EL COMITÉ DE INVESTIGACIONES Y POSGRADOS
DE LA FACULTAD DE CIENCISA E INGENIERÍA
UNIVERSIDAD DE MANIZALES

HACE CONSTAR QUE

Oscar Mario Agudelo Nieto - Código MG1201916399
de la Maestría en Gestión Estratégica de la Información

Cumplió con la presentación y sustentación del trabajo de grado titulado **"*Implementación de un modelo de Machine Learning como estrategia de prevención de deserción estudiantil universitaria en la Universidad de Manizales,*"** para optar al título de Magister en Gestión Estratégica de la Información.

A dicho trabajo se le asignaron como jurados evaluadores a los profesionales ***Juan Pablo Giraldo*** y ***Jairo Pineda***, quienes dieron el concepto de aprobación al presente trabajo.

Para constancia se firma a los a los veintisiete (12) días del mes de mayo de 2023

Néstor Jaime Castaño Pérez
Decano
Facultad de Ciencias de Ingeniería

José Fernando Mejía Correa
Director Posgrados
Facultad de Ciencias e Ingeniería

Identificación de Deserción Temprana en la Universidad de Manizales con Aprendizaje Automático, como parte de la estrategia de prevención

Oscar Mario Agudelo Nieto

Universidad de Manizales
Facultad de Ciencias e Ingeniería
Maestría en Gestión Estratégica de la Información
Manizales, 2023

Identificación de Deserción Temprana en la Universidad de Manizales con Aprendizaje Automático, como parte de la estrategia de prevención

Oscar Mario Agudelo Nieto

Informe final de trabajo de grado presentado como requisito para optar al título de Magíster en Gestión Estratégica de la Información

Director:

M.Sc. Juan Alejandro Trujillo P.

Línea de Investigación:

Gestión y aprovechamiento de la Información

Universidad de Manizales

Facultad de Ciencias e Ingeniería

Maestría en Gestión Estratégica de la Información

Manizales, 2023

Resumen

La presente investigación tiene por objetivo la implementación de dos (2) modelos predictivos enmarcados en la Analítica de datos con el uso de inteligencia artificial alineado a una de sus ramas como es la aplicación de ***machine learning*** (reconocimiento de patrones) aplicada al contexto de la educación universitaria en los programas de pregrado (estudios que requieren ser egresado de enseñanza media, conducentes a un título académico de educación superior), modalidad presencial para las facultades de Ciencias e Ingeniería y Ciencias Contables, Administrativas y Económicas cuya duración es igual o superior a ocho () semestres. Los algoritmos permiten identificar las diversas razones económicas, personales, pedagógicas, familiares, sociales y/o vocacionales que inciden en la decisión de abandonar los estudios superiores (Spady 1970), durante el transcurso de los primeros cuatro (4) semestres por parte de los estudiantes.

Surge esta necesidad de la carencia en la Universidad de Manizales de un método efectivo que identifique anticipadamente o emita alertas tempranas sobre los casos potenciales de estudiantes con la intención de abandonar sus estudios superiores de pregrado presencial. La información analizada corresponde a los periodos transcurridos entre enero 2018 a diciembre 2022.

Para alcanzar los objetivos trazados la investigación se fundamentó en información histórica de la Universidad de Manizales obtenida de los sistemas de información alrededor del ecosistema SIGUM (data sobre matriculas, rendimiento académico, deserción, resultados pruebas saber 11, acompañamiento psicopedagógico, inasistencia y otros). Dichos datos fueron procesados aplicando la metodología CRISP-DM (método probado para orientar trabajos de minería de datos), que consta de las fases: **entendimiento del negocio, entendimiento de los datos, preparación de los datos** basados en las reglas de negocio**, modelado** (desarrollo e Python de algoritmo de aprendizaje supervisado de **Decision Tree** y **Random Forest**)**, evaluación de resultados** y **despliegue**. Que al final

derivan en la implementación de estrategias de mitigación del evento producto de la revisión de resultados, impactos, conclusiones y/o recomendaciones.

Producto de la aplicación de los algoritmos con sus respectivos ajustes durante el proceso de afinamiento, se logró obtener una **exactitud en la predicción** del **93%** de acierto, lo cual permite afirmar que, por cada 100 casos analizados por los modelos, estos predicen 93 de ellos acertadamente. Derivado de lo anterior fue posible identificar las variables de mayor prevalencia que explican el fenómeno y en consecuencia asociar estrategias específicas y efectivas a cada caso en cuestión. Comprender entonces las variables que inciden en la deserción aporta beneficios sociales, económicos y de gobierno a todo el ecosistema académico.

Palabras clave: *Machine Learning*, Deserción temprana universitaria, *Random forest*, Decisión Tree, CRISP-DM.

Contenido

Pag.

Lista de figuras

Lista de tablas

Lista de siglas

IES: Institución de Educación Superior. Según su carácter académico, el Ministerio de Educación Nacional las clasifica en cuatro tipos: (1) Institución Técnica Profesional; (2) Institución Tecnológica; (3) Institución Universitaria / Escuela Tecnológica, y (4) Universidad.

SNIES: Sistema Nacional de Información de la Educación Superior.

SPADIES: Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior.

SIGUM: Sistema de Información General de la Universidad de Manizales.

1 Introducción

La presente investigación se inspira en la necesidad apremiante de la Universidad de Manizales de implementar estrategias efectivas que prevengan la deserción estudiantil temprana en los programas de pregrado en modalidad presencial en las facultades de Ciencias Contables, Administrativas y Económicas y Ciencias e Ingeniería. Este fenómeno, común a la educación universitaria en general, tanto en Colombia como en el resto del mundo, representa un reto para los gobiernos y las instituciones de educación superior.

Algunos datos históricos demuestran que las tasas de abandono más altas se presentan durante los dos primeros años de estudio (deserción temprana), razón por la que durante este periodo son indispensables las estrategias para garantizar la permanencia de los estudiantes. Necesariamente, una intervención oportuna, organizada y coherente producirá impactos positivos para revertir el porcentaje creciente de la deserción, de manera que los matriculados en primer semestre de pregrado logren concluir su proyecto educativo.

El Gobierno de Colombia, a través del Ministerio de Educación Nacional, ha establecido lineamientos para garantizar la educación superior como política nacional; razón por la cual, algunas instituciones de educación superior incluyen en sus objetivos institucionales, planes de apoyo estudiantil que ponen en marcha estrategias para reducir la tasa de deserción; tal es el caso de la Universidad de Manizales con su *Programa de acompañamiento*.

Lo anterior da la pauta para presentar un proceso investigativo orientado a la mitigación del fenómeno, con el fin último de lograr la permanencia de los estudiantes universitarios de pregrado, durante todo el proceso hasta su graduación y, por lo tanto, aumentar el indicador de graduados efectivos de la Universidad de Manizales.

Sin embargo, durante la fase de recolección de la data se evidencio la ausencia de información disponible relacionada con el conjunto total de variables requeridas para que el proceso de predicción fuese más preciso (ver anexo No. 5: Variables explicativas de la deserción universitaria). No obstante, a la situación fue posible recopilar información de treinta y siete (37) variables. Las cuales fueron utilizadas en la construcción de los dos (2) algoritmos de *machine learning* como estrategia para reducir el índice de abandono, y que permita al *Programa de acompañamiento* de la Universidad de Manizales la activación de protocolos de actuación sobre los estudiantes con dicho riesgo.

En particular los algoritmos de predicción utilizados son el ***random forest*** y **decisión tree**. La selección de estos, parte de la revisión de investigaciones previas, nacionales e internacionales, que con el objetivo propuesto obtuvieron indicadores de **exactitud, sensibilidad, f1 y una matriz de confusión** con altos índices de precisión, destacándose el **random forest**.

Ligado a lo anterior, es importante destacar que la base fundamental de la predicción radica en un buen entendimiento del problema, el cual debe evidenciarse en un inventario asertivo del conjunto de variables que describa los diversos fenómenos relacionados con los estudiantes desertores y no desertores. Denota lo anterior la gran relevancia de contar información complementaria (metadatos) alrededor de las características de las variables, dominio de las mismas, reglas de validación y/o de negocio, calidad de los datos, etc.

El proceso investigativo se focaliza en la población de los dos primeros años académicos (el concepto de deserción temprana abarca los primeros cuatro (4) semestres, y es allí donde se presenta el mayor número de deserción universitaria). de las facultades de Ciencias Contables, Económicas y Administrativas, y de Ciencias e Ingeniería, de la Universidad de Manizales durante los periodos enero 2018 a diciembre 2022. En la primera parte del proceso, se analiza la permanencia de los estudiantes que presentaron dificultades académicas en su proceso de adaptación a la vida universitaria, estudiantes catalogados con Rendimiento Académico Insuficiente (RAI)[1] producto de problemas

[1] De acuerdo con el artículo 99 del Reglamento Estudiantil de la Universidad de Manizales, “un estudiante ha obtenido un rendimiento académico insuficiente cuando en un mismo periodo

previos y otras originadas durante el desenvolvimiento universitario, como situaciones deficitarias socio-económicas, familiares, afectivas y tecnológicas. La conjugación de estos factores se considera como desencadenante de la deserción.

La revisión documental, en primer lugar, se enfoca en la verificación de la literatura acerca de estudios previos, que permiten orientar la efectividad de la investigación al contar con el respaldo de producciones científicas que resultaron en experiencias exitosas. Se analiza información relacionada con la deserción estudiantil y la implementación de modelos analíticos predictivos existentes. Posteriormente, se recopila información estadística e histórica sobre deserción, de los programas académicos vinculados a las facultades en estudio entre los años 2018 a 2022 referente a pruebas Saber 11, resultados académicos, matrículas y acompañamiento estudiantil de Bienestar Universitario, etc.

Identificadas claramente las fuentes de datos y sus variables, se ejecutan sobre ellas los correspondientes procesos ETL (extracción, transformación, validación y carga de los datos), los cuales se archivan de acuerdo con sus características en SQL SERVER.

Posteriormente, se hace un análisis exploratorio de los datos para entender mejor, aspectos relacionados con las reglas de negocio y la comprensión de los datos, y así derivar en la creación o selección del modelo clasificatorio más idóneo que permita dar cuenta del fenómeno de deserción temprana estudiantil universitaria. Para esto, se valida la exactitud y precisión de cada uno de los algoritmos seleccionados y concluir cual es el más apropiado de los dos seleccionados para dar respuesta a la siguiente pregunta:

¿Es posible implementar modelos de aprendizaje automático supervisado para identificar estudiantes de pregrado en modalidad presencial de la Universidad de Manizales que estén en riesgo de deserción temprana?

académico pierde más del cincuenta por ciento (50%) de las asignaturas cursadas y validadas o reprueba dos (2) de ellas por segunda vez o una (1) por tercera vez o más" (U. Manizales, 2012, pág. 21).

2 Planteamiento del problema de investigación y su justificación

> Según un estudio realizado por el Banco Mundial, Colombia es el segundo país en América Latina con la mayor tasa de deserción universitaria con un 42% de estudiantes que se retiran de las instituciones en los primeros semestres. (…) De esta forma, el país se encuentra tan solo precedido por Bolivia, donde la tasa alcanza el 48%. Así mismo, de acuerdo con la OCDE, solo el 22% de la población entre los 25 y 34 años tiene un título universitario, mientras que el promedio de los países miembros de la organización es del 38%. (Sectorial, 2020)

Necesariamente, esta situación obliga a los gobiernos a implementar políticas de cobertura nacional no solo aplicables en la educación oficial sino privada, que impliquen estrategias efectivas de intervención oportuna, de manera organizada y coherente, a fin de producir impactos positivos que permitan revertir el porcentaje creciente de la deserción.

En 2012, la Organización para el Desarrollo y la Cooperación Económica – OCDE, publicó el informe *La Educación Superior en Colombia 2012* (OCDE, Banco Mundial, 2012). Allí se destaca el aumento significativo en la cobertura universitaria para programas de pregrado pues, la tasa pasó del 24.4% en el año 2002, al 46% en 2014, datos que trascienden la realidad de la Universidad de Manizales.

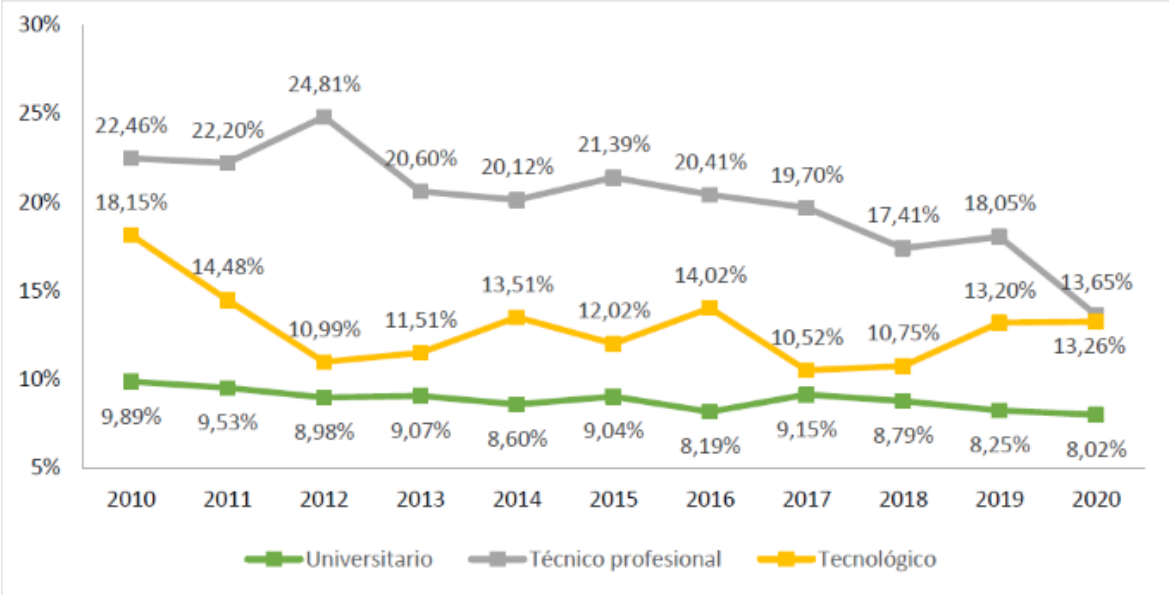
Sin embargo, este crecimiento porcentual de cubrimiento en la educación superior, se ve afectado por los índices, cada vez más preocupantes, de deserción estudiantil universitaria, tasada para el 2013, en un 44.9% de matriculados que no concluyeron su proceso educativo en programas de pregrado e incluso de postgrado (MEN, 2014, p. 1). Según los datos arrojados por el sistema SPADIES, el porcentaje de deserción universitaria para 2016 es del 43.8%. Un 1% para la formación Técnico profesional, 17.1%

para la formación Técnica y Tecnológica, 16. 7% para la formación Tecnológica y 9.0% para la Educación Universitaria. Dese la perspectiva del SPADIES de define el fenómeno de la deserción como: ***Estado de un estudiante que de manera voluntaria o forzosa no registra matrícula por dos o más períodos académicos consecutivos del programa en el que se matriculó; y no se encuentra como graduado, o retirado por motivos disciplinarios. La deserción es el resultado del efecto de diferentes factores como individuales, académicos, institucionales, y socioeconómicos***. (SPADIES, Glosario, 2020).

Según los datos referidos en el informe **Education at the glace** (2016), la deserción universitaria en los países de la OCDE alcanza un 31%. En el ámbito Europeo de Educación Superior (EEES), la tasa de deserción varía de un 20% a un 55%. En el ámbito latinoamericano la deserción se establece entre un 8% a un 48%. En el ámbito colombiano, la deserción se ubica en un 48.8%. Estos indicadores develan que la deserción en la educación universitaria, es un problema que aqueja a todos los sistemas educativos a escala global.

El estudio de la OCDE, en uno de sus apartes, especifica la ausencia de estrategias adecuadamente articuladas e indica que solo, a raíz de las políticas establecidas por el gobierno nacional y canalizadas a través del Ministerio de Educación Nacional, se han establecido lineamientos para garantizar la educación superior como política nacional (OCDE, Banco Mundial, 2012). Es así que, algunas instituciones de educación superior incluyen dentro de sus objetivos institucionales programas de acompañamiento estudiantil para poner en marcha estrategias que reduzcan la tasa de deserción; se incluye en ellas la Universidad de Manizales con su *Programa de acompañamiento estudiantil*.

Para una mayor ilustración ver figura No. 75: Tasa de deserción anual según nivel de formación.



Fuente: Sistema para la Prevención y Análisis de la Deserción en las instituciones de educación superior SPADIES con corte de los datos a julio del 2021.

La deserción universitaria en el ámbito internacional, en el marco de los países miembros de la OCDE se refleja en un 31% mientras que la deserción universitaria en el ámbito colombiano, en su comportamiento se registra ligeramente mayor, en la formación técnica profesional, para el sector oficial es de 39.33% y para el sector privado es del 30.17%; **igualmente, es de resaltar que la educación universitaria registra un menor porcentaje de deserción que para el sector oficial es del 8.57 y para el sector privado es del 9.67%.** En este mismo sentido, los estudios relacionados con la deserción en Colombia muestran que la relación costo - beneficio de la educación, es interferida por la deserción, lo que se ve representado en una inversión social insuficiente que no renta una tasa de retorno, ya que ante el abandono de los estudios se ocupan ámbitos laborales no cualificados reduciendo el desarrollo de la economía y de los territorios, (UNAD, Gutiérrez, Vélez, López, 2020).

En lo referente al comportamiento de la deserción en la Universidad de Manizales según el informe de gestión 2020, el indicador marca un 5.7% cifra por debajo del 9.67% identificado para Colombia.

2.1 Descripción del área problemática

En la actualidad, las universidades colombianas presentan diversos inconvenientes para el seguimiento integral del estudiante, lo que afecta en la deserción estudiantil. Consecuencia de esta situación y en aras de contrarrestar el alto porcentaje de retiro, el Ministerio de Educación Nacional definió lineamientos para articular procesos que conduzcan a reducir la tasa del indicador de deserción descritos en el documento ***Acuerdo nacional para disminuir la deserción en la educación superior MEN, 2013-2014).*** Para esto, concentró la información en el sistema de información SPADIES (Sistema para la Prevención de la Deserción de la Educación Superior):

> Este sistema es la herramienta para hacer seguimiento sobre las cifras de deserción de estudiantes de la educación superior. Con los datos suministrados por las instituciones de educación superior a SPADIES, se identifican y se ponderan los comportamientos, las causas, variables y riesgos determinantes para desertar. Además, con esta información se agrupan los estudiantes de acuerdo con su riesgo de deserción. (SPADIES, 2022)

SPADIES es netamente descriptivo en sus estadísticas y en ninguna de sus funcionalidades presenta datos predictivos ni prescriptivos.

Tal es el caso de la Universidad de Manizales, institución que, a pesar de desarrollar estrategias para impactar sobre la problemática, requiere de una herramienta articulada que le permita identificar y alertar de forma anticipada aquellas situaciones

académicas, sociales, económicas o familiares que puedan derivan en eventos de deserción.

Un primer acercamiento al control de la situación ocurrió en el año 2010, cuando la Universidad de Manizales, en su Plan de desarrollo del Sistema Bien–Ser y Bien–Estar, se propuso implementar el Sistema de Información Gerencial de la Universidad de Manizales -SIGUM, herramienta informática que articulara el flujo de información disponible de cada estudiante, procedente de las facultades, Registro académico, la División financiera y Apoyo estudiantil, para monitorear la evolución durante su estancia en la universidad y desde allí, alertar sobre posibles casos de deserción. Sin embargo, el sistema de información SIGUM registra algunos datos relevantes, pero no todos los que son necesarios para lograr esta detección. La solución tecnológica actual solo aporta análisis descriptivos que en ningún caso predicen el quién, el cómo, el cuándo, el dónde y el por qué un estudiante podría estar en riesgo de abandonar sus estudios de pregrado en modalidad presencial (U. Manizales, 2010).

Posteriormente, la Universidad de Manizales se acoge a los requerimientos del sistema de información SPADIES del Ministerio de Educación Nacional, en busca de un camino complementario que le aporte a la solución del fenómeno. Sin embargo, este sistema de consolidación de datos de diversas universidades del país, también resulta insuficiente para sortear con éxito las necesidades de seguimiento, control y detección temprana de posibles desertores académicos.

Con base en esta situación, se concluye que necesita implementarse una solución tecnológica que efectúe análisis predictivos, basados en la búsqueda de correlaciones, patrones ocultos y útiles que le permitan generar alertas tempranas de deserción estudiantil, para que la Universidad, a través del programa de Acompañamiento estudiantil, ponga en curso estrategias de contención que impacten positivamente los indicadores de reducción del abandono temprano (dos primeros años de estudio).

2.2 Formulación del problema

En esta época de análisis de datos, la predicción está marcando el camino de la ciencia de los datos aplicada a situaciones problema cotidianas y complejas, que requieren tanto del entendimiento del problema como de considerar todas las variables que pueden intervenir. Por tanto, esta tendencia da sustento al proceso de construcción de una solución basada en inteligencia artificial, en su variante de *machine learning* de tipo supervisado, para responder la siguiente pregunta:

¿Es posible implementar un modelo de machine learning, en la Universidad de Manizales, que permita identificar los estudiantes de pregrado en modalidad presencial en riesgo de deserción temprana?

2.3 Justificación

Las altas tasas de deserción en la educación superior han despertado la necesidad de estudiar este fenómeno. Investigaciones señalan que universidades en Europa, Norteamérica y América Latina padecen este fenómeno, y dan cuenta de los altos niveles de estudiantes que no finalizan con éxito sus estudios universitarios. Las cifras son mucho más altas para Latinoamérica que para Europa o Estados Unidos; por ejemplo, mientras que en Bélgica la tasa es del 26.9% y en Estados Unidos, del 24%, en Colombia y Ecuador supera el 40%; y en Costa Rica y Brasil, el 50% (Alban y Mauricio, 2018).

En Colombia, el tema no ha pasado desapercibido para las autoridades competentes. “Uno de los mayores retos en educación superior es continuar disminuyendo la tasa de deserción, que sigue siendo alta y pone en evidencia la dificultad que tienen los jóvenes para permanecer en el sistema, así como para conseguir su graduación” (MEN, 2017). Abordar la deserción es tarea del Plan Decenal de Educación 2016-2026 (MEN, 2017), el Plan Nacional de Desarrollo 2018-2022 (DNP, 2018), Plan Sectorial de Educación 2010-2014 (MEN, 2010), así como de las universidades en Colombia, quienes

específicamente plantean logros, estrategias y acciones a partir de las cuales se espera reducir el índice de deserción y sus consecuentes afectaciones de tipo familiar, social y económico en la sociedad y el Estado.

La Universidad de Manizales ha implementado estrategias para contrarrestar la problemática, al focalizar acciones específicas para mantener un nivel de supervivencia académica adecuado, pero sobre un análisis de datos descriptivos; es decir, analiza el pasado para tratar de construir protocolos para el presente y futuro. A pesar de ello, persisten dificultades de control en la deserción que preocupan en gran medida, por su impacto económico y social.

Bajo estos argumentos, es necesario incrementar la efectividad de las estrategias mediadas por la tecnología, como herramientas de análisis de datos de índole predictivo, con el fin de clasificar a aquellos individuos más propensos a declinar su continuidad dentro de los programas, de manera que la institución pueda actuar para impedir el egreso prematuro.

Con este panorama, el presente trabajo de investigación surge de la necesidad de implementar modelos mediados por tecnología que puedan correlacionar datos producto de un sinnúmero de variables de deserción universitaria; en el caso de este estudio, aplicados específicamente a la Facultad de Ciencias Contables, Económicas y Administrativas, y la Facultad de Ciencias e Ingeniería de la Universidad de Manizales, con la intención de, una vez superada la investigación y confrontados sus resultados, hacer los ajustes e implementaciones necesarios, de manera que pueda aplicarse eficientemente en la Universidad para apoyar el Programa de acompañamiento estudiantil y reducir la tasa de deserción temprana.

Si bien es cierto, sobre deserción estudiantil, sus causas y cómo reducirla se ha estudiado desde hace décadas; no obstante, la aplicación de herramientas predictivas para ayudar a disminuir el fenómeno es reciente (véase Anexo 2). Los resultados de esta investigación sumados a los ensayos que han hecho otras universidades, además de

contribuir con la reducción del índice de deserción en la Universidad de Manizales, redundarán en modelos y herramientas que puedan replicarse y favorecer los índices de otras universidades privadas y públicas.

3 Antecedentes

Como se dijo anteriormente, una condición para alcanzar el objetivo propuesto es entender el problema de la deserción, de manera que la solución planteada pueda ponerse en marcha adecuadamente. Por esto, para iniciar, se revisaron investigaciones que abordaran la deserción estudiantil en Colombia, Latinoamérica y España, en los últimos años, con especial atención en sus causas, de tal manera que llegáramos a determinar una completa selección de las variables que pueden desencadenar la decisión de abandonar los estudios superiores en los dos primeros años de estudio.

Para comenzar a entender la complejidad del problema de la deserción en las universidades colombianas, está la investigación de Diego Barragán y Luceli Patiño (2013), quienes analizan diferentes interpretaciones del fenómeno. Colombia es un país que presenta números bajos de ingreso a la educación superior y, aún más bajos, de graduandos. Aunque las miradas usuales se han orientado a reflexionar y ofrecer mecanismos de intervención en las universidades, en la actualidad es necesario entender cómo universidad, sociedad y estado deben ofrecer las condiciones óptimas para que los jóvenes proyecten su futuro y materialicen sus sueños en un contexto con pocas opciones laborales y constantes crisis económicas y sociales (Barragán y Patiño, 2013). Los estudios sobre deserción se encuentran al día en las agendas educativas porque este fenómeno afecta la calidad del sistema educativo y, ante todo, porque deja vacíos en la política de cobertura educativa colombiana. Los autores concluyen que, aunque existe gran interés en atender el fenómeno, la falencia que más resaltan los diversos estudios

hallados, yace en la ausencia de estrategias y políticas efectivas y replicables que frenen la deserción en las universidades.

Una mirada puntual, ubica el problema en la Corporación Universitaria Lasallista (Londoño, 2013). El estudio describe factores de riesgo personales, académicos, institucionales y socioeconómicos hallados en una muestra conformada por 281 estudiantes activos en diferentes programas de la institución durante el año 2010 y 31 estudiantes que desertaron en 2009. Este análisis descriptivo mostró la distribución de frecuencias y medidas de tendencia central para hallar los factores relacionados con la deserción. Predominaron los factores de riesgo socioeconómicos, institucionales, académicos y personales, donde el socioeconómico fue el más representativo. Hacer evidente este factor encuentra consonancia con otros estudios realizados en Colombia y en Latinoamérica (Londoño, 2013).

Lo anterior se complementa con la propuesta investigativa de Gastón Quintela (2013)**,** quien muestra otra causa determinante de deserción: el factor sociológico como activador de decisiones en educación de los estudiantes y sus familias, respecto a los umbrales educativos mínimos por clase social. Concluye que los entornos sociales marginados por dificultades económicas y por déficit de acceso a tecnologías de la información y comunicaciones son los que requieren mayor apoyo para mantener la posición la estructura social–educativa como estrategia preventiva de la deserción universitaria (Quintela, 2013)**.**

Por su parte, el Ministerio de Educación Nacional también expresa su preocupación e interés respecto a este fenómeno (MEN, 2015). Acorde con el estudio del 2012 hecho por la Organización para el Desarrollo y la Cooperación Económicos-OCDE, sobre el Sistema de Educación Superior Colombiano, el aumento en cobertura es un indicador del avance gubernamental por fortalecer la educación técnica, tecnológica y profesional. Puede apreciarse que la tasa de pregrado paso del 24.4% al 46% entre 2002 y 2014. Estos altos niveles de inclusión y equidad en el sistema implican retos en el mejoramiento de la

calidad, pertinencia, permanencia y graduación de nuevos estudiantes**.** Este documento presenta cifras de los estudiantes de educación superior que no finalizan estudios universitarios. En 2013, la tasa de deserción en programas de pregrado fue del 44.9%, en programas técnicos y tecnológicos llegó al 62.4% y 53.8% respectivamente. En consonancia con lo anterior, para alcanzar las metas fijadas en términos de equidad y crecimiento, es necesario trabajar en estrategias de permanencia y continuidad para los estudiantes que ingresan al sistema de educación superior.

El trabajo de Torres, Acevedo y Gallo (2015), que recoge experiencias latinoamericanas, se enfoca en encontrar las causas de la deserción y sus consecuencias en el sistema educativo. Señalan la deserción como un fenómeno multicasual, en el que, además de las características personales del estudiante, donde el rendimiento académico tiene un gran peso, también intervienen componentes familiares, como la situación económica y educativa de los padres, o por componentes externos relacionados con el contexto de sus comunidades de origen e, incluso, por deficiencias del sistema educativo. Ellos consideran que conocer estas causas permitirá diseñar estrategias institucionales integrales que tengan un verdadero impacto en su disminución.

Además de profundizar en el entendimiento del problema, las causas encontradas servirán de base para determinar las variables de este proyecto, que los autores han dividido en factores individuales y factores regionales y contextuales.

Un estudio para caracterizar la deserción estudiantil en la Universidad de Caldas a partir de los datos registrados en el sistema SPADIES pone en evidencia algunos factores de deserción. Gartner, Dussán y Montoya (2016) encuentran que los estudiantes con mayor índice de deserción son hombres que presentan características como bajo desempeño en las pruebas de Estado, ocupación laboral durante la presentación de las pruebas de Estado, mayores de 25 años, con madre de nivel educativo básica o inferior, y sin vivienda propia.

María Camila Jiménez Mora (Jiménez Mora, 2021), también aborda la problemática a nivel iberoamericano, con el fin de encontrar las variables que influyen en la deserción y permanencia en la educación superior, desde un enfoque multinivel que contempla un contexto social, económico y cultural. En su propuesta también busca si las diferencias entre instituciones de educación superior inciden en el abandono; lo que resulta de valor para este proyecto que se centra específicamente en la Universidad de Manizales.

Un primer intento hacia encontrar la forma de predecir la deserción es la presentada por Andrés Giraldo, Carlos Zapata y Eliana Toro (2008) quienes, a través del proceso de Markov discreto en el estado y continuo en el tiempo, analizaron la probabilidad de ocurrencia de transferencia de estudiantes entre programas de pregrado y de deserción estudiantil en la Universidad Tecnológica de Pereira. Al respecto, son interesantes sus aportes sobre la obtención de índices descriptores de los fenómenos y sobre una forma de análisis que va más allá de los enfoques meramente estadístico-descriptivos.

Como paso intermedio de estos Antecedentes está la investigación de la Universidad de Oviedo, curso 2010-2011 (Bernardo et al., 2015), donde, además de determinar las variables (las cuales escogen teniendo en cuenta la disponibilidad y consistencia de las fuentes), los autores analizan diferentes modelos predictivos a partir de investigaciones previas; estos son: análisis correlacionales, análisis de regresión logística, análisis de supervivencia y minería de datos. Adicionalmente, aplican la primera metodología para determinar el valor predictivo de las variables de rendimiento académico previo, fecha de matrícula, rendimiento académico en el primer semestre de universidad y asistencia a clase. Como conclusión, destacan la importancia de utilizar herramientas predictivas para mitigar la deserción, así como tomar medidas para su prevención. La comparación entre diferentes metodologías predictivas es un aporte interesante para este proyecto por cuanto da luces que apoyan, de alguna manera, la tesis propuesta de utilizar una herramienta tecnológica basada en *Machine learning*, como idónea para predecir los posibles desertores, de manera que pueda revertirse su situación.

Paso seguido, se revisaron varias investigaciones que se concentran en una o varias de estas herramientas predictivas basadas en IA, aplicadas al fenómeno de la deserción universitaria. Al respecto, se encontraron trabajos tanto en Colombia como en otros países pues, se trata de un fenómeno presente en universidades alrededor del mundo, tal cual se había expresado anteriormente. Esta revisión fue de gran ayuda para el proyecto en tanto las investigaciones encontradas dieron luces respecto a las variables y formas de categorizarlas, así como a la aplicación de diferentes técnicas predictivas.

En la búsqueda de herramientas tecnológicas en torno al fenómeno de deserción estudiantil que demarquen el camino propuesto para la Universidad de Manizales, se analizó la solución tecnológica PASPE de la Universidad de la Costa, que es un sistema de caracterización y seguimiento a estudiantes con la finalidad de identificar aquellos alumnos más propensos a desertar (Cómbita, 2014)**.** Este sistema consolida la información disponible del alumno en indicadores que orientan a posteriori (análisis descriptivo sobre datos del pasado) al programa de acompañamiento y seguimiento para la permanencia del estudiante. Identificados los candidatos, se activa el conjunto de acciones alrededor de los servicios que la Universidad de la Costa provee para revertir la decisión de abandonar.

El interés en la problemática de deserción estudiantil universitaria ha aumentado en los últimos años. Tanto en el ámbito internacional como en el nacional y local se han realizado estudios con el propósito de determinar los motivos que causan este fenómeno. Las investigaciones previas de deserción-permanencia describen el impacto económico y social, subrayan las dificultades en los adolescentes, debidas a sus estudios en básica y media asociadas a las variables relacionadas con la trayectoria y desempeño académico. Concluyen que los promedios alcanzados asociados al rendimiento académico y la cantidad de materias que fueron reprobadas durante su permanencia en el colegio y universidad determinan el grado de éxito o fracaso de los estudiantes mexicanos en el examen nacional de ingreso a la licenciatura en ingeniería (Eckert y Suénaga, 2015)

Así mismo, señalan que las investigaciones resaltan la necesidad de aplicar técnicas de análisis de datos asistidos tecnológicamente para analizar el fenómeno de deserción y tratar de implementar acciones antes de que ocurra la decisión de abandonar. El uso de estos modelos beneficia a estudiantes, docentes, padres y gestores de la educación, no sólo para informar sobre la situación de los alumnos cuyo desempeño podría estar asociado con una característica particular (positiva o negativa), sino también como asesoramiento para la toma de decisiones.

Otro aporte para la definición de variables se encuentra en el trabajo aplicado en la Universidad de La Salle (Felizzola, Jaime Arias, Castillo, y Villa, 2018). Los autores indican que la deserción estudiantil es un fenómeno complejo que involucra diversos factores en los ámbitos sociales, económicos, familiares, psicológicos y académicos del estudiante. Estudios previos del Ministerio de Educación Nacional (MEN), indican que algunos de los factores determinantes en la deserción son: estrato, género, nivel educativo de los padres, ingresos económicos de la familia, clasificación según el SISBÉN, número de personas que componen el núcleo familiar, resultados de las Pruebas de Estado Saber 11° y ocupación del joven.

El informe sobre educación superior en América Latina y el Caribe, presentado por la UNESCO (2016), estudia la deserción en esta área geográfica; para esto, promedia las áreas de conocimiento con las tasas de deserción de los países de estas regiones, de donde resultado que el área Salud tiene una tasa de graduación de 54.2%; Administración y comercio, del 49.6%; Derecho, del 49%; Educación, del 48.8%; Ciencias sociales, del 47.4%; Agricultura, del 41.9%; Arte y la Arquitectura, del 40.8%; Tecnología e Ingeniería, 38.5 %; Ciencias básicas, de 36.8%, y el área de Humanidades, de 23.1%.

El propósito de esta investigación fue implementar un modelo predictivo de clasificación sobre la deserción temprana en la Facultad de Ingeniería de la Universidad de la Salle, a través de la aplicación de la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*). Para ello, se revisó la literatura desde 1982 hasta 2017;

se analizaron aplicaciones de *machine learning* y minería de datos para abordar la problemática con métodos como arboles de decisión, redes neuronales artificiales, vectores de soporte de máquina, naive bayes, uniform random, k vecino cercano, regresión logística, entre otros.

Estas técnicas de exploración de datos (matemática y tecnológica) son una alternativa aplicable al estudio del fenómeno de la deserción. El modelo puede ser predictivo (a futuro) o descriptivo (conocimiento previo) (Ethem, 2014).

De la literatura revisada, los autores concluyen que las técnicas más utilizadas que arrojan un mayor porcentaje de certidumbre son los árboles de decisión (Decisión Tree, *Random forest*) y las redes neuronales. De otra parte, señalan que ensambles no necesariamente proporcionan un mayor índice de precisión en los modelos (Felizzola, Jaime Arias, Castillo, y Villa, 2018).

En Chile, Ramírez y Grandon (2017), abordan el fenómeno por medio del software RapidMiner para mejorar la predicción, a partir de árboles de decisión con parámetros optimizados. La clasificación basada en árboles de decisión -CBAD- se basa en el enfoque de “dividir para conquistar”, según el cual, el algoritmo divide en cada paso, los datos en diferentes segmentos, de manera que en cada segmento se refleje, lo mejor posible, una de las clases asociadas a la clasificación en estudio. De aquí resulta un árbol invertido donde, cada nodo prueba un atributo particular de los datos y cada hoja, una decisión para una clase particular.

Las variables que determinó el análisis son el puntaje de ingreso, el promedio de notas y los años de avance en la carrera. Los resultados de la investigación chilena arrojaron un 82,27% de precisión, lo que refuerza nuestra hipótesis respecto a los árboles de decisión.

Por su parte, Miranda y Guzmán (2017), también en Chile, utilizaron los clasificadores de red bayesiana, árbol de decisión y red neuronal, obteniendo 76%, 75% y 83% de acierto, respectivamente. Esta investigación descriptiva se planteó dos hipótesis:

Por un lado, que las condiciones académicas del estudiante al momento de ingresar a la institución, determinan su permanencia o abandono en la carrera, y por el otro, son las condiciones socioeconómicas del estudiante las que la determinan.

Así mismo, señalan que la tarea de aprendizaje ideal para utilizar árboles de decisión es la clasificación, y destacan que, comparados con las redes neuronales o los vectores de soporte, los árboles arrojan resultados inteligibles para las personas y para sistemas semiautomáticos que procesen reglas.

Además de lo señalado, esta investigación presenta los parámetros de evaluación utilizados para cada clasificador; lo que se traduce en un insumo aprovechable para este proyecto.

Pérez, Grandón, Caniupán y Vargas (2018), comparan dos técnicas de predicción: la Regresión logística, utilizada en la UBB y los Árboles de decisión, como propuesta alternativa. Encontraron que dentro de las variables que mejor explican el fenómeno de deserción, encontradas por los algoritmos de regresión logística están la carrera de inscripción, el puntaje de matemáticas de la prueba del Estado y el nivel académico alcanzado en la enseñanza media, mientras que con las variables reflejadas al utilizar los algoritmos de árbol de decisión el orden cambió: en primer lugar, el nivel académico del estudiante en la enseñanza media, el lugar de domicilio del alumno y, por último, el puntaje de matemáticas. Los autores concluyen que los árboles de decisión logran mayor exactitud y precisión en sus predicciones. Así mismo, resaltan la importancia de contar con una base de datos suficiente, depurada y actualizada que permita una mayor precisión y eficiencia de la herramienta, condición que ya habíamos mencionado como necesaria para este proyecto.

En Ecuador, en 2018, se dio otro ejercicio de comparación. En primer lugar, tomaron como factor de análisis los alumnos, las universidades, el contexto académico, así como el social y económico; posteriormente, a través de Regresión logística, Árboles de decisión y Máquina de vectores de soporte si los factores escogidos se relacionan o

pueden contribuir a predecir la deserción (Alban y Mauricio, 2018). Consideraron 10 pruebas para cada algoritmo, se encontraron 11 factores determinantes de deserción y comprobaron que la clasificación a partir de Árboles de decisión es la técnica que presenta mayor precisión (98%). Uno de los aportes de esta investigación tiene que ver con la multiplicidad de las variables escogidas y cómo abarcan diferentes aspectos del estudiante, tanto propios como externos.

Bedregal, Aruquipa y Cornejo (2019), comparan y evalúan la precisión de diferentes técnicas de minería de datos para detectar patrones y predecir el comportamiento académico de los estudiantes. Exponen el rendimiento académico como uno de los factores determinantes de la permanencia o abandono de los estudios superiores y hacen sus pruebas con base en esta categoría y los diferentes indicadores que la determinan, aplicando minería de datos.

La investigación se desarrolla en 4 fases (metodología CRISP-DM): Comprensión del negocio; comprensión de los datos; preparación de los datos, y Modelado.

Los resultados arrojan un 90,24% de identificación correcta, así que queda validada su tesis respecto a la aplicación de las técnicas de minería de datos para predecir la deserción. Su modelo trae como novedad que además de tener en cuenta las notas obtenidas, considera cuántas veces ha llevado la asignatura y cómo ha sido su desempeño dentro del grupo o cohorte.

Viloria y otros (2019), también ponen a prueba las técnicas de minería de datos para abordar esta problemática, en la UAT (México), y comparan la coincidencia con las variables identificadas en la literatura, donde obtienen una correlación superior al 75%. Para esto, crean tres clasificadores que categorizan a los estudiantes, utilizan tres algoritmos (redes bayesianas, redes neuronales y árbol de decisión) y siguen el proceso Knowledge Database Discovery -KKD (selección, limpieza, transformación y proyección de datos).

Cada clasificador tiene sus particularidades y ventajas, de manera que no pueden compararse totalmente. Sin embargo, en la aplicación hecha hay una coincidencia respecto a que beneficios para los estudiantes, como becas y créditos, son variables determinantes de la deserción; así como el puntaje de la UAT, dentro de la categoría de rendimiento académico.

Por su parte, el método de *Random forest* es presentado por Utari, Warsito y Kusumaningrum (2020), quienes, además de presentar el método como beneficioso para reducir la tasa de deserción, muestran que el *random forest* tiene varias ventajas, entre las que están que puede producir menos errores, dar buenos resultados en la clasificación, manejar grandes cantidades de datos de entrenamiento de manera eficiente, así como métodos efectivos para estimar los datos faltantes. Debido a que hay un desequilibrio en los datos a procesar, aplican el método método Synthetic Minority Over Sampling (SMOTE), con el fin de aumentar la precisión del algoritmo. Respeto a esto, aunque señalan el valioso aporte de SMOTE, recomiendan utilizar otras técnicas de muestreo que permitan reducir aún más el desequilibrio encontrado en los datos.

