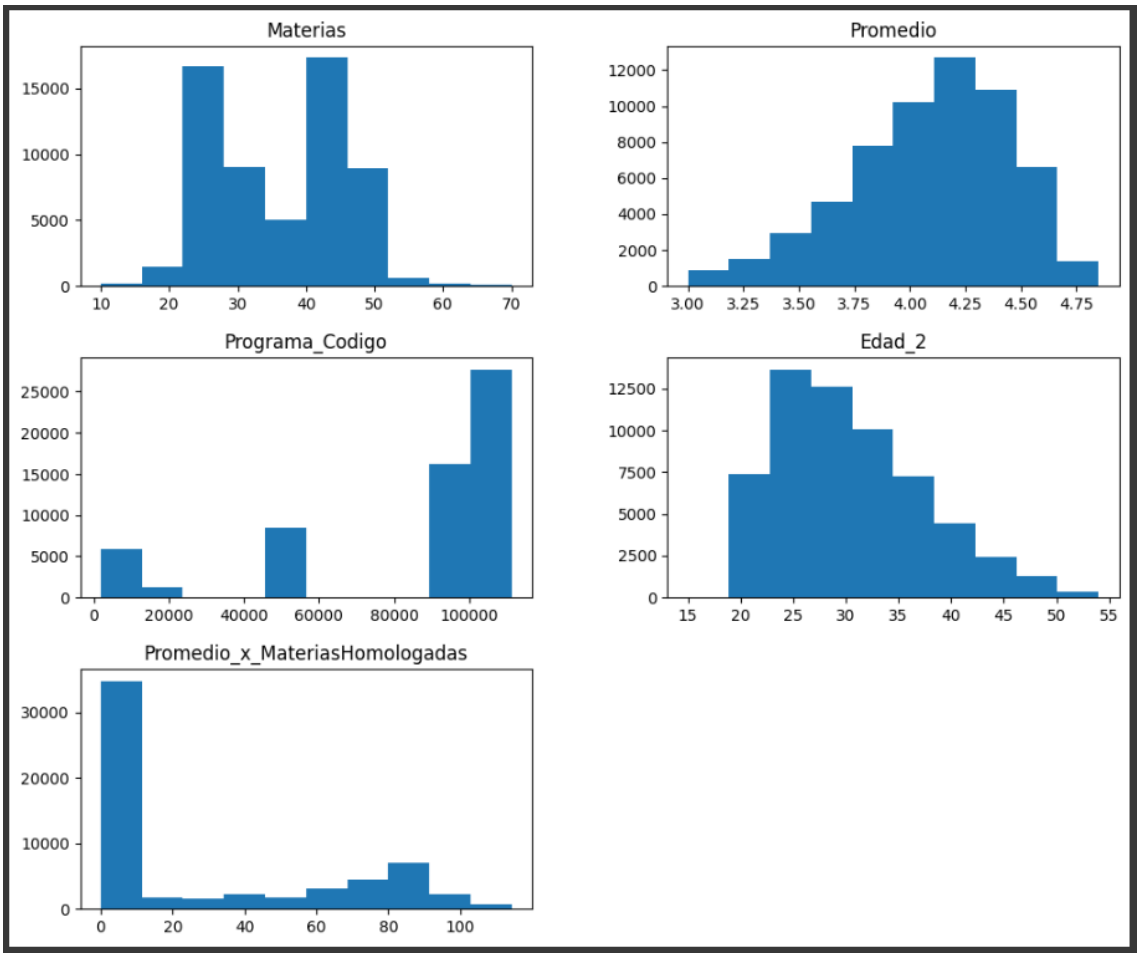
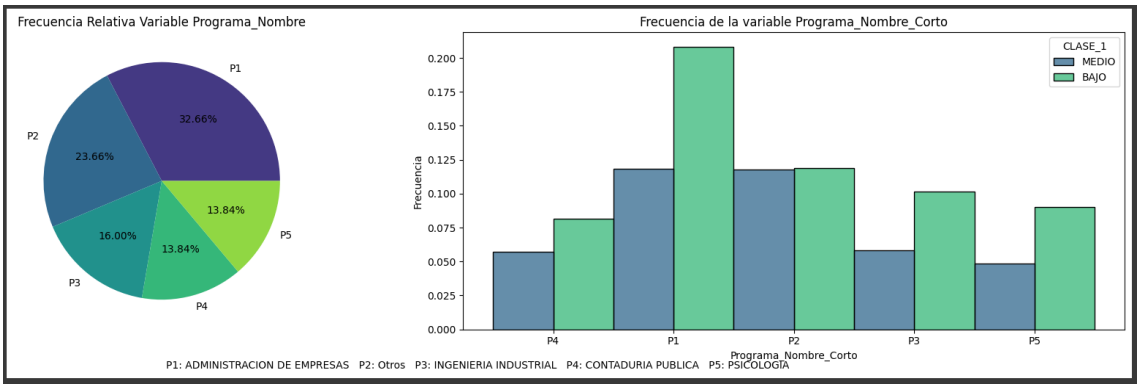
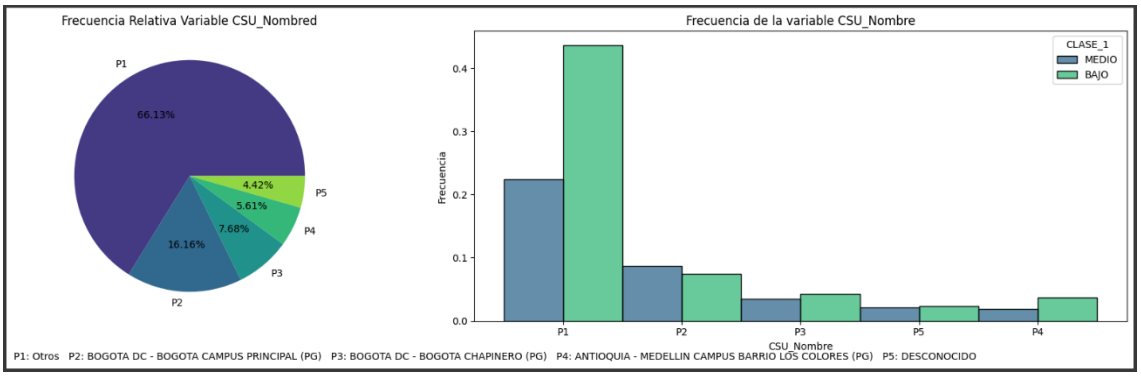
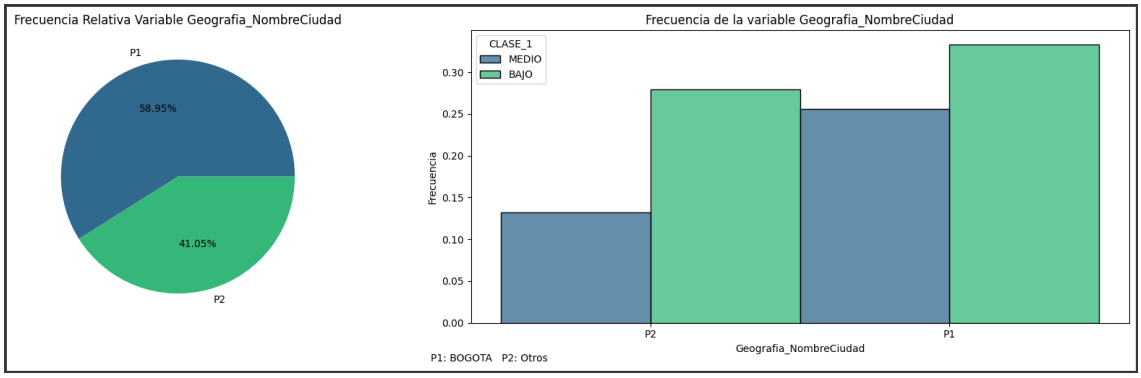


Ilustración 65. *Análisis Univariado de las variables tratadas*
Fuente: Elaboración propia


Frecuencia Relativa Variable Programa_Nombre
P1
P2
P3
P4
P5
32.66%
23.66%
16.00%
13.84%
13.84%
Frecuencia de la variable Programa_Nombre_Corto
CLASE_1
MEDIO
BAJO
Frecuencia
Programa_Nombre_Corto
P1: ADMINISTRACION DE EMPRESAS P2: Otros P3: INGENIERIA INDUSTRIAL P4: CONTADURIA PUBLICA P5: PSICOLOGIA


Frecuencia Relativa Variable CSU_Nombred
P1
P2
P3
P4
P5
66.13%
16.16%
7.68%
5.61%
4.42%
Frecuencia de la variable CSU_Nombre
CLASE_1
MEDIO
BAJO
Frecuencia
CSU_Nombre
P1: Otros P2: BOGOTA DC - BOGOTA CAMPUS PRINCIPAL (PG) P3: BOGOTA DC - BOGOTA CHAPINERO (PG) P4: ANTIOQUIA - MEDELLIN CAMPUS BARRIO LOS COLORES (PG) P5: DESCONOCIDO


Frecuencia Relativa Variable Geografia_NombreCiudad
P1
P2
58.95%
41.05%
Frecuencia de la variable Geografia_NombreCiudad
CLASE_1
MEDIO
BAJO
Frecuencia
Geografia_NombreCiudad
P1: BOGOTA P2: Otros

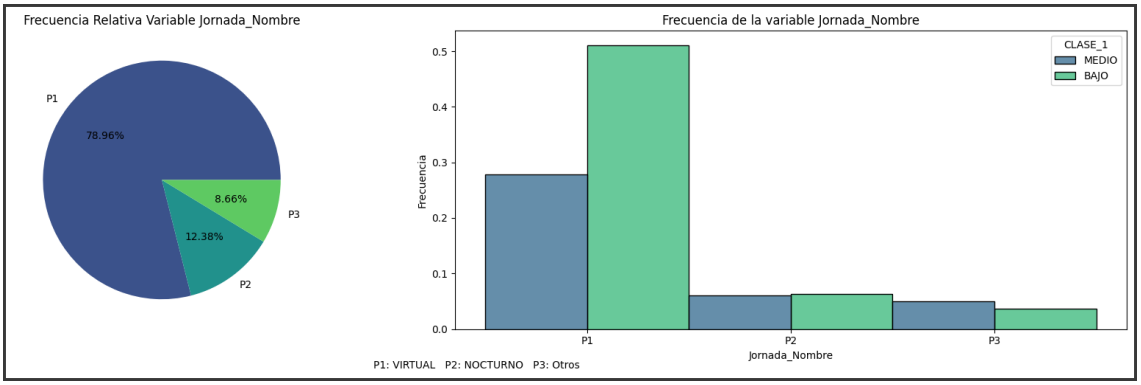

Frecuencia Relativa Variable Jornada_Nombre
P1
P2
P3
78.96%
12.38%
8.66%
Frecuencia de la variable Jornada_Nombre
CLASE_1
MEDIO
BAJO
Frecuencia
Jornada_Nombre
P1: VIRTUAL P2: NOCTURNO P3: Otros

Ilustración 66. *Análisis Univariado de las variables por Frecuencia relativa y Frecuencia*
Fuente: Elaboración propia

7.4. FASE DE MODELADO

7.4.1. SELECCIÓN TÉCNICA DE MODELADO

En esta fase se eligen y prueban técnicas de modelado que ayuden al objetivo propuesto que es predecir la mayor cantidad de estudiantes en clase 1(Bajo), para la presentación de las pruebas saber pro.