Random forest aparece nuevamente, ahora en una investigación que busca desarrollar un modelo predictivo de la deserción universitaria que evalúa 5 algoritmos: árbol de decisión, KNN, *random forest*, redes neuronales y SVM (Guerra, Rivero, Ortiz, Diaz, y Quishpe, 2020). Los autores dividen las variables en tres categorías: personales-cognitivas, académicas-organizacionales y socioeconómicas. Para entrenar cada algoritmo, utilizaron el 60% de los datos, de manera que la prueba se realizó con el 40% restante. Un aporte destacado es la comparación que hacen entre los algoritmos utilizados. En esta investigación en particular (aplicada solo a 20 estudiantes), el árbol de decisión arrojó un 43% de exactitud, mientras que la red neuronal obtuvo un 92%.

Espinoza y Carretero (2020), conscientes de que el análisis predictivo es necesario si se quiere evitar la deserción estudiantil, desarrollan un modelo conceptual a partir de técnicas de aprendizaje automático (IA), que produce alertas preventivas y correctivas, con

el fin de evitar que el estudiante abandone prematuramente sus estudios. Con el uso de una regresión polinominal, que maneja datos segmentados de acuerdo con las diferentes dimensiones del problema, dentro de las que están la situación financiera, preparación secundaria, orientación y rendimiento académico, entre otras, y correlaciona las variables, su modelo predice hasta el 97% de los posibles desertores. Esto gracias, entre otras cosas, al uso de diferentes técnicas para subsanar los desequilibrios entre los datos, como la librería lmblearn.over_sampling, el método SMOTE, el coeficiente de presión y las métricas de regresión (Error absoluto medio "MAE", error cuadrático medio "MSE" y raíz cuadrada del error medio cuadrático "RMSE").

Además de lo enunciado, esta investigación presenta los momentos en los que el modelo no fue eficiente y cómo solucionaron los impases. Su modelo permite identificar el comportamiento de la deserción de los estudiantes, las variables comunes en la población, e identificar la distribución de la deserción según cada subconjunto de datos analizados.

En la Universidad Nacional de Colombia – Sede Medellín, Daniel Zapata Medina (2021) desarrolla una investigación alrededor de los métodos basados en IA para la detección de deserción estudiantil con el fin de crear un método nuevo que, además de incluir los aspectos cognitivos y académicos de los estudiantes, considere también los socio-económicos y personales, con el uso de métricas que posibiliten reunir y relacionar más información de los factores que más influyen, de manera que esto apoye y favorezca la representación de características de entrada a los algoritmos de aprendizaje. Lo anterior, con el fin último de mejorar el rendimiento del clasificador, así como facilitar la interpretación de los resultados de los algoritmos utilizados. Si bien, esta investigación no aplica árboles de decisión ni *random forest*, tiene un valor en cuanto al entendimiento del problema y de las variables que intervienen, así como por la validación de la efectividad de otros modelos de *machine learning*, sobre todo, por desarrollarse en una universidad colombiana.

Otro estudio local, esta vez de la UNAD (Universidad Nacional Abierta y a Distancia), se centra en los estudiantes de primera matrícula de la Escuela de Ciencias Básicas, Tecnología e Ingeniería, desarrollado por Ávila Pérez (2021), con el fin de desarrollar un prototipo de modelo predictivo de la deserción estudiantil utilizando herramientas de minería de datos, donde hace un diagnóstico inicial de los requerimientos del modelo, a partir de diferentes fuentes de información, y evalúa su efectividad al comparar los resultados con históricos de la universidad. Este estudio resalta que es necesario tomar en cuenta las particularidades de cada estudiante para detectar posibles desertores y es ahí donde cobra importancia la utilización de herramientas de Data Mining, pues permiten la extracción de factores particulares de cada individuo. Finalmente, reafirman una de las condiciones que hemos expuesto respecto a la calidad de la información, al concluir que la confiabilidad de lo obtenido depende, en gran medida, de partir de un set óptimo de datos.

Para concluir con las investigaciones nacionales recientes, está la de Chaparro, Cuatindioy y Barrera (2021), de la facultad de Ingeniería de la Universidad de Antioquia. Los autores compararon diferentes algoritmos de clasificación de aprendizaje supervisado, para identificar los perfiles de los posibles desertores, a partir de dos categorías: número de créditos inscritos en el último semestre cursado y semestre en el que se produce la deserción. Los algoritmos estudiados fueron: redes neuronales artificiales, regresión logística multinomial, métodos de ensamble (*random forest*, bagging, boosting) y máquinas de soporte vectorial. Los autores concluyen que el estado del arte les permitió organizar las variables de forma que el desempeño de los modelos fuera excelente; que las medidas de rendimiento de los modelos dependen de las características de las variables a clasificar y del comportamiento de los datos; que los datos arrojados se ajustan a los estándares nacionales debido a que las variables se clasificaron en académicas, institucionales, socioeconómicas e individuales (lineamientos de MinEducación); por último, que los métodos de ensamble, como *random forest*, arrojaron los mejores resultados para la clasificación de estudiantes desertores.

De 2022, tenemos presentes dos investigaciones, una de la facultad de Ingeniería Informática de la Universidad de Trás-os-Montes y Alto Douro (UTAD), de Portugal y la otra, en la facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos, en Perú.

La primera, sobre la deserción de los estudiantes de la UTAD (Moreira da Silva *et al.*, 2022), aplica técnicas de minería de datos a partir de datos existentes en la institución, razón por la cual se limitan a las notas académicas y la edad de ingreso de los estudiantes, pues otros datos de caracterización no son completos, aunque consideran que las notas académicas de alguna manera se ven afectadas por las condiciones de vida. Los autores presentan y analizan cuatro técnicas de *machine learning*, los algoritmos CatBoost, *Random forest*, XGBoost y redes neuronales artificiales. Para medir el desempeño utilizan la métrica AUROC, donde *Random forest* presenta las mejores métricas y solo es superado en recuperación por XGBoost. Lo alcanzado permite concluir que el análisis es posible, aunque los datos de los estudiantes sean escasos. Vale la pena resaltar que uno de los resultados indica que la madurez de los estudiantes tiene una alta influencia tanto en la culminación de los estudios como en el éxito obtenido en las actividades académicas más exigentes. Uno de los aportes del estudio de la UTAD a este proyecto es que sustenta las razones por las cuales *Random forest* es uno de los algoritmos seleccionados.

Finalmente, la investigación realizada en la Universidad Nacional Mayor de San Marcos (Vega *et al.*, 2022), propone vincular un sistema inteligente con un sistema de minería de datos que utilice el algoritmo de árboles de decisión para predecir qué estudiantes están en riesgo de desertar de sus estudios superiores. Además de describir ampliamente las fases del desarrollo de la investigación, los autores señalan las condiciones que promueven deficiencias en la efectividad del algoritmo, debido, por ejemplo, al desequilibrio de los datos o a que unas variables tienen más peso que otras, y presentan las acciones para sopesar estos errores, de manera que se pueda obtenerse una mayor precisión, a partir del uso de la entropía de la información y del paso de variables continuas a discretas, entre otras. Para el entrenamiento del modelo utilizan el 70% de los

datos y para su validación, el 30%, de donde obtienen un 90,34% y un 95,91% de precisión respectivamente. Además de considerar como válido el modelo propuesto, la investigación concluye que el promedio histórico de las calificaciones, el promedio de las calificaciones del último ciclo y el número de créditos aprobados son los factores que más inciden en la decisión de abandonar los estudios superiores, en los estudiantes de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos (Perú). Su aporte a este proyecto está principalmente en la confirmación de la utilización de árboles de decisión como modelo válido para predecir la deserción.

4 Objetivos

4.1 Objetivo general

Aplicar la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para diseñar e implementar un modelo de aprendizaje automático supervisado que permita identificar casos potenciales de deserción estudiantil universitaria en los primeros dos años de estudio (deserción temprana) en los programas de pregrado modalidad presencial de las facultades de Ciencias Contables, Económicas y Administrativas, y Ciencias e Ingeniería de la Universidad de Manizales.

4.2 Objetivos específicos

4.2.1 Todo el conjunto de actividades implementado con las herramientas: SQL Server como motor de la base de datos, herramientas de desarrollo e PL/SQL y Python. Realizar un análisis exploratorio de los datos (EDA) como parte inicial de la metodología CRISP-DM para comprender los patrones, las tendencias y las relaciones presentes en los datos de los estudiantes de las facultades de Ciencias Contables, Económicas y Administrativas, y Ciencias e Ingeniería de la Universidad de Manizales.

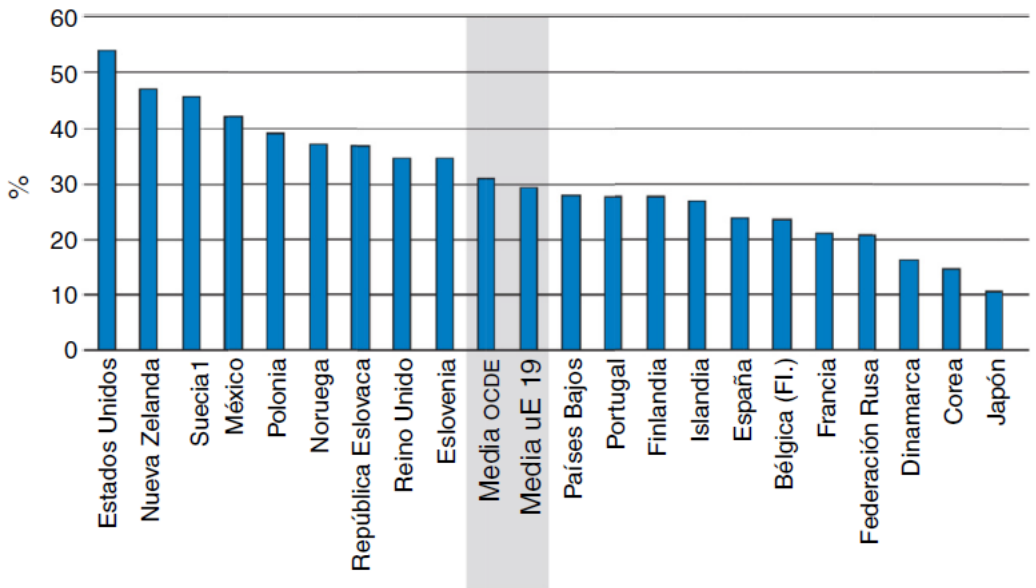
4.2.2 *Implementar al menos dos modelos de aprendizaje automático supervisado que permita identificar casos potenciales de deserción estudiantil universitaria en los primeros dos años de estudio (deserción temprana).*

4.2.3 Desarrollar un tablero de visualización interactivo que presente los resultados del modelo de predicción de forma clara y comprensible para los interesados. El tablero mostrará indicadores clave de rendimiento, métricas del modelo y presentará visualmente los datos obtenidos del modelo para facilitar la interpretación y el uso de los resultados por parte de los administradores y personal académico de la universidad. Este tablero será una herramienta crucial para la toma de decisiones y la implementación de intervenciones para prevenir la deserción estudiantil.

5 Referente Contextual

El fenómeno de la deserción estudiantil universitaria en los programas de pregrado presencial es concebido, tanto en Colombia y Latinoamérica como en Europa y Estados Unidos, como uno de los problemas de mayor relevancia para los gobiernos y las instituciones de educación superior. Las estadísticas internacionales revelan que cerca de un 31% de los estudiantes que ingresan a educación superior no logran graduarse en este nivel (OCDE, 2010) (Véase figura 1).

Figura 1. ***Deserción en programas de pregrado 2008 por país***

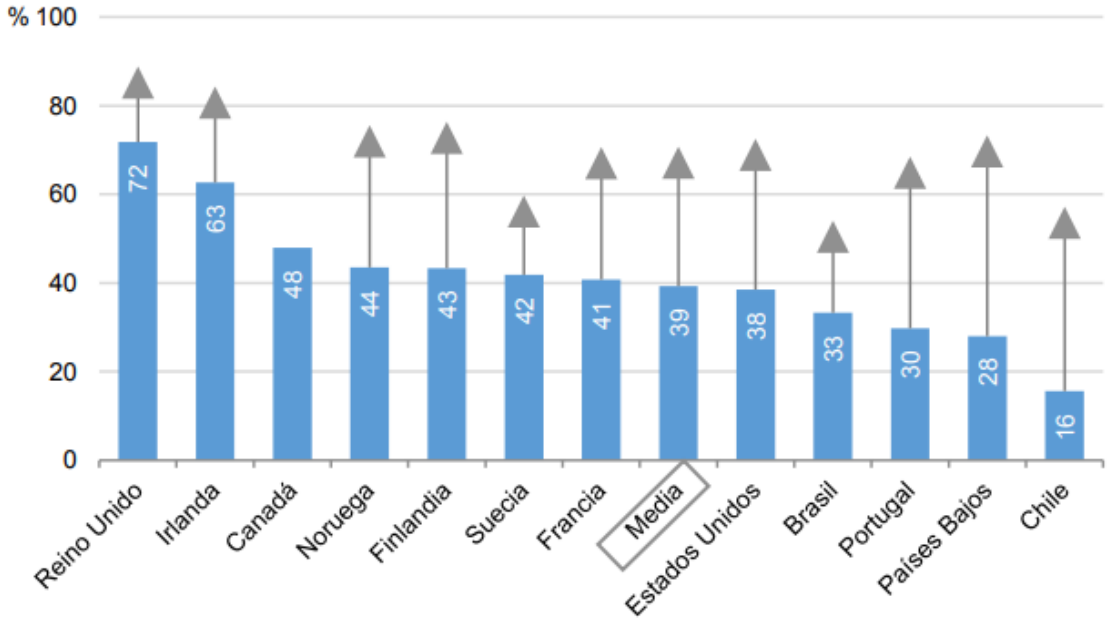


Nota: Tomado de Panorama de la educación 2010. Indicadores de la OCDE (p. 72), por OCDE, 2010, Santillana.

Aunque este panorama es desalentador, cabe anotar que los datos reflejan la deserción de los programas académicos, pero es posible que algunos de estos estudiantes se hayan cambiado de institución o de programa académico, situación que añade aún más complejidad a la intención de determinar el abandono de la educación superior (Fonseca y García, 2016).

En el informe de la OCDE de 2019, respecto a la tasa de finalización de educación terciaria, los datos corroboran lo ya expuesto (OCDE, 2019), que un número importante de estudiantes que ingresan a determinado programa académico no logra graduarse en éste (Véase figura 2). La tasa de finalización expresa “el porcentaje de estudiantes que accede a un programa educativo y se gradúa en él tras un número concreto de años” (p. 28).

Figura 2. ***Tasa de finalización 2017 por país***

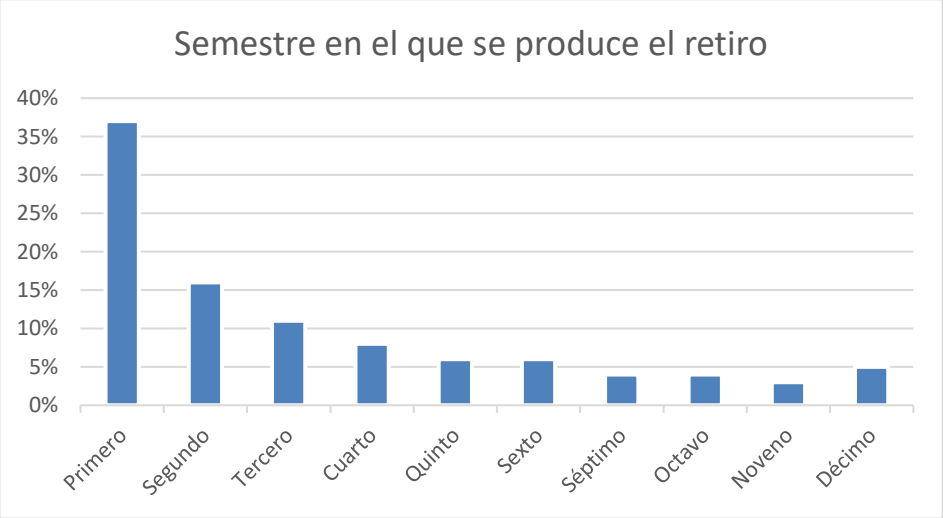


Nota: Tomado de Panorama de la educación 2010. Indicadores de la OCDE (p. 29), por OCDE, 2010, Santillana.

Ahora bien, además de reconocer la magnitud del problema, es importante tener en cuenta en qué momento del recorrido se presenta la deserción con mayor agudeza, con el fin de enfocar los esfuerzos por evitarla con eficacia y eficiencia. De acuerdo a investigaciones previas refieren la deserción estudiantil universitaria muestran un mayor impacto en el primero y segundo año de educación superior. De acuerdo con cifras del Ministerio de Educación Nacional:

> El 37% del total de los estudiantes desertores abandona sus estudios en primer semestre y el 16% en segundo semestre, es decir que más de la mitad de la deserción se concentra en los dos primeros semestres; más aún, el 78% de la deserción tiene lugar en la primera mitad de la carrera. (MEN, 2009, p. 75)

Figura 3. ***Deserción por semestre de abandono***



Nota: Tomado de *Deserción estudiantil en la educación superior colombiana* (p. 75), por MEN, 2009, Ministerio de Educación Nacional

Por su parte, en la página del Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior – SPADIES, que consolida la información

entregada por las instituciones de educación superior del país, pueden observarse datos más recientes. En el reporte sobre deserción con corte a noviembre de 2020, puede observarse, entre otros datos, la tasa de deserción cohorte promedio acumulada, es decir, el porcentaje del conteo acumulado de desertores hasta un semestre determinado, dato que podemos tener en cuenta debido precisamente a que nuestro interés se concentra en la deserción temprana. Como el SPADIES permite ver información particular de cada institución, de acuerdo con el nivel universitario (U) o técnico y tecnológico (TyT), veamos los datos de algunas universidades locales:

Tabla 1. ***Deserción por cohorte Eje Cafetero***

institución	deserción promedio cohorte	
	U	TyT
Universidad de Manizales	27,6%	44,6%
Universidad de Caldas	47,2%	41,1%
Universidad Autónoma de Manizales	33,5%	54,2%
Universidad Católica de Manizales	30,1%	43,9%
Universidad del Quindío	47,7%	50,1%
Universidad Tecnológica de Pereira	40,8%	43,0%
Universidad Católica de Pereira	46,3%	59,9%

Fuente: Adaptado de *Serie de Estadísticas por IES para el periodo 2019* [tabla], por SPADIES (2022), https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/357549:Estadisticas-de-Desercion

Como se ve en la Tabla 1, la deserción es un problema latente en el Eje Cafetero, con datos más preocupantes en unas instituciones que en otras y con una marcada diferencia entre los niveles universitario, y técnico y tecnológico. Lo expuesto, claramente ilustra que deben reforzarse las estrategias para minimizar el problema.

La Universidad de Manizales con el objetivo institucional de minimizar el fenómeno de deserción, implemento el “Programa de acompañamiento - Acceso y permanencia en la Universidad de Manizales” (U. Manizales, 2018), donde están consignadas las políticas, orientaciones y estrategias que se manejan para enfrentar y minimizar este fenómeno.

> Mediante el sistema de acompañamiento y Bienestar (objetivo y subjetivo), se ha logrado que los niveles de deserción sean razonables, potenciando la permanencia de los estudiantes, permanencia que se comprende como *“la capacidad de la Universidad para lograr que los estudiantes culminen su proceso de formación, en los tiempos previstos”.* (p. 4)

La Universidad cuenta con diferentes estrategias que le han permitido apoyar la permanencia de sus estudiantes; se trata de apoyos socio-económicos, promoción a la salud, acompañamiento psicopedagógico y participación estudiantil, principalmente, sumadas a la interacción de las acciones de dependencias, direcciones, coordinaciones y comités con las dinámicas de la Universidad.

La universidad trabaja de manera organizada la promoción y comunicación con los aspirantes a ingresar, el proceso de inscripción y de admisión, la inducción al ingreso a la vida universitaria y el acompañamiento integral; así mismo, realiza la caracterización de los estudiantes nuevos con el fin de tener información que le permita tomar decisiones respecto a las acciones y políticas en procura de potenciar la formación integral y la permanencia de los estudiantes (U. Manizales, 2018). Lamentablemente no toda la información esta ordenada y/o adecuadamente dispuesta para la investigación, por tanto, solo algunos datos de la información referida fueron utilizados en esta investigación.

> La hipótesis principal del Programa que tiene como núcleo el Bienestar, afirma que: “a un mayor y mejor acompañamiento a estudiantes y docentes de la Universidad, le corresponde incrementos manifiestos de bienestar objetivo y subjetivo, que redunda en la potenciación de la formación integral y permanencia

de los estudiantes, logrando que culminen sus estudios en los tiempos previstos de formación". (U. Manizales, 2018, p. 4)

Al ser el *Bienestar*, el núcleo del programa de acompañamiento, las acciones que de allí se desprenden pretenden propiciar niveles de bienestar social, económico y humano en concordancia con la misión de la Universidad y con el sujeto que se pretende formar. Entendemos bienestar como "el sentir de una persona al ver satisfechas todas sus necesidades en materia fisiológica y psicológica, así como contar con expectativas alentadoras que le sustenten su proyecto de vida" (Duarte y Jiménez, 2007, p. 305).

Aunque todos los objetivos del programa de acompañamiento propician la permanencia de los estudiantes, dos de éstos se relacionan más directamente con el fenómeno de la deserción temprana:

- Implementar estrategias preventivas y educativas orientadas a la disminución de los índices de deserción estudiantil.
- Establecer un sistema de seguimiento académico para los estudiantes con el fin de detectar factores de riesgo y contribuir a la identificación y aprovechamiento de sus potencialidades (U. Manizales, 2018, p. 11).

Este último objetivo cobra especial interés, precisamente porque la investigación se ubica en los dos primeros años de estudio, donde el desempeño académico es un factor determinante para permanecer o no en un programa específico. De acuerdo con el MEN (2009), solo el 22% de los estudiantes que desertaron en primer semestre no habían perdido ninguna asignatura, mientras que el 33% aprobó solamente una quinta parte de las materias inscritas; proporción contraria a la observada en los últimos semestres, donde el 67% de quienes desertan en décimo semestre han aprobado todas las asignaturas inscritas (MEN, 2009).

La División de Desarrollo Humano de la Universidad ejecuta diferentes acciones que favorecen condiciones para garantizar una mayor permanencia de los estudiantes. Se trata de:

> Apoyos psicológicos, pedagógicos, psicopedagógicos; acompañamiento a familias, programas de formación integral, de prevención y promoción de la salud, desarrollo de estrategias para la participación; becas y beneficios estudiantiles para la población estudiantil; financiación de matrícula internamente y a través de cooperativas y entidades bancarias; además de las estrategias específicas realizadas por cada programa. (U. Manizales, 2018, págs. 15-16)

Son evidentes el interés, los esfuerzos y las herramientas que la Universidad ha implementado para contener a los estudiantes matriculados hasta su graduación; sin embargo, también es evidente que la detección de casos posibles de deserción se ve limitada porque la información disponible no está unificada, validada ni automatizada en su completitud y, como se ha dicho, las causas de la deserción son variadas y dependen de la combinación de diferentes factores (sociales, vocacionales, económicos y familiares).

Esta automatización de la información y combinación de variables es posible con el desarrollo de una herramienta para predecir los posibles desertores a partir de *machine learning* y, para implementarla, se utilizará información correspondiente al periodo 2018-2022 de estudiantes de primero a cuarto semestre de la Facultad de Ciencias contables, económicas y administrativas, y la Facultad de Ciencias e Ingeniería. Los algoritmos del modelo se entrenarán con el 70% de la totalidad de los datos, y la precisión de la implementación, se verificará con el 30% restante.

6 Referente Normativo y Legal

El abordaje del marco legal dentro del presente trabajo de investigación, se enfoca en dos componentes importante, el *Hábeas data*, pues la herramienta propuesta se basa en el tratamiento de información personal, y la Política de Intervención ante la Deserción Estudiantil.

6.1 Hábeas Data

El habeas data es un recurso jurisdiccional que protege el derecho a la información y a la protección de los datos personales. El término se deriva del vocablo en latín "*habeo, habere*" que se traduce como tener, tomar, exhibir. Respecto a data, existe una disputa léxica; para algunos, proviene de latín "*datum*", singular de data, que significa donativo, presente, oferta, lo que se da; para otros, su origen es anglosajón y significa hechos, cosas conocidas; mientras que en portugués, el data se traduce como documentos, datos, como comúnmente se utiliza en informática (Muñoz de Alba, 2017). En conclusión, pude traducirse como "tener datos presentes".

> En el ámbito latinoamericano, fue la constitución brasileña de 1988, en su Art. 5°, Inc. LXXII, la primera en abordar estos temas, pero, sobre todo, también la primera en "bautizar" constitucionalmente al instituto del Hábeas Data. Dicha norma dispone que: "Se concederá Hábeas Data: a) para asegurar el conocimiento de informaciones relativas a la persona de quien lo pide, que consten en registros o bancos de datos de entidades gubernamentales o de

> carácter público; b) para la rectificación de datos, cuando se dé prefiera hacerlo en proceso reservado judicial o administrativo. (Eguiguren, 1997, p. 295)

De acuerdo con Zambrano, (2004):

> El Hábeas Data tiene por finalidad impedir que se conozca la información contenida en los bancos de datos respecto de la persona titular del derecho que interpone la acción, cuando dicha información esté referida a aspectos de su personalidad que están directamente vinculados con su intimidad y privacidad, no pudiendo entonces encontrarse a la libre disposición del público o ser utilizados en su perjuicio por órganos públicos o entes privados, sin derecho alguno que sustente dicho uso. Se trata, particularmente de información relativa a la filiación política, a las creencias religiosas, la militancia gremial, el desempeño en el ámbito laboral, o académico, entre muchos otros objetivos de protección. (p. 185)

En Colombia, la Constitución Política (CP, 1991), consagra el derecho al acceso a la información (artículo 20), así como garantías y derechos a los titulares de dicha información (artículo 15), como la protección a la intimidad personal, familiar y al buen nombre, o el derecho a conocer, rectificar y actualizar su propia información contenida en bases de datos públicas y privadas.

Según la Corte Constitucional:

> El derecho al habeas data otorga la facultad al titular de datos personales de exigir de las administradoras de esos datos el acceso, inclusión, exclusión, corrección, adición, actualización y certificación de los datos, así como la limitación en las posibilidades de su divulgación, publicación o cesión, de conformidad con los principios que regulan el proceso de administración de datos personales. (Sentencia C-748/11, 2011)

6.2 Política de Tratamiento y Protección de Datos de la Universidad de Manizales

Las Leyes 1266 (2008) y 1581 (2012) y sus decretos regulatorios, establecen una serie de obligaciones para todas las empresas en materia de protección de datos personales; por lo tanto, y para garantizar los derechos consagrados en la Constitución colombiana, la Universidad de Manizales ha adoptado la Política para tratamiento y protección de datos personales (U. Manizales, 2018), la cual puede consultarse en la página de la universidad y cuyo objetivo es:

> Garantizar a los titulares de los datos personales registrados bajo cualquier medio de almacenamiento digital de la Universidad de Manizales, y que no sean públicos, la reserva de la información, incluso después de finalizada su relación con alguna de las labores que comprende su tratamiento, pudiendo solo realizar suministro o comunicación de los mismos cuando ello corresponda al desarrollo de las actividades propias del objeto social de la Universidad de Manizales. (p. 1)

En este orden de ideas y para efectos de esta investigación, nos acogemos y seguimos la Política de tratamiento y protección de datos de la Universidad de Manizales, durante todo el desarrollo del proyecto.

6.3 Política de Intervención ante la Deserción Estudiantil

El Ministerio de Educación Nacional, a través del Decreto 1295 del 20 de abril (2010), establece la responsabilidad de la detección y mitigación de la deserción, en el artículo 6°, numeral 6.5 sobre bienestar universitario, así:

> El modelo de bienestar debe identificar y hacer seguimiento a las variables asociadas a la deserción y a las estrategias orientadas a disminuirla, para lo cual debe utilizar la información del Sistema para la Prevención y Análisis de la

> Deserción en las Instituciones de Educación Superior -SPADIES-, del Ministerio de Educación Nacional. Si se trata de un programa nuevo se deben tomar como referentes las tasas de deserción, las variables y las estrategias institucionales. (Decreto 1295 de 2010)

Asimismo, estable un Acuerdo nacional para disminuir la deserción en Educación superior (MEN, 2012), con el objetivo de fortalecer las estrategias de apoyo, en especial, las dirigidas a los estudiantes con mayor riesgo de desertar. El acuerdo resalta la necesidad de abordar esta problemática integralmente, con la participación de todos los actores involucrados; es decir, las familias, instituciones, entidades territoriales y el gobierno nacional. Entre otros intereses, también destaca la importancia de adelantar procesos de acompañamiento que garanticen el logro académico.

La Universidad de Manizales entra en este proceso de acuerdo con el requerimiento establecido por el MEN, desde el plan de desarrollo del sistema de Bien-Ser y Bien-Estar, en el cual establece un equipo interdisciplinario que se orienta hacia la búsqueda de la formación integral y el ofrecimiento de servicios para el mejoramiento de la calidad de vida, así como a evitar la deserción universitaria en la población estudiantil; para lo cual se establecen tres fases interdependientes: 1) Diseño e implementación de un sistema de información de estudiantes. 2) Análisis y diseño de una propuesta de políticas de Bien-Ser / Bien-Estar, a la luz de la implementación del sistema de acompañamiento a los estudiantes de la Universidad de Manizales. 3) Implementación del Programa de Acompañamiento a Estudiantes (PRAE) en la Universidad de Manizales.

7 Referente Teórico

El referente teórico de este proyecto está dividido en dos partes: 1) Deserción universitaria, conceptualización y variables, y 2) Minería de datos y herramientas de *machine learning* asociadas a la predicción de la deserción estudiantil.

7.1 Deserción universitaria

> Una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica. (MEN, 2009, p. 22)

El concepto de deserción implica que el estudiante abandone el sistema educativo en general. Es decir que no se cuentan como deserción, las transferencias entre instituciones o programas académicos (Tinto, 1987); por consiguiente, enfrentar este fenómeno implica “las acciones de política pública en la vigilancia y armonización de los flujos internos de alumnos, así como en la reducción, si es posible, de la pérdida de estudiantes de las instituciones educativas del sistema nacional de educación superior” (SPADIES, 2014).

Para complementar, Himmel (2002), agrega que la deserción debe valorarse “como el abandono prematuro de un programa de estudios antes de alcanzar el título o grado, y

considera un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore" (págs. 94-95). Se trata de un fenómeno universitario muy complejo, que no puede limitarse a una definición que lo abarque en su totalidad (Tinto, 1987), por lo que es importante para cada institución adoptar aquella más acorde con sus misión e intereses.

Dentro del abordaje de este fenómeno, el perfil vocacional cobra importancia puesto que características personales del estudiante como su personalidad, aptitudes, intereses vocacionales y habilidades, son factores que influyen en el desarrollo académico y, por tanto, pueden determinar la permanencia de un estudiante en un programa específico (Díaz C. *et al.*, 2009). Es más probable tener éxito en el proceso de formación universitaria para quienes conocen su capacidades y habilidades, tienen metas claras y están orientados a conseguirlas, y han escogido un programa acorde con su perfil.

En concordancia con lo anterior, de acuerdo con Zulma Perassi (2009), la deserción puede ser el punto final del fracaso escolar pues, con frecuencia, quienes desertan de la universidad, previamente han alargado su trayecto escolar por reprobar grados y materias, con lo que su autoestima se ha debilitado.

7.1.1 Modelos de Deserción

Dentro de las teorías que se han desarrollado para abordar el tema de la deserción universitaria y estudiantil, se debe de citar los modelos psicológicos de deserción estudiantil. De acuerdo con Himmel (2002), existen características, rasgos de la personalidad y otras variables individuales que distinguen a quienes desertan de quienes culminan su proceso educativo; "la deserción equivale al debilitamiento de las intenciones iniciales, pérdida de la motivación y finalmente desinterés del primer factor motivador" (Díaz C. *et al.*, 2009).

También, pueden considerarse como precursores de la decisión de desertar o continuar, las creencias y actitudes; la deserción puede ocurrir por el debilitamiento de las motivaciones e intenciones iniciales del estudiante (Himmel, 2002; Díaz C., 2008).

De otro lado, están las percepciones del estudiante una vez se encuentra integrado al sistema universitario (Antináis, citado por Himmel, 2002). Se destaca el componente económico y la capacidad para pagar los estudios como un factor determinante y, por tanto, la posibilidad de acceder a becas o subsidios puede alterar su percepción. Por esto, tanto subsidios como becas son herramientas que permiten apalancar la retención de estudiantes (Donoso y Schiefelbein, 2007), así como equilibrar las oportunidades para todos, lo que demuestra que los beneficios estudiantiles si tienen impacto en la retención de estudiantes.

Por su parte, Ethington (1990), elabora un modelo general sobre las conductas de logro con base en la teoría de Jacquelynne Eccles *et al.* (1983), a la que adiciona capacidades como el desempeño, la elección y la perseverancia. Según su postura, el rendimiento académico escolar influye en el rendimiento universitario porque determina tanto el autoconcepto y los valores, como las metas y expectativas de éxito.

Ahora bien, un ingrediente adicional se suma a partir de la teoría del suicidio de Durkheim (1928), superpuesta a la educación superior, con el fin de atraer las influencias externas a las psicológicas, donde la incongruencia normativa y la incorporación social deficiente son factores que incrementan la probabilidad de desertar; es decir que, la falta de integración al entorno universitario es también un factor desencadenante (Spady, 1970).

Estudios de la deserción en Estados Unidos han determinado los siguientes predictores de la deserción estudiantil: género, estado socioeconómico, promedio semestral, calidad de la carrera e integración social y académica (Díaz C., 2008). Para Spady (1970), el medio familiar es una de las principales fuentes de demandas, expectativas e influencias, respecto a la permanencia de los universitarios en determinado programa; de tal manera que, si la fuente es consistente y positiva, el desarrollo social y

académico corresponderá con las expectativas del estudiante y de la institución y, por lo tanto, será más probable su permanencia hasta finalizar el programa educativo.

Para Donoso y Schiefelhein (2007), otro factor que interfiere en la permanencia es la relación costo-beneficio, que tiene que ver tanto con el esfuerzo económico como con el social. Si el estudiante considera que estos esfuerzos no corresponden con los beneficios obtenidos o que obtendrá en el futuro, es probable que busque otras opciones que considere más favorables durante y después de su trayectoria académica.

7.1.2 Deserción según Vincent Tinto

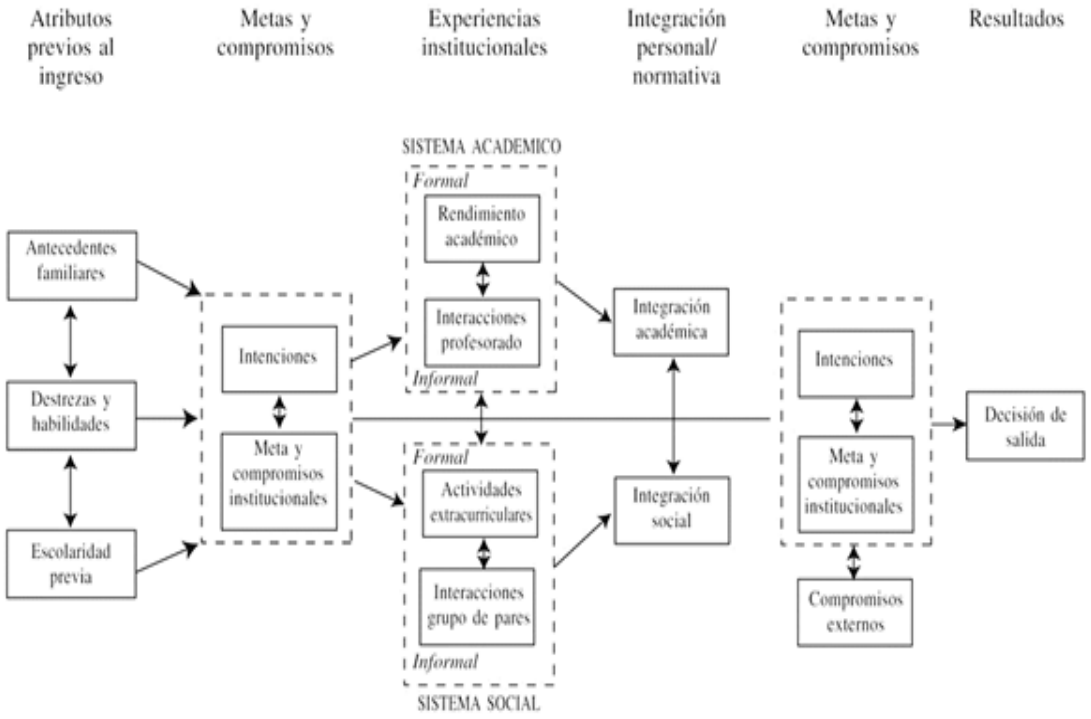
Vincent Tinto (1987) desarrolló un modelo que destaca el rol de las instituciones universitarias en la decisión de desertar, basado en el modelo de Spady (1970), y en la perspectiva organizacional. También agregó elementos de la teoría del intercambio, según la cual, las personas prefieren buscar recompensas, estados emocionales e interacciones, mientras evitan aquellas conductas que implican costos (Nye, 1979); esto, a raíz de la integración académica y social de los estudiantes.

A propósito del modelo de Tinto (1987), Himmel (2002) y Donoso y Schiefelhein (2007), explican que existen variables previas al ingreso a la educación superior que refuerzan o no, la adaptación del estudiante; variables relacionadas con las experiencias vividas alrededor de la educación precedente a la universidad. Se consideran tres tipos de atributos previos al ingreso: 1. Condiciones familiares (nivel socioeconómico y cultural, valores); 2. Competencias y capacidades personales (destrezas), y 3. Escolaridad previa (calidad de la formación previa, experiencias escolares): Estos atributos influyen en la actitud, el compromiso y la intención de lograr la graduación y titulación. La figura 4 describe el modelo de Tinto (1987).

En general, el modelo propuesto por Tinto (1987), indica:

> Si los beneficios percibidos de permanecer en una institución educativa son inferiores a los costos personales, si percibe otras actividades como mayormente recompensadas, o el historial de interacciones del estudiante con el sistema académico y social de la universidad es poco satisfactorio, entonces el estudiante tenderá a desertar de la institución. (Arredondo, 2011, p. 20)

Figura 4. ***Modelo de deserción de Tinto***



Nota: Tomado de “Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social” (p. 17), por V. Tinto (1987, p. 114), tal como aparece en Donoso y Schiefelbein (2007).

De acuerdo con la Figura 4, las expectativas institucionales se ubican en dos sistemas; el académico y el social. El primero se relaciona con la integración académica, a partir del desarrollo intelectual y el rendimiento académico, y repercute en el nivel de compromiso con el propósito de graduarse; mientras que la segunda tiene que ver con la

integración social que se desarrolla mediante la interacción con pares y docentes, y la participación en actividades extraclase, desde donde se conforma el compromiso con la institución universitaria. Entonces, a mayor compromiso con sus logros académicos y con su institución, menor será el riesgo de deserción.

Díaz (2008) señala que las instituciones que han utilizado el modelo de Tinto, ratifican la capacidad predictiva de estas dos categorías en la deserción, aunque no afecten en la misma medida a todos los estudiantes. Al respecto, Pascarella *et al.*, (1986), indican que el efecto de la integración institucional es indirecto, frente al gran peso que representa la integración académica para permanecer o no en un programa determinado.

Por su parte, para Himmel (2002), en diferentes formatos de institución, el modelo de Tinto presenta inestabilidad en la dirección y el peso de sus postulados. Lo cual es confirmado por Tinto, quien indica que el tipo de deserción, voluntaria o involuntaria, determina la forma en la que los factores inciden en ésta (citado por Himmel, 2002).

Como se ha dicho, hay diferentes perspectivas desde las cuales ver la deserción: desde el individuo, sus comportamientos y metas; desde los factores institucionales; o bien, desde factores gubernamentales o nacionales.

Cuando el comportamiento es analizado desde el punto de vista del individuo, desertar tiene un significado diferente para cada persona, pues cada quien tiene diferentes intereses y metas. La perspectiva individual de la deserción debe tener en cuenta las metas iniciales, las cuales pueden no corresponder con las metas de otros o, incluso, con la idea de graduarse; además, durante la trayectoria en la institución, las metas pueden cambiar y dejar de corresponder con el deseo de titularse (Tinto, 1987).

No todos los estudiantes que ingresan a un programa aspiran a completar el programa de estudios; según Tinto (1987), en este grupo pueden encontrarse tres tipos diferentes: quienes tienen metas educativas restringidas; los estudiantes que trabajan, y quienes tienen metas mayores a las que pueden cumplir estudiando en su universidad; este último caso se refiere a los que se inscribieron como forma para transferirse a otras

instituciones y a quienes desean abandonar porque creen que ya han alcanzado sus metas educativas, caso en el que la deserción no podría tomarse como fracaso, pues ese no es el significado que el alumno da a la acción de retirarse (Tinto, 1987).

Respecto a la perspectiva institucional, cabe anotar que no es posible para ninguna institución solucionar todos los casos de deserción; no obstante, hay momentos en los que la interacción alumno/institución puede ser un factor desencadenante. Según Tinto (1987), dos de estos momentos clave son el proceso de admisión y la transición del nivel secundario a la universidad. Durante el proceso de admisión se forman las primeras impresiones tanto del perfil académico como social de la universidad, difundidos a través de medios promocionales. Formar expectativas erróneas puede decepcionar al estudiante, lo que a su vez produciría que desertara; por esto, es importante que el material sobre la institución guie al estudiante a ser realista sobre sus expectativas en la institución.

En el proceso de transición del nivel secundario a la universidad se presentan más casos de retiro, especialmente al finalizar el primer año. Con frecuencia no se trata de salida del sistema de educación superior, sino de cambio de programa o institución. De acuerdo con un estudio de 2005 hecho por la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), de México, “el abandono voluntario ocurre durante los primeros meses posteriores al ingreso a la institución; y cinco de cada diez estudiantes desertan al inicio del segundo año” (Santes, Ramos, Lavoignet, Cruz, y Lara, 2017, p. 1). Durante los dos primeros meses de ingreso los estudiantes experimentan cambios drásticos de ambiente, pues además de novedoso puede ser impersonal y provocar la sensación de estar perdidos (Díaz C. , 2008), lo que justifica la máxima atención de la universidad para evitar la deserción temprana. Como medidas para superar estas primeras semanas, Tinto sugiere que alumnos adelantados sirvan como consejeros, realizar sesiones de orientación y asesoría, fomentar la conformación de grupos, entre otras medidas que, aunque sencillas, producen efectos rápidamente (Tinto, 1987).

En conclusión, conocer los tipos de abandono como las variables que lo desencadenan permite a las instituciones elaborar protocolos de retención adecuados y eficaces para minimizar os índices de deserción (Díaz C., 2008).

7.2 Ciencia de Datos

Las siguientes definiciones acerca de la Ciencia de datos han sido tomadas del libro online IAAR (López, 2017):

> Es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados. Es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y el análisis predictivo.
>
> En general, el proceso que utiliza la **Ciencia de Datos** para explorar el mundo usando datos es el siguiente:
>
> 1. El primer paso consiste en establecer un objetivo de investigación. El propósito principal aquí es asegurarse de que todos los interesados comprendan el *qué, cómo y por qué* del proyecto. Siempre debemos tener bien en claro cuál es la pregunta que queremos responder con la ayuda de los datos.
> 2. El segundo paso consiste en la obtención de los datos. Los datos deben estar disponibles para poder ser analizados. Este paso incluye encontrar los datos adecuados y obtener acceso a los mismos. El resultado de esta etapa suelen ser los datos en su forma cruda, que probablemente necesitarán ser pulidos y transformados antes de que puedan ser utilizados.

3. Ahora que ya tenemos los datos sin procesar, el siguiente paso es prepararlos. Esto incluye la transformación de los datos de una forma cruda a una forma en la que puedan ser utilizados directamente en los modelos. Para poder lograr esto, debemos detectar y corregir diferentes tipos de errores en los datos, combinar datos de diferentes fuentes y transformarlos. Una vez completado este paso, podemos avanzar hacia la visualización de datos y el modelado.
4. El cuarto paso es la exploración de datos. El objetivo de esta etapa es obtener una comprensión profunda de los datos. Buscaremos patrones, correlaciones y desvíos basados en técnicas visuales y descriptivas. Los conocimientos adquiridos en esta fase nos permitirán comenzar con el armado del modelo.
5. Finalmente llegamos al paso principal y más importante: la construcción de modelos. En esta etapa intentamos obtener los conocimientos o hacer las predicciones de acuerdo a los lineamientos establecidos en la primera etapa. Aquí podemos utilizar todas las técnicas y herramientas que nos proporciona el *Machine learning*. El objetivo es obtener el modelo o la combinación de modelos que mejor predicción proporcione.
6. El último paso del proceso de la Ciencia de Datos es presentar los resultados y automatizar análisis. Un buen modelo no sirve de nada si no es utilizado para mejorar la eficiencia y obtener mejores predicciones. En esta última etapa debemos presentar los resultados del análisis a las personas responsables de tomar las decisiones en las organizaciones para que los modelos puedan ser adoptados. (López, 2017)

7.2.1 Machine learning

Nota: Las siguientes definiciones acerca de *Machine learning* han sido tomadas del libro online IAAR (IBERDROLA, 2023):

> El *Machine learning* es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, *dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones* (análisis predictivo).
>
> El término se utilizó por primera vez en 1959. Sin embargo, *ha ganado relevancia en los últimos años debido al aumento de la capacidad de computación y al boom de los datos*. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del *Big Data.*
>
> Los algoritmos de *Machine learning* se dividen en tres categorías, siendo las dos primeras las más comunes:
>
> - Aprendizaje supervisado: estos algoritmos cuentan con un *aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones.* Un ejemplo es un detector de *spam* que etiqueta un *e-mail* como *spam* o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).
> - Aprendizaje no supervisado: estos algoritmos no cuentan con un conocimiento previo**.** *Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera.* Por ejemplo, en el campo del *marketing* se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas.

- Aprendizaje por refuerzo: su objetivo es que *un algoritmo aprenda a partir de la propia experiencia.* Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo a un proceso de prueba y error en el que se recompensan las decisiones correctas. En la actualidad se está utilizando para posibilitar el reconocimiento facial, hacer diagnósticos médicos o clasificar secuencias de ADN. (IBERDROLA, 2023)

7.2.2 Técnicas de análisis de datos para machine learning

De acuerdo con la categoría de los algoritmos (supervisados o no), pueden aplicarse diferentes técnicas de minería de datos, tal como se presenta en la tabla 2.

Tabla 2. ***Clasificación de las técnicas de minería de datos***

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (clustering)
Series temporales	Reglas de asociación
Random forest	Patrones secuenciales

Fuente: Tomado de Minería de datos (p. 51), por B. Beltrán, 2016, Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación.

Para la presente investigación se aplicaron métodos basados en árboles de decisión y *random forest*.

7.2.3 Árboles de decisión (decision trees)

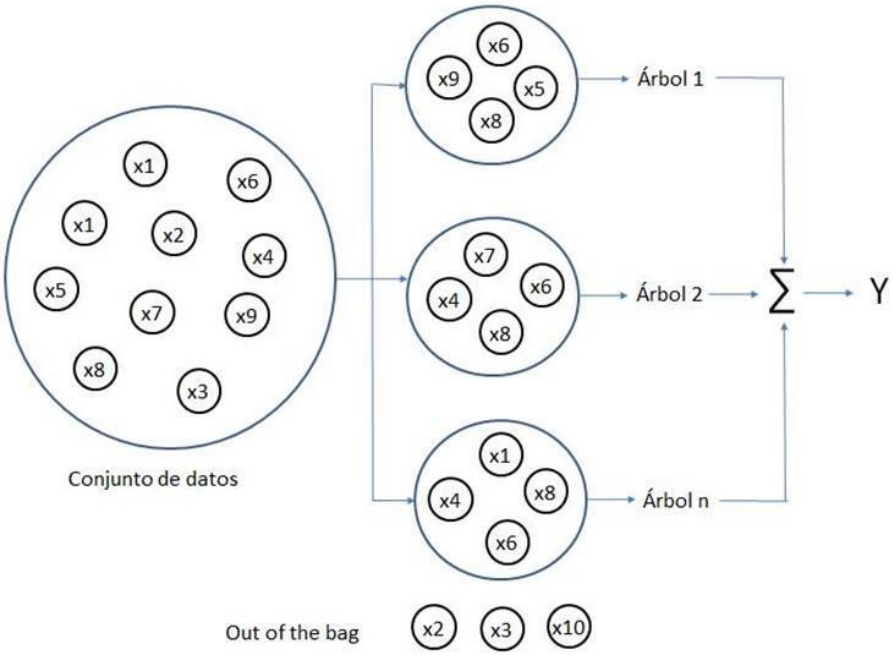
Se utilizan para descubrir reglas y relaciones. Para esto, rompe y subdivide sistemáticamente la información de la base de datos. La construcción del árbol de decisión se basa en los valores de las variables predictoras para partir en dos subconjuntos de observaciones el conjunto de datos (CART, Classification and Regression Tree), o en más de dos (CHAID, Chi Squared Automatic Interaction Detector). El proceso de partición de subconjuntos continúa con la aplicación del mismo algoritmo, “hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta” (Beltrán, 2016, p. 53). De manera que se conforma un árbol cuya raíz es el conjunto de datos íntegro; las ramas corresponden con los conjuntos y subconjuntos, y se llama nodo a cada conjunto donde se hace una partición. Cuando quiere dividirse una población en diferentes segmentos a partir de un criterio de decisión específico, puede utilizarse el método CHAID.

7.2.4 Random forest (Bosque aleatorio)

Este algoritmo de aprendizaje supervisado produce varios árboles de decisión a partir de un conjunto de datos de entrenamiento, con el fin de combinar la predicción obtenida para llegar a un modelo único, que será más robusto que los resultados de cada árbol de decisión por separado (Lizares, 2017). Con este método, se genera un gran número de árboles de decisión, donde cada uno contiene un subconjunto aleatorio de variables y debe crecer hasta lo máximo posible.

Cuando el modelo se entrena para predecir, el resultado corresponde con el promedio de salidas de todos los árboles, tal como se muestra en la Figura 6.

Figura 5. ***Algoritmo Random forest***



Nota: Tomado de “Aplicación de algoritmos *Random forest* y XGBoost en una base de solicitudes de tarjetas de crédito” (Espinoza, 2020, p. 3).

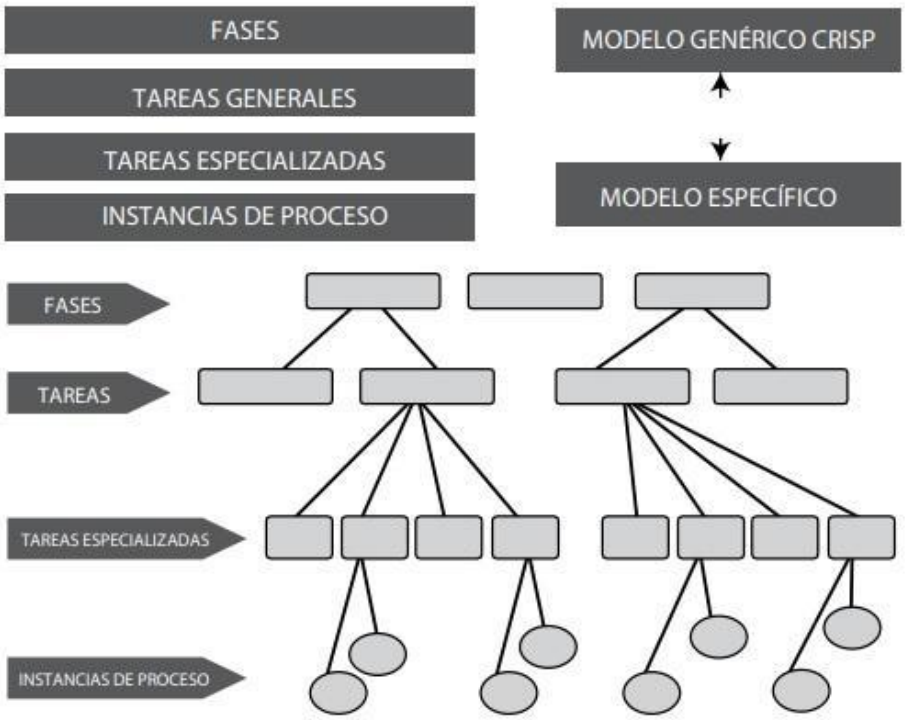
7.2.5 Metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining)

Esta metodología surgió en la década del 90, gracias al desarrollo impulsado por líderes de la industria europea, con el fin de apalancar la evolución de la minería de datos. Una de sus ventajas es que se construyó a partir de experiencias reales sobre la forma en la que las personas hacen proyectos, y no, de manera académica y teórica (Moro *et al.*, 2011). Actualmente, se utiliza con frecuencia en los proyectos de Data Mining.

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining), a partir de la identificación de diferentes niveles de abstracción, organiza un conjunto de

tareas de forma jerárquica, del nivel más general hasta el más específico, en 4 niveles de abstracción: a) fases, b) tareas generales, c) tareas especializadas y d) instancias de proceso (véase Figura 7).

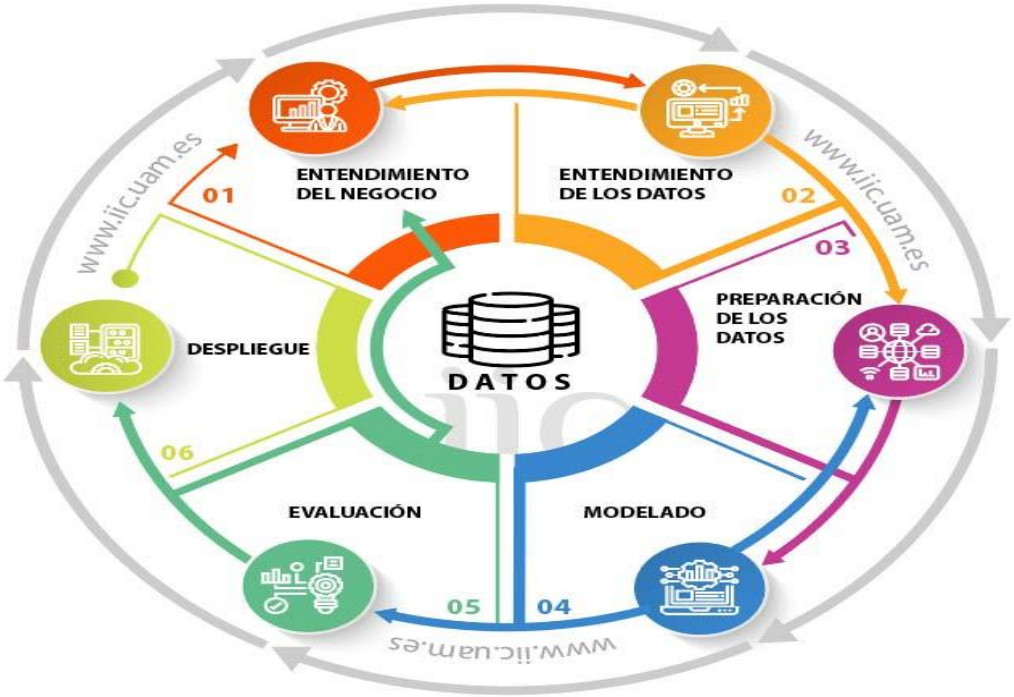
Figura 6. ***Niveles de abstracción de la metodología CRISP-DM***



Nota: Tomado de *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional***,** (Timarán y otros, 2016, p. 14)**.**

Esta metodología se desarrolla en seis fases de interacción iterativa durante el ciclo de vida de los proyectos de minería de datos, así: 1) Comprensión del problema o negocio, 2) Comprensión de datos, 3) Preparación de datos, 4) Modelado, 5) Evaluación del modelo y 6) Implementación del modelo, tal como se presenta en la Figura 8.

Figura 7. ***Fases de la metodología CRISP-DM***



Fuente: instituto de ingenieria del conocimiento, metodologia CRISP-DM

Fase 1 - Entendimiento del negocio o problema: En esta fase requiere entenderse completamente el problema por resolver, para determinar los objetivos y requisitos del proyecto y traducirlos a objetivos técnicos para elaborar el plan de desarrollo.

Fase 2 - Entendimiento de los datos: Se realiza una primera recolección de datos con el fin de lograr un contacto inicial con el problema. Dentro de las tareas de esta fase están la recolección, descripción, exploración y verificación de la calidad de los datos iniciales.

Fase 3 - Preparación de los datos: En esta fase se adaptan los datos a la técnica de minería de datos; para esto, se visualizan los datos y se buscan relaciones entre las

variables. Esta fase interactúa con la fase de modelado porque es necesario procesar los datos de diferentes maneras.

Fase 4 - Modelamiento: Se escoge un modelo específico y adecuado que sea apropiado para el problema, disponga de datos adecuados, cumpla con los requisitos del problema y se tenga un pleno conocimiento pleno de éste. En el presente caso de estudio, por tratarse de un problema de predicción, pueden elegirse árboles de decisión y técnicas *random forest*.

Fase 5 - Evaluación del modelo: Como su nombre lo indica, en esta fase se evalúa el modelo seleccionado interpretando los resultados mediante diferentes herramientas que permitirán saber si se cumplieron los criterios de éxito. Si los resultados son satisfactorios, se da paso a la explotación del modelo.

Fase 6 - Implementación del modelo: Cuando el modelo ha sido construido y validado, el conocimiento adquirido se transforma en acciones dentro del proceso de negocio. En esta fase se planea la implementación, se hace mantenimiento y monitoreo, se efectúa el informe final y se revisa el proyecto. De la mano de la implementación y por medio de herramientas de visualización de datos se construyen tableros, gráficos y otros elementos que facilitan el entendimiento al usuario final de los resultados obtenidos en la predicción.

Con los elementos anteriormente descritos, pretende lograrse una empleabilidad efectiva para construir el sistema de información de analítica predictiva con la finalidad de anticipar el evento de deserción temprana estudiantil en los programas de pregrado presencial de la Universidad de Manizales, para lo cual se desarrollará el proyecto en las facultades de Ciencias contables y económicas y Ciencias e Ingeniería, sometiéndose a prueba los diferentes algoritmos estadísticos y matemáticos que más se acerquen a la efectividad y aplicabilidad del mismo.

7.2.6 Análisis Exploratorio de Datos (EDA)

Nota: Las siguientes definiciones acerca del análisis exploratorio de datos (EDA) sido tomadas de la documentación de IBM analitycs, (2021). Análisis exploratorio de datos. Ibm.com. https://www.ibm.com/es-es/topics/exploratory-data-analysis.

Los científicos de datos utilizan el análisis exploratorio de datos (EDA) para analizar e investigar conjuntos de datos y resumir sus características principales, a menudo empleando métodos de visualización de datos. Ayuda a determinar la mejor manera de manipular los orígenes de datos para obtener las respuestas que necesita, lo que facilita que los científicos de datos descubran patrones, detecten anomalías, prueben una hipótesis o verifiquen suposiciones.

El EDA se utiliza principalmente para ver qué pueden revelar los datos más allá del modelado formal o tarea de prueba de hipótesis y proporciona una mejor comprensión de las variables del conjunto de datos y las relaciones entre ellas. También puede ayudar a determinar si las técnicas estadísticas que usted está considerando para el análisis de datos son apropiadas. Desarrolladas originalmente por el matemático estadounidense John Tukey en la década de 1970, las técnicas EDA continúan siendo un método ampliamente utilizado en el proceso de descubrimiento de datos en la actualidad.

El objetivo principal del EDA es ayudar a analizar los datos antes de hacer suposiciones. Puede ayudar a identificar errores obvios, así como a comprender mejor los patrones dentro de los datos, detectar valores atípicos o eventos anómalos y encontrar relaciones interesantes entre las variables.

Los científicos de datos pueden utilizar el análisis exploratorio para garantizar que los resultados que producen sean válidos y aplicables a los resultados y objetivos de negocio deseados. El EDA también ayuda a las partes interesadas mediante la confirmación de que están haciendo las preguntas correctas. El EDA puede ayudar

a responder preguntas sobre desviaciones estándar, variables categóricas e intervalos de confianza. Una vez que el EDA está completo y se obtienen los conocimientos, sus características se pueden usar para un análisis o modelado de datos más sofisticado, incluyendo el machine learning.

Las funciones y técnicas estadísticas específicas que puede realizar con las herramientas de EDA incluyen:

- Técnicas de agrupación y reducción de dimensiones, que ayudan a crear visualizaciones gráficas de datos de alta dimensión que contienen muchas variables.
- Visualización univariante de cada campo en el conjunto de datos sin procesar, con estadísticas de resumen.
- Visualizaciones bivariante y estadísticas de resumen que le permiten evaluar la relación entre cada variable en el conjunto de datos y la variable de destino que está viendo.
- Visualizaciones multivariante, para mapear y comprender interacciones entre diferentes campos en los datos.
- La agrupación en clústeres de K-medias (K-means clustering en inglés) es un método de agrupación en aprendizaje sin supervisión en el que los puntos de datos se asignan en K grupos, es decir, el número de clústeres, en función de la distancia desde el centroide de cada grupo. Los puntos de datos más cercanos a un centroide específico se agruparán en clústeres en la misma categoría. La agrupación en clústeres de K-medias se utiliza comúnmente en la segmentación del mercado, el reconocimiento de patrones y la compresión de imágenes.
- Los modelos predictivos, como la regresión lineal, utilizan estadísticas y datos para predecir resultados.

Hay cuatro tipos principales de EDA:

- **Univariante no gráfico.** Esta es la forma más simple de análisis de datos, donde los datos que se analizan constan de una sola variable. Dado que es una sola variable, no se ocupa de causas o relaciones. El propósito principal del análisis univariante es describir los datos y encontrar patrones que existen dentro de ellos.
- **Univariante gráfico.** Los métodos no gráficos no proporcionan una imagen completa de los datos. Por tanto, se requieren métodos gráficos. Los tipos comunes de gráficos univariantes incluyen:
    - Diagramas de tallos y hojas, que muestran todos los valores de los datos y la forma de la distribución.
    - Histogramas, un diagrama de barras en el que cada barra representa la frecuencia (recuento) o proporción (recuento/recuento total) de casos para un rango de valores.
    - Diagramas de caja, que representan gráficamente el resumen de cinco números de mínimo, primer cuartil, mediana, tercer cuartil y máximo.
- No gráfico multivariante: los datos multivariantes surgen de más de una variable. Las técnicas del EDA no gráfico multivariante generalmente muestran la relación entre dos o más variables de los datos a través de tabulaciones cruzadas o estadísticas.
- Gráfico multivariante: los datos multivariantes utilizan gráficos para mostrar las relaciones entre dos o más conjuntos de datos. El gráfico más utilizado es un gráfico de barras agrupadas o un gráfico de barras en el que cada grupo representa un nivel de una de las variables y cada barra dentro de un grupo representa los niveles de la otra variable.

Otros tipos comunes de gráficos multivariantes incluyen:

- Diagrama de dispersión, que se utiliza para trazar puntos de datos en un eje horizontal y vertical para mostrar cuánto se ve afectada una variable por otra.
- Gráfico multivariante, que es una representación gráfica de las relaciones entre factores y una respuesta.
- Gráfico de ejecución, que es un gráfico de líneas de datos trazados a lo largo del tiempo.
- Gráfico de burbujas, que es una visualización de datos que muestra varios círculos (burbujas) en un gráfico bidimensional.
- Mapa de calor, que es una representación gráfica de datos donde los valores se representan por color.

7.2.7 Visualización de Datos

La ciencia de datos es un área del conocimiento que combina diferentes disciplinas y cuyo principal objetivo es convertir los datos en valor real; estos pueden ser estructurados o no estructurados, en gran o poca cantidad, estáticos o en reproducción en línea y su valor puede ser dado a manera de predicciones, decisiones automatizadas, modelos entrenados o mediante visualización de datos que entreguen información de importancia (Van der Aalst, 2016). O como lo destaca en su articulo Margaret Rouse, (2022). Visualizacion de datos. Computerweekly.com. https://www.computerweekly.com/es/definicion/Visualizacion-de-datos.

La visualización de datos es utilizada en una amplia variedad de campos, desde pequeños emprendimientos en los diferentes sectores económicos hasta la ciencia y la tecnología. Por ejemplo la representación de las ventas, recaudos y costos en negocios como pastelerías, comidas rápidas, transporte escolar, así como resultados durante procesos de investigaciones climáticas para la prevención de víctimas de los huracanes en Estados Unidos, buscando con esto identificar patrones que originen los diferentes

causas del fenómeno producto de la investigación, hallar las relacione entre las diferentes variables, detección de errores o anomalías en los datos, lo que puede ayudar a mejorar la calidad de la data incorporada en el estudio y por ende mejorar la precisión de los análisis y/o predicciones.

Power BI es una herramienta de análisis y visualización de datos de Microsoft. Permite conectar, modelar y visualizar datos de diversas fuentes en una sola plataforma, lo que facilita la toma de decisiones basadas en datos. Power BI tiene una amplia gama de funciones de análisis y visualización que permiten crear informes interactivos y paneles de control en tiempo real.

Power BI está diseñado para ser fácil de usar y no requiere conocimientos avanzados de programación o informática. Permite conectar y transformar datos de diversas fuentes, incluyendo bases de datos, archivos de Excel, servicios en la nube y aplicaciones empresariales. Una vez que los datos se han conectado, se pueden modelar y transformar para crear relaciones y jerarquías que permiten analizarlos de manera más efectiva.

Power BI también ofrece una amplia gama de opciones de visualización, incluyendo gráficos, tablas, mapas y cuadros de mando interactivos. Los usuarios pueden personalizar los informes y los paneles de control para satisfacer sus necesidades específicas y compartirlos con otros usuarios de la organización.

Power BI se integra con otras herramientas de Microsoft, como Excel y SharePoint, y también tiene una amplia gama de conectores de datos para trabajar con otras herramientas y servicios de terceros. Además, Power BI ofrece capacidades de análisis avanzado, como aprendizaje automático e inteligencia artificial, que permiten a los usuarios obtener información más profunda y detallada de sus datos.

8 Metodología

8.1 Enfoque metodológico

Este proyecto de investigación es descriptivo y predictivo. La investigación es descriptiva, porque trabaja sobre realidades de hecho y su característica fundamental es la de presentar un panorama general de los hechos, al describirlos tal como son observados, es decir la realidad se describe en todos sus componentes principales (Lerma, 2009). En este caso, se trata de la descripción minuciosa y detallada de los datos que puedan analizarse y procesarse para predecir quiénes son los estudiantes de pregrado de las facultades de Ciencias Contables, económicas y administrativas y Ciencias e Ingeniería de la Universidad de Manizales, modalidad presencial, que estarían en riesgo de desertar durante los dos primeros dos años de estudio del programa de educación superior (deserción temprana).

También es una investigación predictiva porque orienta el proceso investigativo hacia el uso del algoritmo *random forest*, para poder determinar las correlaciones e incidencias del conjunto de variables que puedan explicar el fenómeno de la deserción estudiantil en los programas de pregrado, modalidad presencial, de las facultades de la Universidad de Manizales seleccionadas.

El *random forest* es una de las técnicas de clasificación de datos más común para este tipo de problema, ofrece algunas ventajas en relación con otras técnicas de clasificación, como representar el conocimiento extraído a través de reglas de decisión de

fácil entendimiento, producir menos errores, dar buenos resultados en la clasificación y manejar grandes cantidades de datos de entrenamiento de manera eficiente, tal como lo indican varias investigaciones similares (Chaparro *et al.*, 2021; Guerra et. Al., 2020; Moreira da Silva *et al.*, 2022; Utari *et al.*, 2020).

8.2 Tipo de estudio

Se trata de una investigación cuantitativa de tipo correlacional, pues pretende entender y determinar la interacción entre las diferentes variables y grupos de variables que intervienen en el fenómeno de la deserción universitaria. Así mismo, es de tipo explicativo, porque pretende encontrar las causas del problema de estudio (Hernández *et al.*, 2014).

El alcance de este estudio en particular, permitió la identificación y clasificación de las relaciones entre variables, así como determinar en qué condiciones se manifiesta el riesgo de deserción estudiantil y quiénes serían los estudiantes en dicho riesgo, como insumo para que el departamento de Bienestar estudiantil de la Universidad de Manizales pueda activar sus protocolos de contención del fenómeno de manera más eficiente y oportuna, así como desarrollar otras estrategias a partir de lo señalado por el modelo diseñado.

8.3 Diseño de investigación

Esta investigación se desarrolló a partir de la metodología CRISP-DM, específica para proyectos de minería de datos, explicada de manera general en el sexto capítulo de este texto (Referente teórico), con la que se completaron tres fases: 1) Análisis exploratorio de los datos, 2) Generación del modelo de inteligencia artificial clasificatorio y 3) Análisis de los resultados obtenidos.

8.3.1 Fase 1: Análisis exploratorio de los datos

8.3.1.1 Entendimiento del negocio

Se parte de la comprensión de las políticas y protocolos que sigue la Universidad de Manizales para mitigar el fenómeno de deserción universitaria de los estudiantes de pregrado presencial, así como de los datos históricos registrados al respecto. Una primera fuente de información es el SIGUM (Sistema de Información Gerencial de la Universidad de Manizales), donde se articula información de cada estudiante, procedente de las facultades, Registro académico, División financiera y Apoyo estudiantil; sin embargo, este sistema solamente aporta análisis descriptivos del fenómeno (U. Manizales, 2010).

Una segunda fuente es el sistema de información SPADIES del Ministerio de Educación Nacional, que consolida información de un gran número de instituciones de educación superior en Colombia, pero, nuevamente, el alcance es meramente descriptivo.

Finalmente, está el “Programa de acompañamiento - Acceso y permanencia en la Universidad de Manizales” (U. Manizales, 2018), que contiene las políticas, orientaciones y estrategias que se manejan para enfrentar y minimizar este fenómeno, liderado por Bienestar estudiantil. Ciertamente, este departamento tiene un conocimiento amplio del fenómeno y requiere de alguna herramienta que emita alertas tempranas para que pueda llegar a más estudiantes en riesgo de desertar.

8.3.1.2 Entendimiento de los datos

Esta fase inicia con la recolección de datos a partir de los sistemas de información y otras fuentes de datos la Universidad de Manizales, la comprensión de su naturaleza y significado y el análisis de calidad para reconocer su validez.

Para esto se realizaron de manera ordenada las siguientes cuatro actividades:

Actividad 1: Recepción de los archivos de datos en formato de texto (.csv). Los archivos corresponden a: Puntajes saber, hoja de vida académica, historial de matrícula, historial de atención del programa de acompañamiento estudiantil.

Actividad 2: Conversión de los archivos de texto (.csv) a archivos en Excel formato 97-2003.

Actividad 3: Almacenamiento de los archivos en Excel (.xls) en las tablas de las bases de datos stage BD_DESERCION_STG. Se importan sin ninguna transformación. Este proceso técnicamente se conoce como ELT (extracción, carga y transformación de la data).

Actividad 4: Descripción de los datos. Al revisar la totalidad de la data recibida como fuente se determina la categoría de cada una de las variables, al ejecutar sobre ellas validaciones de la calidad, así como al aplicar algunas reglas de ajuste sobre ellas para adaptarlas a los significados apropiados. Los datos residen en SQL Server en la base de datos BD_DESERCION_MRL, mediante conteos, valores únicos, valores comunes, agrupamiento de valores, etc.

Para las variables continuas se construyeron medidas de tendencia central (media, mediana, moda, desviación estándar y distribución por cuartiles). Los análisis exploratorios se enfocaron en análisis univariado, bivariado y multivariado.

8.3.2 Fase 2: Implementación de modelos clasificatorios que permitan explicar la deserción temprana al interior de la Universidad de Manizales.

8.3.2.1 Preparación de los datos

Durante esta etapa del proceso investigativo, se busca ajustar la data a los requerimientos técnicos del modelado de preparación para el algoritmo de *random forest*,

con acciones como agrupar variables existentes, seleccionar las mejores variables para el modelo (feature importance) y derivar nuevas variables en función de las existentes.

8.3.2.2 Modelado

En esta etapa del proceso se prepara y configura el algoritmo *random forest* que va a emplearse para la predicción de la deserción. El algoritmo toma las variables previamente seleccionadas como de mayor importancia para la explicación del fenómeno (feature importance).

Posterior a ello, se construyen las variables X (variable a predecir: desertor o no desertor), y la variable Y (conjunto de variables que explican el fenómeno).

Establecidas plenamente X y Y, se seleccionan el conjunto de datos para entrenamiento, equivalente al 70% de la data, y el conjunto de datos test, equivalente al 30% de la data.

8.3.2.3 Evaluación

La base fundamental de la evaluación de resultados se dirige por la matriz de confusión, la cual muestra la cantidad de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos de la predicción. Posteriormente, se chequean los resultados ofrecidos por la sensibilidad, exactitud, precisión, especificidad y el área bajo la curva.

8.3.3 Fase 3: Construcción del tablero de visualización de resultados en PowewBI que refleje las condiciones de los estudiantes con riesgo de deserción temprana en la Universidad de Manizales.

Para hacer más fácil el entendimiento de los resultados obtenidos se construyó un tablero en PowerBI con diferentes capas que refleje las diferentes relaciones, condiciones

y características existentes entre el conjunto de variables que hacen parte del análisis y el resultado de la predicción de deserción temprana. Dicho tablero permite a los usuarios navegar desde la información general hasta el resultado más particular clasificado por diferentes segmentos, permitiendo incluso quien o quienes cumplen con dicho patrón o patrones de comportamiento.

9 Resultados

9.1 Fase 1: Análisis exploratorio de los datos

9.1.1 Entendimiento del negocio

La Universidad de Manizales es una institución educativa de carácter regional mediante la Resolución No. 2317 de 1992 otorgada por el Ministerio de Educación Nacional de Colombia de índole superior, que,

> …desde los principios de la solidaridad, la pluralidad, la equidad y la justicia social, despliega su acción educativa y cultural articulando los procesos de formación de profesionales críticos, creativos y comprometidos con el país; construcción de conocimiento válido y pertinente; e interacción con el entorno orientada a la promoción del desarrollo humano y social. (U Manizales, 2017)

Articulado a lo anterior el departamento de Bienestar crea el programa de acompañamiento estudiantil bajo la premisa:

> A un mayor y mejor acompañamiento a estudiantes y docentes de la Universidad, le corresponde incrementos manifiestos de bienestar objetivo y subjetivo, que redunda en la potenciación de la formación integral y permanencia de los estudiantes, logrando que culminen sus estudios en los tiempos previstos de formación. (U. Manizales, 2018, p. 4)

De esta premisa se desprende, como objetivo institucional, lograr una reducción importante de la deserción estudiantil en uno de los focos donde mayor índice se presenta, es decir, la deserción temprana, la cual ocurre durante los primeros cuatro (4) semestres de educación.

Así mismo el objetivo particular del departamento de Bienestar Estudiantil es conocer con el mayor nivel de detalle posible, las causas que ocasionan el alto nivel de deserción temprana, para anticiparse a que la situación se haga efectiva y aplicar estrategias específicas guiadas por las causas particulares que la ocasionan.

Una vez comprendido ampliamente el contexto general relevante para la investigación, producto de la serie de sesiones y entrevistas, revisión de documentos, lectura de artículos científicos, exploración de los datos en conjunto con las personas relacionadas con el tema dentro de la Universidad de Manizales, se identificaron elementos clave para tener en cuenta durante el proceso; estos son:

- El conjunto de datos sobre los cuales se ejecuta la predicción con los algoritmos ***random forest*** y ***decisión tree*** provienen de varias fuentes (sistema de información académica SIGUM, sistema financiero y planillas de registro de atención provenientes del programa de acompañamiento estudiantil).
- Utilización de herramientas de desarrollo Open source (código libre), como *Python, anaconda, jupyter, sql server express edition, power bi desktop*, para las etapas de exploración, extracción, transformación, validación y carga de los datos, así como para la implementación del algoritmo de predicción.
- La documentación de mayor relevancia se encontró en inglés, sin desconocer que alguna de la bibliografía encontrada que orientó el proceso, fue hallada en español y estaba relacionada con universidades de prestigio latinoamericano y nacional.
- De los dos algoritmos implementados ***decisión tree*** y ***random forest*** este último fue el seleccionado dada la alta precisión que arrojo para este tipo de predicciones

de tipo clasificatorio; decisión que se apoyó en la revisión de los artículos científicos (precisión del 93%).

9.1.2 Entendimiento de los datos

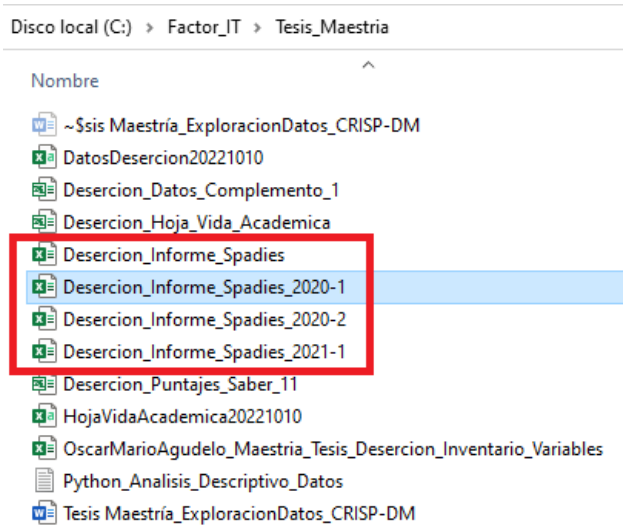
Todo el conjunto de actividades relacionadas con el entendimiento de los datos obtenidos de las diferentes fuentes, se canalizo en primera instancia mediante el uso de **paquetes de extracción ETLs (Import data)** en la herramienta **SQL Server Express Edition V18**. Allí es posible identificar la fuente, delimitar los campos a importar y almacenar en una o más tablas de la base de datos la data recolectada.

Una vez dispuesta toda la información en SQL Server en una base de datos tipo Staging (BD_DESERCION_STG), esta se transforma para ser dispuesta en una base de datos tipo relacional estableciendo las reglas de integridad referencial a la data y garantizar que es coherente. Todo lo anterior es implementado en la base de datos BD_DESERCION_MRL).

Por ultimo y tomando como base el conjunto de reglas de negocio producto del entendimiento de todo el proceso y/o terminología asociada a la deserción temprana en la Universidad de Manizales, se desarrolló el código de preparación y transformación de datos en PL/SQL (Tesis_Maestria_Exploracion_Datos_SQL_Validaciones_Final.sql, Tesis_Maestria_Estructura_Vista_Unifica_data_Desercion_Completa.sql).

Actividad 1: Recepción de los archivos de datos en formato de texto (.csv).

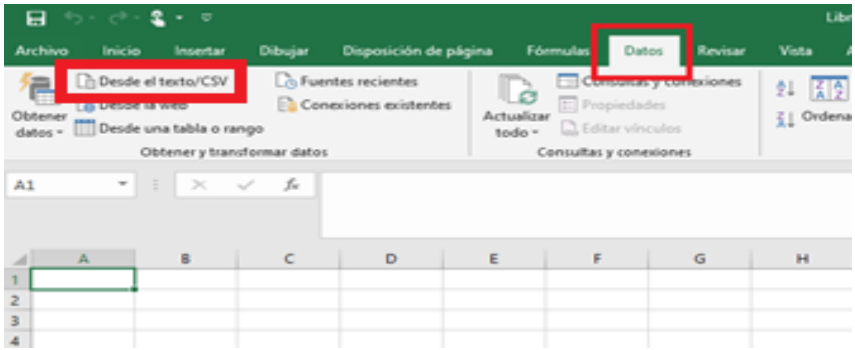
Figura 8. Recepción de archivos de datos



Actividad 2: Conversión de los archivos de texto (.csv) a archivos en Excel formato 97-2003. Para esto se siguieron los siguientes pasos:

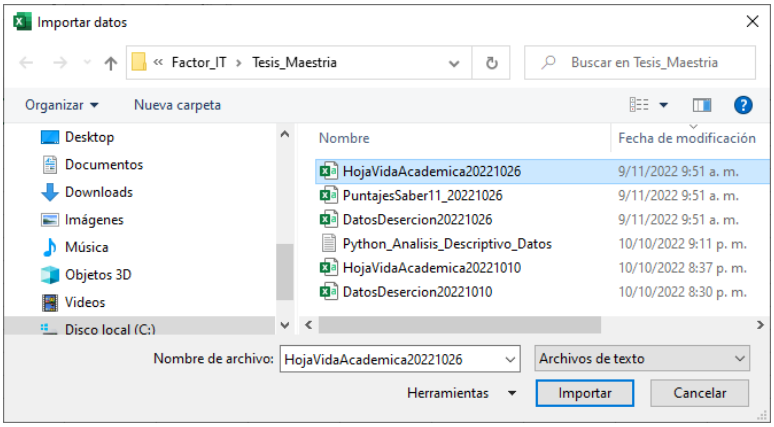
1. Abrir en Excel un archivo nuevo.
2. Seleccionar la pestaña **DATOS** (parte superior del menú Excel); luego, seleccionar el botón **DESDE EL TEXTO/CSV** (parte superior izquierda menú Excel).

Figura 9. Conversión de archivos a Excel



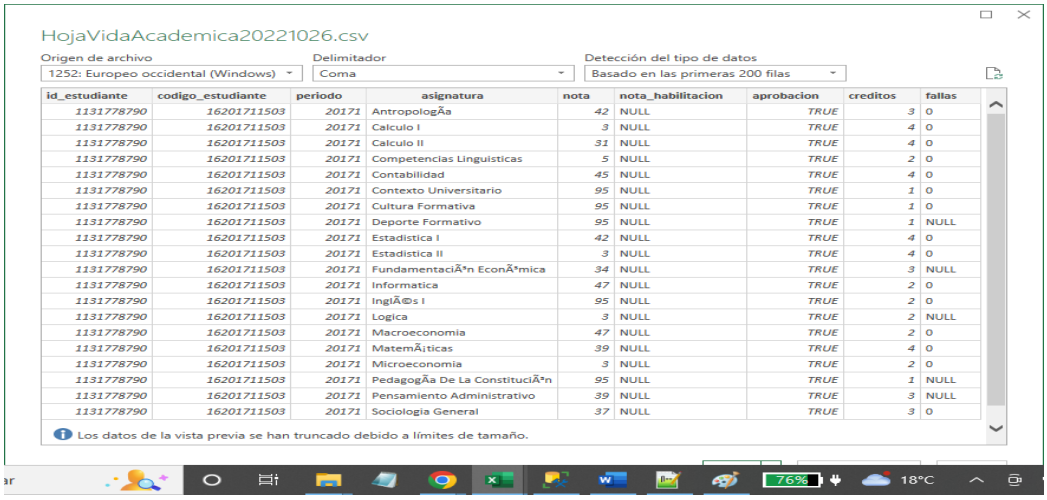
3. Seleccionar uno de los archivos de texto que hacen parte de todo el conjunto de datos de la deserción (HojaVidaAcademicaYYYYMMDD.csv, PuntajesSaber11_YYYYMMDD.csv, DatosDesercionYYYYMMDD.csv); a continuación, pulsar el botón IMPORTAR

Figura 10. Instrucciones de uso



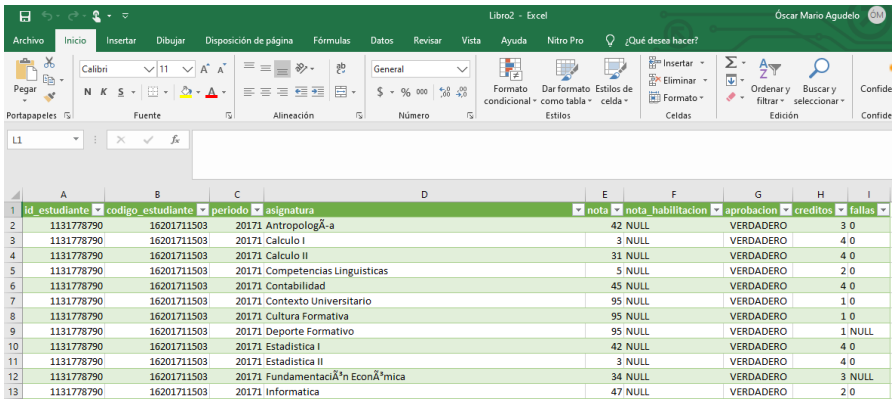
4. Una vez seleccionado el archivo a importar, debe verse la siguiente visualización; por último, presionar el botón CARGAR. Dar un vistazo general a los datos a importar.