Se realizará la división de la base ya depurada y lista en datos de entrenamiento y de prueba, se verificará con apoyo de las matrices de confusión e indicadores de eficiencia sobre los diferentes modelos, evaluando su capacidad para predecir.

7.4.2. DISEÑO Y CONSTRUCCIÓN DE MODELOS

En esta fase se buscará apoyo en la librería pycaret que de manera eficiente evalúa 15 modelos de clasificación y nos muestra sus indicadores principales, con base en ello, se escogerán los 6 con los mejores resultados y luego se realizará el entrenamiento y prueba con estos.

	Description	Value
0	Session id	123
1	Target	CLASE_1
2	Target type	Binary
3	Original data shape	(33358, 40)
4	Transformed data shape	(33358, 40)
5	Transformed train set shape	(23350, 40)
6	Transformed test set shape	(10008, 40)
7	Numeric features	39
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	1c42

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.6800	0.6769	0.9184	0.6892	0.7875	0.1898	0.2204	
ridge	Ridge Classifier	0.6784	0.0000	0.9152	0.6889	0.7861	0.1873	0.2218	
lda	Linear Discriminant Analysis	0.6782	0.6607	0.9067	0.6912	0.7844	0.1932	0.2238	
ada	Ada Boost Classifier	0.6774	0.6609	0.9042	0.6913	0.7835	0.1928	0.2222	
lightgbm	Light Gradient Boosting Machine	0.6773	0.6727	0.8944	0.6941	0.7816	0.1987	0.2254	
lr	Logistic Regression	0.6754	0.6557	0.9120	0.6874	0.7839	0.1803	0.2130	
xgboost	Extreme Gradient Boosting	0.6680	0.6550	0.8575	0.6970	0.7693	0.1990	0.2131	
nb	Naive Bayes	0.6601	0.6217	0.9202	0.6732	0.7776	0.1252	0.1580	
rf	Random Forest Classifier	0.6466	0.6346	0.8046	0.6957	0.7462	0.1746	0.1797	
dummy	Dummy Classifier	0.6457	0.5000	1.0000	0.6457	0.7847	0.0000	0.0000	
knn	K Neighbors Classifier	0.6278	0.5999	0.7802	0.6863	0.7302	0.1382	0.1412	
qda	Quadratic Discriminant Analysis	0.6254	0.6480	0.7251	0.6335	0.6762	0.1797	0.1814	
et	Extra Trees Classifier	0.6246	0.6101	0.7542	0.6920	0.7218	0.1480	0.1494	
dt	Decision Tree Classifier	0.5909	0.5563	0.6752	0.6861	0.6806	0.1117	0.1117	
svm	SVM - Linear Kernel	0.5039	0.0000	0.4922	0.5639	0.4014	0.0242	0.0348	

Ilustración 67. *Resultados de la Librería Pycaret*
Fuente: Elaboración propia

En la anterior ilustración se ve como luego de una comparación de 15 modelos, pycaret muestra los principales modelos que presentan números favorables al trabajar con nuestra base de datos.

Con base en el ejercicio anterior se escogen los siguientes modelos para el modelo predictivo:

1. **Gradient Boosting Classifier**
2. **Ridge Classifier**
3. **Linear Discriminant Analysis**
4. **Ada Boost Classifier**
5. **Light Gradient Boosting Machine**
6. **Logistic Regression**
7. **Extreme Gradient Boosting**

7.4.2.1. DIVISIÓN CONJUNTO DE DATOS

En esta fase se determina que la variable objetivo a predecir CLASE_1 debe estar aparte del conjunto de datos.

Después de lo planteado se realizará la partición de la data en dos bases, una de entrenamiento con el 70% de los registros y otra de prueba con el 30% de los registros.

3	Original data shape	(33358, 40)
4	Transformed data shape	(33358, 40)
5	Transformed train set shape	(23350, 40)
6	Transformed test set shape	(10008, 40)

Ilustración 68. *Resultados de la división de datos.*
Fuente: Elaboración propia

7.4.2.2. MATRICES DE CONFUSIÓN

La matriz de confusión es una herramienta que se utiliza en los modelos de clasificación para evaluar el rendimiento del modelo, a través de una tabla que muestra los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) de un clasificador en relación con las clases verdaderas.

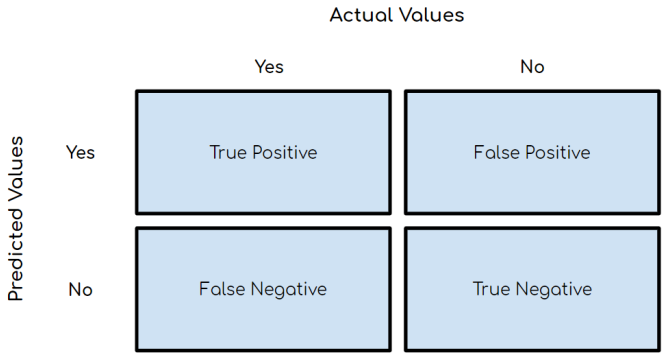


Ilustración 69. *Comprensión de la matriz de confusión y cómo implementarla en Python"*
Fuente: datasource.ai.

7.4.2.2.1. GRADIENT BOOSTING CLASSIFIER

Se analizará la matriz de confusión del modelo Gradient Boosting Classifier, centrándose en la clasificación de los casos pertenecientes a la clase 1, que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

```
Accuracy: 0.6821
Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.24      0.34      2350
           1       0.69      0.92      0.79      4322

    accuracy                           0.68      6672
   macro avg       0.66      0.58      0.57      6672
weighted avg       0.67      0.68      0.63      6672
```

```
El AUC del modelo es: 0.6838932921125955
```

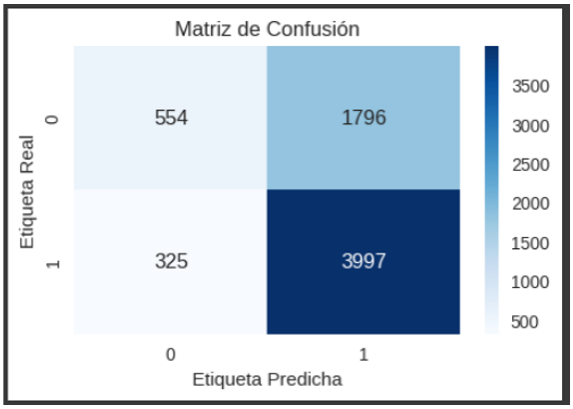


Ilustración 70. *Métricas de rendimiento Gradient Boosting Classifier*
Fuente: Elaboración propia

Los resultados del modelo muestran un nivel de precisión general del 68.21%. Sin embargo, al analizar la clasificación por clases se observa que la precisión para la clase 0 es baja con solo un 34% de f1-score, mientras que para la clase 1 es más alta con un 79% de f1-score. Esto sugiere que el modelo tiene dificultades para predecir la clase 0 correctamente, aunque tiene un mejor desempeño con la clase 1.

7.4.2.2.2. RIDGE CLASSIFIER

Se analizará la matriz de confusión del modelo Ridge Classifier, centrándose en la clasificación de los casos pertenecientes a la clase 1, que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

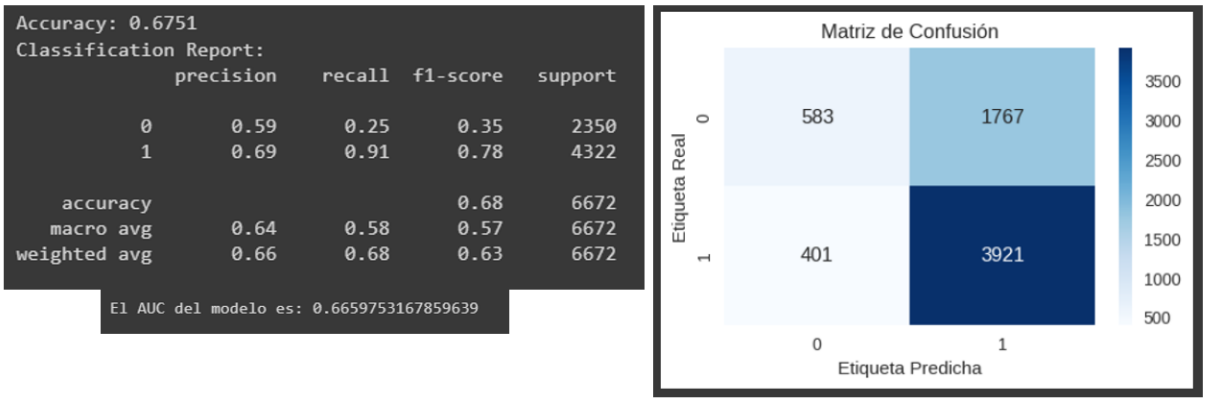


Ilustración 71. *Métricas de rendimiento Ridge Classifier*
Fuente: Elaboración propia

Los resultados del modelo indican una precisión general del 67.51%. Se observa que la precisión para la clase 0 es baja, con un f1-score del 35%, mientras que para la clase 1 es más alta, con un f1-score del 78%. Esto sugiere, que el modelo también tiene dificultades para predecir correctamente la clase 0, pero tiene un mejor desempeño con la clase 1.

7.4.2.2.3. LINEAR DISCRIMINANT ANALYSIS

Se analizará la matriz de confusión del modelo *Linear Discriminant Analysis*, centrándose en la clasificación de los casos pertenecientes a la clase 1, que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

```
Accuracy: 0.6734
Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.26      0.36      2350
           1       0.69      0.90      0.78      4322

    accuracy                           0.67      6672
   macro avg       0.64      0.58      0.57      6672
weighted avg       0.65      0.67      0.63      6672
```

```
El AUC del modelo es: 0.6658125670739512
```

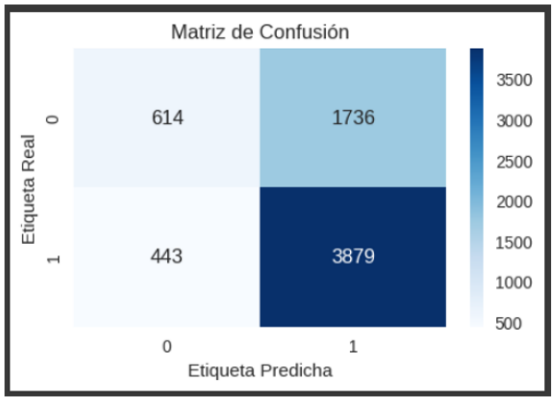


Ilustración 72. *Métricas de rendimiento Linear Discriminant Analysis*
Fuente: Elaboración propia

Los resultados del modelo muestran una precisión global del 67.34%. Sin embargo, al analizar la clasificación por clases, se observa que la precisión para la clase 0 es baja con un f1-score del 36%, mientras que para la clase 1 es más alta con un f1-score del 78%. Esto sugiere que el modelo tiene dificultades para predecir correctamente la clase 0, aunque tiene un mejor desempeño con la clase 1.