Figura 11. Instrucciones para importar archivo



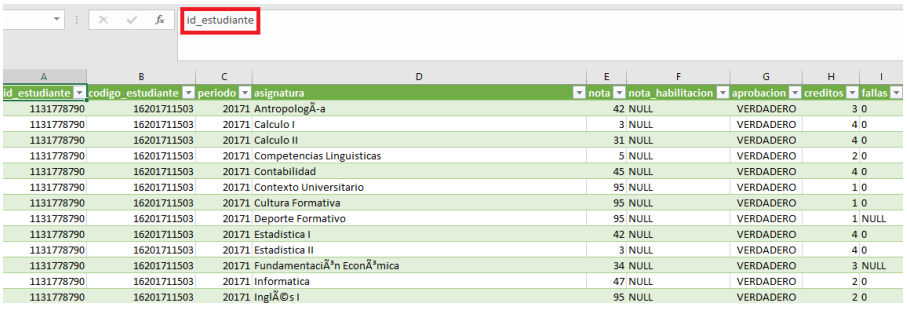
5. El resultado del archivo importado es:

Figura 12. Resultados de la operación



	A	B	C	D	E	F	G	H	I
1	id_estudiante	codigo_estudiante	periodo	asignatura	nota	nota_habilitacion	aprobacion	creditos	fallas
2	1131778790	16201711503	20171	AntropologÃ-a	42	NULL	VERDADERO	3	0
3	1131778790	16201711503	20171	Calculo I	3	NULL	VERDADERO	4	0
4	1131778790	16201711503	20171	Calculo II	31	NULL	VERDADERO	4	0
5	1131778790	16201711503	20171	Competencias Linguisticas	5	NULL	VERDADERO	2	0
6	1131778790	16201711503	20171	Contabilidad	45	NULL	VERDADERO	4	0
7	1131778790	16201711503	20171	Contexto Universitario	95	NULL	VERDADERO	1	0
8	1131778790	16201711503	20171	Cultura Formativa	95	NULL	VERDADERO	1	0
9	1131778790	16201711503	20171	Deporte Formativo	95	NULL	VERDADERO	1	NULL
10	1131778790	16201711503	20171	Estadistica I	42	NULL	VERDADERO	4	0
11	1131778790	16201711503	20171	Estadistica II	3	NULL	VERDADERO	4	0
12	1131778790	16201711503	20171	FundamentaciÃ³n EconÃ³mica	34	NULL	VERDADERO	3	NULL
13	1131778790	16201711503	20171	Informatica	47	NULL	VERDADERO	2	0

6. A continuación, revisar el archivo para ajustarlo así: Eliminar títulos del informe y dejar solamente el nombre de las columnas. Renombrar las columnas con caracteres especiales (incluso la eñe - ñ); eliminar los espacios entre los nombres de las columnas (véase el recuadro resaltado en rojo que indica nombre campo).

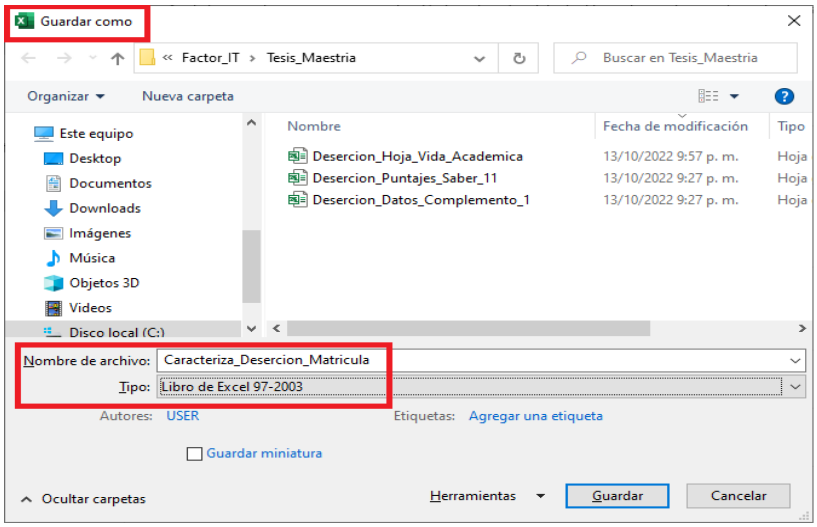


	A	B	C	D	E	F	G	H	I
	id_estudiante	codigo_estudiante	periodo	asignatura	nota	nota_habilitacion	aprobacion	creditos	fallas
	1131778790	16201711503	20171	AntropologÃ-a	42	NULL	VERDADERO	3	0
	1131778790	16201711503	20171	Calculo I	3	NULL	VERDADERO	4	0
	1131778790	16201711503	20171	Calculo II	31	NULL	VERDADERO	4	0
	1131778790	16201711503	20171	Competencias Linguisticas	5	NULL	VERDADERO	2	0
	1131778790	16201711503	20171	Contabilidad	45	NULL	VERDADERO	4	0
	1131778790	16201711503	20171	Contexto Universitario	95	NULL	VERDADERO	1	0
	1131778790	16201711503	20171	Cultura Formativa	95	NULL	VERDADERO	1	0
	1131778790	16201711503	20171	Deporte Formativo	95	NULL	VERDADERO	1	NULL
	1131778790	16201711503	20171	Estadistica I	42	NULL	VERDADERO	4	0
	1131778790	16201711503	20171	Estadistica II	3	NULL	VERDADERO	4	0
	1131778790	16201711503	20171	FundamentaciÃ³n EconÃ³mica	34	NULL	VERDADERO	3	NULL
	1131778790	16201711503	20171	Informatica	47	NULL	VERDADERO	2	0
	1131778790	16201711503	20171	InglÃ©s I	95	NULL	VERDADERO	2	0

7. Seleccionar el botón GUARDAR COMO. Asignar el nombre, siempre con la cadena Caracteriza_Desercion_xxxxxxxx. Reemplazar las xs por la temática del archivo. Para el ejemplo seria:

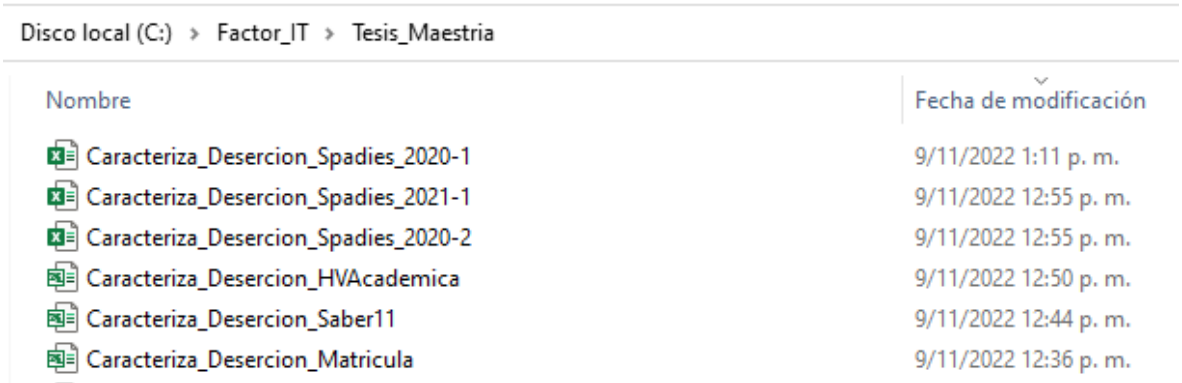
Caracteriza_Desercion_Matricula.xls. El Tipo de archivo es Libro Excel 97 – 2003 (xls).

Figura 13. Instrucciones para guardar



8. Repetir uno a uno los pasos de la ACTIVIDAD 2, para cada uno de los archivos que conforman la temática de deserción temprana universitaria.

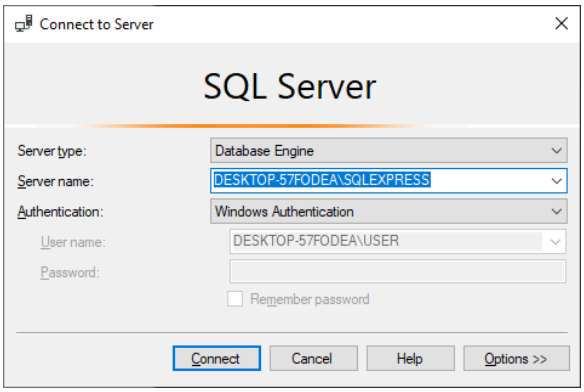
Se anexa ejemplo del inventario de archivos dispuestos para llevarlos a la base de datos (SQL Server).



Actividad 3: Proceso ELT (Almacenamiento de los archivos en Excel (.xls) en las tablas de las bases de datos stage BD_DESERCION_STG). El proceso se realiza dentro del gestor de base de datos SQL Server, con una de sus herramientas de importación de datos, de la siguiente manera:

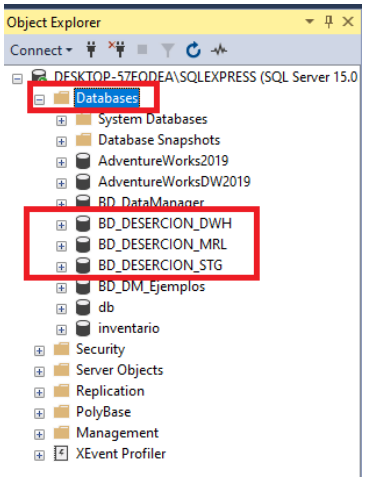
1. Iniciar una sesión en SQL server. Respetar los parámetros que presenta la pantalla y pulsar el botón **CONNECT.**

Figura 14. Imagen de los parámetros SQL



2. Una vez el ingreso al motor SQL Server sea exitoso, ubicar las tres bases de datos dispuestas para construir el almacén de datos (datalake), señaladas en la siguiente gráfica:

Figura 15. Imagen de almacén de datos



Se definieron tres (3) bases de datos:

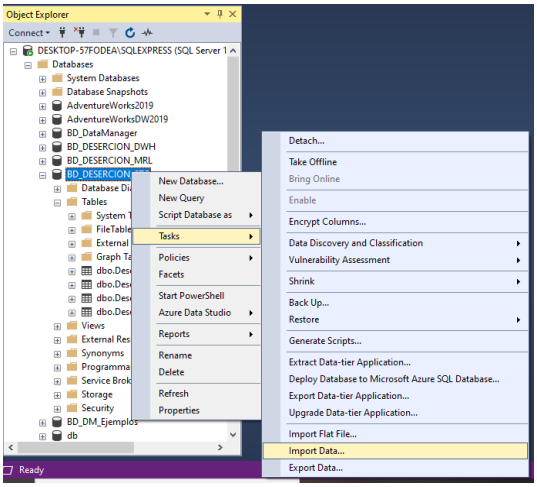
BD_DESERSION_STG: Base de datos de réplica en la cual se almacenan los archivos tal como llegan de la fuente original por cada periodo en proceso; dentro de esta base de datos se aplican algunas reglas de calidad, completitud y consistencia de la data, así como reglas de negocio iniciales.

BD_DESERCION_MRL: Base de datos donde se recrea un modelo relacional aplicando ingeniería inversa a partir de los datos tal como llegan en los archivos de texto, para identificar elementos comunes que unan estos archivos, como el DOCUMENTO y/o CÓDIGO ESTUDIANTE; esto permite tener una caracterización más precisa de todas las variables que afectan al estudiante. En esta fase se aplican reglas de integridad referencial de la data con el fin de eliminar errores de consistencia de la data una vez se lleve al modelo de datawarehouse dentro del datalake.

BD_DESERCION_DWH: Base de datos tipo datawarehouse que almacenará las variables construidas para aplicarlas al modelo de inteligencia artificial y *machine learning*.

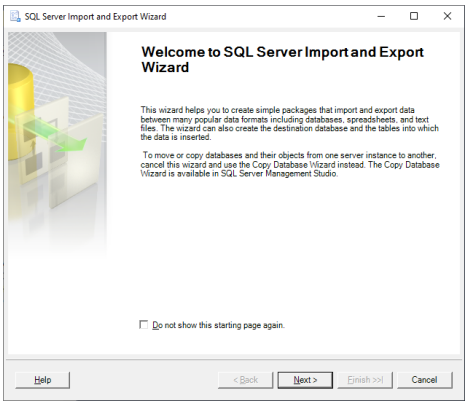
3. Importar todos los archivos en la base de datos réplica sin realizar transformación alguna, así:

Figura 16. Instrucciones de importación



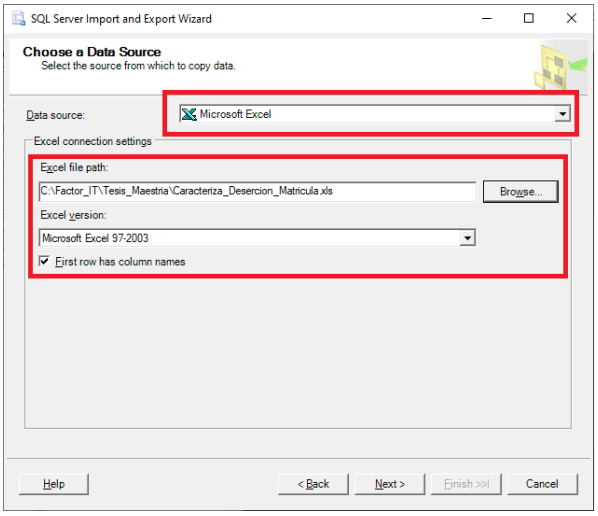
Seleccionar la BD_DESERCION_STG, con el click derecho sobre la BD (sombreada en azul) seleccionar TASKS, luego seleccionar IMPORT DATA.

Figura 17. Imagen del servidor



Solo basta con pulsar el botón NEXT para avanzar al cuadro de Data fuente u origen:

Figura 18. Imagen para avanzar a data fuente



Diligenciar el recuadro anterior de la siguiente forma:

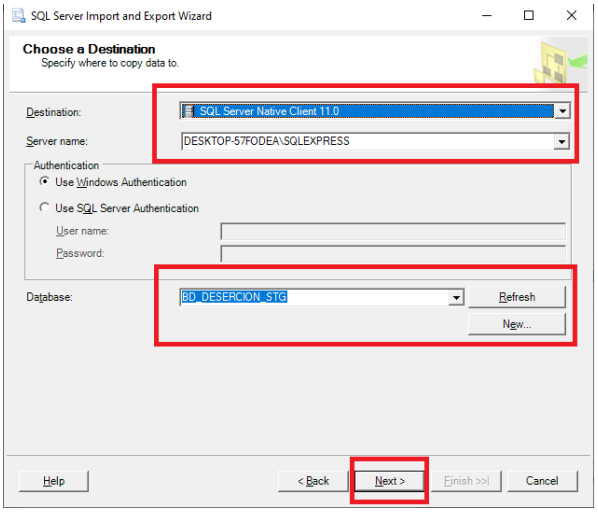
Data source: Seleccionar de la lista de opciones *Microsoft Excel.*

Excel file path: Seleccionar la carpeta fuente y el archivo a importar .xls.

Excel version: *Microsoft Excel 97-2003.*

Por último, presionar el botón NEXT para avanzar al cuadro Data Destino o de llegada de los datos.

Figura 19. Instrucciones para avanzar a Data Destino



Diligenciar el recuadro anterior de la siguiente forma:

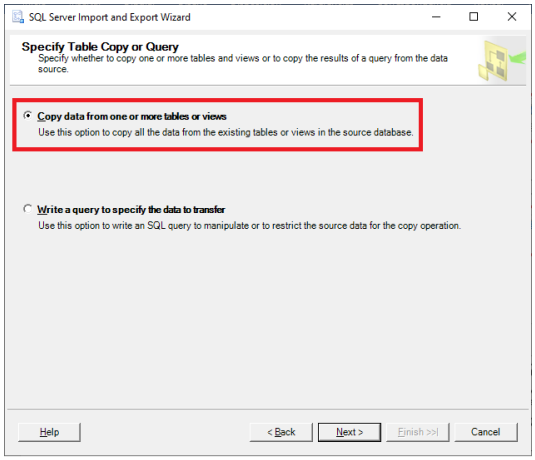
Destination: Seleccionar de la lista de opciones *SQL Server Native Cliente 11.0*

Server Name: No modificar, dejar el servidor por defecto.

Database: Nombre de la base de datos destino. Seleccionar de la lista BD_DESERCION_STG

Por último, presionar el botón NEXT para avanzar al recuadro: Especificar la forma de copiado de los datos.

Figura 20. Instrucción para avanzar al copiado de datos



Seleccionar la opción por defecto (resaltada en un recuadro rojo).

Por último, presionar el botón NEXT para avanzar al recuadro seleccionar nombre de archivo fuente y nombre de la tabla destino, de la siguiente manera:

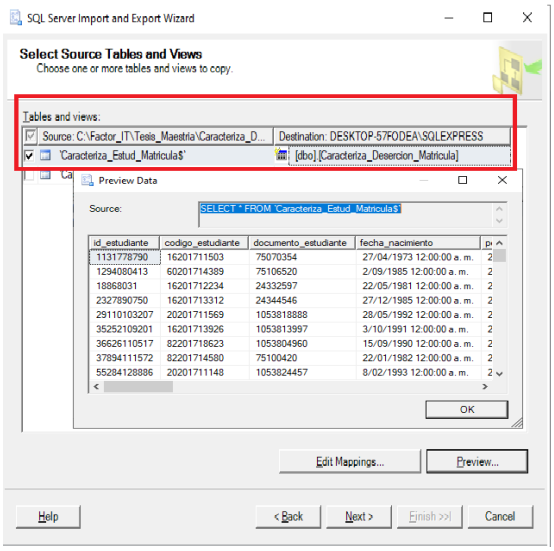
Source: Seleccionar con click la primera opción (véase recuadro en rojo).

Destination: con doble click en esta casilla, reescribir el verdadero nombre de la tabla Caracteriza_Desercion_Matricula (véase recuadro en rojo).

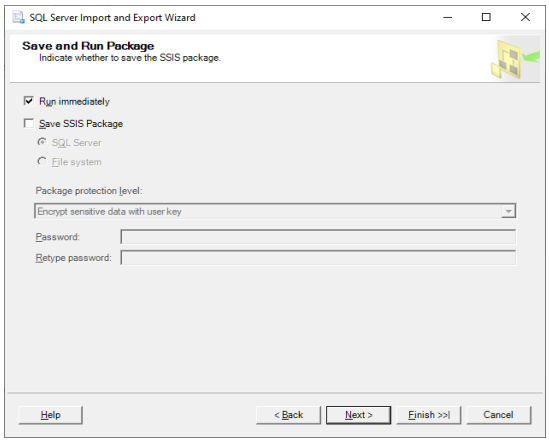
Para verificar que los datos cargaran adecuadamente pulsar el botón PREVIEW.

Si los datos mostrados con la opción PREVIEW son acordes, pulsar el botón NEXT para avanzar al recuadro de ejecutar el proceso de extracción.

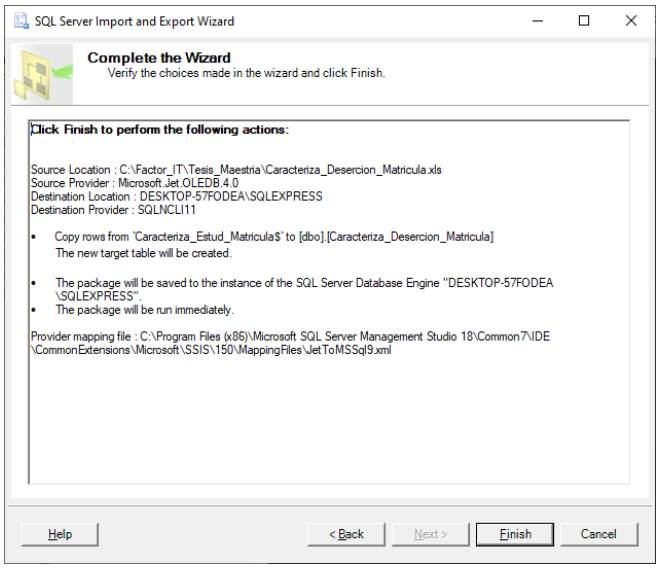
Figura 21. Instrucción para ejecutar la extracción



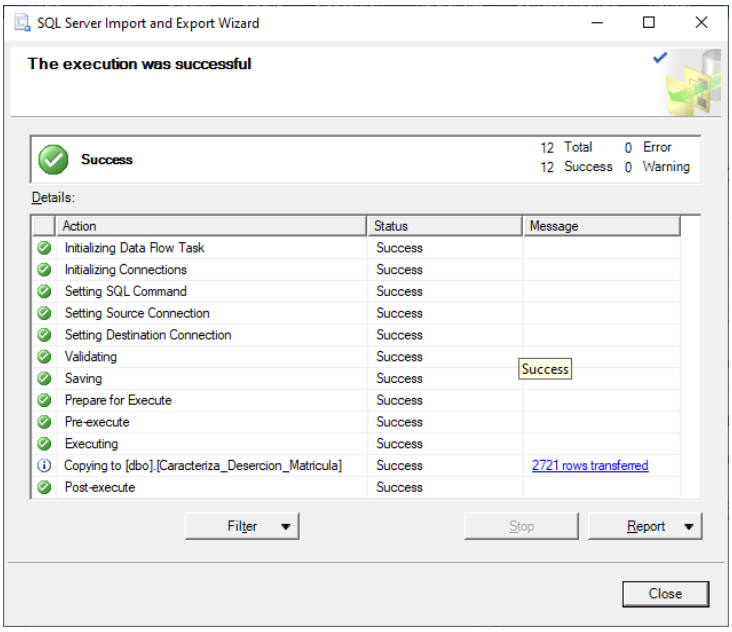
Para ejecutar el paquete de extracción, pulsar el botón NEXT.



Avanzar con NEXT hasta llegar al último recuadro y pulsar FINISH para iniciar el proceso de extracción.

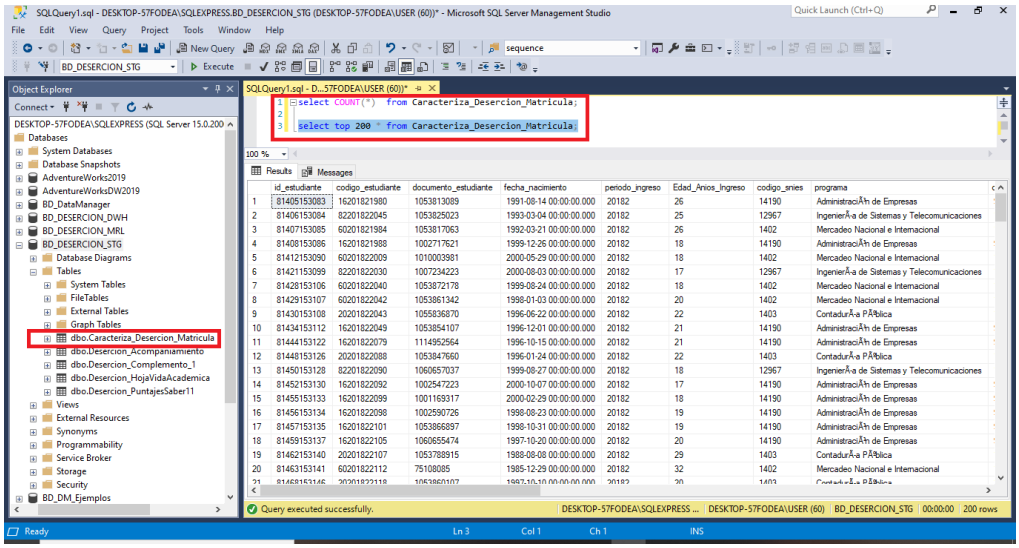


Si todo termina correctamente, debe mostrar el siguiente recuadro con botones en color verde:



Pulsar el botón CLOSE. Ir a la base de datos en SQL Server y verificar que la tabla se haya cargado con los datos de manera correcta, así:

Figura 22. Para ir a base de datos en SQL

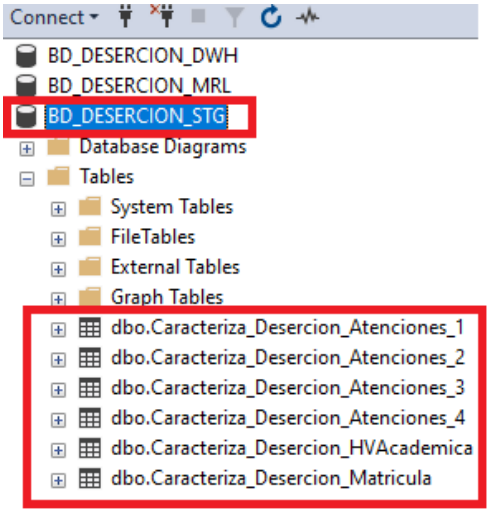


Nota: Ver los recuadros en rojo que indican al lado izquierdo, el nombre de la tabla, y al lado derecho, las sentencias en comandos SQL que verifican los datos extraídos.

4. Repetir el paso a paso del numeral 3 para cada uno de los archivos en Excel, con el fin de convertirlos en tablas dentro de la base de datos BD_DESERCION_STG.

Finalmente, al consultar al lado derecho de la base de datos, deben aparecer las siguientes tablas:

Figura 23. Para consultar



Al finalizar todo el proceso, se obtuvieron las siguientes estructuras de datos fuente que están almacenadas en SQL Server en la base de datos Stage BD_DESERCION_STG: El conjunto de datos está compuesto por 48 variables de tipo categórico y continuo.

Tabla 3. *Estructura de* datos *Caracteriza_Desercion_HVAcademica*

	Caracteriza_Desercion_HVAcademica	
No.	**Variable**	**Descripcion**
1	id_estudiante	Numero aleatorio que identifica el numero del estudiante. Aplica para enlazar con las otas tablas. Valores NULL, cero, negativos o vacios no son validos.
2	codigo_estudiante	Codigo identificacion del estudiante asociado a la carrera a la cual esta matriculado. Un estudiante puede estar matriculado a una o mas carreras. Valores NULL, cero, negativos o vacios no son validos.
3	periodo	Periodo Academico al cual fue registrada la nota. Esta conformado por 4 digitos del año y un digito que indica el primer o segundo semestre cursado de dicho año. Valores NULL, cero, negativos o vacios no son validos.
4	asignatura	Nombre de la asignatura cursada y a la cual se le reporta la nota, Valores NULL o vacios no son validos.
5	nota	Nota con un rango de valores entre 0.0 y 5.0 para materias cuantitativas, y valores de 5.5 (perdida) o 9.5 (aprobada) para materias cualitativas. Valores NULL, negativos o vacios no son
6	nota_habilitacion	Nota habilitacion, idem a la descripcion de NOTA. Asume valores en NULL (validos) si la NOTA fue aprobada por ende no habilita. Valores negativos no son validos si los hubiese
7	aprobacion	Dos valores posibles TRUE / FALSE (Aprobada o Reprobada) en nota normal o en habilitacion. Valores NULL o vacios no son validos.
8	creditos	numero de creditos que equivale a la materia vista: Valores Null, cero (0), negativos o vacios no son validos.
9	fallas	Numero de ausencias que tuvo el estudiante para dicha materia durante el semestre cursado. Valores NULL o numeros negativos no son validos.

Tabla 4. ***Estructura de datos Caracteriza_Desercion_Matricula***

	Caracteriza_Desercion_Matricula	
No.	**Variable**	**Descripcion**
10	id_estudiante	Numero aleatorio que identifica el numero del estudiante. Aplica para enlazar con las otas tablas. Valores NULL, cero, negativos o vacios no son validos.
11	codigo_estudiante	Codigo identificacion del estudiante asociado a la carrera a la cual esta matriculado. Un estudiante puede estar matriculado a una o mas carreras. Valores NULL, cero, negativos o vacios no son validos.
12	documento_estudiante	Documento de identificacion (cedula, tarjeta identidad, pasaporte, cedula de extranjeria) del estudiante.
13	fecha_nacimiento	Formato AAAA-MM-DD de nacimiento del estudiante. Pueden llegar valores NULL, vacios o cero(0), pues de algunos estudiantes es posible no haberse cargado la fecha de nacimiento.
14	periodo_ingreso	Periodo Academico de ingreso al programa. Esta conformado por 4 digitos del año y un digito que indica el primer o segundo semestre cursado de dicho año. Valores NULL, cero, negativos o vacios no son validos.
15	Edad_Anios_Ingreso	Edad en años cumplidos de ingreso a la universidad
16	codigo_snies	Codigo del SNIES (Sistema Nacional de Información de la Educación Superior) del programa al cual se matriculo
17	programa	Nombre del programa academico al cual se matriculo el estudiante. Valores NULL, vacios no son validos.
18	duracion_programa	Duracion en semestres del programa academico al cual se matriculo.
19	genero	Genero del estudiante matriculado (Masculino, Femenino, Otro).
20	estrato_procedencia	Estrato de procedencia del estudiante. Rango de valores permitidos de 0 a 6.
21	estrato_residencia	Estrato de residencia actual del estudiante. Rango de valores permitidos de 0 a 6.
22	estado_civil	Estado civil del estudiante matriculado (Casado, Soltero, union libre, viudo, otro).
23	departamento_procedencia	Nombre del departamento de procedencia del estudiante matriculado.
24	municipio_procedencia	Nombre del municipio de procedencia del estudiante matriculado.
25	aplicacion_encuesta_caracterizacion	SI o No. Se le realizo encuesta de caracterizacion al estudiante al momento del ingreso a la universidad.
26	aplicacion_entrevista	SI o No. Se le realizo entrevista al estudiante al momento del ingreso a la universidad.
27	procedencia_colegio	Tipo de Colegio de procedencia del estudiante matriculado. Publico o Privado. Pueden llegar valores NULL o vacios en este campo-
28	fecha_terminacion_secundaria	Año, mes y dia de finalizacion de secundaria del estudiante matriculado. Formato: AAAA/MM/DD. Pueden llegar valores NULL o vacios en este campo.
29	programa_primera_opcion	Si o No. La eleccion del programa por parte del estudiante matriculado fue su primera opcion ?.
30	ultima_matricula	Ultimo periodo en el cual registro matricula. Esta conformado por 4 digitos del año y un digito que indica el primer o segundo semestre cursado de dicho año. Valores NULL, cero, negativos o vacios no son validos.

Tabla 5. ***Estructura de datos Caracteriza_Desercion_Saber11***

	Caracteriza_Desercion_Saber11	
No.	**Variable**	**Descripcion**
31	id_estudiante	Numero aleatorio que identifica el numero del estudiante. Aplica para enlazar con las otas tablas. Valores NULL, cero, negativos o vacios no son validos.
32	registro	Numero del registro de las pruebas ICFES o SABER 11 con el cual quedaron regstradas en e ministerio de educacion los resultados el estudiante matriculado en la universidad
33	anio_icfes	Año de presentacion de las pruebs ICFES o SABER 11. Formato AAAA
34	tipo_icfes	Descripcion del tipo de ICFES o SABER 11. El tipo define la vigencia de las pruebas y su modelo de calificacion por area o puntaje.
35	area_icfes	nombre del area al cual esta asociado el puntaje obtenido.
36	puntaje_icfes	Puntaje numerico obtenido por cada una de las areas de las pruebas

Tabla 6. ***Estructura de datos Caracteriza_Desercion_Atenciones 1…4***

	Caracteriza_Desercion_Atenciones_1/ 2/3/4	
No.	**Variable**	**Descripcion**
37	Periodo_Semestre	Periodo Academico en el cual registro atencion estudiantil. Esta conformado por 4 digitos del año y un digito que indica el primer o segundo semestre cursado de dicho año. Valores NULL, cero, negativos o vacios no son validos.
38	Sec	Numero secuencial de la atencion brindada durante dicho periodo
39	ApellidosCompletos	Apellidos del estudiante matriculado que recibio atencion
40	NombresCompletos	Nombres del estudiante matriculado que recibio atencion
41	Tipodocumento	Tipo de documento de identificacion del estudiante matriculado (cedula, tarjeta identidad, pasaporte, cedula extranjeria, etc.).
42	Numerodocumento	Numero del documento de identificacion del estudiante matriculado
43	Programa	nombre del programa al cual esta matriculado el estudiante durante la atencion de acompañamiento estudiantil.
44	codigoestudiante	codigo de identificacion del estudiante
45	telefono	Numero telefonico del estudiante que recibe la atencion.
46	E-mail	correo electronico del estudiante que recibe la atencion.
47	serviciosolicitado	Tipo de acompañamiento que solicito el estudiante (psicologico, psicopedagogico, individual o grupal).
48	otroserviciosolicitado	Otro servicio solicitado por el estudiante que solicito el acompañamiento en bienestar.

Para cada una de las tablas creadas en SQL Server en la base de datos stage BD_DESERCION_STG, se realizaron procesos de exploración de la data, así como la corrección de la data inconsistente. Para observar en detalle lo realizado, véase el

documento Tesis_Maestria_Exploracion_Datos_SQL_Validaciones_Final.sql (vinculo: Tesis_Maestria_Exploracion_Datos_SQL_Validaciones_Final.sql)

9.2 Fase 2: Implementación de modelos clasificatorios que permitan explicar la deserción temprana al interior de la Universidad de Manizales

9.2.1 Preparación de los datos

Durante el desarrollo de esta actividades llevaron a cabo uno a uno actividades como lo son la extracción desde SQL Server base de datos BD_DESERCION_MRL mediante la construcción en PL/SQL una vista lógica que reúne toda la integralidad del modelo relacional de los datos dispuestos en un conjunto de tablas para que pueda ser leído como un set de datos (**data frame)** dentro del notebook en anaconda Python. Tesis_Maestria_Estructura_Vista_Unifica_data_Desercion_Completa.sql.

Dentro del código de Python se llevaron a cabo etapas como la preparación del entorno de desarrollo del Notebook el cual consiste en instalar las librerías para el análisis de datos, graficación y los algoritmos de predicción.

Etapa de exploración, preparación de datos, así como la inclusión de variables derivadas de la misma. Para este ítem se implementaron dos (2) sets de datos, el primero para los algoritmos de **decisión tree** y **random forest**; el segundo para la aplicación de la técnica importancia de las características (**feature imporance**). Y por último el desarrollo del código para el entrenamiento y test de los sets de datos para obtener el resultado de la predicción (desertó/No deserto). Resultados que son presentados en la herramienta PowerBI mediante cuadros de visualización de datos que permiten observar los datos fuente versus la predicción cruzando el conjunto de variables categóricas que explican el evento de deserción temprana y la variable objetivo.

Actividad 1: Extracción, carga, validación, transformación y preparación de los datos para alimentar el modelo de predicción:

Después de identificar, verificar y corregir las posibles inconsistencias que se recibieron de los datos fuente, se procede con el proceso preparatorio de los datos que proveerán al modelo de las variables fuente y de aquellas derivadas de otras variables. En el documento Tesis Maestria Exploracion Datos SQL Validaciones Final.sql, puede observarse con detenimiento este proceso.

Producto de la ejecución de cada una de las sentencias referidas en el documento mencionado, se obtiene la siguiente estructura base para el modelo predictivo, la cual fue implementada en una vista lógica dentro de SQL Server Tesis Maestria Estructura Vista Unifica data Desercion Completa.sql, ubicado en la carpeta en conjunto con el documento de tesis. No fue incluido como anexo debido a su tamaño).

Esta actividad vincula, no solo los procesos de extracción de la data y parte del análisis exploratorio de los datos fuente dispuestos en el set de datos. De una manera gráfica permite realizar análisis descriptivo univariados, bivariados y multivariados del comportamiento actual de los datos frente a la situación evaluada (deserción temprana). Esta representación gráfica permite el primer acercamiento de como en la actualidad se perfilan los estudiantes con o sin deserción, elaborando las primeras conclusiones que a la postre confirman lo ya dicho en el artículo Indicadores de deserción universitaria y factores asociados por (Gutiérrez, Vélez Diaz, López, 2021). *En el país el problema de la deserción en el sistema de educación superior obedece fundamentalmente a un problema estructural del sistema educativo, que acumula falencias desde la formación inicial, la formación básica primaria, secundaria y media hasta llegar a la formación universitaria y, en particular, a la formación en los programas de educación.*

Tabla 7. *Estructura de datos vv_df_unifica_data_desercion_dsc*

	vv_df_unifica_data_desercion_dsc		
No.	**Variable**	**Descripcion**	**Categoria**
1	id_estudiante	Numero aleatorio que identifica el numero del estudiante. Aplica para enlazar con las otas tablas. Valores NULL, cero, negativos o vacios no son validos.	
2	cod_estudiante	Codigo identificacion del estudiante asociado a la carrera a la cual esta matriculado. Un estudiante puede estar matriculado a una o mas carreras. Valores NULL, cero, negativos o vacios no son validos.	
3	id_genero	Codigo del genero del estudiante matriculado (Masculino, Femenino, Otro).	1, 2
4	dsc_genero	Descripcion del genero del estudiante matriculado (Masculino, Femenino, Otro).	1-MASCULINO, 2-FEMENINO
5	id_estcivil	Codigo Estado civil del estudiante matriculado (Casado, Soltero, union libre, viudo, otro).	1,2,3,4,5,6
6	dsc_estcivil	Descripcion Estado civil del estudiante matriculado (Casado, Soltero, union libre, viudo, otro).	1-Soltero, 2-Divorciado, 3-Union Libre, 4-No Aplica, 5-Casado, 6-Separado
7	desertor_t	Variable a predecir. 0 -No deserto, 1-Desercion Temprana	0, 1
8	EsDesertor	Descripcion del resultado de la variable desertor_t. 0-NO, 1-SI	0-No, 1-Si es desertor
9	id_estrato_procedencia	Codigo Estrato de procedencia del estudiante. Rango de valores permitidos de 0 a 6.	1,2,3,4,5,6
10	dsc_estrato_procedencia	Descripcion Estrato de procedencia del estudiante. Rango de valores permitidos de 0 a 6.	Estrato 1, Estrato 2, Estrato 3, Estrato 4, Estrato 5, Estrato 6
11	id_estrato_residencia	Codigo Estrato de residencia del estudiante. Rango de valores permitidos de 0 a 6.	1,2,3,4,5,6
12	dsc_estrato_residencia	Descripcion Estrato de residencia del estudiante. Rango de valores permitidos de 0 a 6.	Estrato 1, Estrato 2, Estrato 3, Estrato 4, Estrato 5, Estrato 6
13	id_depto_procedencia	Codigo del departamento de procedencia del estudiante	1,2,3…..25
14	dsc_depto_procedencia	Nombre del departamento de procedencia del estudiante matriculado.	Antioquia, Arauca, Caldas, Huila, Caqueta, Cundinamarca, etc.
15	id_mpio_procedencia	Codigo del municipio de procedencia del estudiante matriculado.	N numero de ciudades
16	dsc_mpio_procedencia	Nombre del municipio de procedencia del estudiante matriculado.	Manizales, Villamaria, Armenia, Bogota, Cucuta, etc.
17	id_tipo_colegio_procedencia	Codigo Tipo de Colegio de procedencia del estudiante matriculado. Publico o Privado. Pueden llegar valores NULL o vacios en este campo.	1,2
18	dsc_tipo_colegio_procedencia	Descripcion Tipo de Colegio de procedencia del estudiante matriculado. Publico o Privado. Pueden llegar valores NULL o vacios en este campo.	1-Publico, 2-Privado
19	fec_finaliza_secundaria	Año, mes y dia de finalizacion de secundaria del estudiante matriculado. Formato: AAAA/MM/DD. Pueden llegar valores NULL o vacios en este campo.	Multiples fechas
20	id_calendario_secundaria	Codigo identifica el periodo que curso la secundaria (A o B)	1,2
21	dsc_calendario_secundaria	Descripcion del calendario curso secundaria (Calendario A o B).	1-Calendario A, 2-Calend B
22	id_programa	Codigo del programa al cual esta matriculado el estudiante durante la atencion de acompañamiento estudiantil.	1,2,3,4,5,6,7
23	dsc_programa	Nombre del programa al cual esta matriculado el estudiante durante la atencion de acompañamiento estudiantil.	1-Contaduria, 2-Mercadeo, 3-Ing. Sistemas, 4-Ing. Analitica de Datos, 5-Ing. Logistica, 6-Adm Empresas, 7-Ing. Seg Informac
24	semestres_duracion_programa	Duracion en semestres del programa academico al cual se matriculo.	8,9,10
25	id_aplico_encuesta_caracterizacion	Codigo de SI o No. Se le realizo encuesta de caracterizacion al estudiante al momento del ingreso a la universidad.	0,1
26	respuesta_caracteriza	Descripcion Si o No. Se le realizo encuesta de caracterizacion al estudiante al momento del ingreso a la universidad.	0-No, 1-Si encuesta caracterizacion.

27	id_aplico_entrevista	Codigo de Si o No. Se le realizo entrevista al estudiante al momento del ingreso a la universidad.	0,1
28	respuesta_entrevista	Descripcion de Si o No. Se le realizo entrevista al estudiante al momento del ingreso a la universidad.	0-No, 1-Si entrevista al ingreso a la universidad
29	id_programa_primera_opcion	Codigo de Si o No. La eleccion del programa por parte del estudiante matriculado fue su primera opcion ?.	0,1
30	respuesta-primeraopcion	Descripcion de Si o No. La eleccion del programa por parte del estudiante matriculado fue su primera opcion ?.	0-No, 1-Si fue el programa elegido la primera opcion
31	edad_anios_ingreso	Edad en años cumplidos de ingreso a la universidad	Multiples edades de ingreso
32	id_grpo_edad	Codigo identifica el grupo etareo que clasifica los estudiantes.	0,1,2
33	grpo_edad	Descripcion identifica el grupo etareo que clasifica los estudiantes.	0-Menores, 1-Jovenes , 2-Adultos
34	puntaje_matemat	Puntaje obtenido en el area de MATEMATICAS de las pruebas del ICFES o SABER 11.	Multiples puntajes, en rangos de 0 a
35	nivel_matemat	Nivel obtenido en el area de MATEMATICAS de las pruebas del ICFES o SABER 11. (1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-	1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado
36	puntaje_lectucri	Puntaje obtenido en el area de LECTURA CRITICA de las pruebas del ICFES o SABER 11.	Multiples puntajes, en rangos de 0 a
37	nivel_lectucri	Nivel obtenido en el area de LECTURA CRITICA de las pruebas del ICFES o SABER 11. (1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado).	1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado
38	puntaje_ingles	Puntaje obtenido en el area de INGLES de las pruebas del ICFES o SABER 11.	1,2,3
39	nivel_ingles	Nivel obtenido en el area de INGLES de las pruebas del ICFES o SABER 11. (1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado).	1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado
40	puntaje_total	Puntaje total obtenido en las pruebas del ICFES o SABER 11, Acumula todas las areas de conocimiento evaluadas.	Multiples puntajes
41	rendi_global	Rendimiento genral en las pruebas ICFES o SABER 11. (1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado).	1-Insuficiente, 2-Minimo, 3-Satisfactorio, 4-Avanzado
42	promedio_total_cuanti	Promedio obtenido en la materias que se califican como cuanitativas. Calificacion de 0.0 a 5.0	Multiples promedios
43	cant_materias_promedio_cuanti	Cantidad de materias cuantitativas tomadas en cuenta para el promedio general	Multiples valores
44	total_fallas_cuanti	Cantidad de fallas o ausencias a clas en las materias cuantitativas	Multiples valores
45	total_creditos_cuanti	Total creditos vistos por el estudiante en las materias cuantitativas	Multiples valores
46	id_tipo_promedio_cuanti	Codigo identifica el tipo de promedio obtenido sobre las materias cuantitativas.	1,2,3
47	tipo_promedio_cuanti	Descripcion del promedio cuantitativo obtenido. (1-Bajo, 2-Medio, 3-Superior).	1-Bajo, 2-Medio, 3-Superior.
48	total_periodos_cursados	Cantidad total de periodos cursados por el estudiante hasta la fecha de corte	Multiples valores
49	total_materias_perdidas_cuanti	Cantidad total de materias cuantitativas perdidas.	Multiples valores
50	total_creditos_perdidas_cuanti	Cantidad total de creditos perdidos sobre materias cuantitativas.	Multiples valores
51	total_fallas_perdidas_cuanti	Cantidad total de fallas o ausencias que presento en las materias cuantitativas.	Multiples valores
52	total_materias_aprobadas_cuanti	Cantidad total de materias cuantitativas aprobadas	Multiples valores
53	total_creditos_aprobadas_cuanti	Cantidad total de creditosaprobados sobre materias cuantitativas.	Multiples valores

54	total_fallas_aprobadas_cuanti	Total fallas o ausencias a clase relacionaas con maerias cuantitativas.	Multiples valores
55	cant_maerias_cualita	Cantidad de materias cualitativas cursadas.	Multiples valores
56	total_fallas_cualita	Cantidad total de fallas o ausencias que presento en las materias cualitativas.	Multiples valores
57	total_creditos_cualita	Cantidad total de creditos vistos sobre materias cualitativas	Multiples valores
58	total_materias_perdidas_cualita	Cantidad total de materias cualitativas perdidas	Multiples valores
59	total_creditos_perdidas_cualita	Cantidad total de creditos perdidos sobre materias cualitativas.	Multiples valores
60	total_fallas_perdidas_cualita	Cantida total de fallas o ausencias sobre materias cualitativas perdidas.	Multiples valores
61	total_materias_aprobadas_cualita	Cantidad total de materias cualitativas aprobadas	Multiples valores
62	total_creditos_aprobadas_cualita	Cantidad total de creditos aprobados sobre materias cualitativas	Multiples valores
63	total_fallas_aprobadas_cualita	Cantidad total de fallas o ausencias de clase sobre las materias cualitativas	Multiples valores
64	total_atenciones_acompaniamiento	Cantidad total de atenciones recibidas por el estudiante en el programa de acompañamiento estudiantil	Multiples valores
65	recibe_acompania	Numero Respuesta (0-No, 1-Si) afirmativa o negativa si recibio o no atencion por parte del programa de acompañamiento estudiantil.	0,1
66	recibe_atencion	Descripion Respuesta (0-No, 1-Si) afirmativa o negativa si recibio o no atencion por parte del programa de acompañamiento estudiantil.	0-No, 1-Si.
67	total_materias_cursadas	Cantidad total de materias cursadas durante los periodos cursados con respecto a la fecha de corte	Multiples valores
68	total_materias_perdidas	Cantidad total de materias perdidas durante los periodos cursados repecto a la fecha de corte.	Multiples valores
69	total_fallas_clase	Cantdad total de fallas o ausencia de clase durante los periodos cursados con respecto a la fecha de corte.	Multiples valores
70	total_creditos_cursados	Cantidad total de creditos cursados durante los periodos academicos y respecto a la fecha de corte	Multiples valores
71	total_creditos_perdidos	Cantidad total de creditos perdidos durante los periodos academicos y respecto a la fecha de corte	Multiples valores
72	ultimo_periodo_cursado	Ultimo perdiodo cursado en la universidad con respecto a la fecha de corte. El periodo esta conformado por el año de 4 digitos AAAA y el periodo de primero o segundo semestre.	Multiples valores
73	periodo_corte	Periodo de corte. Año AAAA y el primer o segundo periodo del año.	20222
74	prc_materias_aprob	Porcentaje de materias aprobadas con el respecto al numero total de materias cursadas	multiples valores, entre 0 y 100%
75	rendi_materias	Porcentaje de rendimiento obtenido sobre las materias materias cursadas	multiples valores, entre 0 y 100%
76	prc_creditos_aprob	Porcentaje de creditos aprobadas con el respecto al numero total de creditos cursadas	multiples valores, entre 0 y 100%
77	rendi_creditos	Porcentaje de rendimiento obtenido sobre los creditos cursados materias cursadas	multiples valores, entre 0 y 100%

El conjunto de datos de la data set está conformado por un total de setenta y siete (77) variables agrupadas entre continuas y categóricas.

Las variables continuas se basan en los resultados obtenidos por el estudiante en la prueba Saber 11. En el Anexo

Variables continuas 1: PUNTAJE_MATEMAT, PUNTAJE _INGLES, PUNTAJE_LECTUCRI y PUNTAJE_TOTAL. En la Tabla 8 se detallan algunas medidas estadísticas de las variables.

Tabla 8. ***Variables continuas 1***

		Puntaje_Matemat	Puntaje Lectura Critica	Puntaje Ingles	Puntaje Total
Muestra Total	**Count**	1.344,000000	1.344,000000	1.344,000000	1.344,000000
Media	**Mean**	53,065952	51,758043	54,146503	262,292969
Desv Standard	**Std**	10,473051	11,258389	8,838265	41,791135
Valor Minimo	**Min**	8,000000	9,000000	9,000000	44,000000
25%	**25%**	46,000000	43,687500	48,000000	231,705000
50%	**50%**	53,000000	51,000000	54,000000	259,000000
75%	**75%**	60,000000	58,000000	60,000000	290,250000
Valor Maximo	**Max**	97,000000	100,000000	91,000000	434,000000

Figura 24. Histogramas estudiantes por rendimiento saber 11 área de matemáticas

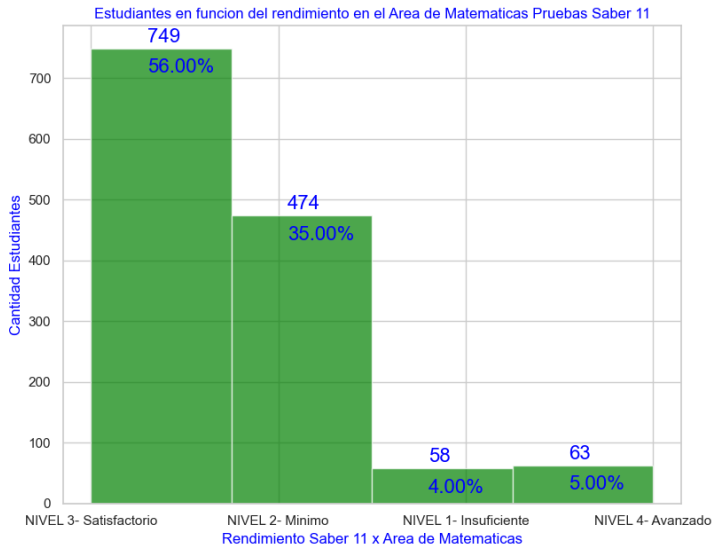


Figura 25. Histogramas estudiantes por rendimiento saber 11 área de ingles

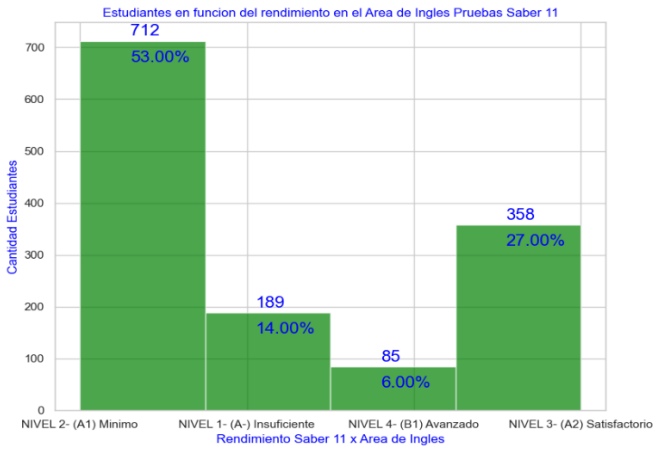
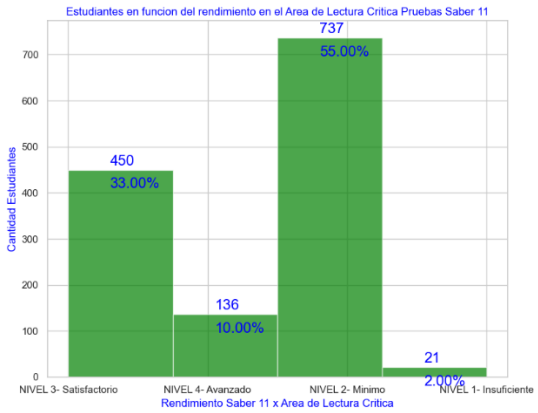


Figura 26. Histogramas estudiantes por rendimiento saber 11 área de lectura critica



Puede concluirse que la media aritmética de los puntajes de las áreas de las pruebas saber 11 que mayor incidencia tienen para las facultades en estudio, como **matemáticas, lectura crítica** e **inglés,** se ubica por encima del 50% de rendimiento, que corresponde con el Nivel 3 – Satisfactorio.

Adicionalmente, la desviación estándar de 10.47 nos permite identificar que la mayor concentración de los datos está entre 42.00 y 53.50 de promedio de los puntajes en las tres áreas en estudio.

Finalmente, el puntaje global obtenido como media aritmética (para llegar a este valor se suman todas las áreas de conocimiento evaluadas por el ICFES) está ubicado en el Nivel 2 – Medio, lo que permite expresar que la preparación previa del 75% de la población no es apropiada para iniciar el ciclo de vida académico superior.

Para complementar la información sobre las pruebas Saber 11, véase el Anexo 3 de este documento y los documentos del ICFES dispuestos en conjunto con el documento de tesis, que explican de manera detallada, qué se evalúa en cada área y cuál es el alcance de cada uno de los niveles alcanzados según el puntaje obtenido.

Variables continuas 2: MATERIAS_CURSADAS, MATERIAS_PERDIDAS, CREDITOS_CURSADOS, CREDITOS_PERDIDOS. En la Tabla 9 se detallan algunas medidas estadísticas de las variables.

Tabla 9. ***Variables discretas***

		Materias Cursadas	Materias Perdidas	Creditos Cursados	Creditos Perdidos
Muestra Total	**Count**	1.344,000000	1.344,000000	1.344,000000	1.344,000000
Media	**Mean**	16,365327	2,296131	36,461310	5,227679
Desv Standard	**Std**	8,895924	2,970209	23,325020	6,877198
Valor Minimo	**Min**	1,000000	0,000000	2,000000	0,000000
25%	**25%**	9,000000	0,000000	17,000000	0,000000
50%	**50%**	15,000000	1,000000	32,000000	2,000000
75%	**75%**	22,000000	4,000000	52,000000	9,000000
Valor Maximo	**Max**	59,000000	15,000000	155,000000	37,000000

Figura 27. Rendimientos estudiantes en función de las materias aprobadas

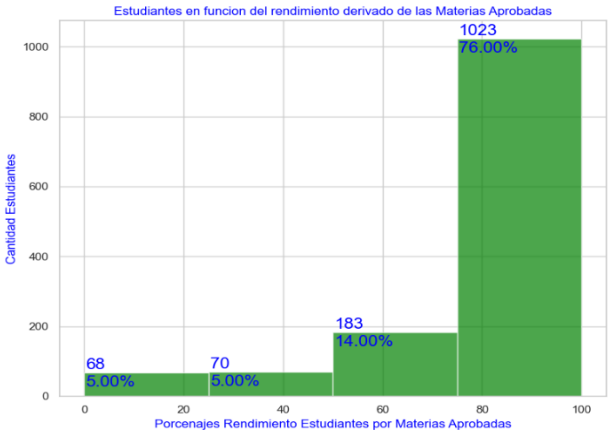
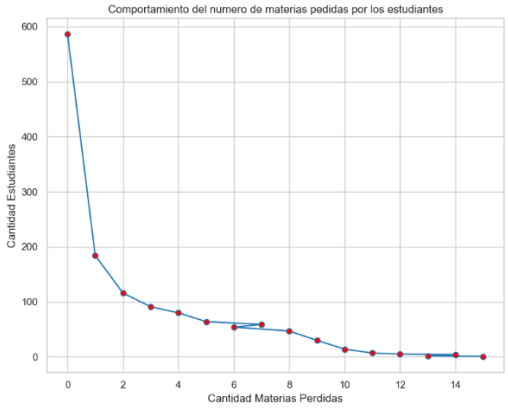


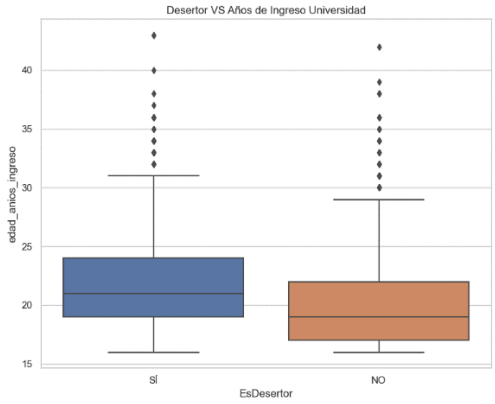
Figura 28. Estudiantes clasificados por número de materias perdidas



Puede concluirse que un 24% de los estudiantes durante su estancia en la universidad aprobaron hasta el 80% de las materias, frente a un 76% de los estudiantes que superan el 80% de las materias cursadas durante los primeros cuatro (4) semestres en la universidad.

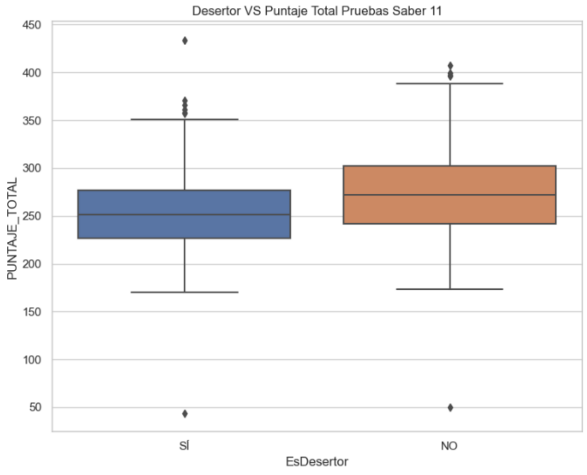
Un total de 580 estudiantes (43.15%) de los matriculados no perdieron ninguna materia del programa cursado y el restante número de ellos (784) perdió al menos una materia durante los primeros cuatro (4) semestres en la universidad.

Figura 29. Distribución del género estudiante vrs variable objetivo (Desertor).



La grafica de cajas y bigotes permite concluir que el 50% de la población que deserto se encuentra entre los 19 y 24 años de edad y presentan valores atípicos a partir de los 31 años. De igual manera el 50% de la población que NO deserto está concentrada entre los 17 y 22 años, con valores atípicos a partir de los 29 años.

Figura 30. Distribución del puntaje pruebas saber 11 estudiantes vrs variable objetivo (Desertor).



La grafica de cajas y bigotes permite concluir que el 50% de los estudiantes desertores obtuvieron puntajes en las pruebas saber 11 entre 240 y 275 puntos, con valores atípicos inferiores a 1 y superiores a 425 puntos. Así mismo el 50% de los NO desertores se concentra entre los 245 y 303 puntos en las pruebas saber 11, con valores atípicos inferiores a 50 y superiores a 405 puntos.

Variables categóricas: En un alto porcentaje las variables producto de esta investigación se clasifican como categóricas. Por tanto, en las gráficas que se muestran a continuación se exploran y analizan algunas de ellas.

Tabla 10. ***Variables categóricas 1***

		Genero	Estado Civil	Estrato Procede	Estrato Reside	Depto Procede	Mpio Procede
Muestra Total	**Count**	1.344	1.344	1.344	1.344	1.344	1.344
Vlrs unicos	**Unique**	2	6	6	6	21	71
Mas Repite	**Top**	MASCULINO	SOLTERO(A)	Estrato 3	Estrato 3	CALDAS	MANIZALES
Frecuencia	**Freq**	847	1.261	654	672	1.252	1.036

Figura 31. Cantidad de estudiantes según el género

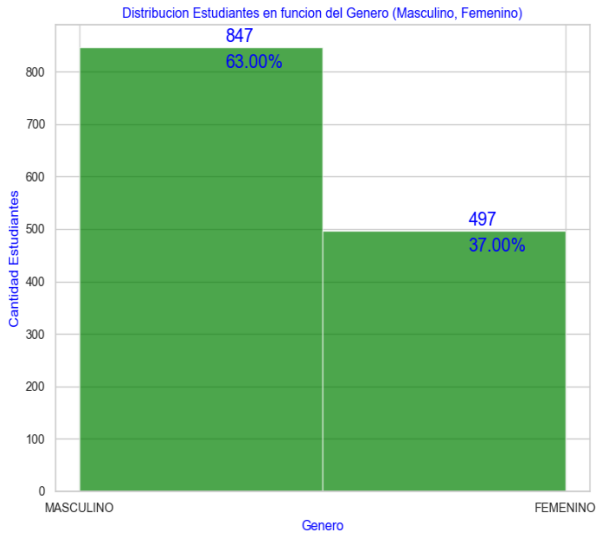
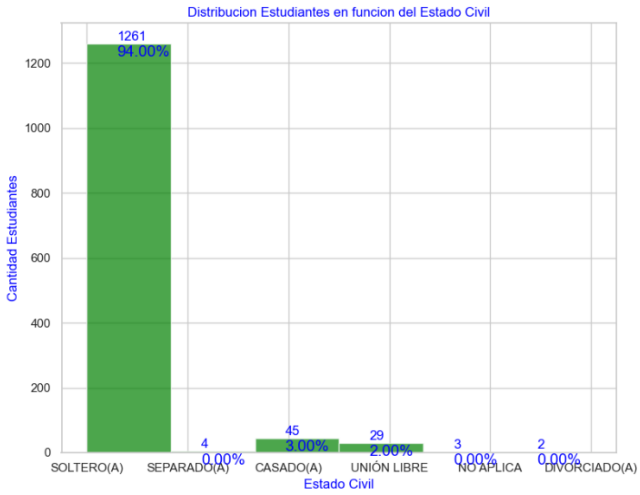


Figura 32. Cantidad estudiantes según su estado civil



De acuerdo con la muestra total de 1.344 estudiantes, puede inferirse que los hombres tienen mayor acceso a la educación superior respecto a las mujeres pues, por cada 100 personas que ingresan a las facultades en estudio, solo 36 son mujeres.

Así mismo, solo el 7% de la población que ingresa a la educación superior pertenece a los 5 estados civiles diferentes a 'soltero'.

La población predominante en las facultades de estudio pertenece al estrato 3 y proviene del departamento de Caldas y de la ciudad de Manizales.

Tabla 11. ***Variables categóricas 2***

		Colegio	Fue la Primera	Grupo	Rendi Global	Rendimiento	Recibio	Ultimo Periodo
		Procedencia	Opcion	Etareo	Pruebas Saber	Ingles	Acompañamiento	Cursado
Muestra Total	Count	1.344	1.344	1.344	1.344	1.344	1.344	1.344
Vlrs unicos	Unique	2	2	3	4	4	2	11
Mas Repite	Top	PUBLICO	NO	JOVENES	2.-Minimo	A2 -Minimo	Sin Acompañam	20221
Frecuencia	Freq	1.121	697	941	1.154	1.036	1.066	640

El comportamiento de estas variables indica que un 83.35% de la muestra total proviene de colegios privados; un 52% está cursando un pregrado que no era de su preferencia (primera opción) y que el 70% de la muestra corresponde a personas jóvenes.

Surge una conclusión preocupante que reafirma un criterio anteriormente señalado; se trata del bajo nivel del desempeño académico previo y las habilidades en segunda lengua, lo que puede ser causal del abandono académico.

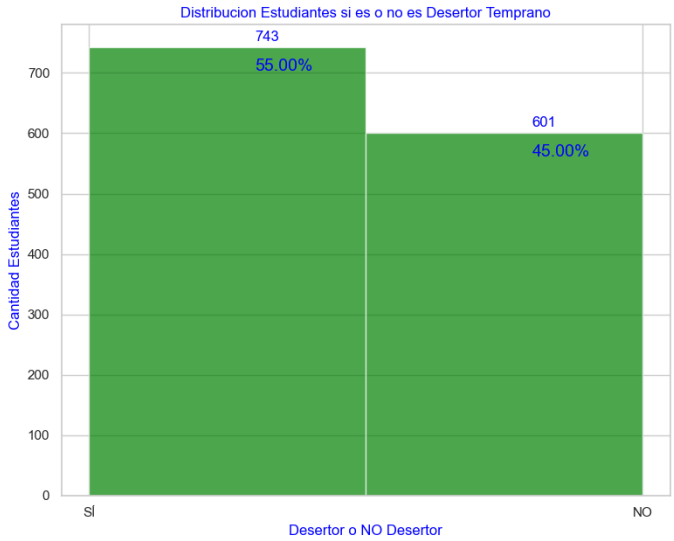
Finalmente, las cifras demuestran que solo el 11% de la población recibe acompañamiento psicopedagógico del área de Bienestar universitario.

Comprendiendo la variable objetivo: A partir de la Tabla 12 puede inferirse que de las dos alternativas: **SI Desertó** o **NO desertó** (No incluye reingresos), el valor más preponderante es **SI**, el cual se establece en un **55.28% = 743 estudiantes** del total de matriculados analizados (1.344 estudiantes). Sin embargo, cabe resaltar que el mayor valor de deserción está condicionado por la fecha de corte en la generación de los datos (inicios de diciembre de 2022), puesto que el mayor número de desertores se presentó en el último periodo cursado, que estuvo condicionado por la pandemia y por el aplazamiento de la fecha de matrículas y pagos; este valor descendió de 10 a 20%, lo cual ubica la deserción temprana real en un 30 a 35%.

Tabla 12. ***Variable Objetivo Deserción***

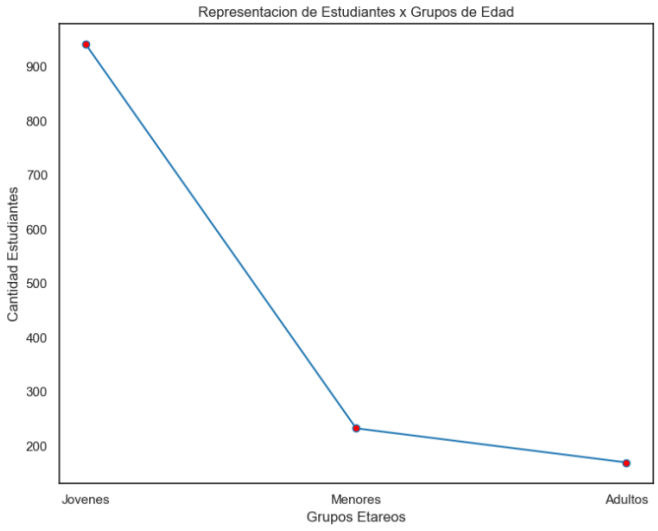
		Deserto o NO Deserto
Muestra Total	**Count**	1.344
Vlrs unicos	**Unique**	2
Mas Repite	**Top**	SI
Frecuencia	**Freq**	743

Figura 33. Distribución del estado del estudiante desertor o No desertor



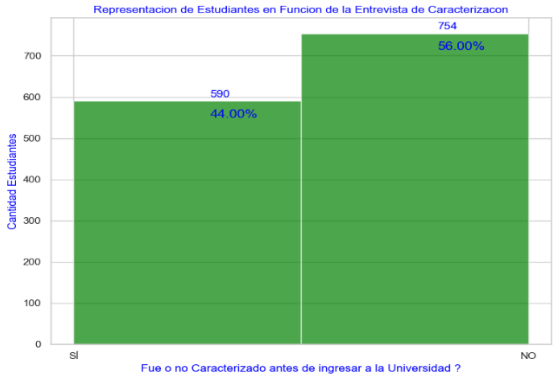
Análisis gráfico univariado de variables:

Figura 34. Cantidad de estudiantes según el grupo etario



La gráfica de grupos etarios confirma que la población universitaria prevaleciente son los jóvenes; un 14% son menores de edad y solo un 7% son adultos que ingresan a las facultades de este estudio.

Figura 35. ***Caracterización previa al ingreso de los estudiantes***

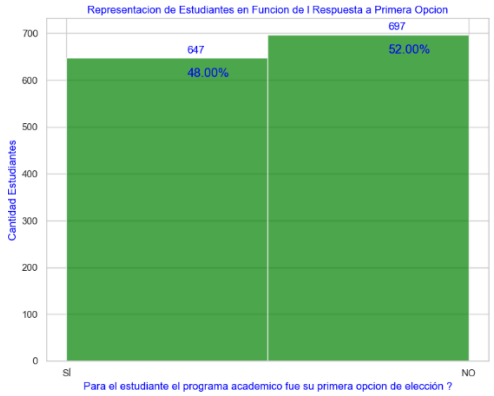


De acuerdo con la gráfica, el proceso de caracterización de estudiantes previo a su ingreso a la universidad se aplica solo en el 44% de los casos. Esta situación permite inferir que es probable que existan estudiantes que ingresan a un programa para el cual no tienen las habilidades previas necesarias.

Figura 36. ***Entrevista previa al ingreso de los estudiantes***

Similar a la situación anterior, la gráfica evidencia que el proceso de entrevista de estudiantes previo al ingreso a la universidad se aplica solo al 48% de la población.

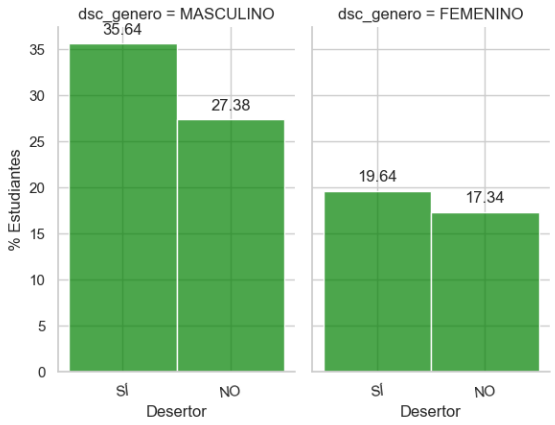
Figura 37. Estudiantes matriculados en su primera opción de programa académico



La grafica evidencia que el 48% de la población matriculada seleccionado como primera opción el programa académico, mientras que el 52% restante lo tenían como segunda opción.

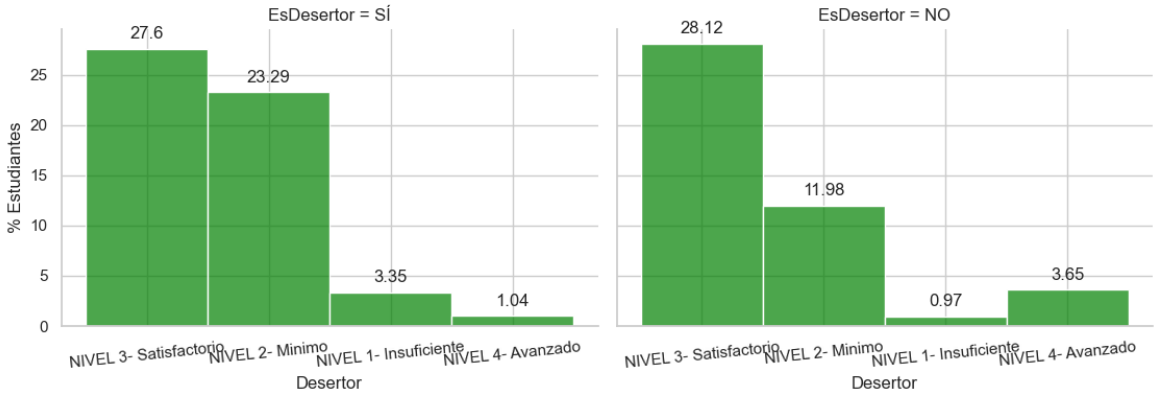
Análisis gráfico bivariado de variables: Por medio de la evaluación gráfica que a continuación se expresa, se analizan los comportamientos y/o relaciones de las diversas del modelo predictivo y la variable objetivo **desertor**.

Figura 38. Comportamiento Desertores vs. Genero



La gráfica muestra que el 35.64% de hombres y el 19.64% de mujeres presentan deserción temprana para un total de desertores del 55.38% (744 personas) del número total de 1.344 estudiantes.

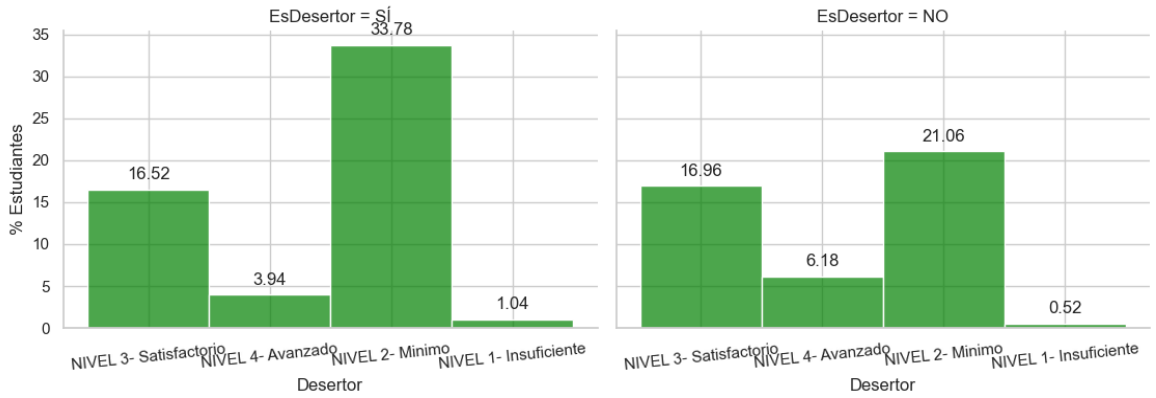
Figura 39. Comportamiento Desertores vs. Rendimiento Matemáticas Saber 11



La gráfica permite deducir que el mayor porcentaje 92.08% (685 personas) de deserción sobre el número total de desertores (744), está concentrado en un rendimiento satisfactorio y mínimo en las pruebas saber 11 en el área de las matemáticas. Así mismo,

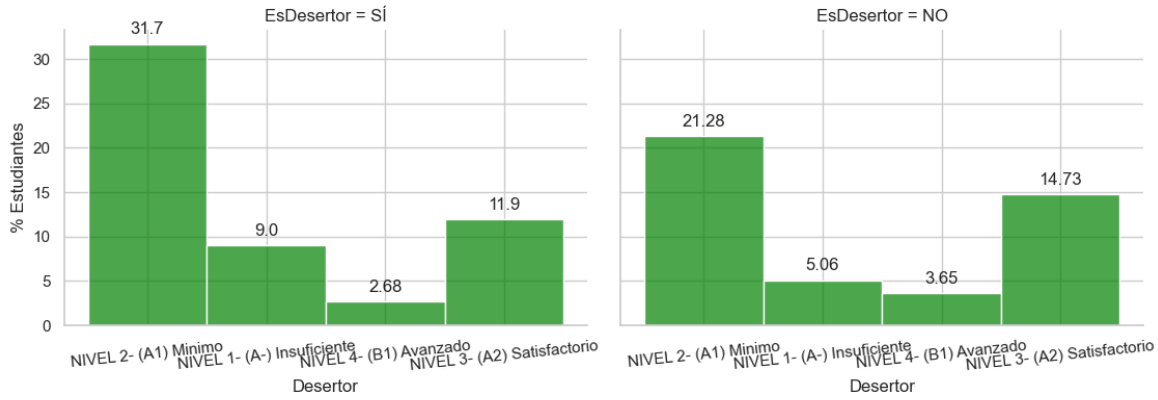
demuestra que aproximadamente 370 estudiantes con rendimiento satisfactorio en matemáticas desertan tempranamente de la universidad.

Figura 40. Comportamiento Desertores vs. Rendimiento Lectura Critica Saber 11



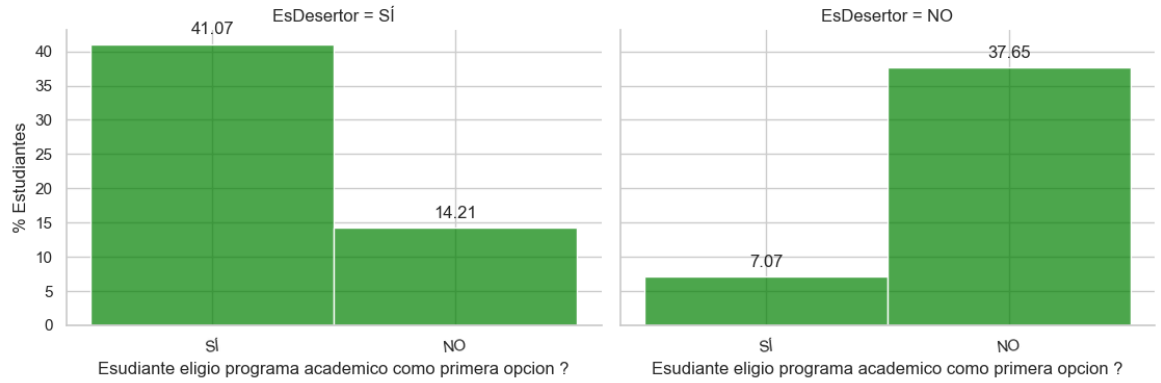
La gráfica permite deducir que el mayor porcentaje 90.99% (677 personas) de deserción sobre el número total de desertores (744), está concentrado en un rendimiento satisfactorio y mínimo en las pruebas saber 11 en el área de lectura crítica. Así mismo, demuestra que aproximadamente 222 estudiantes con rendimientos Satisfactorio desertan tempranamente de la universidad. Cabe resaltar que preocupa el nivel bajo en lectura crítica con la cual ingresan los estudiantes a la universidad, dado que del total de estudiantes analizados 1.344 un total de 750 (54.84%) desertores y no desertores presentan este bajo nivel.

Figura 41. Comportamiento Desertores vs. Rendimiento Inglés Saber 11



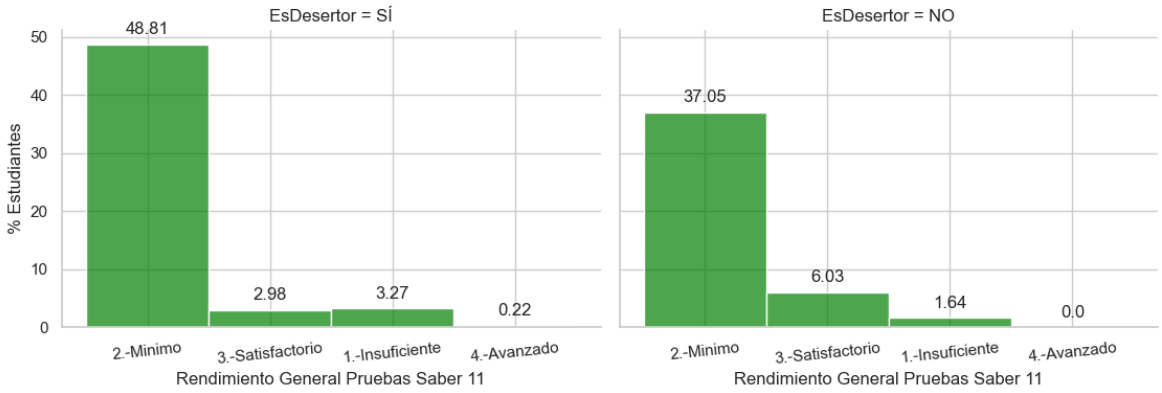
La gráfica permite inferir que la mayor tendencia de deserción se presenta con un nivel de inglés entre mínimo e insuficiente, es decir de 744 estudiantes desertores un total de 548 de ellos su nivel es bajo en esta área temática de las pruebas saber 11.

Figura 42. Comportamiento Desertores vs. El programa académico primera opción



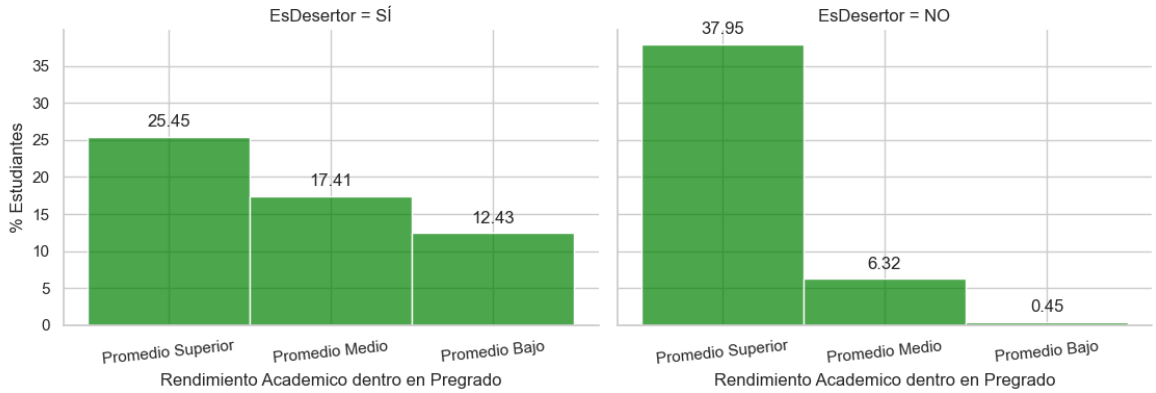
La gráfica permite concluir que, a pesar de haberse matriculado en el programa académico escogido como primera opción, 552 (74.29%) estudiantes de los 744 desertores eligieron este programa como el que más les atraía.

Figura 43. Desertores vs. El rendimiento global en Saber 11



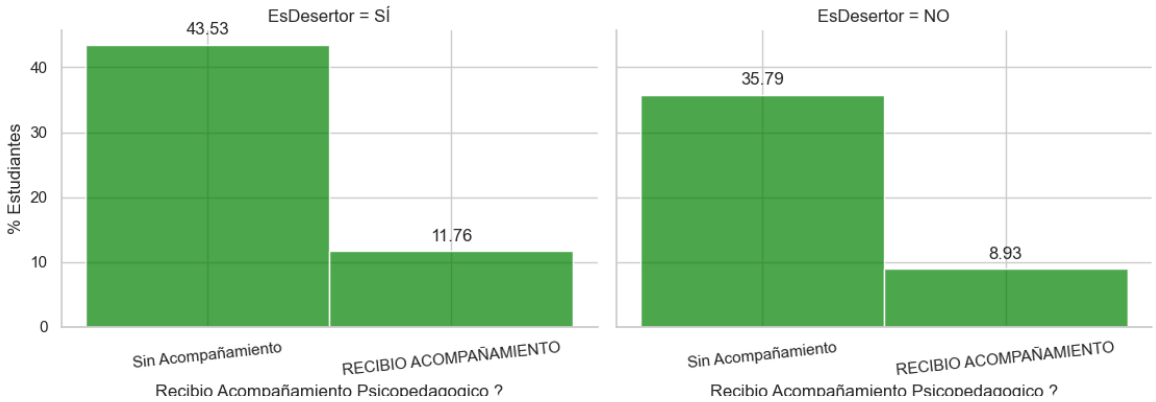
La gráfica da a entender que del total de estudiantes 1.344, un total de 1.153 de ellos desertores o no presentan rendimientos generales mínimos en las pruebas saber 11, es decir que el 85.86% de ellos ingresan a la universidad con competencias académicas por debajo de lo esperado o requerido para asumir un programa académico de pregrado. De igual manera un total de 656 (88.29%) estudiantes de un total de 744 desertores presentan este comportamiento.

Figura 44. Desertores vs. Nivel del Rendimiento académico en el programa



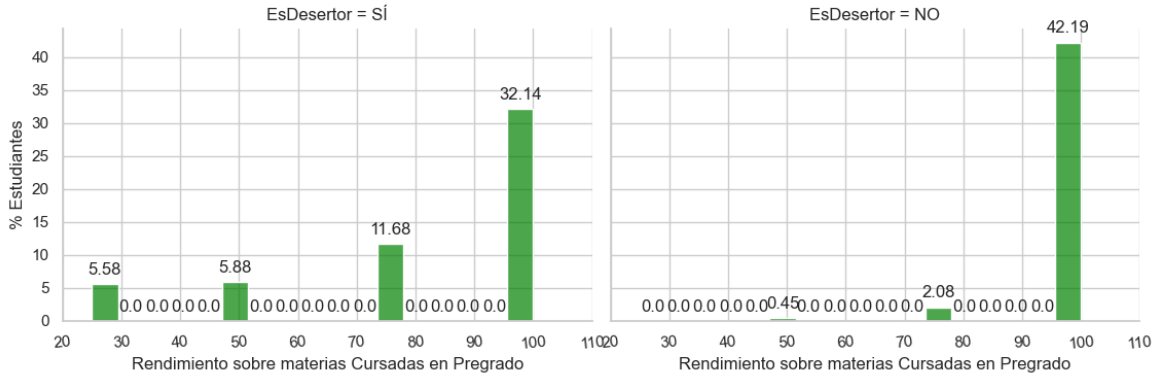
La gráfica evidencia que a pesar que el 46% (373) estudiantes que desertaron presentaron un rendimiento académico en el programa de pregrado superior, y un total de 576 (77.52%) estudiantes desertaron del programa a pesar que su rendimiento académico infiere la aprobación de los periodos o semestres cursados.