7.4.2.2.4. ADA BOOST CLASSIFIER

Se analizará la matriz de confusión del modelo *Ada Boost Classifier*, centrándose en la clasificación de los casos pertenecientes a la clase 1 que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

```
Accuracy: 0.6814
Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.27      0.37      2350
           1       0.69      0.91      0.79      4322

    accuracy                           0.68      6672
   macro avg       0.65      0.59      0.58      6672
weighted avg       0.66      0.68      0.64      6672
```

```
El AUC del modelo es: 0.6731072100190022
```

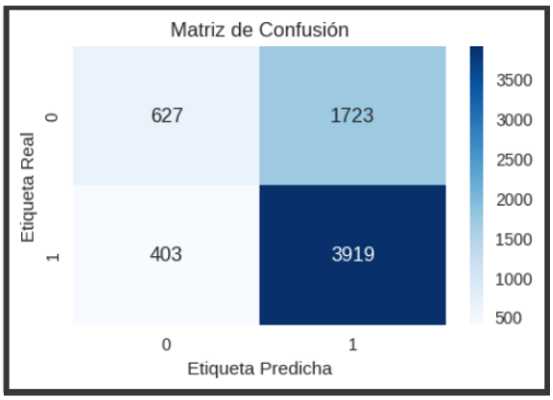


Ilustración 73. *Métricas de rendimiento Ada Boost Classifier*
Fuente: Elaboración propia

El modelo presenta una precisión global del 68.14%. Al desglosar la clasificación por clases se observa que la precisión para la clase 0 es del 37%, mientras que para la clase

1 es del 79%. Esto sugiere que el modelo tiene dificultades para predecir correctamente la clase 0, aunque tiene un mejor desempeño con la clase 1.

7.4.2.2.5. LIGHT GRADIENT BOOSTING MACHINE

Se analizará la matriz de confusión del modelo *Light Gradient Boosting Machine,* centrándose en la clasificación de los casos pertenecientes a la clase 1 que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

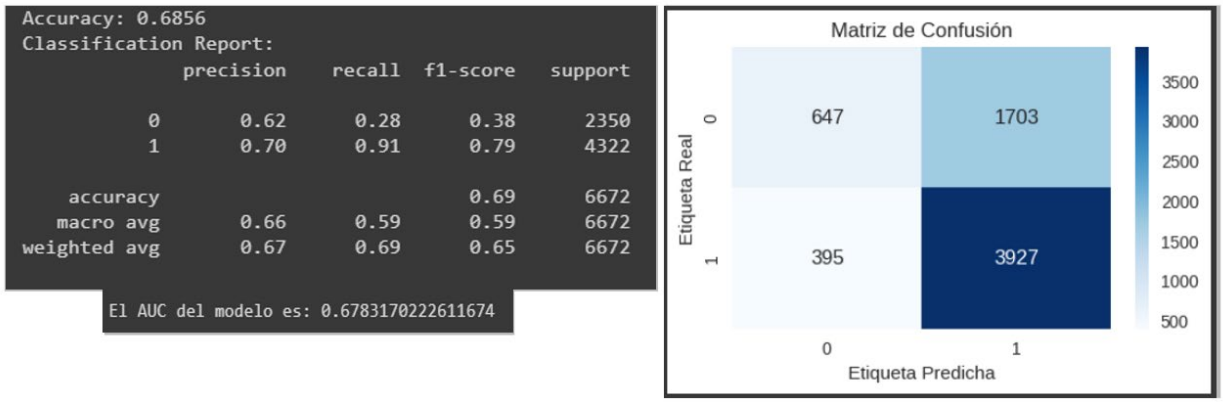


Ilustración 74. *Métricas de rendimiento Light Gradient Boosting Machine*
Fuente: Elaboración propia

El modelo actual alcanza una precisión global del 68.56%. Al analizar la clasificación por clases se observa que la precisión para la clase 0 es del 38%, mientras que para la clase 1 es del 79%. En resumen, el modelo muestra un rendimiento moderado, con una precisión general mejorada y un desempeño más equilibrado entre las clases.

7.4.2.2.6. LOGISTICREGRESSION

Se analizará la matriz de confusión del modelo *LogisticRegression*, centrándose en la clasificación de los casos pertenecientes a la clase 1 que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

```
Accuracy: 0.6709
Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.21      0.31      2350
           1       0.68      0.92      0.78      4322

    accuracy                           0.67      6672
   macro avg       0.64      0.57      0.55      6672
weighted avg       0.65      0.67      0.62      6672

El AUC del modelo es: 0.646941378597379
```

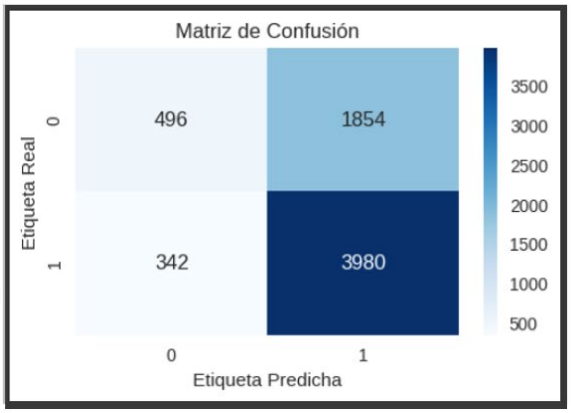


Ilustración 75. *Métricas de rendimiento LogisticRegression*
Fuente: Elaboración propia

El modelo actual alcanza una precisión global del 67.09%. Al desglosar la clasificación por clases se observa que la precisión para la clase 0 es del 31%, mientras que para la clase 1 es del 78%. Esto indica que el modelo tiene dificultades para predecir correctamente la clase 0, mostrando un rendimiento particularmente bajo en términos de recall y f1-score para esta clase. En general, el modelo muestra un rendimiento moderado.

7.4.2.2.7. EXTREME GRADIENT BOOSTING

Se analizará la matriz de confusión del modelo *Extreme Gradient Boosting*, centrándose en la clasificación de los casos pertenecientes a la clase 1 que representan los estudiantes que no tienen el mejor desempeño en las pruebas. Se evaluarán los indicadores de desempeño ajustados específicamente hacia esta clase, con el objetivo de comprender cómo el modelo predice y clasifica estos casos en comparación con su estado real.

```
Accuracy: 0.6740
Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.32      0.41      2350
           1       0.70      0.86      0.77      4322

    accuracy                           0.67      6672
   macro avg       0.63      0.59      0.59      6672
weighted avg       0.65      0.67      0.65      6672

El AUC del modelo es: 0.6652154735297882
```

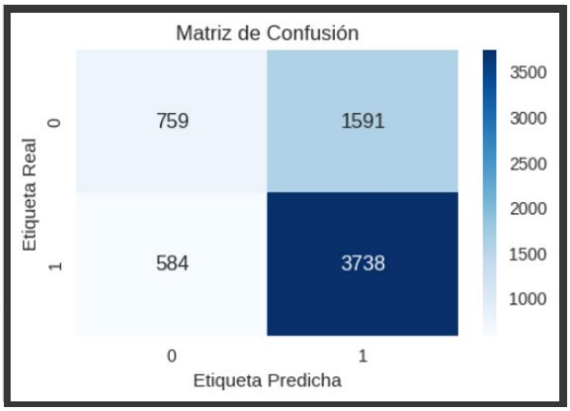


Ilustración 76. *Métricas de rendimiento Extreme Gradient Boosting*
Fuente: Elaboración propia

El modelo actual alcanza una precisión global del 67.40%. Al desglosar la clasificación por clases, se observa que la precisión para la clase 0 es del 41%, mientras que para la clase 1 es del 77%. Esto indica que el modelo tiene dificultades para predecir correctamente la clase 0, aunque tiene un mejor desempeño con la clase 1.

7.4.2.3. CURVA DE ROC Y ÁREA BAJO LA CURVA AUC

La Curva ROC (*Receiver Operating Characteristic*) y el Área bajo la Curva (AUC) son herramientas usadas para evaluar el rendimiento de los modelos de clasificación en *machine learning*.

La curva ROC es un gráfico que muestra la relación entre la tasa de verdaderos positivos (TPR), también conocida como sensibilidad, y la tasa de falsos positivos (FPR), también conocida como 1 - especificidad.

En el eje x (horizontal) se representa la tasa de falsos positivos (FPR), mientras que en el eje y (vertical) se representa la tasa de verdaderos positivos (TPR).

Cada punto en la curva representa un umbral de clasificación diferente.

Una curva ROC ideal se acercaría al vértice superior izquierdo del gráfico, lo que indicaría una alta sensibilidad y una baja tasa de falsos positivos.

Área bajo la Curva (AUC):

El área bajo la curva (AUC) es una métrica que cuantifica el rendimiento global de un modelo de clasificación.

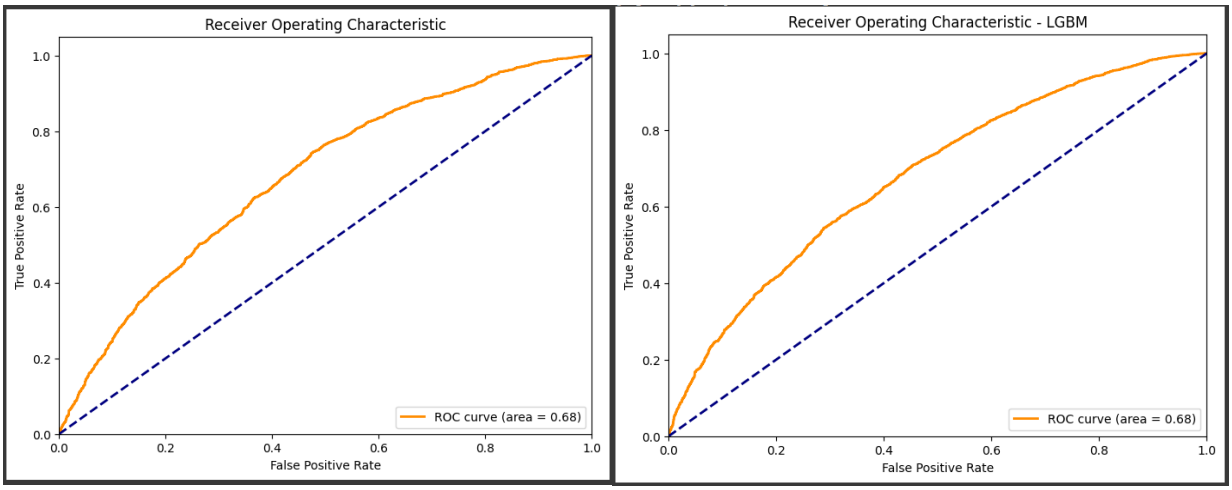
Representa la capacidad del modelo para distinguir entre las clases positiva y negativa.