Figura 45. Desertores vs. Haber recibido acompañamiento



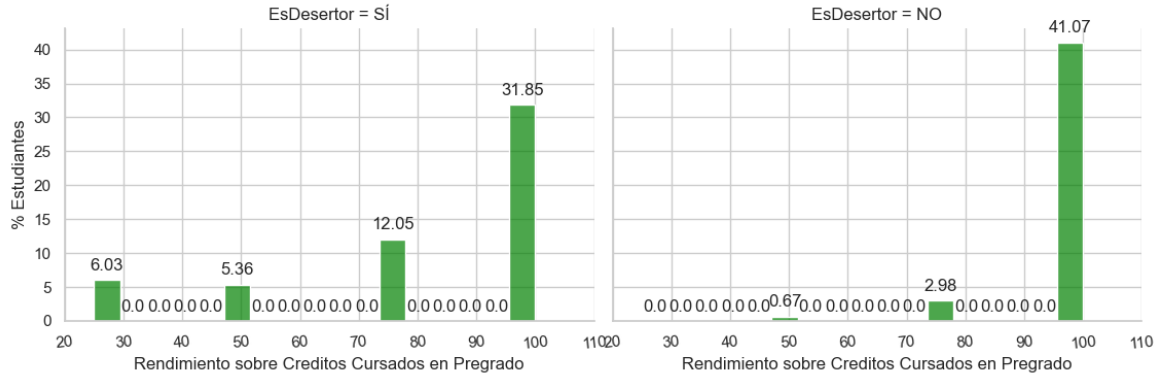
Por la gráfica puede inferirse que sin importar si los estudiantes recibieron o no acompañamiento del área de Bienestar estudiantil, el 79.32% de los estudiantes (1.066) entre desertores y no desertores no recibieron acompañamiento psicopedagógico por parte de la universidad. De los estudiantes desertores (744) el 78.73% (585) de ellos no recibieron acompañamiento.

Figura 46. Desertores vs. Rendimiento Estudiante por materias cursadas



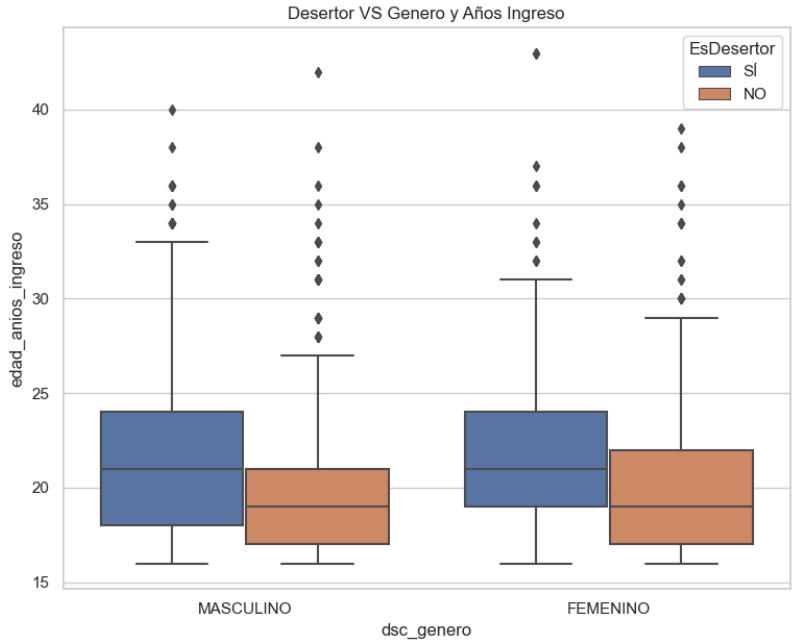
La gráfica permite concluir que el 59% (438) de los estudiantes desertores cursaron el total de las asignaturas y el 79.27% (5.909 de todos los desertores cursaron arriba del 75% e las materias del pregrado.

Figura 47. Desertores vs. Rendimiento Estudiante por créditos cursados



La gráfica permite concluir que respecto al rendimiento sobre el número de créditos cursados se presenta un comportamiento similar al análisis anterior, sin embargo, hay una leve tendencia a disminuir en quienes cursaron el total de las materias de pregrado y que son desertores.

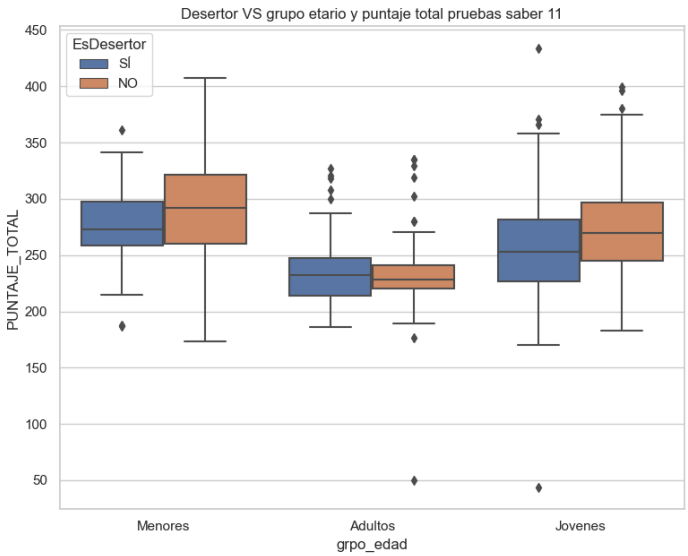
Figura 48. Desertores vs género y edad años de ingreso a la universidad del estudiante



La gráfica evidencia que: El 50% de hombres desertores oscilan entre los 17 y 24 años de edad de ingreso, con valores atípicos de 30 años y más. El 50% de hombres NO desertores oscilan entre los 17 y 21 años, con edades atípicas de 27 años y más.

El 50% de las mujeres desertoras oscilan entre los 19 y 24 años de edad de ingreso a la universidad, con edades atípicas de 31 años y más. El 50% de mujeres NO desertoras oscilan entre 17 y 21 años, con edades atípicas de 30 años y más.

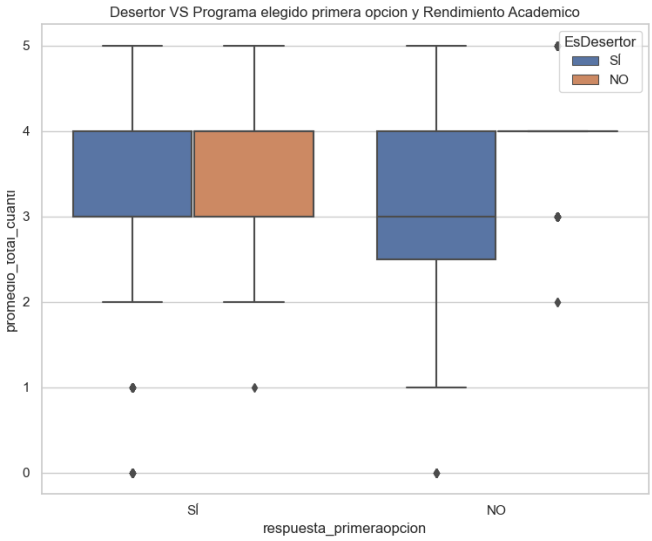
Figura 49. Desertores vs grupo etario, puntaje total pruebas saber 11



La gráfica permite concluir: El rendimiento en las pruebas saber es mayormente significativo en los jóvenes y menores de edad, caso contrario con los adultos cuyo rendimiento esta entre 210 y 250 puntos. Cabe destacar el mayor grupo de desertores con aceptables rendimientos en las pruebas saber está conformado por menores de edad y jóvenes con puntajes entre 235 y 300 puntos.

Dentro del grupo de adultos es notorio que el 50% de los desertores figuren con puntajes entre 230 y 250 puntos, que son considerados puntajes de aceptables para abajo. No sobra enumerar que en todas las categorías los puntajes altos y/o destacados dentro de las pruebas saber 11 son valores atípicos mayoritariamente de 350 puntos hacia arriba.

Figura 50. Desertores vs Programa elegido primera opción y rendimiento académico



La gráfica permite concluir: El porcentaje mayoritario de desertores del programa al cual ingresaron se encuentra que al momento de ingresar a la universidad el programa no fue su primera elección.

De aquellos desertores que si eligieron el programa como primera opción el 50% de ellos se fueron y el 50% restante el programa fue su segunda opción.

9.2.2 Modelado

Future importance (importancia de la característica de permutación: Obtención de las variables más relevantes que explican el modelo. Mediante este proceso se seleccionaron las variables más relevantes para ejecutar el algoritmo predictor de *Random forest*. La implementación de este algoritmo fue elaborada bajo el lenguaje de programación PYTHON en la plataforma ANACONDA.

La relevancia del future impotance radica en la calificación de todas las variables del set de datos de prueba bajo un modelo ya entrenado de regresión o de

clasificación como lo son el decision tree y el random forest. Permitiendo entonces ordenar de mayor a menor calificación las variables que mejor expliquen el evento, en este caso de la deserción temprana universitaria en la Universidad de Manizales. Seleccionado entonces la variable objetivo como parámetro de *metric for measuring performance* (métrica para medir el rendimiento) que será usado para calcular la calidad del modelo. Para el caso de modelos de clasificación que se usa en la investigación arroja resultados para los indicadores de exactitud, precisión y recuperación.

Lista de variables:

Tabla 13. ***Conjunto de variables del modelo predictivo***

		CONJUNTO DE VARIABLES DEL MODELO PREDICTOR	
No.	**Variable**	**Descripcion de la Variable**	
1	**ultpercur**	Ultimo periodo academico cursado frente a la fecha de corte	**Importancia**
2	**primeropc**	Si la Carrera seleccionada fue la primera opcion para el estudiante	0.089661
3	**entrevisto**	Si el estudiante previo al ingreso a la univesidad fue entrevistado	0.088077
4	**matper**	Cantidad de materias perdidas frente al total de materias cursadas	0.056567
5	**rendicreditos**	Nivel de rendimiento obtenido sobre los creitos totales vrs los creditos aprobados durante los periodos cursados	0.043635
6	**caracterizo**	Si el estudiante previo al ingreso a la universidad se caracterizo en funcion de sus habilidades	0.034398
7	**Perdio_Materias**	Si o No perdio materias el estudiante durante los periodos cursados	0.033967
8	**NivPromNotas**	Nivel de ubicación del su rendimiento academico en funcion del promedio obtenido en las materias cuantitativas durante todos los periodos cursados	0.029835
9	**fallas**	Cantidad de fallas o ausencias a clase durante los periodos cursados	0.026282
10	**matcur**	Cantidad de materias cursadas durante los periodos cursados	0.017400
11	**credicur**	Numero total de creditos cursados durante los periodos cursados	0.016077
12	**crediper**	Cantidad de creditos perdidos durante el numero de periodos cursados	0.014890
13	**Rendi_Academico**	Nivel de rendimiento academico general sobre el total de periodos cursados	0.013613
14	**pericur**	Cantidad e periodos cursados	0.011750
15	**rendimat**	Nivel de rendimiento en el area de matematicas durante la presentacion de las pruebas saber 11	0.011414
16	**estreside**	Estrato de residencia del estudiante durante los periodos cursados	0.005990
17	**duracionprog**	Duracion de semestres del programa academico cursado	0.005954

18	**programa**	programa que esta cursando dentro de las facultades de ciencias adminitrativas y economicas o Ciencias e Ingenieria	0.004762
19	**NivIngles**	Nivel de rendimiento en el area de ingles en la presentacion de las pruebas saber 11	0.004705
20	**NivMatemat**	Nivel de rendimiento en el area de matematicas en la presentacion de las pruebas saber 11	0.004026
21	**estprocede**	Estrato de procedencia del estudiante	0.003695
22	**recibioacomp**	Si el estudiante durante los periodos cursados recibio o no acompañamiento psicopedagogico por parte de bienestar universitario	0.002756
23	**genero**	Genero del estudiante MASCULINO, FEMENINO, OTRO	0.002395
24	**NivLectuCri**	Nivel de rendimiento en el area de lectura critica en la presentacion de las pruebas saber 11	0.001955
25	**mpioprocede**	municipio de procedencia	0.001898
26	**dptoprocede**	Departamento de procedencia del estudiante	0.001495
27	**ausencias**	Si presento ausencias o no durante el numero de periodos cursados	0.001471
28	**grupoedad**	Grupo etareo sobre el cual se segmento al numero de estudiantes	0.001373
29	**tipcoleg**	tipo de colegio PRIVADO o PUBLICO	0.000708
30	**NivSaber11**	Nivel de rendimiento en el GLOBAL en la presentacion de las pruebas saber 11	0.000297
31	**escivil**	Estado civil del estudiante	0.000105
32	**calendacoleg**	calendario de la educacion media. CALENDARIO A o B	0.000000

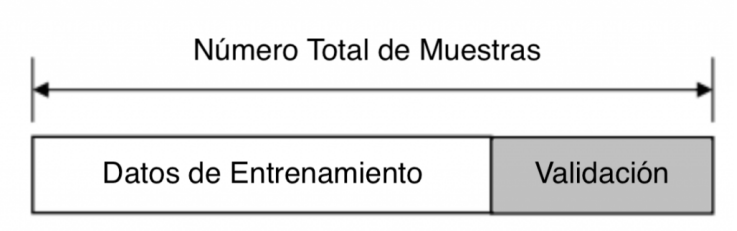
Construcción de las variables X y Y:

Una vez obtenido el resultado del **future importance** donde se obtuvieron las variables más explicativas del fenómeno (Desertor temprano SI/No). Se determina que serán seleccionadas las variables que estén por encima del **0.099%** de explicación de fenómeno (ver tabla 13) para preparar a **X** y **Y**, requisitos indispensables para la ejecución de los modelos predictivos. Es decir que **X** estará conformada por las quince (15) variables mejor calificadas (ver tabla 13) y **Y** es equivalente a la variable objetivo del proyecto (Es Desertor Si o No = desertor_t). La variable **Y**, fue debidamente transformada en = para No Desertor y en 1 para Desertor, sin embargo, para efectos de visualización de resultados se transforma el valor en Si o No.

Creación de Test y Train

El total del dataset de datos es dividido en dos conjuntos, el **Testeo** equivalente al **30%** de la data y el **Entrenamiento** equivalente al **70%**.

Figura 51. Partición, entrenamiento y testeo

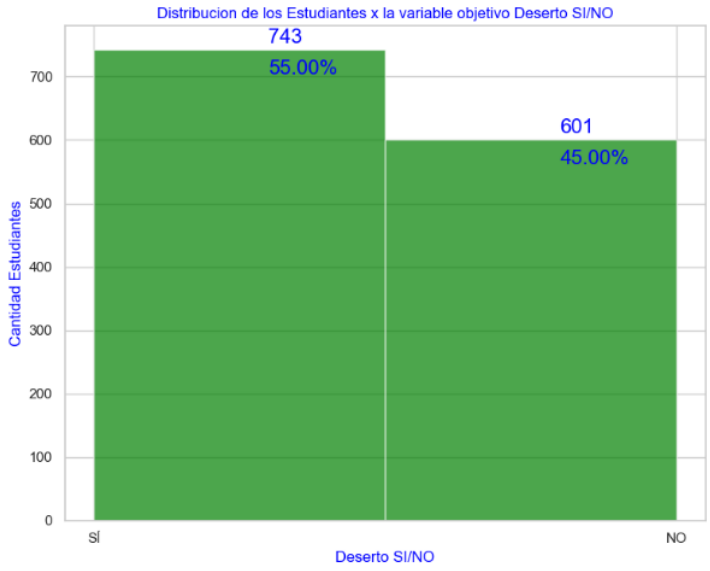


70% **30%**

Nota: Tomado de Aprende *Machine learning*. Enlace: https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

Balancear la variable objetivo

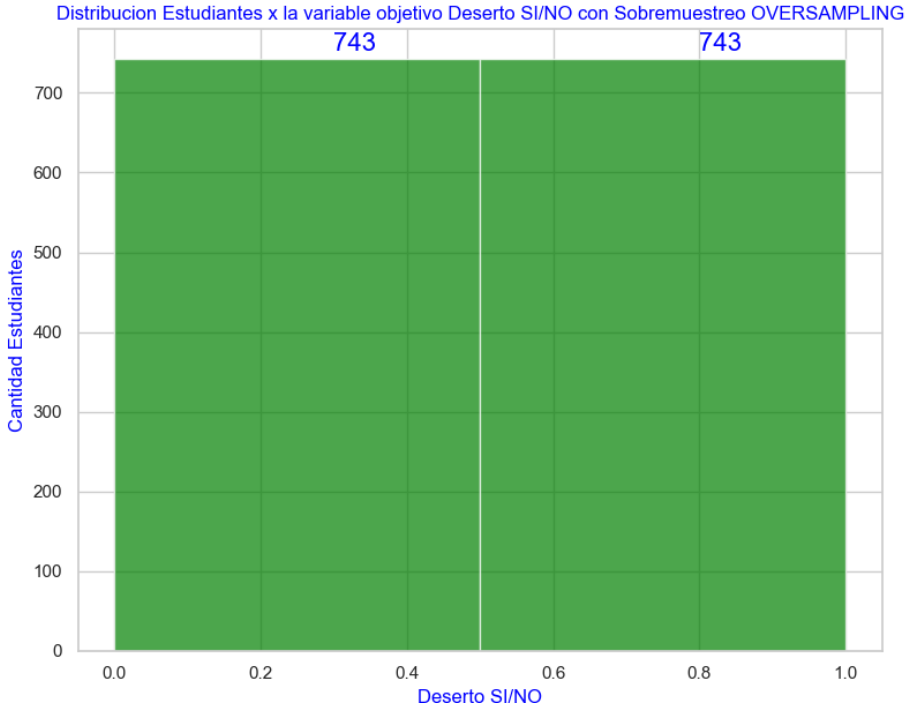
Figura 52. Comparación entre estudiantes desertores y no desertores



Al evaluar los valores de la variable dependiente **EsDesertor** (Si o No), puede observarse que los valores correspondientes al **Si** equivalen al **55%** y los valores del **NO** equivalen al **45%**; por tanto, para obtener una mayor precisión con el modelo, se determinó balancear la variable objetivo Desertor con el método **overdersampling (sobremuestreo).** Esta técnica puede visualizarse con detalle en la figura 54, en la cual la barra azul de menor porcentaje es igualada como se visualiza en las barras rojas al lado derecho de la imagen.

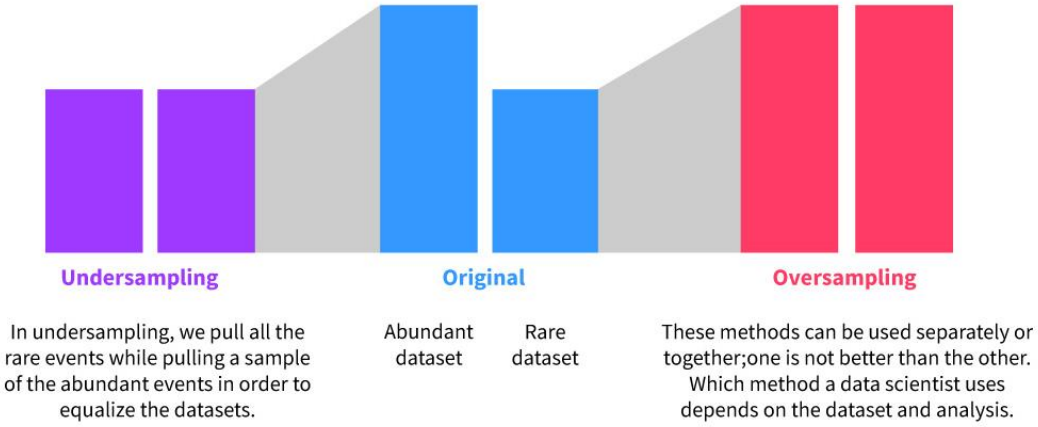
A pesar que el modelo no presenta diferencias significativas entre la distribución de la variable objetivo (55% si desertaron y 45% no desertaron), el aplicar esta técnica sobre las muestras equilibro ambos resultados a un 50% insertando nuevos casos No deserción hasta igualar la muestra (ver figura 73).

Figura 53. Distribución de Estudiantes Desertores después del sobremuestreo



OverSamplig: Duplicación de muestras de la clase minoritaria para balancear el modelo (véanse en la figura 26 los recuadros en rojo). Es empleada cuando se tiene una baja proporción de casos minoritarios en clasificaciones binomiales.

Figura 54. Sobremuestreo / Oversampling



Nota: Tomado de What is Undersampling (MastersInDataScienc, 2022)

Después de ejecutar el **OverSampling** se ejecuta nuevamente el **future importance**, lo que mejora el modelo en un 10% y da origen a una nueva clasificación de las variables /Véase la Tabla 14).

Tabla 14. Segunda clasificación de importancia de variables

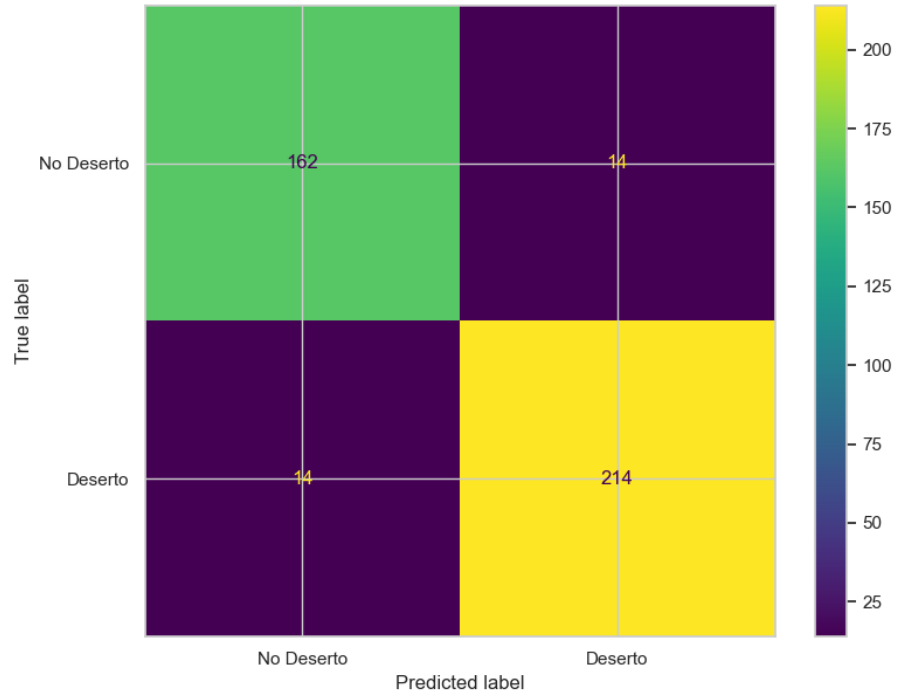
Con el número de variables seleccionadas se corren de nuevo los modelos decisión tree y *random forest*.

		CONJUNTO DE VARIABLES DEL MODELO PREDICTOR	
No.	**Variable**	**Descripcion de la Variable**	
1	**ultpercur**	Ultimo periodo academico cursado frente a la fecha de corte	**Importancia**
2	**primeropc**	Si la Carrera seleccionada fue la primera opcion para el estudiante	0.089661
3	**entrevisto**	Si el estudiante previo al ingreso a la univesidad fue entrevistado	0.088077
4	**matper**	Cantidad de materias perdidas frente al total de materias cursadas	0.056567
5	**rendicreditos**	Nivel de rendimiento obtenido sobre los creitos totales vrs los creditos aprobados durante los periodos cursados	0.043635
6	**caracterizo**	Si el estudiante previo al ingreso a la universidad se caracterizo en funcion de sus habilidades	0.034398
7	**Perdio_Materias**	Si o No perdio materias el estudiante durante los periodos cursados	0.033967
8	**NivPromNotas**	Nivel de ubicación del su rendimiento academico en funcion del promedio obtenido en las materias cuantitativas durante todos los periodos cursados	0.029835
9	**fallas**	Cantidad de fallas o ausencias a clase durante los periodos cursados	0.026282
10	**matcur**	Cantidad de materias cursadas durante los periodos cursados	0.017400
11	**credicur**	Numero total de creditos cursados durante los periodos cursados	0.016077
12	**crediper**	Cantidad de creditos perdidos durante el numero de periodos cursados	0.014890
13	**Rendi_Academico**	Nivel de rendimiento academico general sobre el total de periodos cursados	0.013613
14	**pericur**	Cantidad e periodos cursados	0.011750
15	**rendimat**	Nivel de rendimiento en el area de matematicas durante la presentacion de las pruebas saber 11	0.011414

9.2.3 Evaluación

Como resultado posterior a los ajustes al algoritmo de *Random forest*, se obtienen los resultados de la predicción en la **matriz de confusión** en la figura 27, Resultados Matriz de confusión *Random forest*, pueden observarse cada una de las cuatro (4) posibilidades.

Figura 55. Resultados Matriz de confusión Random Forest



La matriz de confusión del *random forest* arroja una exactitud del 93%, lo cual expresa que, por cada 100 eventos predichos, 93 son verdaderos y 7 se predicen erradamente. En la gráfica, de un total de 404 eventos 376 son acertados y 28, desacertados.

Figura 56. Matriz de confusión para interpretar resultados

		Clasificación	
		Positivo	**Negativo**
Verdad Terreno	**Positivo**	Positivos Ciertos	Negativos Falsos
	Negativo	Positivos Falsos	Negativos Ciertos

Para interpretar de forma apropiada la matriz de confusión, los resultados deben interpretarse de la siguiente maneara:

- Ciertos positivos: El clasificador predice una muestra como positiva y realmente es positiva.
- Falsos positivos: El clasificador predice una muestra como positiva, pero de hecho es negativa.
- Ciertos negativos: El clasificador predice una muestra como negativa y realmente es negativa.
- Falsos negativos: El clasificador predice una muestra como negativa, pero de hecho es positiva.

Ligadas a los resultados de la matriz de confusión, vienen una serie de métricas que dan la validez al modelo; éstas son: **exactitud** (accuracy: fracción de predicciones correctas)**, sensibilidad** (recall: fracción de positivos identificados correctamente por el modelo entre todos los positivos reales)**, precisión** (fracción de elementos clasificados correctamente como positivo entre todos los que el modelo ha clasificado como positivos) y **especificidad** (F1 score: combina las métricas **Precision** y **Recall** para arrojar la media armónica)**.** Sus valores pueden apreciarse en la Tabla 15, expuesta a continuación.

Tabla 15. Métricas, exactitud, sensibilidad, precisión, especificidad Random forest

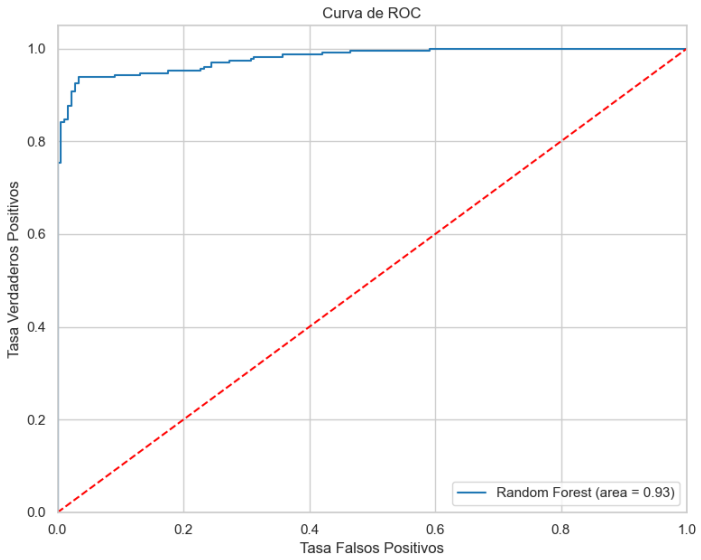
Metricas Random Forest	
Accuracy	0,930693
Precision	0,938596
Recall	0,938596
F1	0,938596
RocAuc	0,929526

Las métricas expuestas en la gráfica del *random forest* arrojan medidas que superan el 90%, lo cual permite concluir que las variables y el modelo elegido son adecuados para el tipo de problema en estudio.

Por último, se incluye una representación gráfica de la sensibilidad frente a la especificidad del sistema clasificador binario (Desertó Si o No), que también puede interpretarse como la proporción de verdaderos positivos frente a la proporción de falsos positivos.

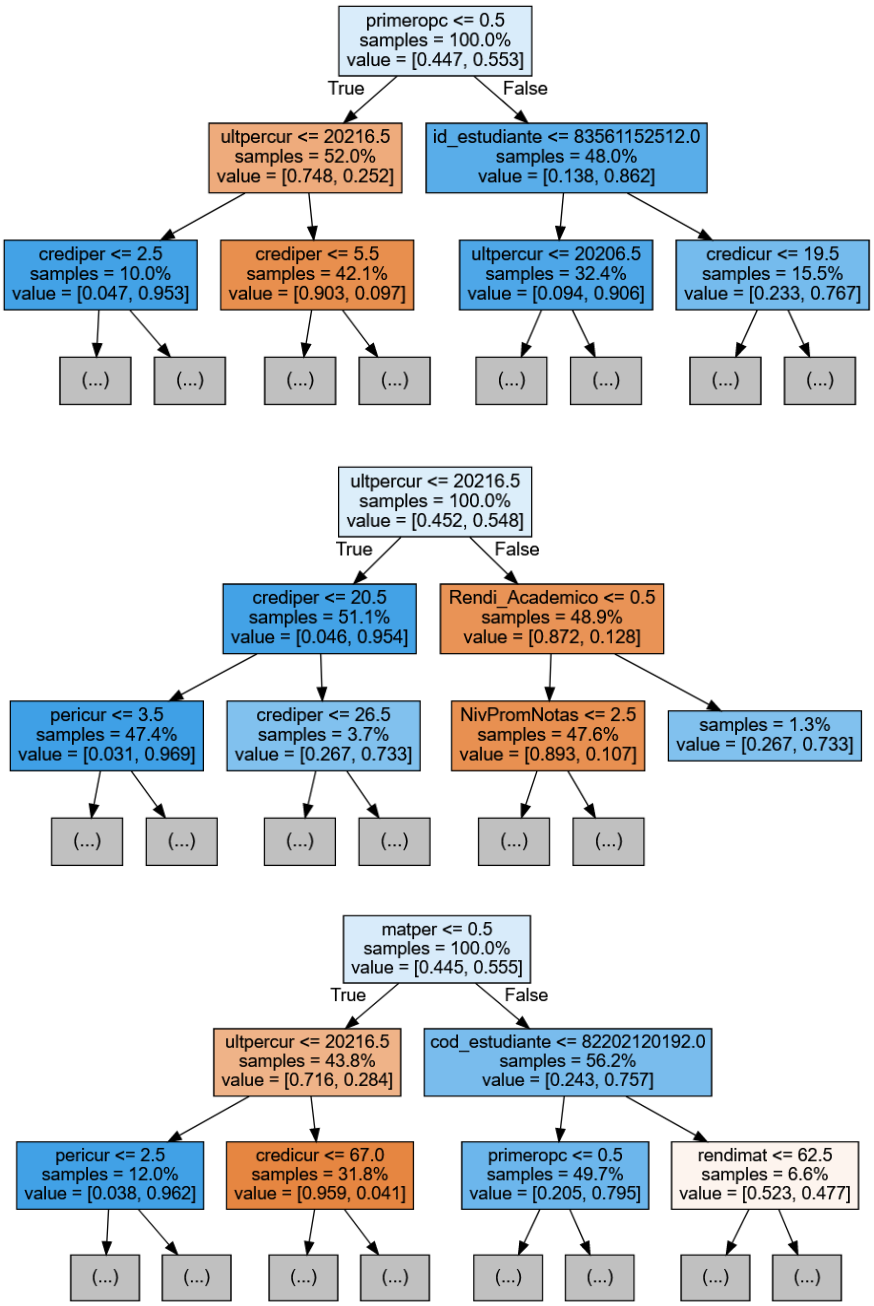
En resumen, **la Curva de ROC** proporciona herramientas para seleccionar los modelos más óptimos y descartar modelos no apropiados para la clase de problema planteado o predicción a obtener.

Figura 57. Área bajo la curva de Random forest



El indicador de área bajo la curva de este modelo arrojó un resultado de 93%, lo que indica que el modelo es capaz de clasificar los verdaderos desertores con una probabilidad de 93 por cada 100 estudiantes evaluados.

Figura 58. Arboles de decisión resultantes de la predicción por Random Forest

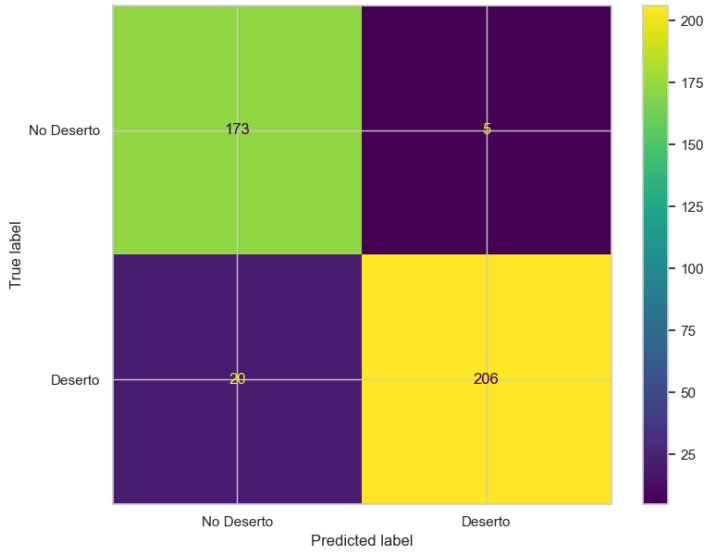


Producto del *future importance* del árbol de decisión, el resultado es el siguiente: Nombre de la variable y el valor de importancia.

ultpercur	0.913494
rendicreditos	0.047225
credicur	0.017325
matcur	0.011643
pericur	0.010313

Se ejecuta la técnica de **Oversampligng** para balancear la variable objetivo para el modelo de Decision Tree. Se emplea idéntica justificación y técnica empleada en el modelo de *Random forest*.

Figura 59. Resultados matriz de confusión *Decisión Tree*



La matriz de confusión producto del algoritmo Decisión tree arroja una exactitud del 93.8%, lo cual expresa que, por cada 100 eventos predichos, 94 son verdaderos y 6 se predicen erradamente. En la gráfica, del total de 404 eventos, 379 son acertados y 25,

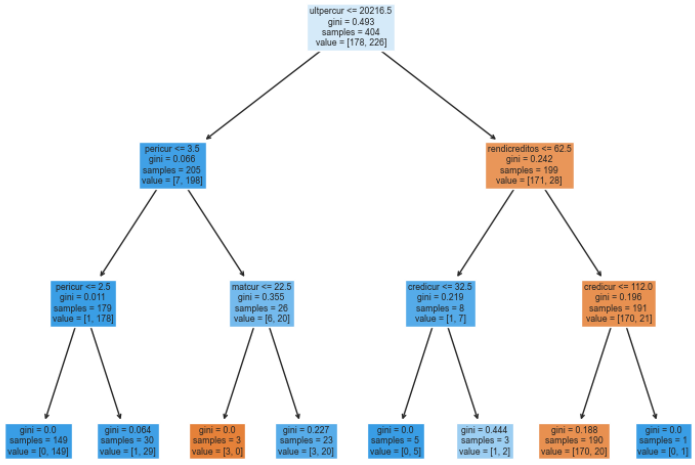
desacertados. Esos niveles fueron producto de la aplicación de herramientas de corrección de desbalance de datos como el oversampling (remuestreo).

Tabla 16. Métricas, exactitud, sensibilidad, precisión, especificidad Decisión Tree

Metricas Decision Tree	
Accuracy (Exactitud)	0,938118
Precision	0,976303
Recall (Sensibilidad)	0,911504
F1 Score	0,942791
RocAuc	0,941707

Las métricas expuestas en la gráfica de Decisión tree arroja medidas superiores al 90%, lo cual permite concluir que las variables y el modelo elegido son adecuados para el tipo de problema en estudio.

Figura 60. Arboles de decisión resultantes de la predicción por Decision Tree



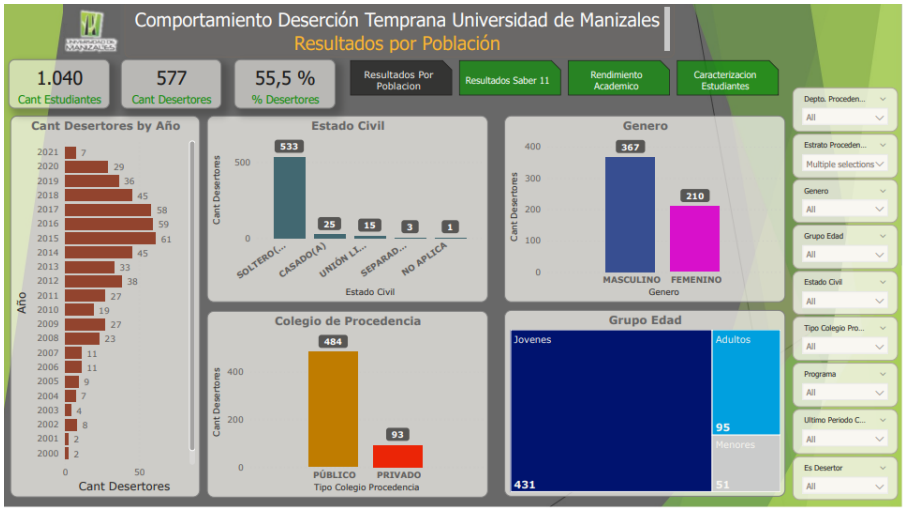
9.3 Fase 3: Construcción del tablero de visualización de resultados en PowewBI que refleje las condiciones de los estudiantes con riesgo de deserción temprana en la Universidad de Manizales

Los cuadros de visualización de los datos relacionados con la variable objetivo construidos en PowerBI Desktop, permiten a las personas interesadas en la clasificación de los datos fuente y los datos de la predicción, efectuar consultas a ellos cruzando la exploración por las diferentes variables categóricas en función de la variable objetivo (Es desertor o No), como por ejemplo: Estudiantes desertores clasificados por estrato, genero, municipio procedencia, rendimiento académico, etc.

La distribución actual de los cuadros de exploración es la siguiente:

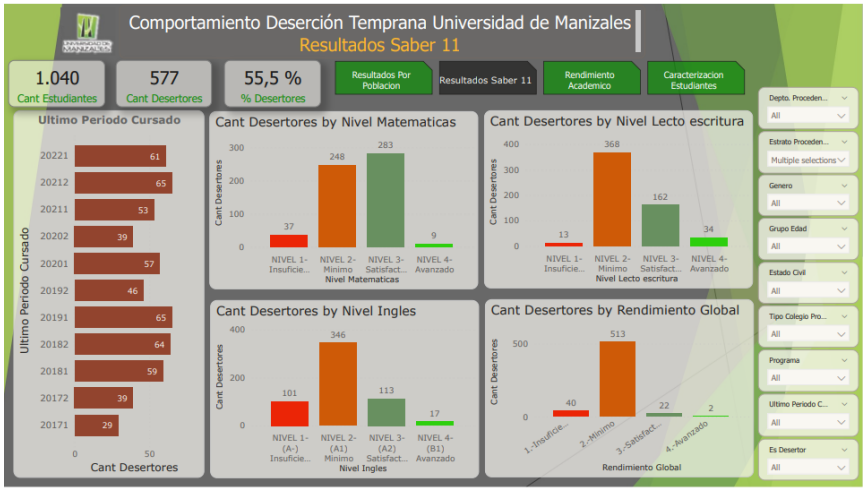
- La primera hoja en Power Bi se denomina **Resultados por Población.** Allí se pueden ver los gráficos de cantidad de estudiantes **general, estado civil, genero, colegio procedencia** y grupo de edad, estos datos pueden filtrarse (ventana a la derecha del cuadro resultados) por departamento procedencia, estrato, genero, grupo etario, estado civil, tipo colegio, programa, último periodo cursado, tipo estudiante (deserto o no deserto).

Figura 61. Deserción de acuerdo con características de población



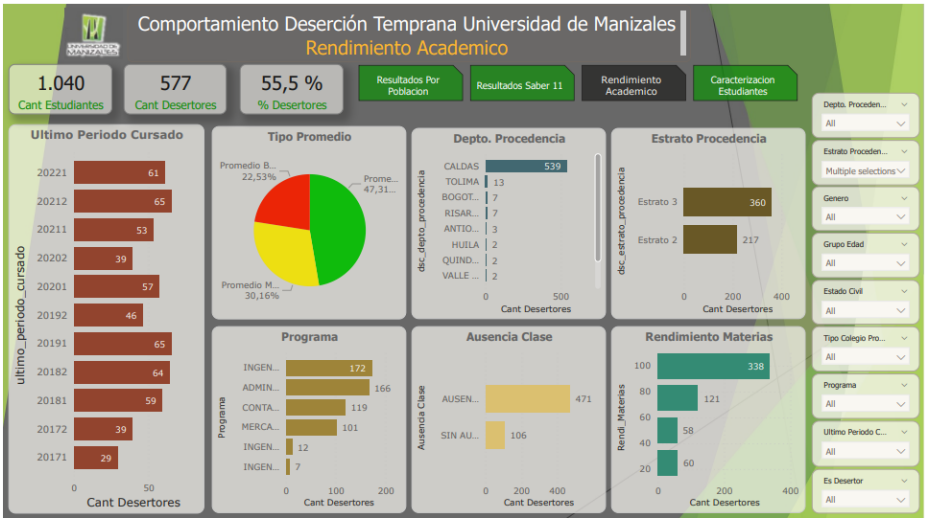
- La segunda hoja en Power Bi se denomina **Resultados Saber 11.** Allí se pueden ver los gráficos de cantidad de estudiantes por su rendimiento académico en las pruebas saber 11 previo al ingreso a la universidad clasificado por **nivel matemáticas, nivel lecto escritura, nivel inglés, rendimiento global, último periodo cursado** y grupo de edad, estos datos pueden filtrarse (ventana a la derecha del cuadro resultados) por departamento procedencia, estrato, genero, grupo etario, estado civil, tipo colegio, programa, último periodo cursado, tipo estudiante (deserto o no deserto).

Figura 62. Deserción resultados de la Prueba Saber 11



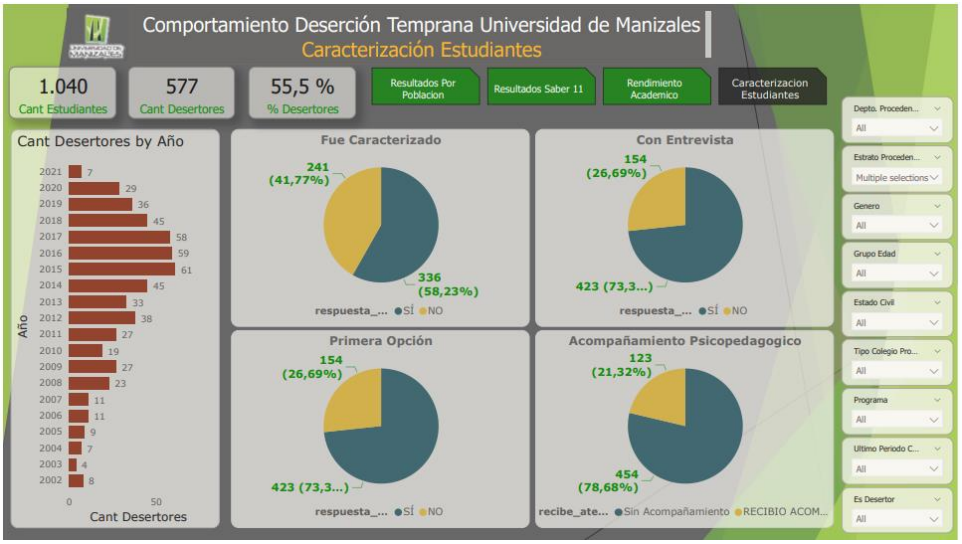
- La tercera hoja en Power Bi se denomina **Rendimiento académico.** Allí se pueden ver los gráficos de cantidad de estudiantes por su rendimiento académico en durante la permanencia en la universidad previo al ingreso a la universidad clasificado por **último periodo cursado, tipo de promedio obtenido en el rendimiento académico, departamento de procedencia, estrato de procedencia, programa académico, ausencia clase, rendimiento en materia cursadas**, estos datos pueden filtrarse (ventana a la derecha del cuadro resultados) por departamento procedencia, estrato, genero, grupo etario, estado civil, tipo colegio, programa, último periodo cursado, tipo estudiante (deserto o no deserto).

Figura 63. Deserción de acuerdo con el rendimiento académico



- La cuarta hoja en Power Bi se denomina **Caracterización estudiantes** Allí se pueden ver los gráficos de cantidad de estudiantes por su estado de caracterización al ingreso previo a la universidad clasificado por **úSi fue o no caracterizado previo al ingreso, Si se le realizo entrevista previo al ingreso, si el programa académico seleccionado fue su primera opción, si durante la estancia en la universidad recibió acompañamiento psicopedagógico institucional**, estos datos pueden filtrarse (ventana a la derecha del cuadro resultados) por departamento procedencia, estrato, genero, grupo etario, estado civil, tipo colegio, programa, último periodo cursado, tipo estudiante (deserto o no deserto).

Figura 64. Deserción de acuerdo con la caracterización del estudiante



- La quinta hoja en Power Bi se denomina **Detalle.** Este cuadro en particular genera el detalle estudiante por estudiantes de quienes cumplen el filtro seleccionado. Esto permite a la institución conocer al detalle de quienen son los posibles estudiantes que desertaran, permitiendo a esta actuar con anticipación a que suceda el evento.

Procedimiento para descargar la carpeta que contiene la visualización

En la carpeta compartida del proyecto en OneDrive **PRY_GRADO_MAESTRIA_OSCARMARIOAGUDELO_JURADOS,** subcarpeta Desercion Power BI Visualizacion Resultados, subcarpeta para ver los resultados de la predicción Desercion Power BI Visualizacion Prediccion. Para que el usuario pueda consultar y navegar en cada uno de los cuadros de visualización de datos, es necesario descargarla a un equipo local, luego ingresar a la carpeta local y ejecutar el archivo powerbi **Dashboard Universidad de Manizales.pbix,** previo a esto verificar que en la maquina local tenga instalado la versión desktop de powerBi.

Figura 65. Detalle de estudiantes desertores

Como resultado de la predición bajo los algoritmos de DecisionTree y Random forest se generaron los resulaods de los mismo en un archivo tipo CSV, los cuales pueden ser leidos desde POWER BI o desde Excel y se construyeron las siguiente figuras resultantes de la prediccion sobre 404 elementos de la data fuente, arrojando el siguiente resultado:

- Desertores: 228
- NO desertores: 176

A simple vista puede suponerse que el **56.43%** del total de estudiantes que se predijo abandonarian el programa, sin embargo un elemento relevante muestra una tendencia alta para aquellos estudiantes que se predijeron como Desertores, lo cumplen con solo un periodo cursado. A lo cual se profundiza sobre el hecho y se valida con el departamento de Tecnologias de la Informacion de la Universidad de Manizales y se concluye que la fecha de corte para la data fuente ocurrio en los primeros dias de diciembre, y a esta fecha aun a se tenia pendiente por matricular a un alto porcentaje de estudiantes de la Universidad. Para corroborar lo anterior se cruzaron las predicciones de desercion de los desertores de primer periodo cursado en el mes de febrero del año siguiente y se concluyo

que de la lista de los 112 desertores 101 de ellos efectivamente estaban matriculados para iniciar un segundo periodo.

Figura 76. Análisis de resultados de la predicción deserción temprana

	A	B	C	D	E	F
1	**Tabla Dinamica de Consulta Resultados de la Prediccion Desercion Temprana**					
2						
3	EsDesertor	Si				
4	ultimo_periodo_cursado	(Todas)				
5	total_materias_perdidas_cuanti	(Todas)				
6	fec_finaliza_secundaria	(Todas)				
7	dsc_tipo_colegio_procedencia	(Todas)				
8	dsc_depto_procedencia	(Todas)				
9	dsc_estrato_procedencia	(Todas)				
10	total_fallas_cuanti	(Todas)				
11	recibe_atencion	(Todas)				
12						
13	**Total_Estudiantes**	**Etiquetas de columna**				
14	**Peridos Cursados**	**1.-Insuficiente**	**2.-Minimo**	**3.-Satisfactorio**	**4.-Avanzado**	**Total general**
15	1	4	102	3	3	112
16	2	3	52	2		57
17	3	1	30	1		32
18	4	3	21	3		27
19	**Total general**	**11**	**205**	**9**	**3**	**228**

Para el conjunto de resultados producto de la predicción del modelo de machine learning se implementó bajo power bi, el mismo número de hojas de visualización.

Figura 77. Predicción deserción de acuerdo con características de población

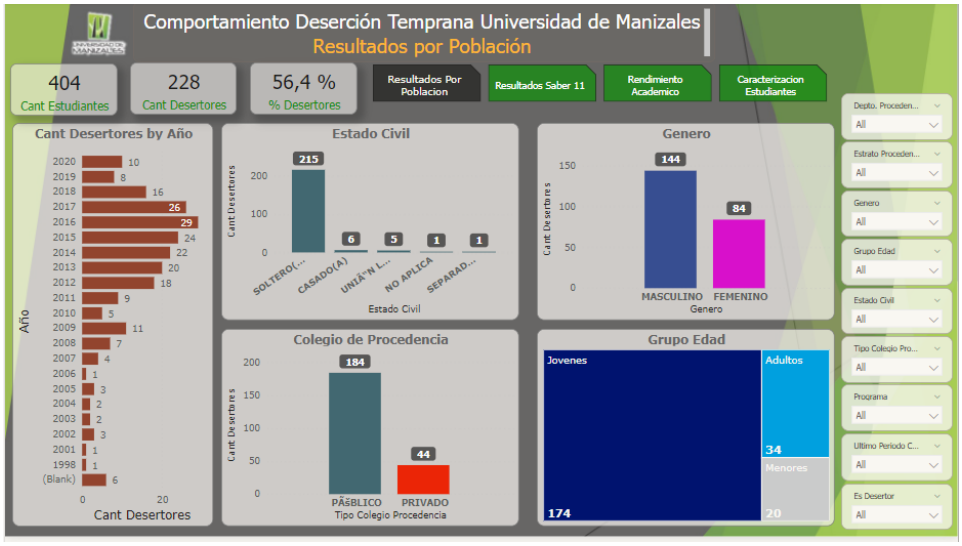


Figura 78. Predicción deserción resultados de la Prueba Saber 11

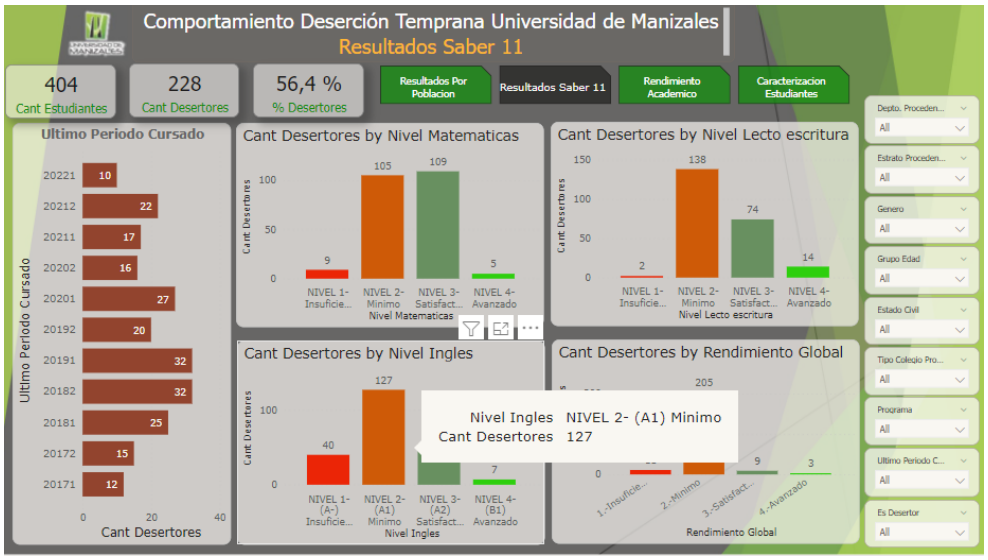


Figura 79. Predicción deserción de acuerdo con el rendimiento académico

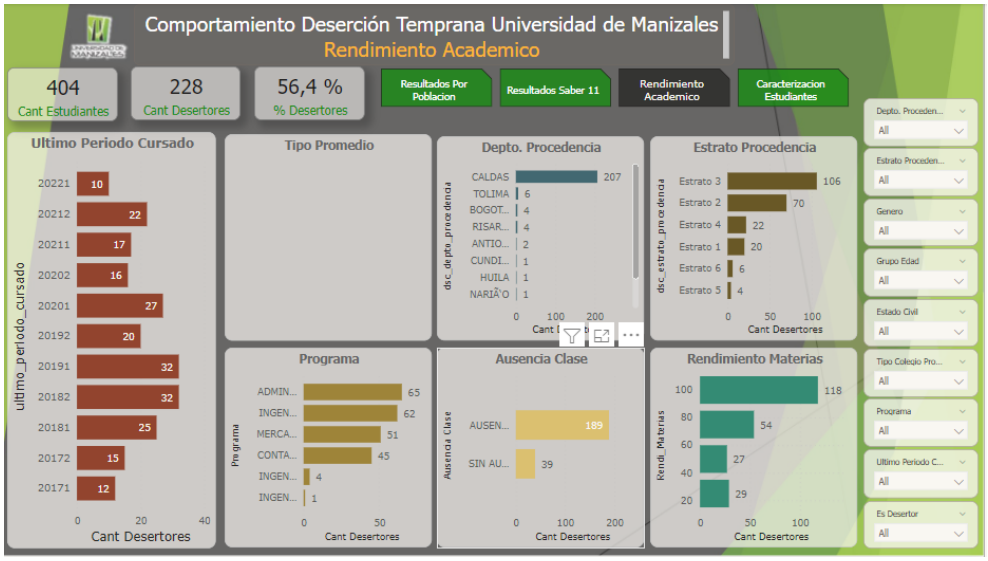


Figura 80. Predicción deserción de acuerdo con la caracterización del estudiante

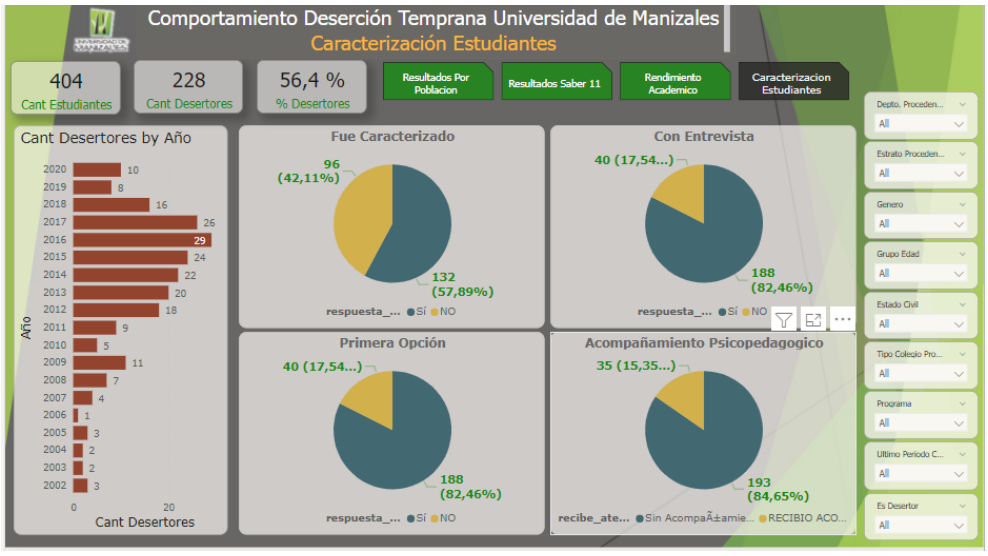


Figura 81. Predicción detalle de estudiantes desertores

10 Impactos

10.1 Impactos Sociales

Identificar de manera precisa y anticipada las circunstancias que detonan la deserción temprana universitaria en los programas de pregrado modalidad presencial en las facultades de estudio permite que los estamentos intervinientes (estado, universidad, estudiante y familia) puedan articular acciones proactivas preventivas para mitigar el riesgo de deserción escolar. Gracias a esta identificación oportuna, pueden diseñarse y aplicarse estrategias específicas para cada tipo de caso, así como invertir recursos económicos y humanos en actividades que efectivamente impacten en el estudiante, tales como: asistencia económica, alimentaria y nutricional, pedagógica, psicológica, en salud mental, orientación profesional, etc.).

De acuerdo Guadalupe Izquierdo y Reina Mestanza (2017),

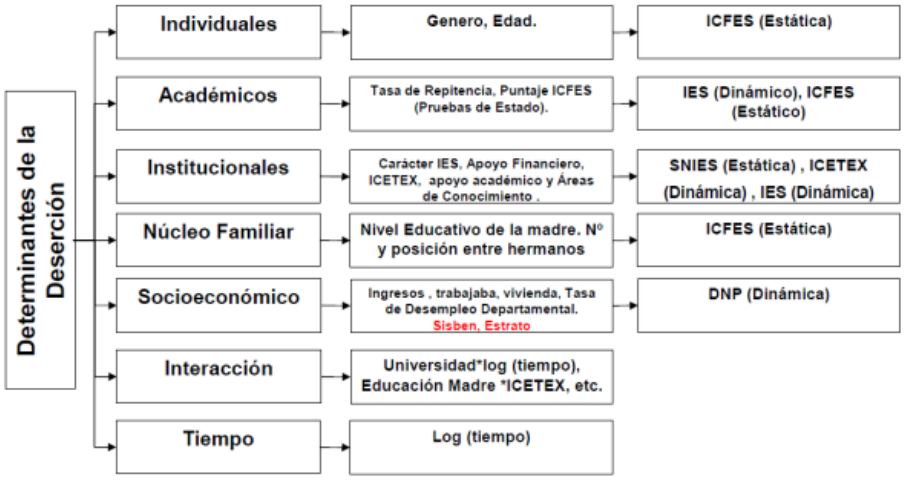
> Es innegable que siempre tendrá mayor riesgo de desertar un estudiante que llega con diferentes carencias (económicas, académicas o intelectuales) a un mundo universitario que no tiene los medios de acompañamiento necesarios para llevarlo de la mano a la culminación de una carrera profesional por lo que se debería tomar en cuenta los diferentes factores predisponentes y poner más atención a las necesidades de los universitarios. (págs. 18-19)

Por su parte, Salim Chalela, Alejandro Valencia, Gustavo Ruiz y Marcela Cadavid (2020), describen:

El interés por contener el problema (deserción) y buscar alternativas de solución obedece a las variadas causas que llevan a la presencia de este fenómeno en los sistemas educativos. La literatura existente devela detonantes asociados a situaciones financieras complejas, inadecuada orientación profesional y vocacional, patrones propios de los sistemas culturales de las sociedades en que están inmersos los estudiantes, dificultades de estos para rendir académicamente e incluso, elementos más estructurales de cada individuo como sus condiciones psicológicas y la calidad de las relaciones existentes dentro de la familia. Por tanto, la deserción estudiantil se ha constituido en un tema prioritario de la agenda global, pues afecta a todos los países independientemente de su nivel de crecimiento económico y desarrollo humano, impactando en promedio al 50% de la población estudiantil en su dimensión social, económica e individual (Isaza, Enríquez y Pérez-Olmos, 2016). Adicionalmente, esto disminuye las posibilidades de inserción laboral y la remuneración de quienes no lograron culminar la educación superior (Solé-Moro, Sánchez-Torres, Arroyo-Cañada y Argila-Irurita,2018).

Así, en países como Colombia las tasas de deserción para el año 2015 se encontraron en el 46,1%, cifra muy similar a las de otros países de la región como México y Argentina -con un 43%-; Venezuela y Chile -con un 53%- y Costa Rica con un 62%. Este hecho devela que más del 50% de los estudiantes que ingresan al sistema de educación superior procrastinan en su intento por continuar con niveles de formación universitaria (Isaza, Enríquez y Pérez-Olmos, 2016). (p. 105)

Figura 66. Determinantes de la deserción universitaria



Nota: Tomado de ¿Cómo funciona el Spadies? (MEN, 2016)

10.2 Impactos económicos

Es indudable la relación existente entre los impactos sociales y económicos del evento de la deserción universitaria; en especial, la deserción temprana, puesto que del total de estudiantes que abandonan sus estudios superiores el mayor porcentaje al rededor del 75%, corresponde a deserción temprana. En el caso particular de la Universidad de Manizales, se estima que los ingresos que dejan de percibirse por abandono escolar temprano oscilan entre los 1.200 y 1.300 millones de pesos por semestre, en las facultades objeto de estudio, situación que impacta el flujo de caja de la Universidad y, por consiguiente, limita los recursos de inversión en proyectos institucionales.

Adicionalmente, respecto al impacto en el estudiante desertor y su familia, no se tienen cifras de los valores invertidos por el estudiante y su familia ni del estimado que pierden al no continuar con sus estudios, con su consabido impacto en la economía personal y familiar.

11 Conclusiones

Producto del proceso de investigación fue posible comprobar que al seguir de manera estricta los pasos asociados a la metodología dispuesta para procesos analíticos de información predictiva como CRISP-DM, y de la mano de un efectivo proceso de recolección, depuración y preparación de la información disponible en la Universidad de Manizales con influencia en la **variable objetivo** (deserción), puede implementarse un modelo predictivo basado en algoritmos de ***machine learning*** que deriva en la aplicación de estrategias articuladas a las características o perfiles de los estudiantes con riesgo de desertar tempranamente modalidad presencial en las facultades de Ciencias Contables, Económicas y Administrativas y Ciencias e Ingeniería y minimizar dicho riesgo.

El modelo implementado en Python utilizó dos tipos de algoritmos que ofrecen un buen desempeño para problemas de este tipo (binarios de dos salidas posibles, deserto o no deserto): los algoritmos de ***Random forest*** y **Decisión Tree**, los cuales arrojaron métricas de exactitud, sensibilidad y precisión superiores al 90%, lo cual permite inferir que de cada 100 estudiantes evaluados el algoritmo predice 93 correctos y solo 7 incorrectos.

De la información dispuesta por la Universidad, las variables que en su orden de importancia explican mejor el fenómeno de la deserción temprana fueron: **Último periodo escolar cursado, si el programa fue o no la primera opción elegida, cantidad de materias pérdidas durante su estancia en el programa, si previo al ingreso a la universidad fue entrevistado y caracterizado**, así como su **rendimiento académico** tanto durante su periplo en la universidad, como su **desempeño previo en las pruebas**

Saber 11, especialmente en las áreas que prevalecen en las facultades de estudio, es decir, matemáticas, lectura crítica e inglés.

Para brindar mayor información a los estamentos de la universidad involucrados en la reducción del riesgo de la deserción temprana, ligado a los algoritmos, se implementó un cuadro de control en **PowerBI** que guía en detalle **el quién, el cuándo y el porqué** del abandono de uno o más estudiantes. Detallar las particularidades del estudiante que se intervendrá permite llegar a él con más conocimiento de causa y con estrategias particulares personalizadas para atender su situación.

Producto de la evaluación descriptiva y predictiva durante la exploración de los datos y aplicación de los algoritmos se concluyó:

- El **61% (820 alumnos)** de toda la población incluida en la muestra (**1.344 estudiantes**) presentan un rendimiento en las pruebas saber 11 en el área de **matemáticas** entre satisfactorio y avanzado, el restante **39%** un rendimiento mínimo e insuficiente.
- El **67%** de toda la población muestra presenta un rendimiento en las pruebas saber 11 en el área de **Ingles** entre mínimo e insuficiente, el restante **33%** un rendimiento satisfactorio y avanzado.
- El **57%** de toda la población muestra presenta un rendimiento en las pruebas saber 11 en el área de **Lectura Critica** entre mínimo e insuficiente, el restante **43%** un rendimiento satisfactorio y avanzado.
- Dentro de la población desertora, destaca que su nivel global de las pruebas saber 11 se encuentra entre los 230 y 275 puntos de un total de 500 puntos. Mientras que la población NO desertora mayoritariamente se encuentra en el rango de 240 y 305 puntos rendimiento global pruebas saber 11.
- Variables importantes al momento de el estudiante realizar su proceso previo de ingreso a la Universidad de Manizales como lo son a)Fueron caracterizadas

al momento del ingreso SI/NO), b) Presento entrevista previo al ingreso (SI/NO), c) si el programa académico cursado fue la primera opción elegida (SI/NO), d) si recibió acompañamiento psicopedagógico durante la estadía en el programa académico (SI/NO) presenta un comportamiento entre la población que deserto entre el **44% y 48%** expresaron que nunca realizaron dicha prueba. Así mismo al tratar de acceder a datos actualizados o al menos relacionados con la población que aportaran a los modelos predictivos, NO fue posible obtenerlos debido a que ningún sistema los tenía almacenados y de aquellos que se conserva alguna información estaba dispuesta en planillas en Excel en el departamento de bienestar estudiantil. Por supuesto no están integrados a los sistemas de información asociados al ecosistema SIGUM.

- A continuación, se enumeran algunas preocupantes cifras obtenidas, como. a) el **50%** de los desertores presentan rendimientos en matemáticas satisfactorio, el **61%** de los desertores tienen rendimiento en lectura crítica mínimo, en ingles el **57.36%** rendimiento mínimo.

  Con respecto al rendimiento global tanto en pruebas saber 11 del **88.26%** como en rendimiento académico **88.70%** durante la estadía del estudiante en el programa su rendimiento fue de mínimo y satisfactorio.

- Del total de las predicciones **404** estudiantes, de ellos un total de **228** fueron marcados por el modelo **random forest** como desertores potenciales, sin embargo, profundizando en la exploración de la predicción se encuentra que **112** estudiantes corresponden a posibles abandonos durante el primer periodo cursado, pero que están condicionados por la fecha de corte al momento de obtener la muestra total que corresponde a la primera semana de diciembre, momento en el cual no todos los estudiantes están matriculados para el siguiente periodo académico. Para comprobar esta afirmación, se validaron los 112 estudiantes en el mes de febrero del año siguiente contra la lista de

matriculados a la fecha y esta lista se redujo en un **90.08%** (**101 alumnos**), lo indica que el porcentaje de abandono o deserción es del **28.71%**, lo cual es notorio frente a la media latinoamericana y nacional que supera el **35%**.

Lo importante de este proceso reafirma el concepto de que el problema educativo nacional es estructural y transversal al gobierno nacional, universidades y comunidad universitaria. Demuestra también que estrategias de ayuda sociales, económicas, familiares, psicopedagógicas apropiadamente articuladas a cada caso redunda en beneficios económicos por la continuidad del estudiante en un programa académico superior aumentando la productividad nacional si no también en lo social y familiar dado que redunda en la calidad de vida económica, emocional y familiar del entorno del estudiante que al final obtiene el título de pregrado.

Se puede deducir entonces que el algoritmo de predicción aplicado a tiempo activa la alarma para el accionar apropiado del departamento de Bienestar Estudiantil de la Universidad de Manizales aplicando las estrategias actuales y las nuevas que se deriven del perfilamiento obtenido de dichos modelos predictivos.

12 Recomendaciones

Como resultado del proceso investigativo y con el ánimo de obtener mejores resultados durante la ejecución de los modelos predictivos ***random forest*** *y* ***decisión tree***, se recomienda a la Universidad de Manizales articular los diferentes procesos de bienestar universitario y los sistemas de información para automatizar el registro de variables complementarias categóricas de las que aún no se tiene registro, las cuales complementan las causas que ocasionan la deserción universitaria temprana en pregrado modalidad presencial en las facultades de estudio. Así mismo, estas variables aportan al direccionamiento especifico de estrategias de retención del estudiante antes que el evento ocurra.

Complementando lo anterior y producto del análisis a los resultados obtenidos de la investigación se sugiere revisar, ajustar e implementar políticas de mejora del rendimiento académico de los estudiantes al inicio de su periplo por la universidad en las áreas de matemáticas, inglés y lectoescritura. Es notorio que los resultados no son favorables como lo denotan resultados de las pruebas saber 11.

Así mismo desplegar el modelo de predicción articulado con las estrategias de reducción del índice de deserción y paulatinamente integrarle las nuevas variables que se vayan recolectando.

A continuación, se enumeran un conjunto de variables que se deben involucrar en los procesos de automatización de información al interior de la Universidad de Manizales que posteriormente se convertirán en insumo para los algoritmos de predicción de la

deserción temprana en los programas de pregrado, las cuales permitirán un mayor grado de precisión y exactitud a las predicciones y por consiguiente ser más asertivos al momento de implementar y aplicar estrategias de retención de estudiantes antes de que el evento ocurra:

Tabla 17. Inventario variables pendientes por registrar con efectos en la deserción

Nro	Lista de Variables porrecopilar de los estudiantes en la Universidad de manizales para la prediccion de desercion universitaria temprana en los programas de pregrado modelidad presencial
1	Numero de Hermanos del estudiante
2	Nivel educativo de la madre (Primaria, Secundaria, Tecnico o Tecnologico, Pregrado, Postgrado)
3	Madre cabeza de hogar (SI/NO)
4	La Madre Trabaja (SI/NO)
5	Tipo de trabajo de la madre (Formal / Informal)
6	Nivel educativo del padre (Primaria, Secundaria, Tecnico o Tecnologico, Pregrado, Postgrado)
7	Padre cabeza de hogar (SI/NO)
8	El Padre Trabaja (SI/NO)
9	Tipo de trabajo del padre (Formal / Informal)
10	El Estudiante trabaja durante el curso del programa (SI/NO)
11	Nivel de ingresos familiares del estudiante (Bajo bajo, bajo, medio, medio alto, alto)
12	Tipo de Vivienda (propia o en arrendo)
13	Procedencia (Urbana o Rural)
14	Barrio de Origen
15	Nivel del SISBEN
16	Numero de personas que conformar el grupo familiar
17	Posicion del estudiante dentro del grupo de hermanos (primero, segundo, tercero, ultimo, unico hijo, etc).
18	El estudiante recibe apoyo economico (SI/NO)

19	El estudiante recibe apoyo psicopedagogico (SI/NO)
20	El estudiante recibe apoyo Academico (SI/NO)
21	El estudiante recibe orientacion Vocacional (SI/NO)
22	El estudiante recibe apoyo economico familiar (SI/NO)
23	Tasa de repitencia del estudiante (SPADIES)
24	Por cada periodo cursado clasificarlo si estuvo o esta en mora financiera o por pago (SI/NO)
25	Cual fue la via de acceso a la Universidad de Manizales (Matricula tradicional, becado, pilo paga, etc).
26	Dependencia economica de un tercero (SI/NO)
27	El estudiante tiene personas a cargo (Numero de personas).
28	El estudiante realizo estudios preuniversitarios (SI/NO)
29	Modalidad del bachillerato (Presencial, semipresencial, a distancia)
30	El estudiante tiene buenos habitos de estudio (SI/NO)
31	consumo de sustancias sicoactivas
32	como distribuye el tiempo
33	migrante o no
34	Numero de hijos
35	tienen pareja ?
36	tiene alguna incapacidad o limitacion fisica / mental (Ver encuesta de caracterizacion)
37	El progrma lo curso en que jornada (Diurna / Nocturna).
38	Estudiante desplazado por algun tipo de violencia
39	Estudiante esta clasificado como RAI (Rendimiento academico insuficiente)

Los resultados de esta investigación se dirigen, no solo a la Universidad de Manizales, sino a todo el ecosistema educativo, con el fin de unificar esfuerzos que redunden en el aprovechamiento de los recursos para disminuir el efecto del abandono escolar universitario, donde una tarea primordial es el mejoramiento de la calidad de vida del entorno familiar del estudiante.

A. Anexo 1: Glosario

Algoritmo: En Ciencias de la Computación, un algoritmo es una secuencia lógica, finita y con instrucciones que forman una fórmula matemática o estadística para realizar el análisis de datos.

Análisis exploratorio de datos (EDA): En esencia es una herramienta para explorar un conjunto de datos. Su objetivo es efectuar análisis de los mismos, el EDA puede emplearse para la visualización o representación gráfica de los datos, limpieza y transformación de los datos, paso importante en cualquier análisis de datos.

Análisis Predictivo (AP): El análisis predictivo pertenece al área de la Analítica Empresarial y trata de utilizar los datos para determinar qué puede pasar en el futuro. La AP permite determinar la probabilidad asociada a eventos futuros a partir del análisis de la información disponible (presente y pasada). También permite descubrir relaciones y patrones útiles entre los datos que normalmente no son detectados con análisis menos sofisticados. Técnicas como la minería de datos (data mining) y los modelos predictivos son utilizados.

Analytics: Es la forma de capturar informaciones, procesarlas y analizarlas para que se conviertan en insights.

Apoyos Académicos: Es una de las variables que aparece en el módulo de consulta de la base de datos del SPADIES. Se considera apoyo académico una tutoría, monitoria o nivelación que recibe el estudiante por parte de la IES. Los apoyos académicos se brindan en muchas ocasiones para evitar la deserción académica.

Ausencia Intersemestral: Proporción de estudiantes que estando matriculados en un semestre (t) son clasificados como ausentes en un período (t+1).

Big Data: Big Data es la expresión utilizada para designar un conjunto de datos tan grande que es difícil trabajar con los medios habituales (bases de datos). Se suele decir que el Big Data responde a las tres V: volumen, variedad y velocidad.

Clasificación Examen de Estado: Teniendo en cuenta que se cuenta con información de los puntajes del Examen de Estado desde el 1998-1 hasta hoy, se estandarizó la información, así: se tomó al estudiante con el máximo puntaje y se igualó a 100 y el estudiante con menor puntaje se igualó a 1, se hizo una regla de tres simple y todos los puntajes de los estudiantes quedaron entre 1 y 100. Se consideró como Alto un puntaje por encima de 61 puntos, y como Bajo un puntaje por debajo de 61 puntos.

Cohorte: Semestre en el cual el estudiante fue registrado como primíparo.

Código SNIES del Programa: Existen dos códigos SNIES: el consecutivo y el código de veintiún dígitos. Cualquiera de los dos es válido para el SPADIES. Al diligenciar la información en la base de datos, debe escribirse adelante una comilla para que el sistema considere la información ingresada como tipo texto y no como un número que debe aproximar.

Deep Learning: Lleva a cabo el proceso de *Machine learning* usando una red neuronal artificial que se compone de un número de niveles jerárquicos. En el nivel inicial de la jerarquía, la red aprende algo simple y luego envía esta información al siguiente nivel. El siguiente nivel toma esta información sencilla, la combina, compone una información un poco más compleja, y lo pasa al tercer nivel, y así sucesivamente.

Deserción: Estado de un estudiante que de manera voluntaria o forzosa no registra matrícula por dos o más períodos académicos consecutivos del programa en el que se matriculó; y no se encuentra como graduado o retirado por motivos disciplinarios.

La deserción es el resultado del efecto de diferentes factores como individuales, académicos, institucionales, y socioeconómicos.

Desertor Programa: Estudiante que no se matricula en el mismo programa académico durante dos o más períodos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios.

Desertor de la Institución de Educación Superior: Estudiante que no se matricula en una Institución de Educación Superior durante dos o más períodos académicos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios.

Desertor Sistema: Estudiante que no se matricula en ningún programa académico de ninguna Institución de Educación Superior durante dos o más períodos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios.

Deserción Anual: Porcentaje de estudiantes desertores identificados en t+2 que estuvieron matriculados en el periodo t.

Deserción por Cohorte: Porcentaje acumulado de estudiantes de una cohorte que no ha registrado matricula por dos o más períodos consecutivos en un programa académico de una Institución de Educación Superior hasta un semestre determinado. Es decir, el número acumulado de desertores de una cohorte hasta un semestre determinado, sobre los primíparos de esa cohorte.

Deserción Periodo SPADIES 2.8: Proporción de estudiantes que estando matriculados dos semestres atrás son clasificados como desertores un año después. Matemáticamente se toma el número de desertores en t sobre los matriculados no graduados en t-2.

Deserción Promedio Acumulada: Proporción de estudiantes de todas las cohortes que no ha registrado matrícula por dos o más períodos académicos consecutivos en un programa de una Institución de Educación Superior hasta un semestre determinado. Es decir, el conteo acumulado de desertores hasta un semestre determinado de todas

las cohortes que tienen hasta ese semestre, sobre la totalidad de primíparos de dichas cohortes.

Enfoque de Historia de Vida: Forma de abordar un problema que se centra en el análisis de una secuencia de eventos y transiciones para cada individuo. En el caso del SPADIES, este enfoque implica el seguimiento de cada uno de los estudiantes a lo largo de su permanencia en la Institución de Educación Superior.

Entropía de la información: En el ámbito de la teoría de la información. la entropía, también llamada entropía de la información y entropía de Shannon (en honor a Claude E. Shannon), mide la incertidumbre de una fuente de información. ... El concepto entropía es usado en termodinámica, mecánica estadística y teoría de la información.

Exactitud o accuracy: Fracción de predicciones que el modelo realiza correctamente. Se representa como un porcentaje o un valor entre 0 y 1. Es una métrica adecuada cuando el conjunto de datos es balanceado, esto es, cuando el número de etiquetas de cada clase es similar.

F1 score: Combina las métricas Precision y Recall para dar un único resultado. Esta métrica es la más apropiada cuando tenemos conjuntos de datos no balanceados. Se calcula como la media armónica de Precision y Recal. La fórmula es F1 = (2 * precision * recall) / (precision + recall). Se usa media armónica y no simple. porque la media armónica hace que si una de las dos medidas es pequeña (aunque la otra sea máxima), el valor de F1 score sea pequeño.

Graduado: Estudiante que ha recibido el grado por parte de la Institución de Educación Superior como muestra de la culminación de su ciclo académico. Un estudiante que termina materias, pero que no ha obtenido el título es un egresado no graduado y puede ser catalogado como desertor de acuerdo con el criterio de deserción.

Machine learning: Método analítico que permite que un sistema, por sí mismo —sin intervención humana y en forma automatizada—, aprenda a descubrir patrones, tendencias y relaciones entre los datos, y gracias a dicho conocimiento, en cada interacción con información nueva, ofrece mejores perspectivas. Esta competencia inherente para aprender de los datos, sitúa a *Machine learning* como una expresión de la Inteligencia Artificial.

Matriculados SPADIES: Dado el enfoque de historia de vida de los estudiantes, los matriculados en SPADIES son los estudiantes que se pueden rastrear como primíparos. Es decir, en el primer período de reporte de una Institución de Educación Superior los matriculados y los primíparos serán los mismos.

Metodología CRISP-DM: Método empleado para el desarrollo de proyectos de minería de datos. El estándar incluye un modelo y una guía, estructurados en **seis fases**, algunas de estas fases son bidireccionales, lo que significa que algunas de ellas permiten revisar parcial o totalmente las fases anteriores.

Las fases son: **Comprensión del negocio** (Objetivos y requerimientos desde una perspectiva no técnica), **Comprensión de los datos** (Familiarizarse con los datos teniendo presente los objetivos del negocio), **Preparación de los datos** (Obtener la vista minable o dataset), **Modelado** (Aplicar las técnicas de minería de datos a los dataset), **Evaluación** (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio) y **Despliegue** (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización).

Período: Semestre de matrícula del estudiante.

Período de Ingreso: Cohorte a la cual pertenece el estudiante.

Precisión: Métrica determinada por la fracción de elementos clasificados correctamente como positivo entre todos los que el modelo ha clasificado como positivos. La fórmula es VP / (VP + FP). El modelo de ejemplo tendría una precisión de 1 / (1 + 1) = 0.5. En

el caso del modelo que siempre predice la etiqueta positiva, la precisión del modelo es 1 / (1 + 9) = 0.1; este modelo tiene una sensibilidad máxima, pero una precisión muy pobre. Razón por la que se necesitan las dos métricas para evaluar la calidad real del modelo.

Primíparo: Estudiante registrado por primera vez en un programa académico de una Institución de Educación Superior.

RAI: Rendimiento Académico Insuficiente. Cuando un estudiante, “en un mismo periodo académico pierde más del cincuenta por ciento (50%) de las asignaturas cursadas y validadas o reprueba dos (2) de ellas por segunda vez o una (1) por tercera vez o más” (Reglamento Estudiantil, artículo 99) (U. Manizales, 2012, p. 21).

Recall o sensibilidad: Indica la proporción de ejemplos positivos que están identificados correctamente por el modelo entre todos los positivos reales. Es decir, VP / (VP + FN).

Retirado: Estudiante que se ausenta de un programa académico de una Institución de Educación Superior en un período por motivos disciplinarios (no académicos). Incurren en faltas disciplinarias graves que implican su expulsión.

Tasa de Graduación Acumulada: Proporción de estudiantes de todas las cohortes que se ha graduado de un programa académico de una Institución de Educación Superior hasta un semestre determinado. Es decir, el conteo acumulado de graduados hasta un semestre determinado de todas las cohortes que tienen hasta ese semestre, sobre el total de primíparos de dichas cohortes.

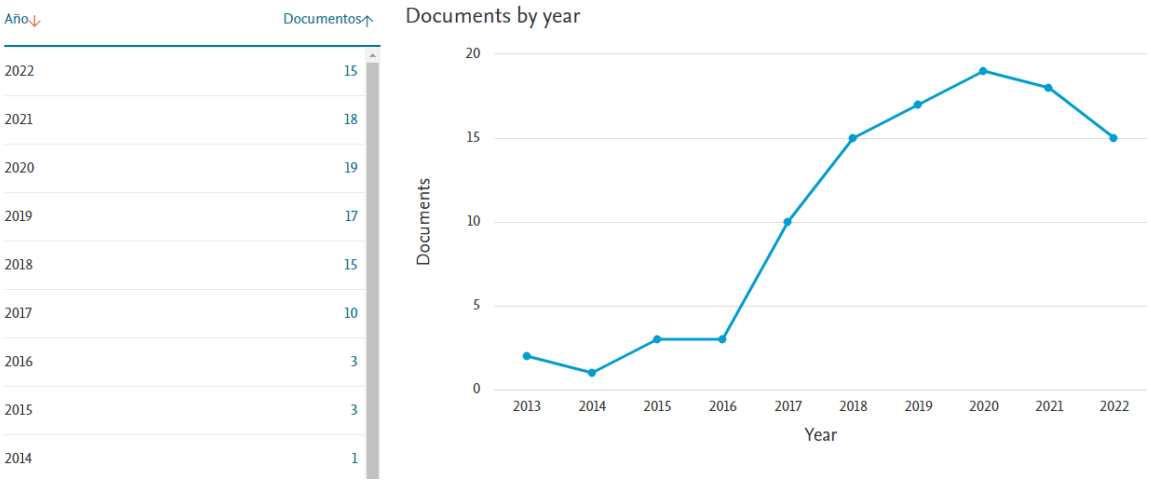
Tasa de Supervivencia: Proporción de estudiantes en cada semestre que permanecen matriculados luego de haber sido primíparos de un programa académico de una Institución de Educación Superior.

B. Anexo 2: Análisis bibliométrico

Las siguientes gráficas corresponden con el análisis bibliométrico hecho a partir de la búsqueda en la base de datos Scopus y las herramientas de análisis tanto de Scopus como de VOSViwer.

La búsqueda se hizo a partir de la categoría **predicción deserción universitaria**, en todos los documentos de los últimos 10 años (desde 2013). Se encontraron 103 documentos que correspondieron con los intereses de esta investigación. Como podrá verse en los gráficos, en los últimos 5 años se concentra alrededor del 80% de esta producción que tuvo su mayor pico en 2020.