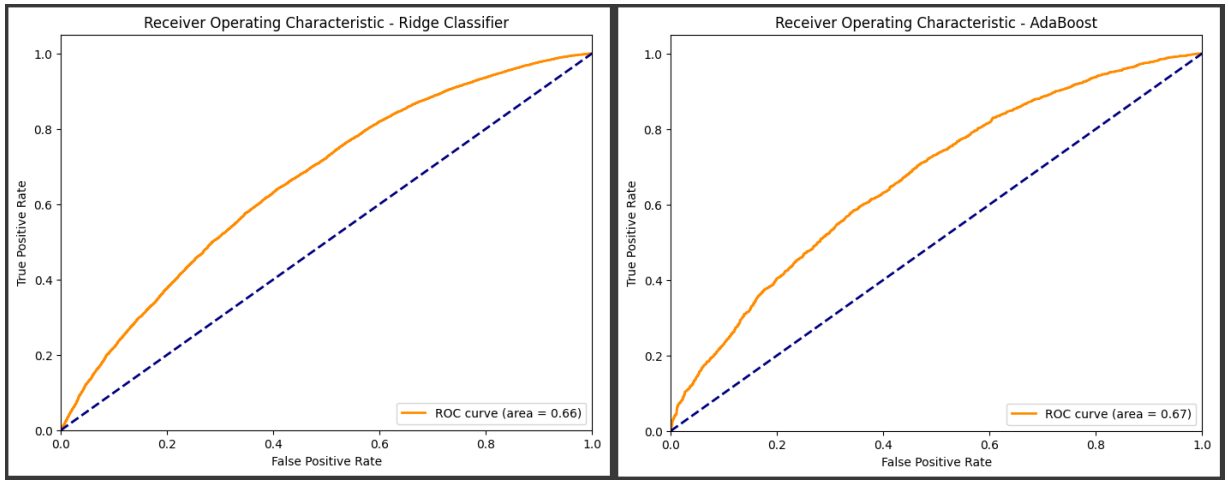
Un valor de AUC cercano a 1 indica un modelo muy efectivo, donde la mayoría de las veces la clase positiva tiene una probabilidad más alta de ser clasificada correctamente que la clase negativa.

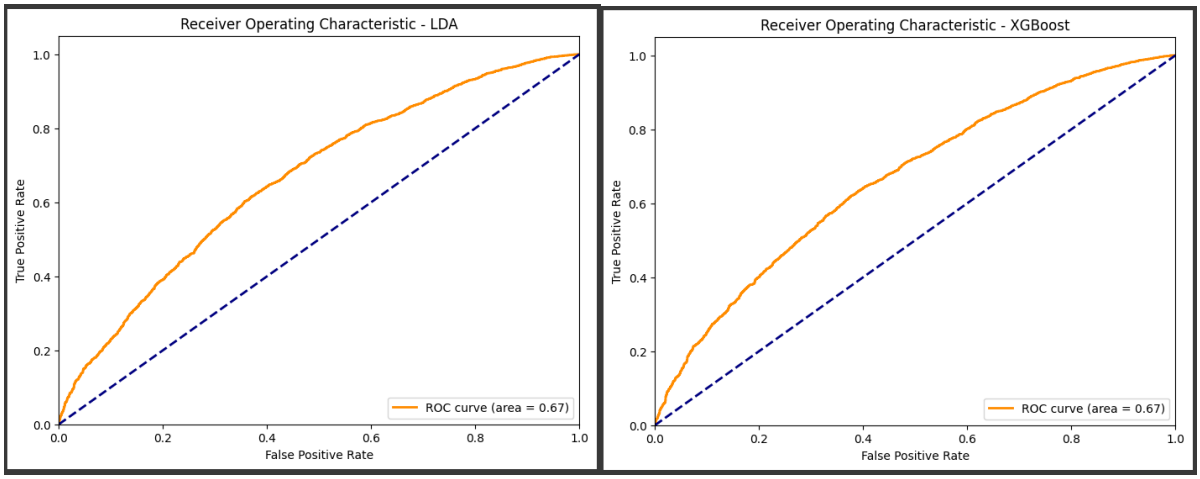
Un valor de AUC cercano a 0.5 indica que el modelo tiene un rendimiento similar al azar, mientras que un valor inferior a 0.5 indica un rendimiento peor que el azar.

Cuanto mayor sea el valor de AUC, mejor será el rendimiento del modelo.

En resumen, la Curva ROC y el Área bajo la Curva AUC son herramientas útiles para evaluar la capacidad de un modelo de clasificación para discriminar entre clases y para comparar diferentes modelos entre sí.


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.68)
Receiver Operating Characteristic - LGBM
True Positive Rate
False Positive Rate
ROC curve (area = 0.68)


Receiver Operating Characteristic - Ridge Classifier
True Positive Rate
False Positive Rate
ROC curve (area = 0.66)
Receiver Operating Characteristic - AdaBoost
True Positive Rate
False Positive Rate
ROC curve (area = 0.67)


Receiver Operating Characteristic - LDA
True Positive Rate
False Positive Rate
ROC curve (area = 0.67)
Receiver Operating Characteristic - XGBoost
True Positive Rate
False Positive Rate
ROC curve (area = 0.67)

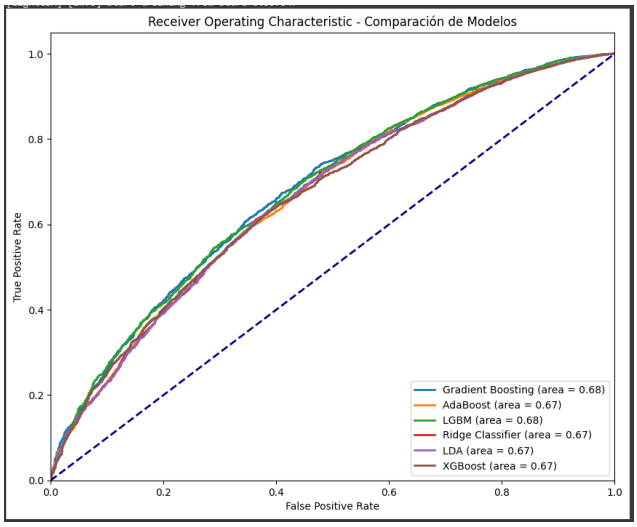


Ilustración 77. *Gráficos de la curva ROC por los principales modelos*
Fuente: Elaboración propia

Se resume de la ilustración 77 que los modelos presentan una leve diferencia, sin embargo, los modelos Gradient Boosting y LGBM presenta un mejor rendimiento frente a los otros modelos usados.

7.4.3. EVALUACIÓN DE MODELOS

Con base en los 7 modelos desplegados se realiza la tipificación de los resultados, para analizar las métricas de cada modelo y ver su desempeño con base en los objetivos propuestos en este trabajo y escoger el mejor modelo para la solución del problema.

7.4.3.1. INDICADORES DE DESEMPEÑO

- **Precisión:** En el contexto de la clasificación, la precisión es una medida que evalúa la proporción de predicciones correctas realizadas por un modelo de aprendizaje automático. Específicamente, la precisión se refiere a la proporción de muestras clasificadas correctamente como positivas (verdaderos positivos) en comparación con el número total de muestras clasificadas como positivas (verdaderos positivos + falsos positivos). Cuanto más precisos sean los resultados falsos, más confianza tendrá en las buenas predicciones del modelo.

- **Recall:** También conocida como sensibilidad o integridad, es una medida que evalúa la capacidad de un modelo de aprendizaje automático para identificar correctamente todas las características de una clase. Se calcula como la proporción de muestras positivas identificadas correctamente (verdaderos positivos) en comparación con el número total de eventos positivos en los datos (verdaderos positivos + falsos negativos). Una mayor recuperación significa que es más probable que el modelo

capture todos los aspectos de la clase de interés, lo cual es especialmente importante en los casos en los que descartar buenas muestras puede resultar de gran beneficio.

- **Score F1:** Es una medida de la precisión de un modelo de aprendizaje automático que combina precisión y recuperación en un solo valor. Se calcula como el mejor promedio de precisión y recuperación, proporcionando una evaluación equilibrada del rendimiento del modelo en términos de precisión e integridad. Una puntuación F1 más alta indica un mejor equilibrio entre precisión y recuperación, lo que significa que el modelo funciona mejor al clasificar los mejores casos.

- **Exactitud:** es una métrica para evaluar un modelo de aprendizaje automático que mide el porcentaje de predicciones correctas realizadas por el modelo sobre el número total de casos. Se calcula como la proporción de casos positivos y negativos correctamente clasificados (verdaderos positivos más verdaderos negativos) sobre el número total de casos. La precisión proporciona una estimación general de la eficacia de un modelo para clasificar todos los casos, independientemente de la clase, y es particularmente útil cuando las clases están equilibradas en los datos.

- **El área bajo la curva (AUC)**: es una métrica utilizada para evaluar un modelo de clasificación binaria. Se calcula como el área bajo la curva ROC (*Receiver Operating Characteristic*), que representa el número de verdaderos positivos (devoluciones) en comparación con la proporción de falsos positivos (1 - especificidad) en diferentes umbrales de decisión. Un valor de AUC más alto indica un mejor rendimiento del modelo en la clasificación binaria, donde un AUC de 1 representa un modelo perfecto y 0,5 representa un rendimiento aleatorio. AUC es una métrica útil para comparar modelos y evaluar su capacidad para discriminar entre clases positivas y negativas.

Modelo	Precisión (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precisión (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Exactitud (Accuracy)	AUC
Gradient Boosting Classifier	63,0%	23,6%	34,3%	69,0%	92,5%	79,0%	68,2%	68,4%
Ridge Classifier	59,0%	25,0%	35,0%	69,0%	91,0%	78,0%	67,5%	66,6%
Linear Discriminant Analysis	58,0%	26,0%	36,0%	69,1%	90,0%	78,0%	67,3%	66,6%
Ada Boost Classifier	61,0%	27,0%	37,0%	69,0%	91,0%	79,0%	68,0%	67,3%
Light Gradient Boosting Machine	62,0%	28,0%	38,0%	70,0%	91,0%	79,0%	68,6%	67,8%
Logistic Regression	59,0%	21,0%	31,0%	68,0%	92,0%	78,0%	67,0%	64,7%
XGBoost	57,0%	32,0%	41,0%	70,0%	86,0%	77,0%	67,4%	66,5%

Ilustración 78. *Tabla de resultados de los modelos seleccionados.*
Fuente: Elaboración propia

En el desarrollo de este trabajo el objetivo es identificar los estudiantes que no van a tener un buen rendimiento en la presentación de las pruebas Saber Pro, esto con el ánimo de disminuir el riesgo de que el promedio global de la institución siga estando por

debajo del promedio nacional. Con base en lo anteriormente expuesto los indicadores del modelo que son más representativos para conseguir el objetivo son el *Recall* y F1-*Score*, centrándonos en los que van a tener un rendimiento bajo (Clase 1).

Para determinar cuál modelo tiene un mejor rendimiento, se suma la Exactitud y el AUC y se divide en dos, los resultados se comparan y con ello se determina que modelo presenta un mejor rendimiento**.**

Modelo	Exactitud (Accuracy)	AUC	Rendimiento
Gradient Boosting Classifier	68,2%	68,4%	68,3%
Ridge Classifier	67,5%	66,6%	67,1%
Linear Discriminant Analysis	67,3%	66,6%	67,0%
Ada Boost Classifier	68,0%	67,3%	67,7%
Light Gradient Boosting Machine	68,6%	67,8%	68,2%
Logistic Regression	67,0%	64,7%	65,8%
XGBoost	67,4%	66,5%	67,0%

Ilustración 79. *Tabla de rendimiento de los modelos*
Fuente: Elaboración propia

Validando las cifras los modelos de los modelos, de acuerdo con los resultados de las métricas de rendimiento son el Gradient Boosting Classifier y el Light Gradient Boosting Machine son las que presentan un mejor rendimiento, adicionalmente si revisamos los resultados identificando la clase 1 que es la de interés, en estos dos modelos tiene un buen comportamiento de predicción y Recall, particularmente para esta clase.

7.4.3.2. VARIABLES MÁS REPRESENTATIVAS EN LOS MODELOS

A continuación, se mostrará gráficamente las variables más importantes tenidas en cuenta para cada uno de los modelos implementados.

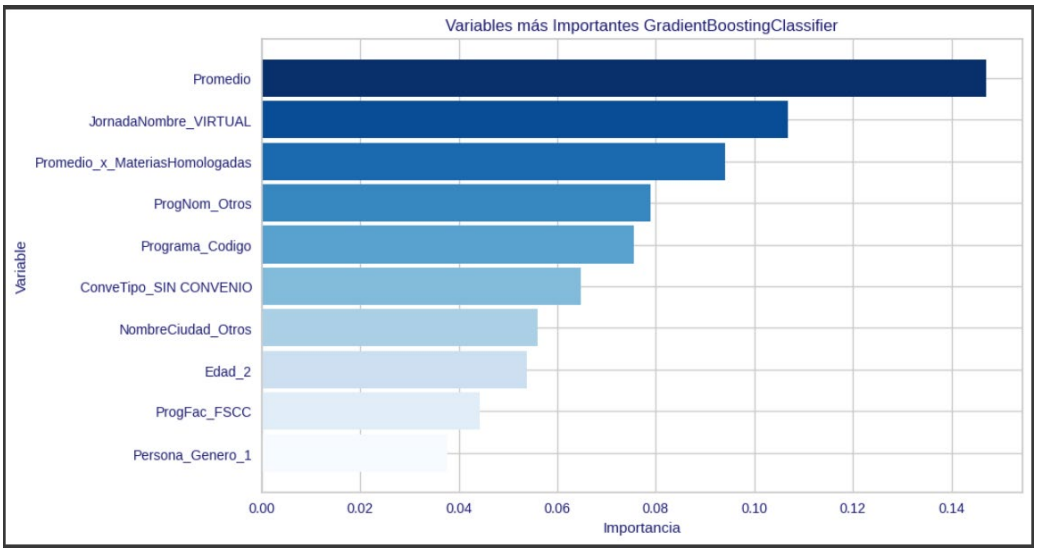


Ilustración 80. *Variables representativas en el modelo Gradient Boosting Classifier*
Fuente: Elaboración propia

En la ilustración 80, se observan las variables que el modelo *Gradient Boosting Classifier* determina importantes para predecir la variable objetivo. Para este modelo el promedio, la modalidad Virtual y la nueva variable que combina Promedio x Materias Homologadas del estudiante, son variables determinantes.

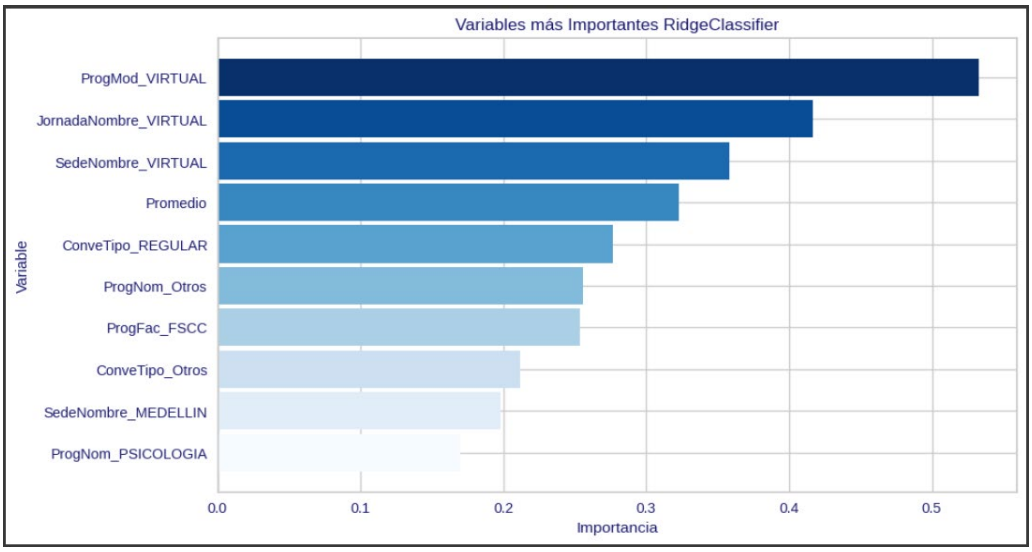


Ilustración 81. *Variables representativas en el modelo Ridge Classifier*
Fuente: Elaboración propia

En la ilustración 81, se observan las variables que el modelo *Ridge Classifier* determina importantes para predecir la variable objetivo. Para este modelo la modalidad virtual como sede, jornada y modalidad del estudiante, son variables determinantes.

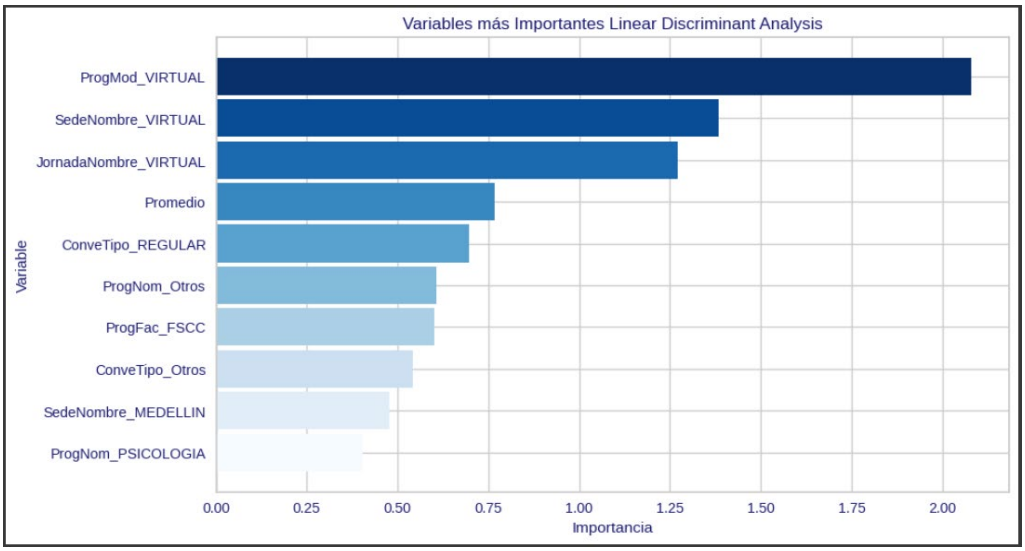


Ilustración 82. *Variables representativas en el modelo Linear Discriminant Analysis*
Fuente: Elaboración propia

En la ilustración 82, se observan las variables que el modelo *Linear Discriminant Analysis* determina importantes para predecir la variable objetivo. Para este modelo el promedio, la modalidad Virtual y la nueva variable que combina Promedio x Materias Homologadas del estudiante, son variables determinantes.

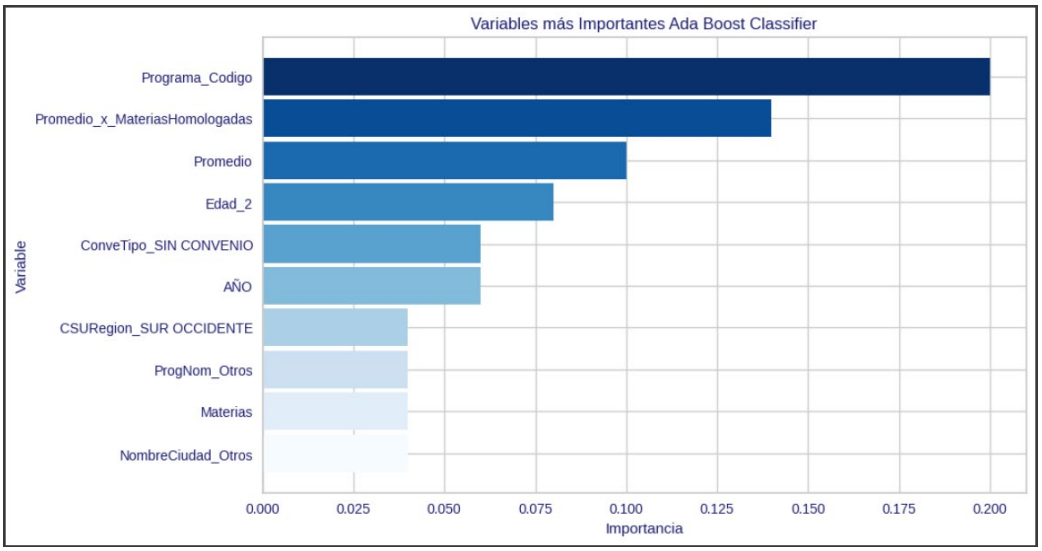


Ilustración 83. *Variables representativas en el modelo Ada Boost Classifier*
Fuente: Elaboración propia

En la ilustración 83, se observan las variables que el modelo *Ada Boost Classifier* determina importantes para predecir la variable objetivo. Para este modelo el Código del programa, la nueva variable que combina Promedio x Materias Homologadas y el promedio del estudiante, son variables determinantes.

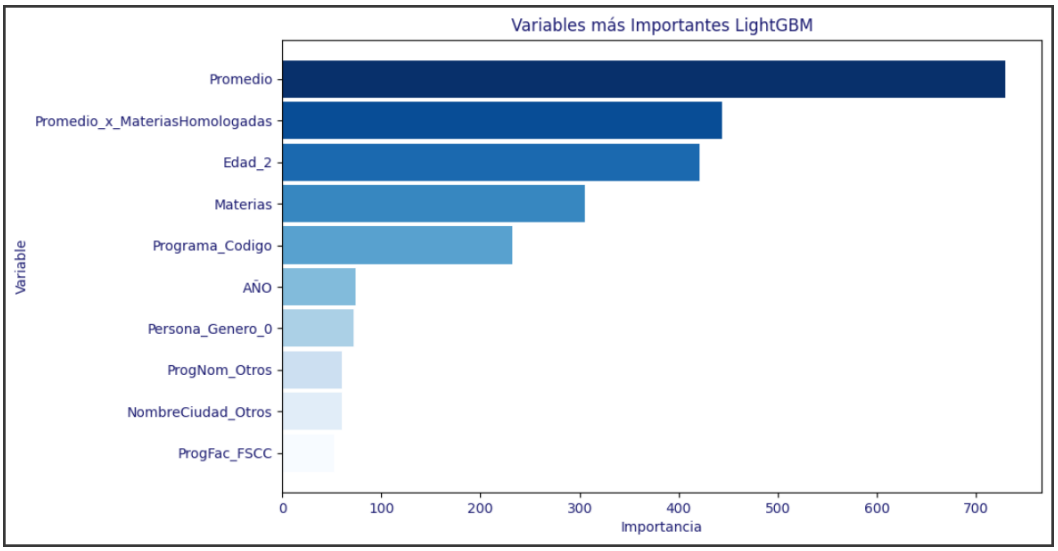


Ilustración 84. *Variables representativas en el modelo Light Gradient Boosting Machine*
Fuente: Elaboración propia

En la ilustración 84, se observan las variables que el modelo *Light Gradient Boosting Machine* determina importantes para predecir la variable objetivo. Para este modelo el Promedio, la nueva variable que combina Promedio x Materias Homologadas y la edad del estudiante, son variables determinantes.

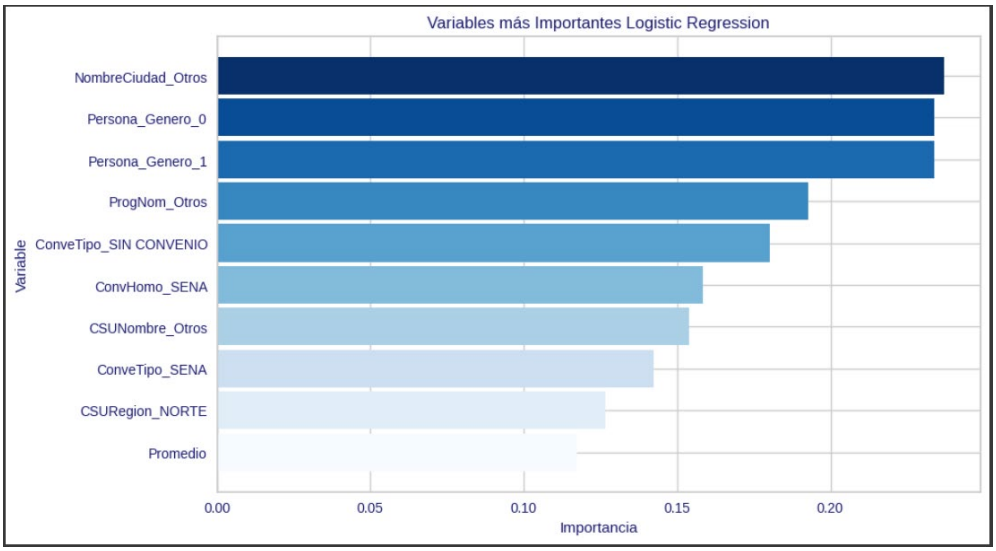


Ilustración 85. *Variables representativas en el modelo Logistic Regression*
Fuente: Elaboración propia

En la ilustración 85, se observan las variables que el modelo *Logistic* determina importantes para predecir la variable objetivo. Para este modelo la variable otras Ciudades, genero 0 y genero 1 del estudiante, son variables determinantes.

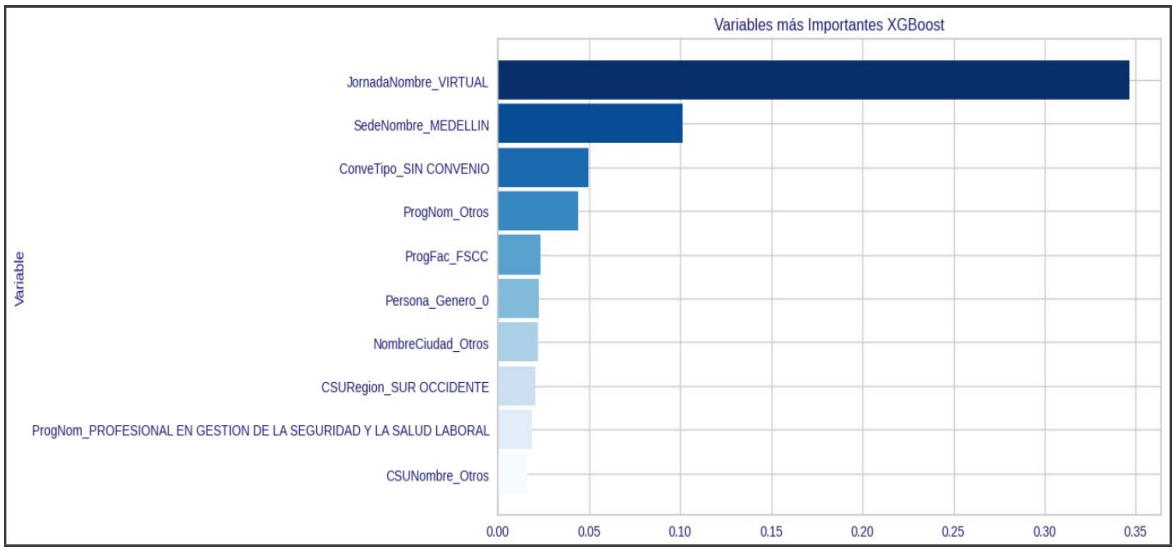


Ilustración 86. *Variables representativas en el modelo XGBoost*
Fuente: Elaboración propia

En la ilustración 86, se observan las variables que el modelo *XGBoost* determina importantes para predecir la variable objetivo. Para este modelo la variable jornada virtual, Sede Medellín y Sin convenio del estudiante, son variables determinantes.

7.4.3.3. HIPERPARÁMETROS

De acuerdo con los indicadores de rendimiento de los modelos planteados, se sigue trabajando con los dos mejores y se procede a realizar la búsqueda de los hiperparámetros óptimos para los dos modelos, buscando mejorar el rendimiento de estos.

Se establece la siguiente configuración inicial para explorar los hiperparámetros óptimos en cada modelo:

Parámetro	Valores Asignados	Descripción
n_estimators	100, 200, 300	Número de estimadores en un conjunto de aprendizaje en conjunto
max_depth	3,5,7	Profundidad máxima de cada árbol de decisión en el conjunto de aprendizaje en conjunto. Controla la complejidad del modelo y puede prevenir el sobreajuste.
learning_rate	0.01 , 0.1 , 0.3	Tasa de aprendizaje que controla la contribución de cada árbol al proceso de optimización

Ilustración 87. *Hiperparámetros*
Fuente: Elaboración propia

Luego de usar esta configuración con cada modelo, se obtienen los siguientes resultados:

Gradient Boosting Classifier:

```
Mejor resultado de la búsqueda en cuadrícula: 0.683344 utilizando {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}
```

Light Gradient Boosting Machine:

```
Mejor resultado de la búsqueda en cuadrícula: 0.684334 utilizando {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}
```

Con base en los resultados obtenidos se entrena los dos modelos con los mejores hiperparámetros encontrados y evalúa su precisión en un conjunto de datos de prueba, con lo cual se generaron los siguientes resultados:

Gradient Boosting Classifier:

```
Precisión en el conjunto de prueba: 0.6864508393285371
```

Light Gradient Boosting Machine:

```
Precisión en el conjunto de prueba: 0.6877997601918465
```

De acuerdo con los resultados obtenidos, se calculan las métricas de evaluación del rendimiento del modelo *Light Gradient Boosting Machine,* generando los siguientes resultados:

```
Precisión: 0.6899706542378733
Exhaustividad (Recall): 0.9248033317908376
F1-score: 0.7903114186851211
Exactitud (Accuracy): 0.6821043165467626
AUC: 0.683834611635669
```

Con base en la anterior ilustración, se puede determinar lo siguiente:

Precisión (*Precision*):

La precisión del modelo es del 68.99%, lo que indica que aproximadamente el 69% de las predicciones positivas realizadas por el modelo son correctas. En el contexto de la predicción del desempeño de los estudiantes en las pruebas SABER PRO, esto significa que alrededor del 69% de los estudiantes clasificados como "mal desempeño" por el modelo realmente obtienen un resultado bajo en las pruebas.

Exhaustividad (*Recall*):

La exhaustividad del modelo es del 92.48%, lo que indica que el modelo identifica correctamente alrededor del 92.48% de todos los casos de estudiantes con mal desempeño en las pruebas SABER PRO. En otras palabras, el modelo tiene una capacidad muy alta para detectar a los estudiantes que realmente tienen un bajo desempeño en las pruebas.

F1-*score*:

El valor F1-score del modelo es del 79.03%, lo que representa un equilibrio entre precisión y exhaustividad. Esto significa que el modelo logra una combinación adecuada de precisión y exhaustividad en la predicción del desempeño de los estudiantes en las pruebas.

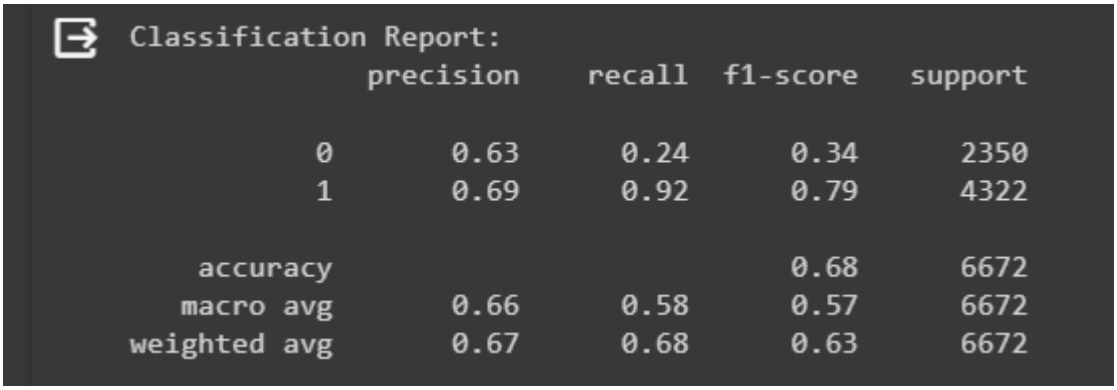
Exactitud (*Accuracy*):

La exactitud del modelo es del 68.21%, lo que indica que alrededor del 68.21% de todas las predicciones realizadas por el modelo son correctas. Esta métrica evalúa la capacidad general del modelo para predecir correctamente tanto los casos positivos como los negativos.

AUC (*Area Under the Curv*e):

El AUC del modelo es del 68.38%, lo que sugiere que el modelo tiene una capacidad moderada para distinguir entre estudiantes con buen desempeño y mal desempeño en las pruebas SABER PRO. Un valor de AUC cercano a 1 indica un mejor rendimiento del modelo en la clasificación.

En general, los resultados muestran que el modelo *LightGBM* tiene un buen rendimiento en la identificación de estudiantes con mal desempeño en las pruebas SABER PRO, con una alta exhaustividad y un F1-score equilibrado.



Classification Report:

	precision	recall	f1-score	support
0	0.63	0.24	0.34	2350
1	0.69	0.92	0.79	4322
accuracy			0.68	6672
macro avg	0.66	0.58	0.57	6672
weighted avg	0.67	0.68	0.63	6672

El informe de clasificación revela el rendimiento del modelo *LightGBM*, destacando los siguientes puntos:

Precisión (*Precision*):

El modelo logra una precisión del 69% para la clase 1 (mal desempeño). Esto indica que aproximadamente el 69% de las predicciones realizadas por el modelo, identificando casos de mal desempeño, son precisas y correctas. Es importante observar que el modelo tiene una tasa de precisión razonablemente alta en la identificación de estudiantes con un bajo desempeño.

Exhaustividad (*Recall*):

La exhaustividad para la clase 1 es del 92%, lo que significa que el modelo captura correcta y aproximadamente el 92% de todos los casos de mal desempeño. Esta alta exhaustividad refleja la capacidad del modelo para identificar la gran mayoría de los estudiantes que realmente tienen un mal desempeño en las pruebas SABER PRO. Es un aspecto muy positivo, ya que garantiza que la mayoría de los estudiantes en riesgo sean identificados.

F1-*score*:

El F1-score para la clase 1 es del 79%, lo que indica un buen equilibrio entre precisión y exhaustividad en la predicción de mal desempeño. Esta medida compuesta refleja la capacidad del modelo para predecir con precisión casos de mal desempeño, mientras minimiza los falsos positivos y negativos.

Exactitud (*Accuracy*):

La exactitud general del modelo es del 68%, lo que significa que alrededor del 68% de todas las predicciones realizadas por el modelo son correctas. Aunque esta métrica considera el rendimiento del modelo en todas las clases, es importante destacar que la precisión y la exhaustividad para la clase 1 son altas, lo que contribuye significativamente a la precisión general del modelo.

En resumen, el modelo *LightGBM* luego de ajustar los hiperparámetros demuestra un rendimiento alentador en la identificación de estudiantes con mal desempeño en las pruebas SABER PRO. La alta precisión, exhaustividad y F1-score para la clase 1 son indicadores positivos de la capacidad del modelo para identificar correctamente a los estudiantes en riesgo y proporcionar intervenciones tempranas. Este análisis resalta los aspectos positivos del modelo, especialmente en relación con la clase 1.

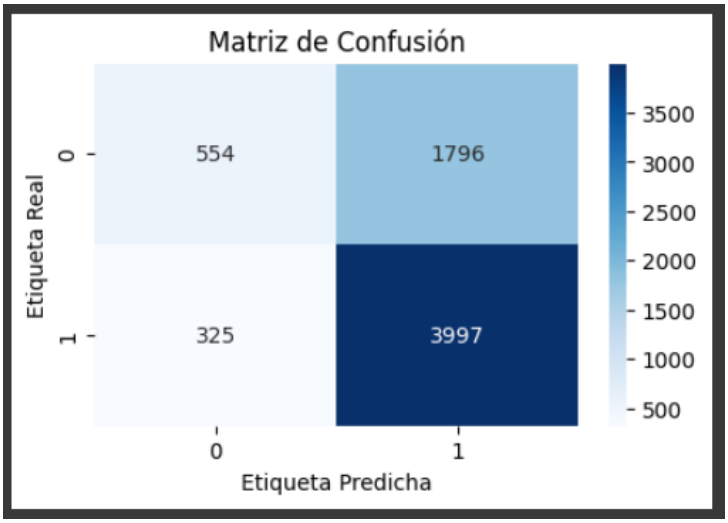


Ilustración 88. *Matriz de Confusión*
Fuente: Elaboración propia

La matriz de confusión revela la capacidad del modelo *LightGBM* para predecir el desempeño de los estudiantes en las pruebas SABER PRO. En este caso, se observa lo siguiente:

- Verdaderos Positivos (TP): 3997
- Falsos Positivos (FP): 1796
- Verdaderos Negativos (TN): 554
- Falsos Negativos (FN): 325

Interpretación:

Verdaderos Positivos (TP):

El modelo identificó correctamente a 3997 estudiantes que realmente obtuvieron un mal desempeño en las pruebas SABER PRO. Este resultado muestra la capacidad del modelo para detectar efectivamente a los estudiantes en riesgo de bajo rendimiento académico.

Falsos Positivos (FP):

Aunque el modelo clasificó erróneamente a 1796 estudiantes como tener mal desempeño cuando en realidad no lo tenían, este valor representa una oportunidad para mejorar el modelo. Sin embargo, el hecho de que el modelo haya identificado correctamente a un gran número de estudiantes con mal desempeño es importante y sugiere que el modelo tiene una capacidad prometedora para identificar a los estudiantes en riesgo y principal objetivo de esta investigación.

Verdaderos Negativos (TN):

Aunque el número de verdaderos negativos es bajo en comparación con los falsos positivos, el modelo logró identificar correctamente a 554 estudiantes que obtuvieron un buen desempeño en las pruebas. Esto indica que el modelo también es capaz de reconocer el buen desempeño de algunos estudiantes.

Falsos Negativos (FN):

Los falsos negativos representan una oportunidad de mejora para el modelo, ya que indican que el modelo no identificó correctamente a 325 estudiantes que obtuvieron un mal desempeño en las pruebas. Sin embargo, este número es significativamente menor que los verdaderos positivos, lo que sugiere que el modelo tiene una alta capacidad para identificar efectivamente a los estudiantes en riesgo.

En resumen, la matriz de confusión muestra que el modelo *LightGBM* tiene una capacidad importante para predecir el desempeño de los estudiantes en las pruebas SABER PRO, con una notable cantidad de verdaderos positivos. Aunque hay margen para mejorar en la reducción de los falsos positivos y falsos negativos, el modelo demuestra una capacidad sólida para identificar a los estudiantes en riesgo de bajo rendimiento académico, lo que lo convierte en una herramienta muy importante para la identificación temprana de estudiantes que pueden necesitar intervención adicional.

7.4.3.4. CURVA DE ROC Y AUC MODELO FINAL

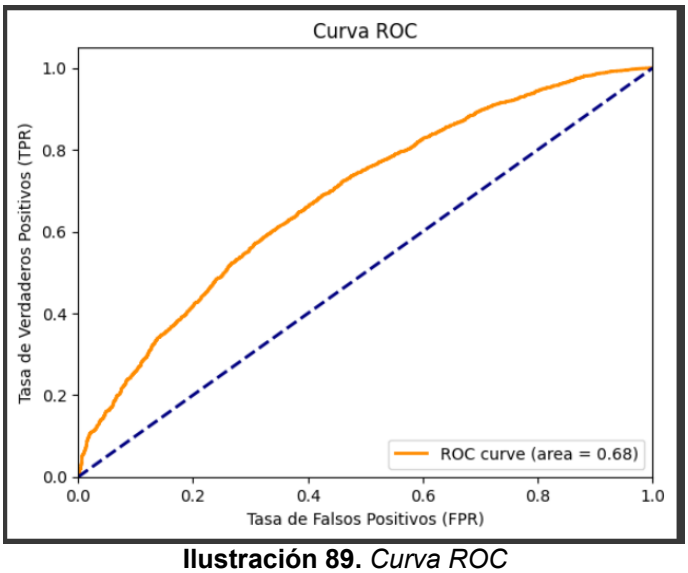


Ilustración 89. *Curva ROC*
Fuente: Elaboración propia

De acuerdo con la gráfica el modelo *LightGBM* muestra una excelente capacidad de discriminación en la predicción del desempeño de los estudiantes en las pruebas SABER PRO, como lo demuestra su curva ROC por encima de la línea diagonal.

Con un AUC (Área Bajo la Curva ROC) de 0.68, el modelo exhibe una destacada habilidad para distinguir entre estudiantes con buen y mal desempeño. Este valor indica un rendimiento significativamente mejor que el azar y sugiere una fiabilidad considerable en las predicciones del modelo.

La curva ROC presenta una pendiente ascendente pronunciada, lo que sugiere que el modelo es eficaz para identificar tanto a los estudiantes con buen desempeño como a aquellos con mal desempeño, sin comprometer la precisión en ninguna de las dos clases.

En conclusión:

El modelo *LightGBM* muestra una capacidad robusta para predecir el desempeño de los estudiantes en las pruebas SABER PRO, lo que sugiere su utilidad como herramienta para identificar y apoyar a aquellos en riesgo de un bajo rendimiento.

El modelo es capaz de identificar con precisión a la mayoría de los estudiantes con buen y mal desempeño, lo que lo convierte en una herramienta valiosa para la toma de decisiones en el ámbito educativo.

Recomendaciones:

Se sugiere utilizar el modelo *LightGBM* para identificar y ofrecer apoyo temprano a los estudiantes identificados como en riesgo de un mal desempeño en las pruebas SABER PRO.

Es importante llevar a cabo un análisis más detallado de los errores del modelo para identificar oportunidades de mejora y optimización en su desempeño.

El objetivo principal sigue siendo mejorar el desempeño de los estudiantes en las pruebas SABER PRO, y el modelo *LightGBM* puede ser una herramienta valiosa para lograr este propósito.

8. CRONOGRAMA DE TRABAJO

Con base en la metodología CRISP-DM, escogida para el desarrollo de este proyecto, se plantean las siguientes actividades a desarrollar, por semanas en un total de 16 semanas equivalente a 4 meses.

FASES / CRISP-DM	ACTIVIDAD	Semanas
Entendimiento del Negocio	Disección y articulación conceptual del proyecto	1
	Análisis holístico del ecosistema y definición de metas analíticas	1
	Construcción de una propuesta de proyecto detallada y estructurada	1
Comprención de los datos	Extracción y acumulación de conjuntos de datos preliminares	1
	Caracterización y perfilado avanzado de datos	1
	Inspección profunda y descubrimiento de patrones en los datos	1
	Evaluación y aseguramiento de la integridad de datos	1
Preparación de los datos	Curación y refinamiento selectivo de datasets	1
	Depuración y normalización de datos	2
	Armonización y fusión de fuentes de datos	1
Modelamiento	Estrategia y selección de algoritmos con base en evidencia	1
	Prototipado y simulación de escenarios de evaluación	1
	Desarrollo iterativo de soluciones predictivas	3
	Validación y ajuste de modelos mediante benchmarks	2
Evaluación	Auditoría de rendimiento y efectividad	1
	Meta-análisis del ciclo de modelado	1
	Síntesis y presentación de hallazgos analíticos	1
Despliegue	Estrategia y cronograma para la implementación operativa	1
	Diseño de operaciones continuas y protocolos de contingencia	1
	Compilación y entrega de documentación ejecutiva y técnica	1

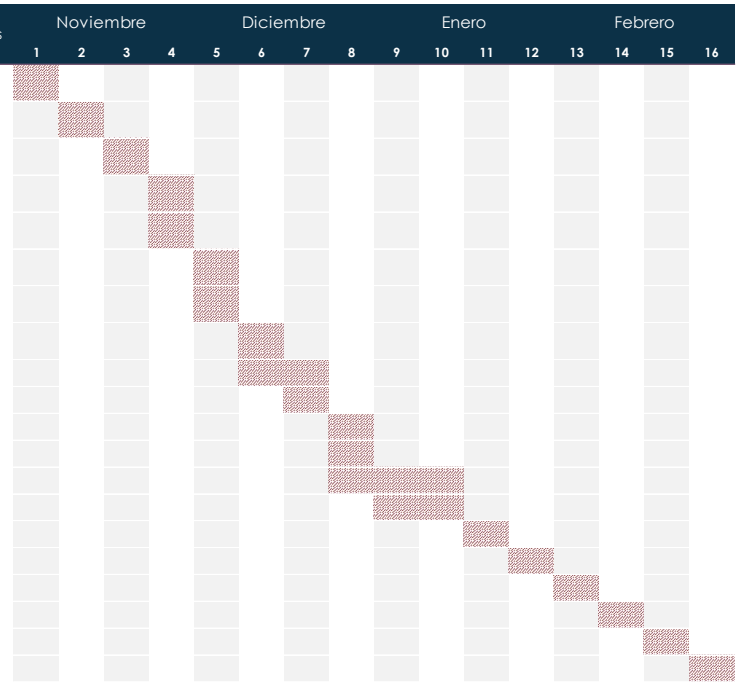


Ilustración 90. *Cronograma*
Fuente: Elaboración propia

9. PRESUPUESTO

En la siguiente ilustración se encuentra el presupuesto proyectado para la ejecución del proyecto durante los 4 meses planteados para su ejecución.

Inicio: 01-Noviembre de 2023
Finalización: 30 de Febrero de 2024

Variable	Cantidad de horas invertidas	Descripción		Precio
Honorarios Estudiante	80	Horas dedicadas al proyecto	$	12.800.000,00
Horario de Tutor	5	Horas dedicadas total al acompañamiento del proyecto	$	1.500.000,00
Costo de Herramientas	30	Herramienta de Power BI versión pro	$	70.000,00
			$	14.370.000,00

Ilustración 91. *Presupuesto*
Fuente: Elaboración propia

10. CONCLUSIONES

Este proyecto de tesis ha sido un fortalecimiento intelectualmente gratificante y esencial en la formación académica como científico de datos, permitiendo sumergirse profundamente en el complejo y fascinante mundo del aprendizaje automático supervisado. A lo largo de esta investigación, se obtuvo una comprensión integral de cómo los modelos de *machine learning* pueden transformar y optimizar los procesos operativos, convirtiéndose en fuentes cruciales de *insights* para la toma de decisiones en niveles gerenciales.

Incorporando la rigurosa metodología CRISP-DM, se inició esta travesía analítica con una exploración meticulosa de los datos, empleando una variedad de técnicas visuales para descifrar el comportamiento subyacente de las variables en estudio. Esta fase descriptiva fue fundamental para establecer una base sólida sobre la cual ejecutar las siguientes etapas del proyecto.

Se procedió con el preprocesamiento de los datos, una fase crítica donde se depuraron y prepararon los datos para su análisis posterior, asegurando que cada atributo se alineara correctamente con los objetivos de los modelos predictivos. La selección y transformación cuidadosa de las características aseguró la relevancia y calidad de los datos alimentados a los algoritmos de clasificación.

El generar una herramienta que pueda apoyar a la parte estratégica de la universidad, como el tablero de *Power* BI, integrando los resultados a nivel nacional de todas las universidades que presentaron las pruebas, y luego entregar un análisis más segmentado de los estudiantes propios de la universidad. Permite a los directivos rápidamente identificar donde se pueden tener oportunidades de mejora y con un impacto alto dentro de los segmentos que más requieren apoyo para tener mejores resultados en las próximas pruebas.

La construcción y evaluación de varios algoritmos permitió medir su rendimiento y seleccionar el más adecuado para la predicción del rendimiento de los estudiantes en las pruebas Saber Pro. El modelo *Light Gradient Boosting Machine* destacó por su alta eficacia, mostrando una impresionante capacidad de discernimiento y una significativa exhaustividad, particularmente para identificar aquellos estudiantes en riesgo de bajo rendimiento.

El estudio ha demostrado que el modelo *Light Gradient Boosting Machine* (*LightGBM*) ofrece un desempeño robusto para predecir el rendimiento de los estudiantes en las pruebas Saber Pro. Este modelo sobresale en su habilidad para identificar a aquellos estudiantes con un probable bajo rendimiento, mostrando un balance óptimo entre precisión y *recall*, y un F1-*Score* destacable. La alta exhaustividad (92.48%) es particularmente relevante en el contexto educativo, donde es crítico identificar a todos los estudiantes que podrían requerir intervenciones adicionales para mejorar su desempeño. Adicionalmente, las variables más representativas identificadas por el

modelo *LightGBM* incluyen el promedio del estudiante, la combinación del promedio con materias homologadas y la edad, lo que sugiere que estos factores son significativos para predecir el rendimiento académico en las pruebas estandarizadas.

En conclusión, esta tesis no solo ha fortalecido la comprensión de los modelos de *machine learning* y su aplicabilidad en entornos reales, sino que, también ha reforzado la convicción de que el correcto aprovechamiento de estas tecnologías puede propiciar un cambio significativo en el sector educativo. Los modelos predictivos, cuando se utilizan de manera ética y eficiente, tienen el potencial de servir como herramientas poderosas para mejorar los resultados de los estudiantes y, en última instancia, fortalecer el rendimiento institucional en el escenario académico nacional.

La creación de un instrumento analítico avanzado, en forma de un tablero de mando interactivo basado en *Power BI*, representa un avance significativo para la estrategia de gestión educativa de la universidad. Este sistema consolida y compara los datos de rendimiento de diversas instituciones de nivel superior a escala nacional, ofreciendo un enfoque comparativo valioso. Además, proporciona una desagregación detallada del desempeño de los estudiantes de la propia universidad. Tal enfoque facilita a los líderes académicos y administrativos la identificación precisa de áreas susceptibles de mejora y la implementación de intervenciones enfocadas en segmentos estudiantiles específicos que requieren mayor soporte. Con este recurso, es posible agilizar el proceso de toma de decisiones y maximizar la eficacia de las políticas educativas, allanando el camino para elevar los estándares académicos en futuras evaluaciones.

Se recomienda la implementación del modelo *LightGBM* para la identificación temprana de estudiantes con riesgo de bajo rendimiento en las pruebas Saber Pro. Sin embargo, se debe prestar especial atención a la optimización de hiperparámetros y a la interpretación ética y responsable de los resultados del modelo, con el objetivo de mejorar las intervenciones educativas y estrategias de apoyo para estos estudiantes. Además, es esencial considerar la exhaustividad del modelo y trabajar en la reducción de falsos positivos para evitar etiquetar incorrectamente a los estudiantes y garantizar intervenciones precisas.

Es importante reconocer las limitaciones inherentes a cualquier modelo predictivo, incluyendo la variabilidad de los resultados según las características demográficas y académicas de la población estudiantil. Los errores de predicción deben ser analizados detenidamente para mejorar continuamente la eficacia del modelo.

El propósito último de esta investigación es contribuir a mejorar el desempeño global de los estudiantes en las pruebas Saber Pro. Se subraya la importancia de integrar el modelo *LightGBM* dentro de un enfoque holístico que abarque medidas educativas y acompañamiento adecuado, facilitando así el mejoramiento continuo y sostenible del rendimiento académico de los estudiantes.

Es importante estar monitoreando sistemáticamente la capacidad predictiva del modelo implementado y a realizar ajustes periódicos en función de los cambios en los patrones

de datos y las tendencias emergentes. Este proceso iterativo es esencial para mantener la relevancia y eficacia del modelo a lo largo del tiempo.

Por último, el tablero de *Power BI* y el modelo predictivo creado en este proyecto, se recomienda ser utilizados como parte de un enfoque integral que incluya intervenciones educativas y estrategias de apoyo adicionales para maximizar su efectividad y beneficios para los estudiantes.

Estas estrategias deben ser a corto, mediano y largo plazo, lo anterior, por que el modelo tiene la capacidad de predecir a los estudiantes cercanos a presentar las pruebas, sino también, tiene la capacidad de ver el desempeño en los estudiantes que ya hubieran cursado el 50% de su programa académico o incluso los que están en su 30%. Para estos últimos se recomienda combinarlo con un modelo que ayude a predecir la deserción.

11. TRABAJOS FUTUROS

De acuerdo con algunas dificultades que se presentaron en el desarrollo del proyecto, es importante trabajar en conseguir algunas variables que, con base en estudios presentados por el ICFES, son determinantes en la educación, como el estrato socioeconómico, el nivel de ingresos de la familia y los niveles de escolaridad de padre y madre; Desde este mismo enfoque se recomienda adicionar algunas variables externas que pueden afectar el segmento educativo como por ejemplo las variables de acceso a dispositivos móviles y acceso a internet.

Se recomienda estar pendientes de nuevos modelos de aprendizaje automático emergentes o menos convencionales, que puedan ofrecer perspectivas novedosas o mejoras en el desempeño de las métricas clave.

12. REFERENCIAS BIBLIOGRÁFICAS

[1] ICFES, *Valor Agregado y Aporte Relativo,* Bogotá.

[2] Colombia. Congreso de la República, *Ley 1324,* 2009.

[3] Colombia. Presidencia de la República, *Decreto 869,* 2010.

[4] OCDE, *Resultados PISA,* 2012-2018.

[5] OCDE, *Programme for International Student Assessment (PISA) Results from PISA 2018,* 2019.

[6] D. Ortiz, E. Gómez and N. Arias, "Resultados en Saber Pro de estudiantes de modalidad presencial y virtual en dos universidades colombianas.," *Revista Academia y Virtualidad,* vol. 8, no. 2, pp. 100-111, 2015.

[7] E. León, «Diplomado en Minería de Datos (Universidad Nacional de Colombia),» 2018. [En línea]. Available: http://disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/.

[8] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education,* vol. 113, p. 177–194, 2017.

[9] A. I. Oviedo y J. Jiménez, «Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO,» *Revista Politécnica,* vol. 15, nº 9, pp. 128-140, 2019.

[10] O. Sifuentes, «Modelos predictivos de la deserción estudiantil en una universidad privada peruana,» *Industrial Data,* vol. 21, nº 2, pp. 47-52, 2018.

[11] E. D. Parra Vargas, «La educación superior en Colombia: Una mirada a los conceptos de calidad y evaluación,» *Boletín Virtual,* vol. 4, nº 9, pp. 95-103, 2015.

[12] V. Galán, «Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario,» 2016. [En línea]. Available: https://e-archivo.uc3m.es/rest/api/core/bitstreams/714c5452-962e-44cf-993f-ebb3088d4aa5/content.

[13] OCDE, «Organización para la Cooperación y el Desarrollo Económicos. PISA 2015,» 2016.

[14] OCDE, «PISA 2018 Results (Volume I): What Students Know and Can Do,» 2018.

[15] S. Celis, L. Moreno, P. Poblete, J. Villanueva and R. Weber, "Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería.," *Revista Ingeniería de Sistemas,* vol. XXIX, 2015.

[16] C. G. Loja Rodas, *Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca,* U. d. Cuenca., Ed., Cuenca: Universidad de Cuenca. Facultad de Ingeniería, 2019.

[17] J. R. Quinlan, «Induction of Decision Trees.,» *Machine Learning,* vol. 1, nº 1, pp. 81-106, 1986.

[18] R. Timarán, J. Caicedo and A. Hidalgo, "Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en

las pruebas Saber 11°," 2019. [Online]. Available: https://doi.org/10.19053/20278306.v9.n2.2019.9184.

[19] Instituto de Ingeniería del Conocimiento, *La metodología CRISP-DM en ciencia de datos.*

[20] Universidad de Cuenca, «Del rendimiento AT,» s.f. [En línea]. Available: https://dspace.ucuenca.edu.ec/bitstream/123456789/33486/1/Trabajo%20de%20titulaci%C3%B3n.pdf.

[21] ICFES, «Pruebas saber Pro,» 2024. [En línea]. Available: https://www.icfes.gov.co/web/guest/resultados_del_examen_saber_pro.