Figura 67. Producción de documentos por año



Año	Documentos
2022	15
2021	18
2020	19
2019	17
2018	15
2017	10
2016	3
2015	3
2014	1

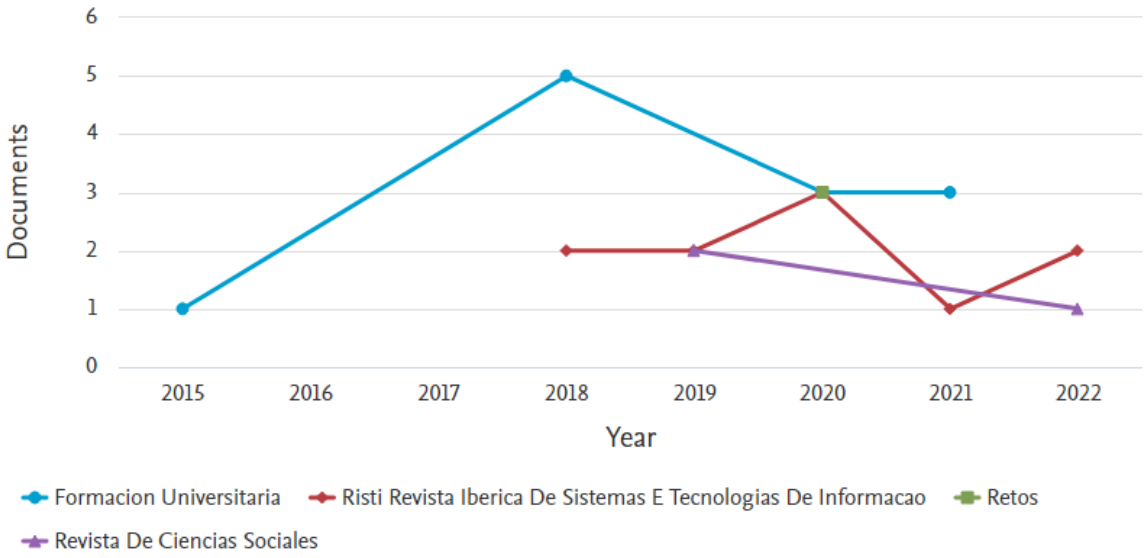


Figura 68. Documentos por año por recurso

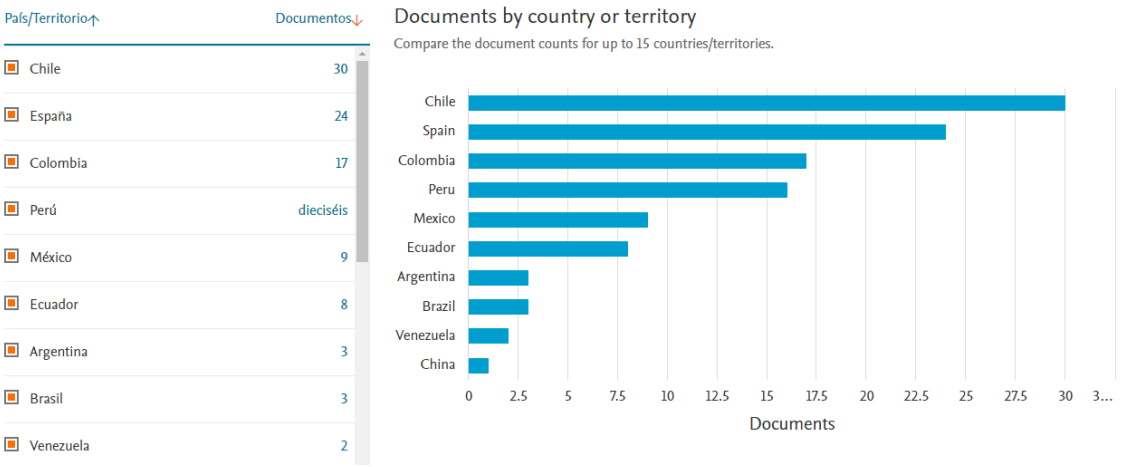


Figura 69. Documentos según país de origen

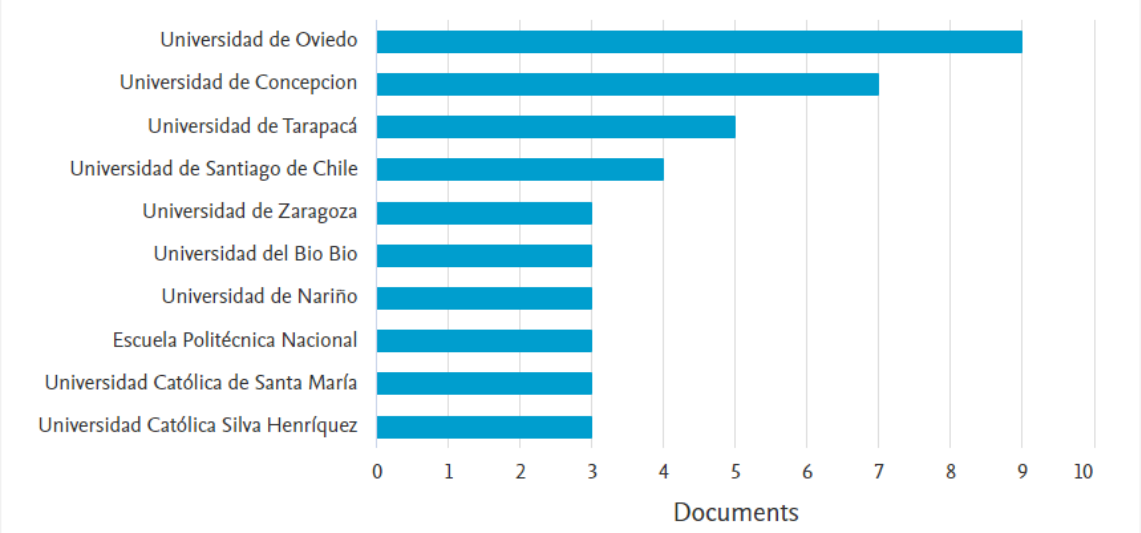


Figura 70. Institución de procedencia

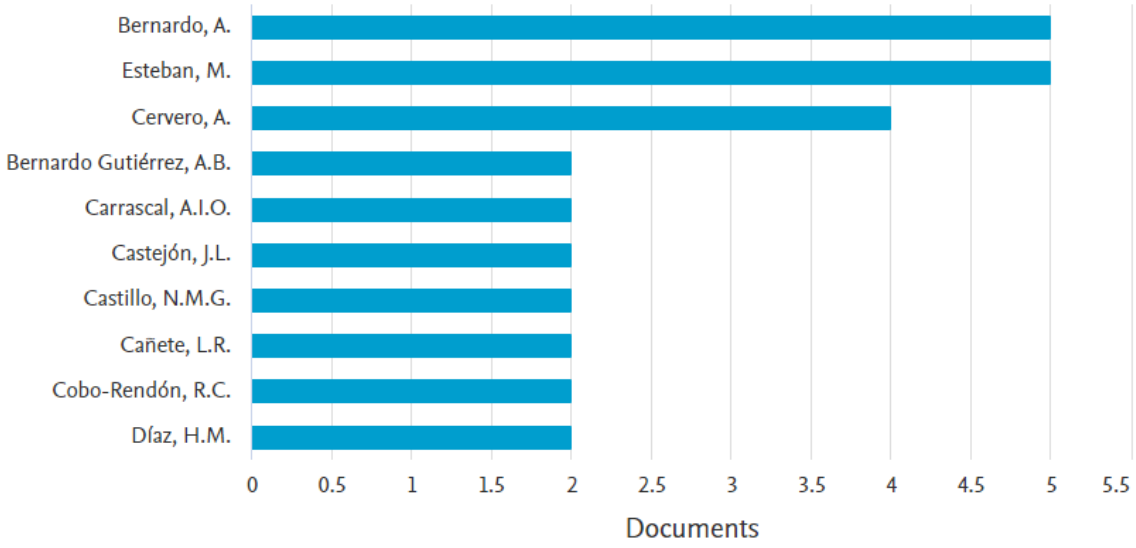


Figura 71. Autores con dos escritos o más

Tipo de Documento	Documentos
Artículo	77
Documento de sesión	21
Revisar	4
Capítulo del libro	1

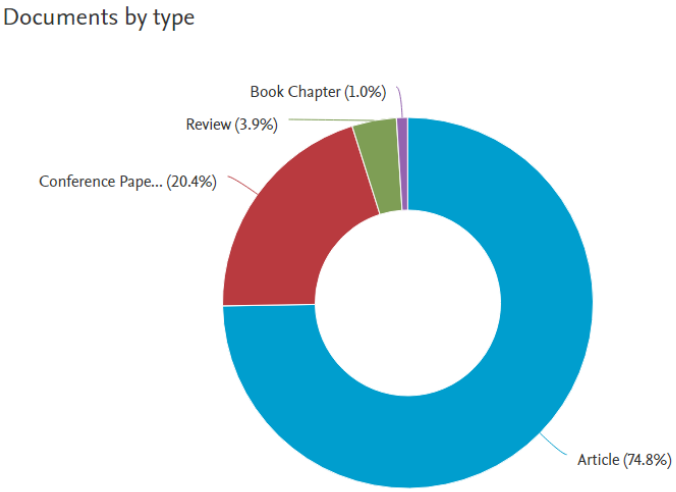
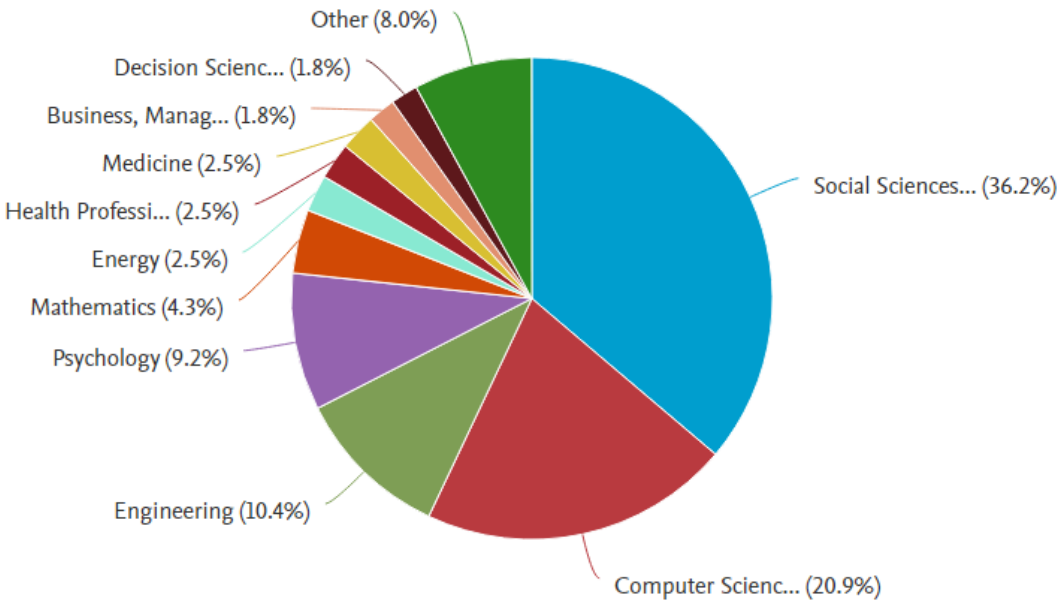


Figura 72. Tipos de documentos



12.1.1 Distribución de acuerdo con el área de estudio

A continuación, dos gráficos adicionales relacionados con las palabras clave de los documentos encontrados. Los colores indican la frecuencia de aparición de las palabras clave: el rojo corresponde con las más calientes; luego, el verde, el azul y el amarillo.

Figura 73. Palabras clave más frecuentes

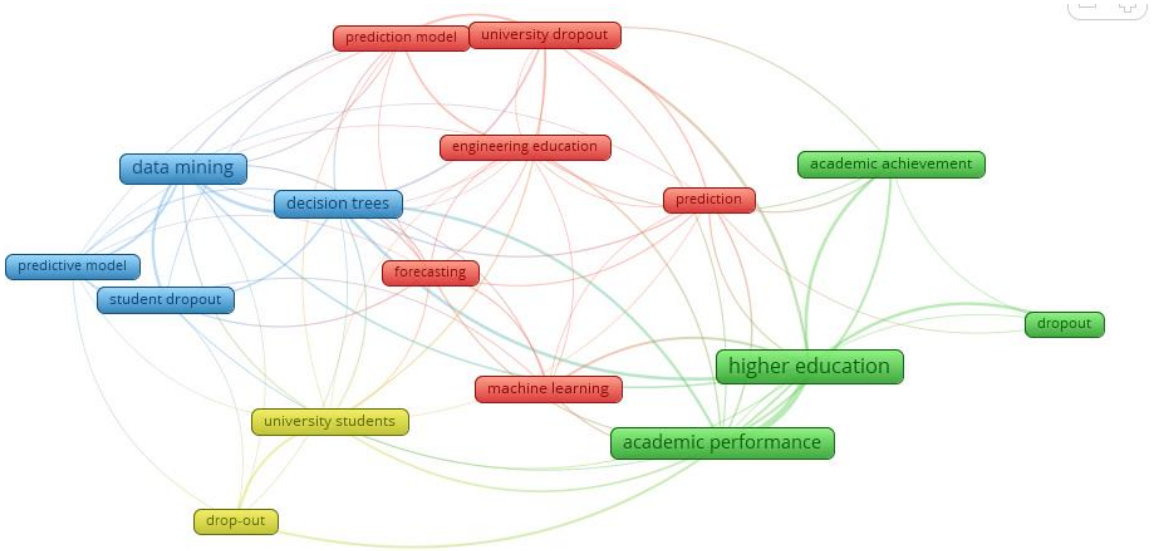
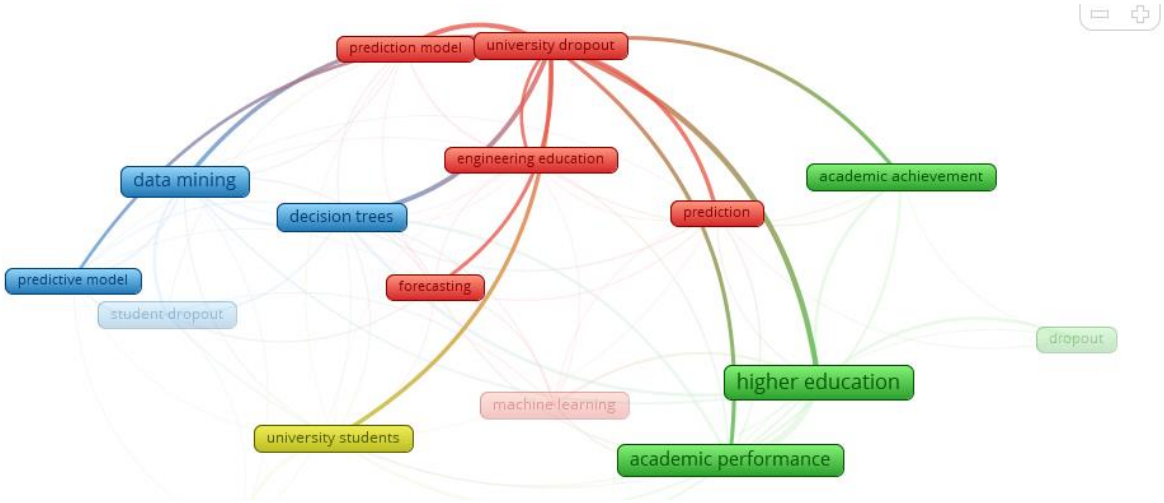


Figura 74. Relación entre palabras clave



C. Anexo 3: Pruebas Saber 11

A continuación, se presenta una reseña general de las pruebas de Estado, Saber 11. Esta información ha sido tomada de https://www.icfes.gov.co/acerca-del-examen-saber-11%C2%B0

Saber 11° está compuesto por cinco pruebas: Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés.

área	preguntas
Matemáticas	50
Lectura Crítica	41
Sociales y Ciudadanas	50
Ciencias Naturales	58
Inglés	55

. Los niveles de desempeño son una descripción cualitativa de las habilidades y conocimientos que se estima ha desarrollado el evaluado, y tienen el objetivo de complementar el puntaje numérico obtenido. De igual manera, permiten agrupar a los estudiantes en 4 niveles (1, 2, 3 y 4).

• El puntaje de cada prueba va de 0 a 100 puntos sin decimales.

• Los niveles de desempeño definidos son: Insuficiente, Mínimo, Satisfactorio y Avanzado; para todas las áreas, excepto inglés.

• Los niveles de desempeño definidos para inglés son de menor a mayor: A-, A1, A2, B1, B+.

• Los niveles de desempeño definen lo que sabe y sabe hacer el estudiante.

• El puntaje global va de 0 a 500 puntos sin decimales, la media teórica son 250 puntos, lo que significa que por encima de 250 empieza a ser un puntaje positivo, pero si deseas estar entre los mejores y obtener una beca, tu puntaje debe superar los 360 puntos.

. **Historia**: Inicialmente, el examen utilizaba una escala de puntuación que oscilaba entre 100 y 400 puntos y se basaba en gran medida en la memoria y los conceptos. Las materias evaluadas incluían biología, química, física, estudios sociales, dominio del español, matemáticas y una materia optativa que podía ser un idioma extranjero o una prueba de razonamiento lógico.

1-INSUFICIENTE **2**-MÍNIMO **3**-SATISFACTORIO **4**-AVANZADO

LECTURA CRÍTICA	
Nivel	*puntaje*
1	0 a 35 puntos
2	36 a 50 puntos
3	51 a 65 puntos
4	66 A 100 puntos

CIENCIAS NATURALES	
Nivel	*puntaje*
1	0 a 40 puntos
2	41 a 55 puntos
3	56 a 70 puntos
4	71 a 100 puntos

MATEMATICAS	
Nivel	*puntaje*
1	0 a 35 puntos
2	36 a 50 puntos
3	51 a 70 puntos
4	71 a 100 puntos

SOCIALES Y CIUDADANAS	
Nivel	*puntaje*
1	0 a 40 puntos
2	41 a 55 puntos
3	56 a 70 puntos
4	71 a 100 puntos

INGLÉS	
Nivel	*puntaje*
A-	0 a 47 puntos
A1	48 a 57 puntos
A2	58 a 67 puntos
B1	68 a 78 puntos
B+	79 a 100 puntos

- Fórmula del puntaje final:

Véase el documento circular que define el cálculo. R (Resultado convertido) P (Puntaje o promedio de las pruebas estudiante)

Desde	*hasta*	*puntos*	*fórmula*
	1999-II	0 - 400 puntos	R = 5(P - 100) / 3
2000-I	2011-II	0 - 100 puntos por cada área	R = 5 Pr (Round a 0 decimales entero Promedio de la sumatoria de todas la áreas)
2012-I	2014-I	0 - 500 puntos	R = P
2014-II	en adelante	0 - 500 puntos	R = P

Reducción para aplicación

Desde	*hasta*	*Aplicación de la reducción*
	1999	R = 5 (P - 100) / 3
2000	2011	R = 5 PR (Round a 0 decimales entero Promedio de la sumatoria de todas la áreas)
2012	En adelante	R = P

Catalogación de los resultados del ICFES sobre la base de 500 puntos

Desde	*hasta*	*nivel*
0	<= 250 puntos	'BAJO'
> 250	<= 350 puntos	MEDIO'
> 350	<= 425 puntos	'SUPERIOR'
> 425		'MUY SUPERIOR'

Para ampliar la información, véanse el documento ubicado en el enlace:

Niveles_Desempeño_Pruebas_Saber11_Homologacion_Puntajes_al_2023.pdf

También puede consultar:

- Modelo de conversión de pruebas ICFES a puntaje de 00 a 500 puntos vigentes, actualmente en la Universidad del Magdalena, Acuerdo Académico No. 1º de 2019: https://www.unimagdalena.edu.co/UnidadesOrganizativas/Dependencia/2002
- Clasificaciones en las pruebas del ICFES: https://www.youtube.com/watch?v=4M3q0abhDwA

D. Anexo 4: Código en PL/SQL exploración de datos, código en Python modelos predictivos, código cuadros de visualización en PowerBi

A continuación, se presenta la información correspondiente a los códigos fuente de desarrollo de cada uno de los elementos que permiten predecir la deserción temprana en la universidad de Manizales.

- Código fuente desarrollado en PL/SQL para la extracción, exploración y transformación de datos:
  Tesis_Maestria_Exploracion_Datos_SQL_Validaciones_Final.sql
  Tesis_Maestria_Estructura_Vista_Unifica_data_Desercion_Completa.sql
- Código fuente desarrollado en Python para el análisis exploratorio de datos (EDA) y la implementación de los algoritmos de deserción:
  Desercion_Temprana_Universidad_Manizales.ipynb
- Código fuente desarrollado en PowerBI para la visualización de los datos producto de la predicción:
  Desercion_Power_BI_Visualizacion_Resultados
  Desercion_Power_BI_Visualizacion_Prediccion
- Cuadro en Power BI publicado en la WEB

https://app.powerbi.com/view?r=eyJrIjoiN2ZkMDg5ZjctYTZjMC00ZjZiLWFjYWUtMGZmYTZjMzE0YmQ0IiwidCI6IjRmMWUwNDRkLTNkNzAtNDk5MC1iMjZhLWI5NWYwYzY0MmUxYSIsImMiOjR9

E. Anexo 5: Variables explicativas de la deserción educación superior

Variables para el Analisis de la desercion temprana estudiantil en la Universidad de Manizales		
Muestra: Estudiantes matriculados Facultas de Ciencias Contables Economicas y Administrativas		
Nro	**Descripcion de la Variable - Periodos requeridos 2017, 2018, 2020, seguimientos hasta x 8 semestres**	
	Factores por los que se agrupan las variables: Academicos, Personales, Economicos, Institucionales	**Estado**
1	Documento de identificacion	Disponible
2	Fecha de Nacimiento	Disponible
3	Fecha de Ingreso a la Universidad de Manizales	Disponible
4	Edad al momento de presentar las pruebas de estado SABER-11	calculada
5	Edad en años al momento de ingresar a un programa de la Universidad de Manizales	Disponible
6	Periodo o Cohorte	Disponible
7	Codigo SNIES del Programa	Disponible
8	Nombre del Programa	Disponible
9	Duracion en semestres del programa	Disponible
10	Genero (Masculino, Femenino, Otro)	Disponible
11	Numero de Hermanos del estudiante	No Disponible
12	Nivel educativo de la madre (Primaria, Secundaria, Tecnico o Tecnologico, Pregrado, Postgrado)	No Disponible
13	Madre cabeza de hogar (SI/NO)	No Disponible
14	La Madre Trabaja (SI/NO)	No Disponible
15	Tipo de trabajo de la madre (Formal / Informal)	No Disponible
16	Nivel educativo del padre (Primaria, Secundaria, Tecnico o Tecnologico, Pregrado, Postgrado)	No Disponible
17	Padre cabeza de hogar (SI/NO)	No Disponible
18	El Padre Trabaja (SI/NO)	No Disponible
19	Tipo de trabajo del padre (Formal / Informal)	No Disponible
20	El Estudiante trabaja durante el curso del programa (SI/NO)	No Disponible

21	Nivel de ingresos familiares del estudiante (Bajo bajo, bajo, medio, medio alto, alto)	No Disponible
22	Fecha de presentacion de los examenes de estados SABER-11	Disponible
23	Resultados de las pruebas de estado clasificado por componente (Lectura critica, Matematicas y razonamiento	calculada
24	Estrato social (1,2,3,4,5,6)	Disponible
25	Estado Civil (Soltero, Casado, Separado, Viudo, Otro)	Disponible
26	Tipo de Vivienda (propia o en arrendo)	No Disponible
27	Procedencia (Urbana o Rural)	No Disponible
28	Ciudad de Origen	No Disponible
29	Barrio de Origen	No Disponible
30	Nivel del SISBEN	No Disponible
31	Numero de personas que conformar el grupo familiar	No Disponible
32	Posicion del estudiante dentro del grupo de hermanos (primero, segundo, tercero, ultimo, unico hijo, etc).	No Disponible
33	El estudiante recibe apoyo economico (SI/NO)	No Disponible
34	El estudiante recibe apoyo psicopedagogico (SI/NO)	Disponible
35	El estudiante recibe apoyo Academico (SI/NO)	Disponible
36	El estudiante recibe orientacion Vocacional (SI/NO)	Disponible
37	El estudiante recibe apoyo economico familiar (SI/NO)	No Disponible
38	Al estudiante se le realizo induccion al ingreso a la Universidad de Manizales (SI/NO)	Disponible
39	Al estudiante se le realizo entrevista de caracterizacion al ingreso a la Universidad de Manizales (SI/NO)	Disponible
40	Al estudiante se le realizo entrevista de ingreso a la Universidad de Manizales (SI/NO)	Disponible
41	El estudiante es desertor (SI/NO)	calculada
42	Fecha de la desercion	calculada
43	Periodo o Cohore de la desercion	calculada
44	Tasa de repitencia del estudiante (SPADIES)	No Disponible
45	Rendimiento academico por periodo o semestre por materia (Notas)	calculada
46	Numero de creditos inscritos en cada periodo	Disponible
47	Numero de credito aprobados en cada periodo	Disponible
48	Numero de credito reprobados en cada periodo	calculada
49	Promedio obtenido en cada periodo	calculada
50	Ausentismo por materia cursada (fallas de asistencia)	Disponible
51	Por cada periodo cursado clasificarlo si estuvo o esta en mora financiera o por pago (SI/NO)	No Disponible
52	Procedencia de su escuela en la secundaria (Oficial / Privado)	Disponible
53	Jornada academica del programa por cada periodo (** Puede cambiar durante el curso de los semestres)	No Disponible
54	El estudiante se matriculo al programa en el primer o segundo semestre del año o periodo	calculada
55	Cual fue la via de acceso a la Universidad de Manizales (Matricula tradicional, becado, pilo paga, etc).	No Disponible
56	El estudiante inicio estudios una vez finalizada su secundaria (SI/NO)	calculada
57	Fecha de Terminacion de la secundaria	Disponible
58	Fecha de Inicio del pregrado	Disponible
59	El programa el cual esta cursando fue su primera opcion (SI/NO)	Disponible
60	Experiencia academica anterior del estudiante (Positiva / Negativa)	No Disponible

61	Dependencia economica de un tercero (SI/NO)	No Disponible
62	El estudiante tiene personas a cargo (Numero de personas).	No Disponible
63	El estudiante realizo estudios preuniversitarios (SI/NO)	No Disponible
64	Modalidad del bachillerato (Presencial, semipresencial, a distancia)	No Disponible
65	El estudiante tiene buenos habitos de estudio (SI/NO)	No Disponible
66	consumo de sustancias sicoactivas	No Disponible
67	como distribuye el tiempo	No Disponible
68	migrante o no	No Disponible
69	estados civil	Disponible
70	Numero de hijos	No Disponible
71	tienen pareja ?	No Disponible
72	tiene alguna incapacidad o limitacion fisica / mental (Ver encuesta de caracterizacion)	No Disponible
73	Ingreso a la Universidad como menor de edad	calculada
74	Estudiante desplazado por algun tipo de violencia	No Disponible
75	Estudiante esta clasificado como RAI (Rendimiento academico insuficiente)	No Disponible

Referencias bibliográficas

Alban, M. y Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in Ecuador. IEEE Global Engineering Education Conference (EDUCON), 1238-1242.

Arredondo, J. (2011). *Perfil de ingreso y deserción de los estudiantes de psicología en la Universidad de Sonora.* Universidad de Sonora.

Ávila Pérez, M. L. (2021). *Modelo De Predicción De Deserción Estudiantil, Apoyado En Tecnologías De Data Mining, En Un Curso De Primera Matrícula De La Escuela ECBTI De La UNAD (Tesis de maestría).* Universidad Nacional Abierta y a Distancia, Escuela de Ciencias Básicas, Tecnología e Ingeniería, Maestría en Gestión de Tecnología de Información.

Barragán, D., y Patiño, L. (2013). Elementos para la comprensión del fenómeno de la deserción universitaria en Colombia. Más allá de las mediciones. *Cuadernos Latinoamericanos de Administración, IX(16)*, 55-66.

Bedregal, N., Aruquipa, D., y Cornejo, V. (2019). Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. *risti Revista ibérica de Sistemas y Tecnologías de Información*, 592-604.

Beltrán, B. (2016). *Minería de datos.* Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación.

Bernardo, A. B., Cerezo, R., Núñez, J. C., Tuero, E., y Esteban, M. (2015). Predicción del abandono universitario: Variables explicativas y medidas de prevención. *Revista Fuentes, 16, junio*, 63-83.

Breiman, L. (2001). *Random forests. Machine learning, 45*, 5-32. Obtenido de https://link.springer.com/article/10.1023/A:1010933404324

Chalela, S., Valencia, A., Ruiz, G., y Cadavid, M. (2020). Factores psicosociales y familiares que influyen en la deserción en estudiantes universitarios en el contexto de los países en desarrollo. *Revisla Lasallista de Investigación, 17(1)*, 103-115.

Chaparro, J., Cuatindioy, J., y Barrera, N. (2021). Análisis comparativo de técnicas de clasificación para determinar la deserción estudiantil de la facultad de ingeniería de la Universidad de Antioquia, Colombia. *Revista Espacios, 42(7)*, 63-81.

Combita, H. (2014). *Plataforma Tecnológica para Disminuir la Deserción Estudiantil en la Universidad de la Costa.* Obtenido de Redclara.net: https://documentos.redclara.net/bitstream/10786/761/1/87-21-3-2014-Plataforma%20Tecnol%C3%B3gica%20Para%20Disminuir%20la%20Deserci%C3%B3n%20Estudiantil.pdf

CP. (1991). *Constitución Política de Colombia.* Asamblea Nacional Constituyente: https://pdba.georgetown.edu/Constitutions/Colombia/colombia91.pdf.

Data and Beyond. (s.f.). Split of Train and Test Data. Obtenido de Data and Beyond: https://dataandbeyond.wordpress.com/2017/08/24/split-of-train-and-test-data/

Decreto 1295 de 2010. (2010). *Por el cual se reglamenta el registro calificado de que trata la Ley 1188 de 2008 y la oferta y desarrollo de programas académicos de educación superior.* Presidente de la República de Colombia: Diario Oficial No. 47687, 21 de abril.

Díaz Arévalo, J., y Pérez García, R. (2002). *Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información no trivial a partir de datos en los sistemas de abastecimiento de agua potable.* Obtenido de ResearchGate: https://www.researchgate.net/publication/338052689_Estado_del_arte_en_la_utilizacion_de_tecnicas_avanzadas_para_la_busqueda_de_informacion_no_trivial_a_partir_de_datos_en_los_sistemas_de_abastecimiento_de_agua_potable

Díaz, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Revista Estudios Pedagógicos, 34(2), Universidad Austral de Chile*, 65-86.

Díaz, D., Morales, M., y Amador, L. (2009). Perfil vocacional y rendimiento escolar en Universitarios. *Revista Mexicana de Orientación Educativa, 6(16)*, 20-23.

DNP. (2018). *Plan Nacional de Desarrollo 2018-2022. Pacto por Colombia, pacto por la equidad.* Obtenido de Departamento Nacional de Planeación: https://www.dnp.gov.co/DNPN/Paginas/Plan-Nacional-de-Desarrollo.aspx

Donoso, S., y Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social. *Estudios Pedagógicos, 33(1)*, 7-17.

Duarte, T., y Jiménez, R. (2007). Aproximación a la teoría del bienestar. *Scientia et Technica, XIII(37), diciembre, Universidad Tecnológica de Pereira*, 305-310.

Durkheim, E. (1928). *El suicidio. Estudio de sociología.* Editorial Reus S.A.

Eccles, J., Adler, T., Futterman, R., Goff, S., Kaczala, C., Meece, J., y Midgley, C. (1983). Expectancies, values, and academic behaviors. En J. Spence (Ed.), *Achievement and achievement motivation* (págs. 75-146). Freeman.

Eckert, K., y Suénaga, R. (2015). Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Formación Universitaria, 8(5) La Serena*, 3-12.

Eguiguren, F. (1997). El Hábeas Data y su desarrollo en el Perú. *Derecho PUCP, (51)*, 291-310.

Espinoza-Aguirre, C., y Carretero-Pérez, J. (2020). Predictive data analysis techniques applied to dropping out of university studies. *LVI Latin American Computing Conference (CLEI)* (págs. 512-521).

Espinoza-Zúñiga, J. (2020). Aplicación de algoritmos *Random forest* y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología; XXI (3), julio-septiembre*, 1-16.

Ethem, A. (2014). *Introduction to machine learning.* The MIT press: https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf

Ethington, C. (1990). A psychological model of student persistence. *Research in Higher Education, 31(31)*, 279-293.

Fayyad, U., Piatestky-Shapiro, G., y Smyth, P. (1996). The kdd Process for Extracting Useful Knowledge from Volumes of Data. *Comunications of the acm, 39(11)*, 27-34.

Felizzola, H., Jaime Arias, Y., Castillo, A. M., y Villa, F. (2018). *Modelo de predicción para la deserción temprana en la Facultad de Ingeniería de la Universidad de La Salle.* Obtenido de ACOFI - Asociación Colombiana de Facultades de Ingeniería: https://acofipapers.org/index.php/eiei/article/view/451/447

Fonseca, G., y García, F. (2016). Permanencia y abandono de estudios en estudiantes universitarios: un análisis desde la teoría organizacional. *Revista de la Educación Superior, 45(179)*, 25–39.

Gartner, L., Dussán, C., y Montoya, D. (2016). Caracterización de la deserción estudiantil en la Universidad de Caldas período 2009-2013. Análisis a partir del sistema para la prevención de la deserción de la educación superior SPADIES. *Revista Latinoamericana de Estudios Educativos, 12(1)*, 132-158.

Giraldo, A., Zapata, C., y Toro, E. (2008). Modelo probabnilístico para los fenómenos de transferencia entre programas de pregrado y de deserción estudiantil. *Scienctia et Technica, XIV(39), Universidad Tecnológica de Pereira*, 212-217.

Guerra, L., Rivero, D., Ortiz, A., Diaz, E., y Quishpe, S. (2020). Modelo de predicción de la deserción universitaria mediante analítica de datos: Estrategia para la sustentabilidad. *risti Revista Ibérica de Sistemas y Tecnologías de Información, E35*, 38-47.

Himmel, E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior. *Revista Calidad de la Educación, 17, Consejo Superior de Educación. Ministerio de Educación, Chile*, 94-95.

IBERDROLA. (2023). Qué es el '*Machine learning*'. Obtenido de IBERDROLA: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

Izquierdo, G., y Mestanza, R. (2017). El aprendizaje consciente y la formación del ser humano. Retos de la Ciencia, 1(2), 15-21.

Jiménez Mora, M. C. (2021). *Abandono y permanencia en educación superior: un análisis multinivel para Iberoamérica.* Universidad Nacional de Colombia - Sede Medellín, Facultad de Ciencias Humanas y Económicas.

Ley 1266 de 2008. (2008). *Por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países y se...* Congreso de la República de Colombia: Diario Oficial, No. 47219, 31 de diciembre.

Ley 1581 del 2012. (2012). *Por la cual se dictan disposiciones generales para la protección de datos personales.* Congreso de Colombia: Diario Oficial, No. 48587, 18 de octubre.

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico.* Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística.

Londoño, L. F. (2013). Factores de riesgo presentes en la deserción estudiantil en la Corporación Universitaria Lasallista. *Revista Virtual Universidad Católica del Norte, (38), febrero-mayo*, 183-194.

López, R. (2017). Ciencia de Datos. Obtenido de IAAR book: https://iaarbook.github.io/datascience/

MastersInDataScienc. (2022). What Is Undersampling? . Obtenido de Master's in Data Science: https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/

MEN. (2009). *Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención.* Ministerio de Educación Nacional: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-

254702_libro_desercion.pdf. Obtenido de Ministerio de Educación Nacional: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

MEN. (2010). *Plan Sectorial 2010-2014. Prosperidad para todos.* Obtenido de Ministerio de Educación Nacional: https://www.mineducacion.gov.co/1759/w3-printer-407780.html

MEN. (2012). *Acuerdo nacional para disminuir la deserción en Educación superior.* Obtenido de Ministerio de Eduación Nacional: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_archivo_pdf_politicas_estadisticas.pdf

MEN. (2014). *Datos generales 2013.* Obtenido de Ministerio de Educación Nacional: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_archivo_pdf_estadisticas_2013.pdf

MEN. (2016). ¿Cómo funciona el SPADIES? Obtenido de SPADIES Sistema para la Prevención de la Deserción de la Educación Superior: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254668.html?_noredirect=1#:~:text=EL%20SPADIES%20centraliza%20informaci%C3%B3n%20proveniente,no%20dentro%20del%20trayecto%20acad%C3%A9mico.

MEN. (2015). *Estrategias para la permanencia en educación superior: Experiencias significativas.* Ministerio de Educación Nacional: https://www.mineducacion.gov.co/1759/articles-356276_recurso.pdf.

MEN. (2017). *Plan Nacional Decenal de Educación 2016-2026. El camino hacia la calidad y la equidad.* Obtenido de Ministerio de Educación Nacional: https://www.mineducacion.gov.co/1780/articles-392871_recurso_1.pdf

Miranda, M., y Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación Universitaria, 10(3)*, 61-68.

Molina, L. C. (2002). *Data mining: torturando a los datos hasta que confiesen.* Obtenido de Business inteligence: https://www.businessintelligence.info/resources/assets/dss/molina-torturando-datos.pdf

Moreira da Silva, D., Solteiro Pires, E., Reis, A., de Moura Oliveira, P., y Barroso, J. (2022). Forecasting Students Dropout: A UTAD University Study. *Future Internet, 76*, 1-14.

Moro, S., Laureano, R., y Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the crisp-dm Methodology. *Proceedings of European Simulation and Modelling Conference -ESM'2011* (págs. 117–121). EUROSIS-ETI Publication.

Muñoz de Alba, M. (2017). *Habeas Data.* Obtenido de Instituto de Investigaciones Jurídicas, UNAM: https://repositorio.unam.mx/contenidos/5018634

Nye, F. (1979). Choice, Exchange, and the Family. En W. Burr, R. Hill, I. Nye, y I. Reiss (Edits.), *Contemporary Theories about the Family. General Theories/ Theoretical Considerations* (págs. 1-41). The Free Press.

OCDE (2010). *Panorama de la educación 2010. Indicadores de la OCDE.* Santillana.

OCDE (2019). *Panorama de la educación Indicadores de la OCDE 2019.* Obtenido de Ministerio de Educación y Formación Profesional, España: file:///D:/Downloads/19884.pdf

OCDE, Banco Mundial. (2012). *La Educación Superior en Colombia 2012.* Obtenido de OCDE: https://www.oecd.org/education/skills-beyond-school/Evaluaciones%20de%20pol%C3%ADticas%20nacionales%20de%20Educaci%C3%B3n%20-%20La%20Educaci%C3%B3n%20superior%20en%20Colombia.pdf

Pascarella, E., Smarta, J., y Ethington, C. (1986). Long-term persistence of two-year college students. *Research in Higher Education. 24(1)*, 47-71.

Perassi, Z. (2009). ¿Es la evaluación causa del fracaso escolar? *Revista iberoamericana de educación, 50, mayo-agosto*, 65-80.

Pérez, A., Grandón, E. E., Caniupán, M., y Vargas, G. (2018). *Análisis Comparativo de Técnicas de Predicción para Determinar la Deserción Estudiantil: Regresión Logística vs Árboles de Decisión.* Obtenido de Universidad del Bío Bío, Departamento de Sistemas de Información: https://dsi.face.ubiobio.cl/mcaniupan/pdfs/desercion_cam_ready.pdf

Quintela, G. (2013). Deserción universitaria, una aproximación sociológica al proceso de toma de decisiones de los estudiantes. *Sociedad Hoy, (24), enero-junio*, 83-106.

Ramírez, P., y Grandon, E. (2017). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados. *Formación Universitaria, 11(3)*, 3-10.

Santes, G., Ramos, I., Lavoignet, B., Cruz, F., y Lara, C. (2017). *Factores de deserción escolar en estudiantes de Enfermería de una universidad pública.* Obtenido de Revista Portales Médicos: https://www.revista-portalesmedicos.com/revista-medica/desercion-escolar-estudiantes-de-enfermeria/3/

Sectorial. (2020). *Deserción Universitaria, ¿Moda en las Nuevas Generaciones o Limitantes de la Educación Superior?* Obtenido de Sectorial, Análisis, monitoreo y evaluación de sectores: https://www.sectorial.co/articulos-especiales/item/296882-deserci%C3%B3n-universitaria,-%C2%BFmoda-en-las-nuevas-generaciones-o-limitantes-de-la-educaci%C3%B3n-superior

Sentencia C-748/11. (2011). *Proyecto de Ley estatutaria de hábeas data y protección de datos personales.* Corte Constitucional de Colombia: https://www.corteconstitucional.gov.co/relatoria/2011/c-748-11.htm.

SPADIES. (2014). *Informe Determinantes de la deserción. Informe mensual sobre el soporte técnico y avance del contrato para garantizar la alimentación, consolidación, validación y uso de la información del SPADIES .* Universidad de los Andes, Facultad de Economía, Centro de Estudios sobre Desarrollo Económico CEDE.

SPADIES. (2022). *Sistema de Información SPADIE S.* Obtenido de SPADIES Sistema para la Prevención de la Deserción de la Educación Superior - MinEducación: https://www.mineducacion.gov.co/sistemasinfo/spadies/

Spady, W. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 19(1)*, 109-121.

Timarán, S., Hernández, I., Caicedo, S., Hidalgo, A., y Alvarado, J. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En S. Timarán, I. Hernández, S. Caicedo, A. Hidalgo, y J. Alvarado, *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (págs. 63-86). Ediciones Universidad Cooperativa de Colombia.

Tinto, V. (1987). *El abandono de los estudios superiores: una nueva perspectiva de las causas del abandono y su tratamiento.* Universidad Nacional Autónoma de México, Asociación Nacional de Universidades e Instituciones de Educación Superior.

Torres, J., Acevedo, D., y Gallo, L. (2015). Causas y consecuencias de la deserción y repitencia escolar: una visión general en el contexto Latinoamericano. *Cultura Educación y Sociedad, 6(2)*, 157-187.

U. Manizales. (2010). *Plan de desarrollo del Sistema Bien-Ser y Bien-Estar, 2010. Revista 4.* Universidad de Manizales.

U. Manizales. (2012). *Reglamento Estudiantil.* Obtenido de Universidad de Manizales: https://umanizales.edu.co/wp-content/uploads/2014/12/Acuerdo-No_09_dic_12-Reglamento-Estudiantil.pdf

U Manizales. (2017). Misión, Visión y Principios. Obtenido de Universidad de Manizales: https://umanizales.edu.co/mision-vision-y-principios/

U. Manizales. (2018). *Política de tratamiento y protección de datos de la Universidad de Manizales.* Obtenido de Univerisdad de Manizales: https://umanizales.edu.co/wp-content/uploads/2018/05/politica-de-privacidad.pdf

U. Manizales. (2018). *Programa de acompañamiento - Acceso y permanencia en la Universidad de Manizales.* Universidad de Manizales.

UNESCO. (2016). *Informe sobre la Educación Superior en América Latina y El Caribe.* Caracas, Venezuela.

Utari, M., Warsito, B., y Kusumaningrum, R. (2020). Implementation of Data Mining for Drop-Out Prediction using *Random forest* Method. *8th International Conference on Information and Communication Technology (ICoICT)* (págs. 1-5).

Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S., y Pretell, J. (2022). Intelligent System to Predict University Students Dropout . *iJOE, 18(7)*, 27-43.

Viloria, A., Garcia Padilla, J., Vargas-Mercado, C., Hernández-Palma, H., Orellano Llinas, N., y Arrozola David, M. (2019). Integration of Data Technology for Analyzing University Dropout. *Procedia Computer Science, 155*, 569-574.

Zambrano, F. (2004). *Constitución de la República de Venezuela.* Editorial Atenea.

Zapata Medina, D. (2021). *Método para la Detección de Estudiantes en Riesgo de Deserción, Basado en un Diseño de Métricas y una Técnica de Minería de Datos (Tesis de maestría).* Universidad Nacional de Colombia - Sede Medellín, Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión.