APLICACIÓN DE LA EDM (EDUCATIONAL DATA MINING) PARA IDENTIFICAR LAS CAUSALES DE BAJO RENDIMIENTO DE LOS ESTUDIANTES DEL PAE PERIODO 2019 – I

LEYDI VIVIANA ASTUDILLO HERRRERA

UNIVERSIDAD DEL CAUCA

FACULTAD DE CIENCIAS CONTABLES, ECONÓMICAS ADMINISTRATIVAS

PROGRAMA DE ADMINISTRACIÓN DE EMPRESAS

POPAYÁN – CAUCA

2019

APLICACIÓN DE LA EDM (EDUCATIONAL DATA MINING) PARA IDENTIFICAR CAUSALES DE BAJO RENDIMIENTO DE LOS ESTUDIANTES DEL PAE PERIODO 2019 - I

LEYDI VIVIANA ASTUDILLO HERRERA

Director Académico:

Magister MARTHA LUCIA ACOSTA RANGEL

UNIVERSIDAD DEL CAUCA

FACULTAD DE CIENCIAS CONTABLES, ECONÓMICAS Y ADMINISTRATIVAS

ADMINISTRACIÓN DE EMPRESAS

POPAYÁN - CAUCA

2019

DEDICATORIA

En primer lugar, agradezco a mi abuela **Aura Rosero** por su gran dedicación, amor y confianza en mí; por creer desde siempre, que este sueño sería posible. A ella que, aunque ya no nos acompaña físicamente, siempre le recordaré.

A mi madre **Elizabeth Herrera**, porque gracias a todos sus esfuerzos, su amor, dedicación y apoyo, es que hoy soy quien soy. Es ella quien siempre me acompañó, me brindó sus consejos y tuvo absoluta fe en mí; nunca dejó que me rindiera y siempre me brindó su apoyo incondicional ante las dificultades que se presentaron durante mi proceso académico. Por nunca dejar que me rindiera.

AGRADECIMIENTOS

Agradezco a la Universidad del Cauca por abrirme sus puertas para formarme profesionalmente, también a mis profesores por su entrega y dedicación, por compartir sus conocimientos y permitir el desarrollo de mi formación educativa profesional gracias a ellos.

En especial quiero agradecer a mi asesora académica **Martha Lucia Acosta Rangel**, por permitirme contribuir con este proyecto y demostrar mis capacidades, también por su entrega al proyecto a pesar de las adversidades, por compartir conmigo sus conocimientos y guiarme y motivarme en el proceso.

RESUMEN.

El objetivo de esta práctica profesional (pasantía) fue encontrar, a partir de EDM o Educational Data Mining (Minería de datos en la educación), las causales de bajo rendimiento de los estudiantes del Programa de Administración de Empresas de la Universidad del Cauca, este proceso se desarrolló, desde el 10 de junio del 2019, hasta el 28 de noviembre del mismo año.

Durante la práctica se desarrollaron todas las etapas del proceso de EDM (Recopilación de datos, selección y limpieza de datos, Data Mining, interpretación y evaluación), para encontrar cuáles son las causales de bajo rendimiento, teniendo en cuenta datos de SIMCA y resultados ICFES

Como resultado, en esta práctica se determinó cuáles eran las causales que influyen en el bajo rendimiento académico de los estudiantes del Programa de Administración de Empresas.

Palabras claves: EDM, Data Mining, bajo rendimiento académico.

ABSTRACT:

The objective of this professional practice (internship) was to find, from EDM or Educational Data Mining, the low performance causes of students of the Business Administration Program of the University of Cauca, this process took place, from June 10, 2019, until November 28 of the same year.

During the practice, all the stages of the EDM process (Data collection, data selection and cleaning, Data Mining, interpretation and evaluation) were developed, to find out what are the causes of low performance, taking into account SIMCA data and ICFES results

As a result, in this practice it was determined what were the causes that influence the low academic performance of the students of the Business Administration Program.

Keywords: EDM, Data Mining, low academic performance.

Contenido

TABLA DE GRAFICAS

TABLA DE ILUSTRACIONES

TABLA DE IMÁGENES

TABLA DE TABLAS

INTRODUCCIÓN

La fortaleza de las organizaciones no se basa en la cantidad de datos que almacena, sino en la forma en la que se interpretan y finalmente aportan al proceso de toma de decisiones.

La minería de datos, pretende hallar patrones que permitan agrupar las grandes bases de datos que se almacenan en las organizaciones para convertir esa información en conocimiento. En los últimos años, los avances en la informática y tecnologías de la información, han expandido los datos disponibles para investigadores y profesionales, en una amplia variedad de dominios.

Desde el gobierno y la tecnología. Hasta la logística empresarial y el deporte profesional, se han creado grandes depósitos de datos. Esta corriente ha llegado al campo educativo, ahí la minería de datos busca utilizar estas bases de datos para mejorar y comprender, tanto estudiantes como maestros, y desarrollar enfoques informáticos que combinen datos y teoría, para transformar la práctica en beneficio de los estudiantes.

Las instituciones de educación superior cuentan con sistemas de información, donde se registran datos socio económicos, información personal y académica. Estos sistemas, aportan información general y netamente académica. Lo que para la Universidad del Cauca aplica por medio de SIMCA (Sistema Integrado de Matrícula y Control Académico). Este trabajo, pretende aplicar técnicas de minería de datos, sobre la información que hallamos en SIMCA e incluyendo también los datos que aporta el ICFES, con el fin de obtener un modelo predictivo que permita conocer de antemano qué estudiantes se encuentran en riesgo de bajo rendimiento académico y posible deserción, para el programa de Administración de Empresas, teniendo en cuenta los patrones de conducta y el entorno del estudiante.

CAPÍTULO I: CONTEXTUALIZACIÓN DEL TRABAJO

En esta primera parte del informe final de la de práctica profesional, se pretende dar a conocer a través del mismo, el trabajo realizado. Este capítulo está dividido en cuatro secciones: el primero, consta de una introducción en la cual se expuso de forma general el contenido del presente; como segundo ítem, se realiza una descripción y definición del problema analizado; posteriormente, se lleva a cabo el planteamiento de una justificación; y por último, se hace un bosquejo de los objetivos desarrollados a lo largo del práctica profesional.

1. PLANTEAMIENTO DEL PROBLEMA

El bajo rendimiento de los estudiantes que realizan sus estudios de pregrado en el programa de administración de empresas, es preocupante. El 13% de los estudiantes que cursan el nuevo pensum, se encuentran atrasados con respecto a las materias que deberían estar cursando teniendo en cuenta su periodo de ingreso y en el caso de las personas que siguen con el antiguo pensum, la situación es de mayor cuidado, pues el 68% de dichos estudiantes, no se encuentran al día.

Además del 7% de los estudiantes de todos los semestres, que se encuentran en casos especiales, entendiendo estos casos como estudiantes que tienen materias en R2 y R3.

La Universidad del Cauca cuenta con una gran cantidad de datos, sin embargo, el solo hecho de registrar y almacenar datos, no basta para crear estrategias que disminuyan el porcentaje de estudiantes que se encuentran en bajo rendimiento académico. Hace falta una herramienta que ayude a condensar dichos datos y nos muestren patrones más asertivos a tener en cuenta para mejorar el rendimiento académico de los estudiantes. Herramienta, que aún no se implementa en el programa de Administración de Empresas.

1.1 definición del problema

"Poco conocimiento sobre las causales de bajo rendimiento académico de los estudiantes del Programa de Administración de Empresas"

Pregunta: ¿Cuáles son las causales que determinan el bajo rendimiento académico en los estudiantes activos en el programa de Administración de Empresas de la Universidad del Cauca en el periodo 2019-I?

2. JUSTIFICACIÓN

Podemos reconocer como factor clave a la hora de hablar de calidad en las instituciones de educación superior, al rendimiento académico. Es sumamente importante, debido a que es un indicador que permite una aproximación a la realidad educativa.

Un tema a tener en cuenta, es el bajo rendimiento académico, este es un problema con múltiples causas y repercusiones en el que están implicados, de acuerdo a Pérez (2007) factores de diversa índole, entre los que se destaca el bienestar psicológico del estudiante, factores docentes o educativos y familiares, entre otros. Igualmente, debemos reconocer que el bajo rendimiento académico, afecta no sólo a los estudiantes, sino también a sus familias, teniendo en cuenta consecuencias económicas justificadas en un aumento de costos de matrícula y manutención. Un atraso en el desarrollo normal de los tiempos de estudio, que en el caso del programa de Administración de Empresas es entre diez y once semestres (teniendo en cuenta la modalidad de opción de grado que elija el estudiante). Limita sus posibilidades de vinculación laboral según lo previsto y retarda de manera considerable el logro de metas pos graduables e inclusive familiares.

Teniendo en cuenta lo anterior podemos notar la importancia de identificar los factores relacionados con el bajo rendimiento académico de los estudiantes del programa de Administración de Empresas de la Universidad del Cauca, pues con esta información se pueden plantear estrategias efectivas, dirigidas a mejorar el rendimiento académico de los

estudiantes, y también, a contribuir en el proceso de mejora de la calidad del programa. Se plantea entonces, encontrar solución a este problema, usando las herramientas de la minería de datos que es definida como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos, Fayyad, Piatetsky- Shpiro & Smith (1996).

Este tema de estudio en la educación es un área de investigación que ha evolucionado en los últimos años y se desarrolla a la par de otras tecnologías o aplicaciones.

Las contribuciones de la minería de datos en la educación han sido utilizadas para obtener una mejor comprensión del proceso educativo con el principal objetivo de proporcionar a los docentes e investigadores recomendaciones para el mejoramiento del proceso de enseñanza - aprendizaje. El objetivo de la EDM (Educational Data Mining), es aplicar la minería de datos a los sistemas tradicionales de enseñanzas, en particular a los sistemas de gestión de contenidos de aprendizaje.

3. OBJETIVOS

3.1. Objetivo General

Identificar causales de bajo rendimiento de los estudiantes activos en el periodo 2019-I, aplicando la EDM, basados en los datos estadísticos de SIMCA y el ICFES; del programa de Administración de Empresas de la Universidad del Cauca.

3.2. Objetivos Específicos

- ✓ Construir un conjunto de datos de estudiantes, que contemplen información de tipo socioeconómico e historial académico.
- ✓ Identificar los diferentes factores que afectan el rendimiento académico del estudiante en el pregrado.
- ✓ Diseñar el modelo predictivo utilizando técnicas de minería de datos.

CAPÍTULO II: CONTEXTUALIZACIÓN TEÓRICA

En el presente capítulo se explica lo concerniente al marco teórico donde se evidencia el uso de los diferentes conceptos y teorías que sirvieron de base para el desarrollo del trabajo. Marco contextual que nos permitirá conocer aspectos de naturaleza interna de la Organización y el marco legal el cual determina la validez y el análisis del proyecto.

4. MARCO TEÓRICO

El marco teórico está constituido por un conjunto de teorías, enfoques teóricos, investigaciones y antecedentes que se consideran válidos para el encuadre correcto de la investigación que se quiere realizar (Santalla, 2003).

Para el desarrollo del proyecto nos basaremos en las siguientes teorías las cuales apoyan y guían el proceso del mismo.

El rendimiento académico hace referencia a la evaluación del conocimiento adquiriendo en el ámbito escolar. Un estudiante con buen rendimiento académico es aquel que obtiene calificaciones positivas en los exámenes que debe rendir a lo largo del curso. En otras palabras, el rendimiento académico es una medida de las capacidades del alumno, que expresa lo que éste ha aprendido a lo largo del proceso formativo. También supone la capacidad del alumno para responder a los estímulos educativos. En este sentido, el rendimiento académico está vinculado a la aptitud (Maradona &Calderón, 2004).

Las calificaciones determinan la tipificación de aprobación o no de las materias vistas durante un periodo académico.

Escudero (1990), aclara que “las calificaciones son una medida de resultados de la enseñanza, pero no estrictamente de su calidad, pues están condicionadas, no sólo por la calidad de los estudiantes, sino por el criterio y el rigor personal del profesor a la hora de diseñar la enseñanza y valorar y calificar el aprendizaje y el rendimiento académico”.

Pero por el otro lado, con respecto al ***bajo rendimiento*** escolar, en la mayoría de las sociedades, se centra en el alumno y se contempla también la acción de otros agentes como las condiciones sociales, la familia o la propia escuela (Shapiro, 2011).

Cuando hay una discrepancia entre la potencialidad del estudiante y su rendimiento, o cuando no han adquirido en el tiempo previsto, de acuerdo al programa establecido y sus capacidades intelectuales, los resultados que se esperan de él, se habla de ***bajo rendimiento académico***, (Fueyo, 1990).

Se considera estudiantes en ***bajo rendimiento académico***, a todo aquel que haya perdido alguna asignatura que curse en calidad de repitente por primera vez (Reglamento estudiantil Unicauca).

para efectos de este estudio se utilizó el concepto de la Universidad del Cauca, establecido en el reglamento estudiantil; Sin embargo, se tuvo en cuenta también como bajo rendimiento, a las personas que no habían seguido el plan de estudios según lo estipulado por el programa. Así un estudiante que según su ingreso debería estar en sexto semestre y en SIMCA aparece como estudiante de quinto semestre, también está en bajo rendimiento.

EDM (Educational Data Mining) es una disciplina emergente, preocupada por desarrollar métodos para explorar los datos únicos y cada vez más a gran escala que provienen de entornos educativos y usar esos métodos para comprender mejor a los estudiantes y los entornos en los que aprenden. También se puede definir como la adaptación del proceso de minería de datos aplicado al campo de la educación. Por eso, en este documento, a partir de este momento, se hablará de minería de datos o Data mining, para abordar el concepto y el proceso que se debe seguir para llegar al conocimiento a partir de este.

El descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso, de acuerdo con Agrawal& Srikant (1994), Chen, Han & Yu (1996) y Han & Kamber

(2001), consiste en extraer patrones en forma de reglas o funciones, a partir de los datos. Para que el usuario los analice, tarea que implica generalmente pre procesar los datos, hacer minería de datos (data mining) y presentar resultados.

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una importante área y una oportunidad para obtener mayores ganancias. Fayyad, Piatetsky- Shpiro & Smyth (1996) definen la ***minería de datos*** como "el proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos".

Según Caridad (2008), el concepto de ***minería de datos*** representa una idea que ha venido madurando durante muchos años, en el sentido de recorrer grandes bases de datos para recuperar información conceptual de interés e inferir nueva información útil. La ***minería de datos*** es un proceso posterior, destinado a lograr un mejor conocimiento de la información disponible, aumentar beneficios o ventas, y disminuir pérdidas; es decir, tiene un objetivo distinto al que ha motivado la recogida y almacenamiento de información.

Liao, Chen y Deng (2010) concluyen que la ***minería de datos*** es el proceso de descubrimiento significativo del conocimiento del cliente (patrones, asociaciones, cambios y estructuras de grandes cantidades de datos almacenados en bases de datos), desde un enfoque que integra la minería matemática, de datos tradicional, y técnicas evolutivas con un sistema multi agente.

Riquelme, Ruiz y Gilbert (2006), al igual que Liao, Chu y Hsiao (2012) definen las tareas más comunes de la ***minería de datos***:

1. ***Clasificación:*** clasifica un dato dentro de una de las clases categóricas predefinidas.

2. ***Regresión:*** el propósito de este modelo es hacer corresponder un dato con un valor real de una variable.
3. ***Clustering:*** se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, y distintos a los registros de otro clúster.
4. ***Generación de reglas:*** aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos.
5. ***Resumen o sumarización:*** estos modelos proporcionan una descripción compacta de un subconjunto de datos.
6. ***Análisis de secuencias:*** se modelan patrones secuenciales, como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del proceso, o extraer e informar de la desviación y tendencias en el tiempo.

Ur-Rahman y Harding (2012) explican que la tecnología de ***minería de datos*** proporciona flexibilidad para explotar la información desde múltiples formatos o bases de datos, tales como bases de datos relacionales, almacenamiento de datos y bases de datos transaccionales, etc. Esto es complementado por PhridviRaj y GuruRao (2014), en tanto comentan que la diferencia entre los datos en las bases de datos y un almacén de datos es que, en las primeras, los datos están en la forma estructurada; mientras que en segundo pueden o no estar presentes de tal manera. Muhammad, Mohamudally y Babajee (2013) explican que el foco de la extensa investigación en el campo de la ***minería de datos*** está en el desarrollo y mejora de los algoritmos existentes, así como en la evaluación de los conocimientos descubierto como un proceso de un solo paso o de múltiples pasos, desde diferentes enfoques (como el álgebra relacional y la teoría de la información, entre otros).

De acuerdo con Peña-Ayala (2014), la ***minería de datos*** es un proceso dedicado a escanear enormes repositorios de datos para generar información y descubrir conocimiento.

Según Pérez-Palacios (2014), la ***minería de datos*** es una parte importante de un proceso más amplio conocido como descubrimiento de conocimiento en bases de datos (KDD en inglés). El objetivo principal de la ***minería de datos*** consiste en la extracción de información oculta de un conjunto de datos. Esto puede ser alcanzado por el análisis automático o semiautomático de gran cantidad de datos, lo que permite la extracción de patrones desconocidos. Estos pueden ser grupos de registros de datos (análisis clúster), inusuales registros (detección de anomalías) y las dependencias entre datos (asociación reglas). Por lo tanto, los patrones pueden ser vistos como un resumen de los datos de entrada, y se pueden utilizar para su posterior análisis.

La minería de datos forma parte de un proceso más amplio que en principio fue llamado como KDD (Knowledge Data Mining), pero dentro de este gran proceso, la actividad más importante a desarrollar es la de minería de datos y finalmente el proceso total terminó quedándose con el nombre de Minería de Datos.

La minería de datos es un proceso que pretende hallar patrones dentro de un conjunto de datos que ya tienen las organizaciones y no están siendo explotados como se debería.

5. MARCO CONTEXTUAL

5.1. Nombre de la Organización:

Universidad del Cauca

5.1.1. Ubicación:

Se encuentra localizado en la Ciudad de Popayán, Cauca.

5.1.2. Área o dependencia:

Facultad de Ciencias Contables Económicas y Administrativas

5.1.3. Sector económico de actividad: Educación, actividad que atiende a satisfacer las necesidades individuales hacia la creación de sociedades del aprendizaje que otorguen oportunidades de educación a toda la población.

La Universidad del Cauca es un ente universitario autónomo del orden nacional vinculado al Ministerio de Educación, con régimen especial, personería jurídica, autonomía académica, administrativa y financiera y patrimonio independiente.

Fue creada el 24 de abril de 1827 mediante decreto dictado por el presidente de la República Francisco de Paula Santander, en desarrollo de la Ley del 18 de mayo de 1826.

Se instaló el 11 de noviembre de 1827 y su nacionalización fue ratificada mediante la Ley 65 de 1964.

El Programa de Administración de Empresas pertenece a la Facultad de Ciencias Contables, Económicas y Administrativas de la Universidad del Cauca y fue creado mediante Resolución No 112 del 20 de diciembre de 1989, expedida por el Honorable Consejo Superior.

Este Programa Académico tiene énfasis en el área de Desarrollo Empresarial y mediante Resolución número 2031 de 24 de marzo de 2010 se otorgó acreditación de alta calidad al Programa de Administración de Empresas de la Universidad del Cauca.

Mediante la resolución 13153 del 16 de octubre de 2012, expedida por el Ministerio de Educación Nacional y registrada en el Sistema Nacional de Información de la Educación Superior (SNIES), se le otorgó el Registro Calificado a este Programa de la Universidad del Cauca, por el término de 7 años.

5.1.4. Contextualización de la situación académica de los estudiantes del PAE

Para el inicio del periodo de 2019-1, el número de estudiantes activos en el Programa de administración de empresas era de 496 estudiantes.

Los estudiantes que ingresaron en el periodo 2018-1, apenas se encuentran cursando (en caso de que estén al día) tercer semestre, no nos sirven para el estudio, pues no han cursado más de dos semestres y por eso, no es posible que haya perdido una materia más de dos veces, que es una de las clases que se investigará en este estudio. Algo parecido ocurre con los estudiantes que ingresaron al periodo 2018-2, que se encuentran en segundo semestre y 2019-1 que iniciaron primer semestre.

El número de estudiantes de ingreso 2018-1 es de 44

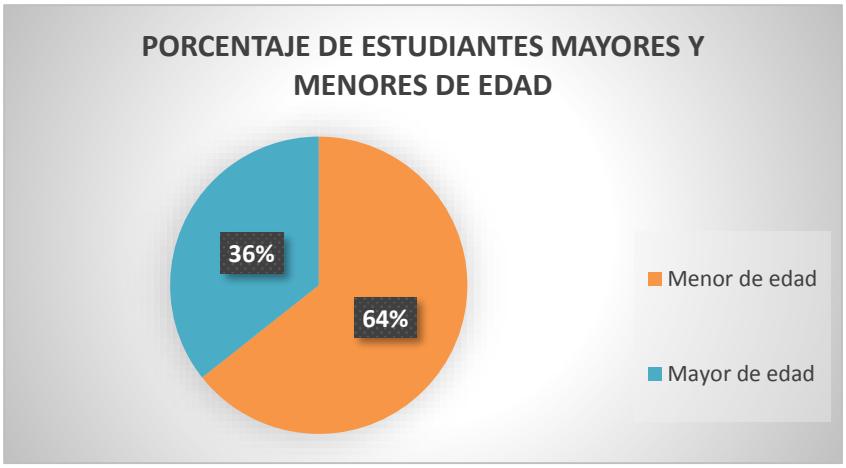
El número de estudiantes de ingreso 2018-2 es de 35

El número de estudiantes de ingreso 2019-1 es de 45

Para un total de 124 estudiantes

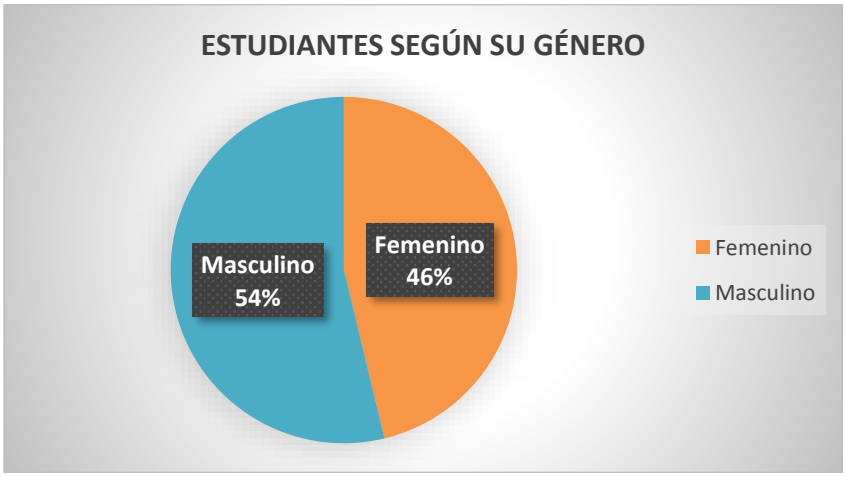
En ese sentido, el número de estudiantes sobre los cuáles se realizará el proceso de EDM será las 496 iniciales menos los 124 que no cumplen con las condiciones para el mismo, para un total de **372 estudiantes**. De estos estudiantes, según datos de SIMCA:

- ***Edad de ingreso:*** El 36% de los estudiantes del programa de administración de empresas, ingresaron siendo menores de edad. Mientras que el 64% restante, ya eran mayores de edad al ingresar a la universidad, como lo muestra el siguiente gráfico.



Grafica 1 Porcentaje de estudiantes mayores y menores de edad

- ***Estudiantes según su género:*** Los estudiantes del programa de sexo femenino es el 46%, mientras que los estudiantes de género masculino es el 54%.



Grafica 2 Estudiantes según su género

- ***Estudiantes según su lugar de procedencia:*** El 90% de los estudiantes del PAE, proceden del departamento del Cauca. El 10% restante, se reparte entre estudiantes

procedentes de los departamentos de Nariño, Putumayo, Huila y Valle del Cauca respectivamente.

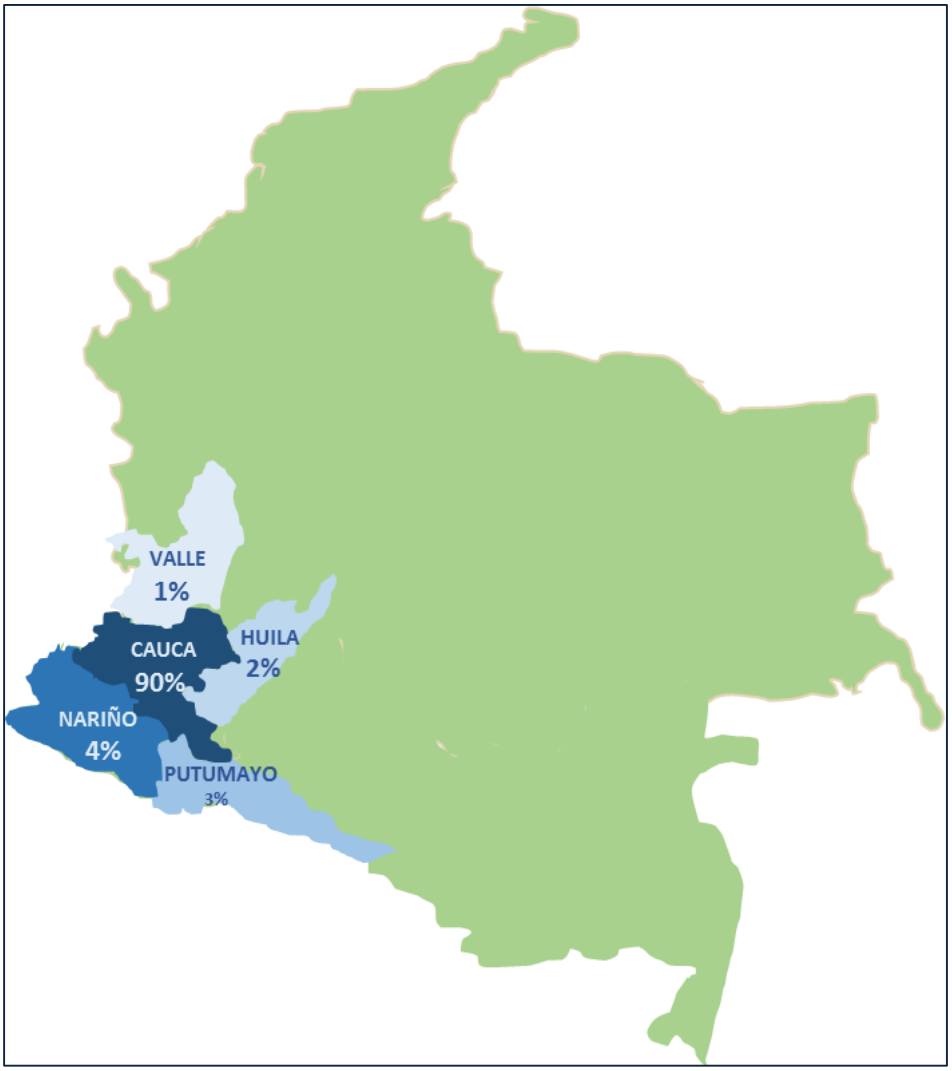


Ilustración 1 • Estudiantes según su lugar de procedencia

Dado que el departamento del que procede la gran mayoría de estudiantes es el departamento del Cauca, es interesante conocer los municipios de los cuales llegan. Se encuentra, como es apenas lógico, el municipio de Popayán en primer lugar, con un 80% de estudiantes que ingresan a estudiar Administración de Empresas en la Universidad del Cauca; seguido a este, se encuentran los municipios de Timbio, Piendamó, Bolivar y Balboa, que si bien no representan grandes porcentajes de estudiantes, sí cabe resaltar que estos dos últimos municipios se encuentran a más de dos horas de la ciudad de Popayán y aun así, el número de estudiantes que proceden de dichos municipios, es muy cercano al

número de estudiantes que proceden de municipios como Piendamó o Timbio que se encuentran a pocos minutos de la ciudad de Popayán.

También vemos estudiantes que proceden de los municipios de Inzá, El Tambo, Cajibio, Caldono, La Vega, Morales, Patía, Argelia, Jambaló, Lopez de Micay, Mercaderes, Rosas y Toribio, lo que nos da como resultado 19 municipios en total; Es decir, el programa de Administración de Empresas de la Universidad del Cauca, forma estudiantes provenientes del 45% de los municipios del Cauca.

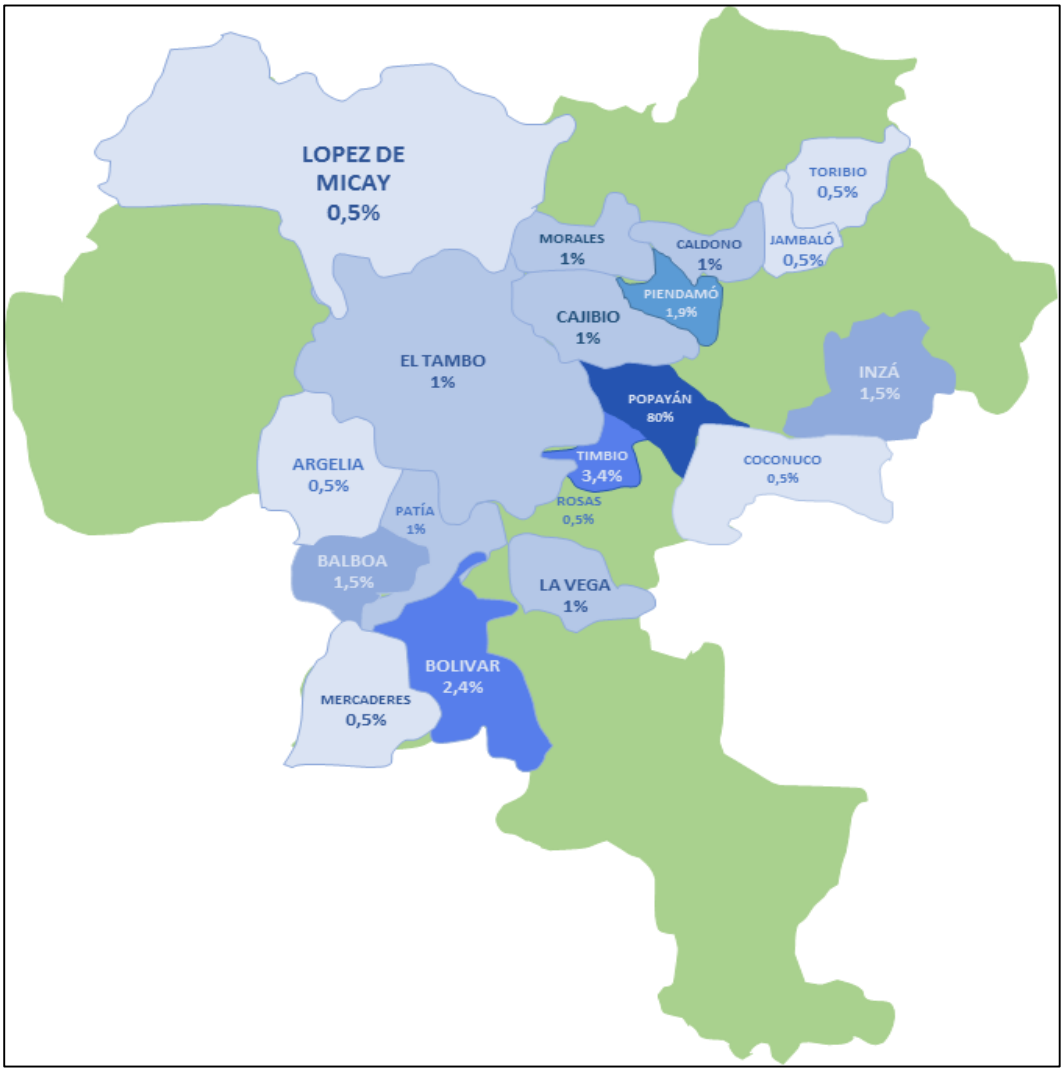
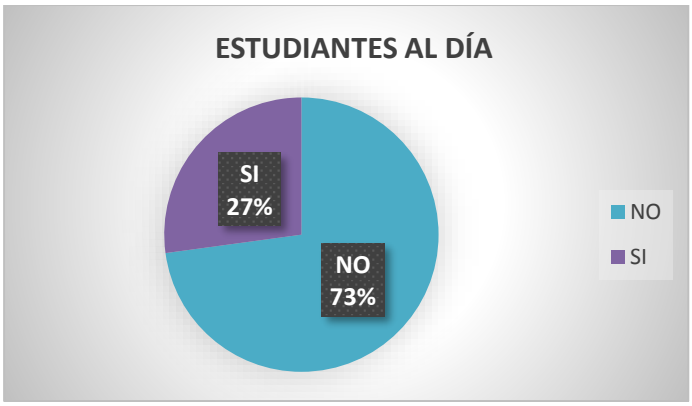


Ilustración 2 • Estudiantes según el municipio del que proceden

- ***Está al día:*** Se pudo notar, según los datos de SIMCA, analizando una a una la situación académica de cada estudiante que, el 73% de los estudiantes del PAE no se encuentran

al día con las materias que deberían estar cursando para este periodo, teniendo en cuenta su periodo de ingreso.



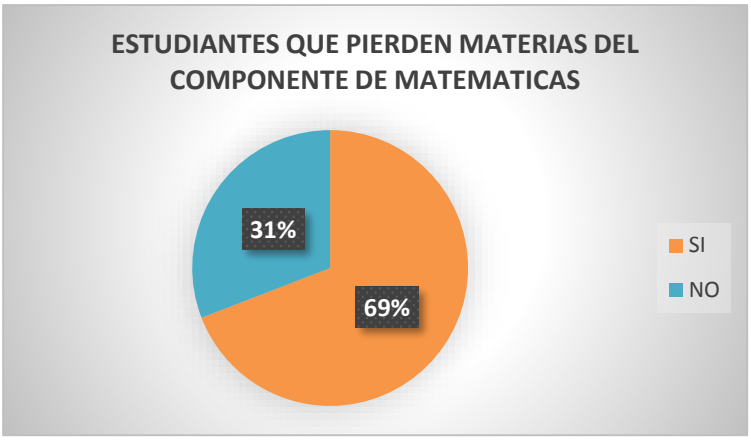
Grafica 3 Estudiantes al día

- ***Materias que han perdido más de dos veces o más de R2:*** Después de las materias del componente de matemáticas, estas son las materias que los estudiantes pierden más, teniendo en cuenta que, para estos casos, los estudiantes han perdido estas materias más de dos veces.

MATERIA	PORCENTAJE
CONTABILIDAD DE COSTOS	14%
ADMINISTRACION I	9%
FUNDAMENTOS DE ECONOMIA	9%
MACROECONOMIA	9%
ADMINISTRACION II	7%
ANALISIS DE SECTORES ECONOMICOS	7%
LABORATORIO DE INFORMATICA I	7%
MICROECONOMIA I	7%
AREAS DE LA EMPRESA	5%
FINANZAS I	5%
INVESTIGACION DE OPERACIONES	5%
TECNOLOGIA	5%
DERECHO COMERCIAL	2%
ETICA PROFESIONAL	2%
GERENCIA EMPRESARIAL	2%
HACIENDA PUBLICA	2%
INVESTIGACIÓN DE MERCADOS	2%
MICROECONOMIA II	2%

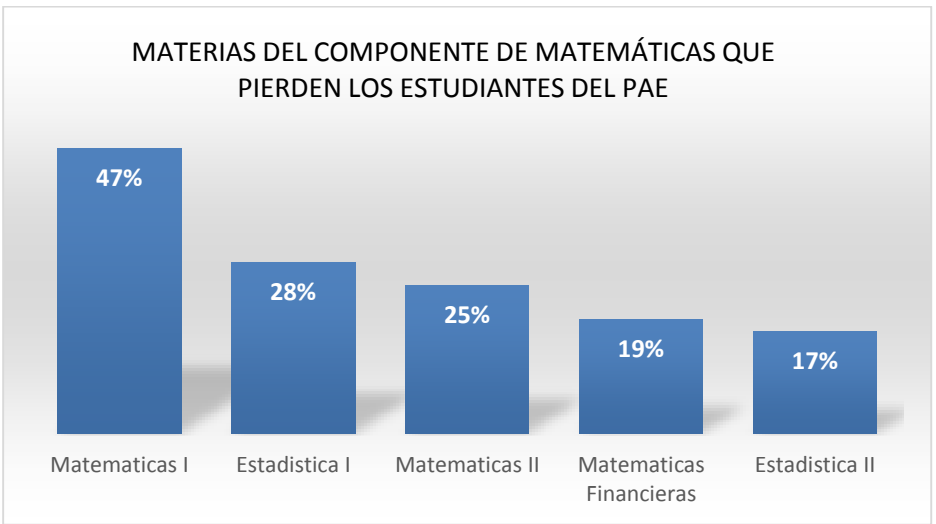
Tabla 1 Materias que pierden más de dos veces

- ***Estudiantes que pierden materias del componente de matemáticas***: Como lo muestra el siguiente gráfico, el porcentaje de estudiantes que no aprueba materias del componente de matemáticas es bastante significativo. 69% es el número de estudiantes que, dentro de su proceso académico hasta el periodo actual, han perdido al menos una de las materias del componente de matemáticas.



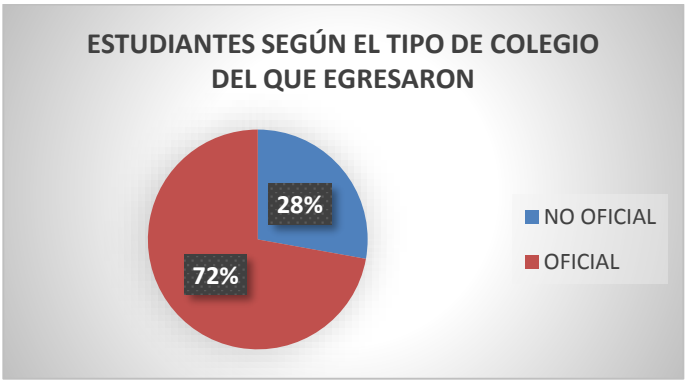
Grafica 4 Estudiantes que pierden materias del componente de matemáticas

- ***Materias del componente de matemáticas que pierden los estudiantes del PAE:*** El siguiente gráfico, representa los porcentajes de estudiantes que pierden cada materia del componente de matemáticas. Como se puede observar, la materia que más pierden los estudiantes es Matemáticas I con un 47% de los estudiantes que no aprueban, seguida de Estadística I con un 28%, en tercer lugar, tenemos la materia de Matemáticas II y finalmente encontramos Matemáticas financieras, junto con Estadística II.



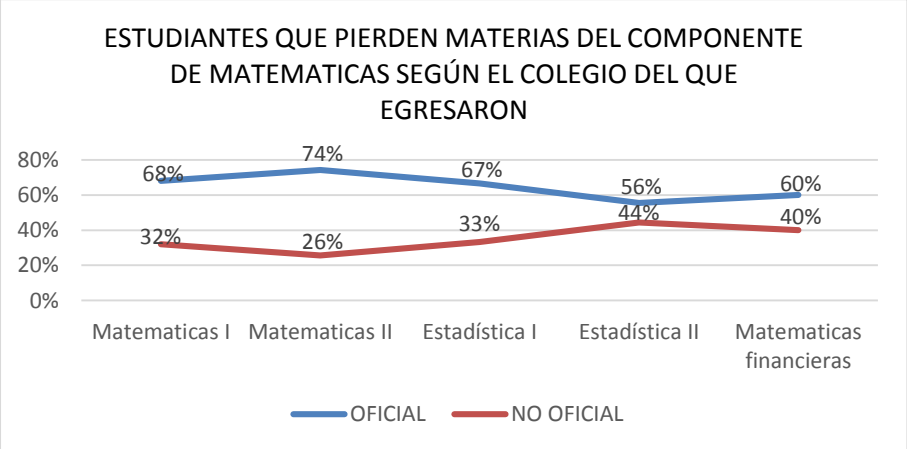
Grafica 5 Materias del componente de matemáticas que pierden los estudiantes del PAE

- ***Estudiantes según el tipo de colegio del que egresaron:*** El 72% de los estudiantes del PAE, provienen de colegios de carácter oficial (público), lo cual marca una notable diferencia con respecto a los estudiantes que provienen de colegio de carácter no oficial (privados) que sólo suman el 28% de los estudiantes del PAE.



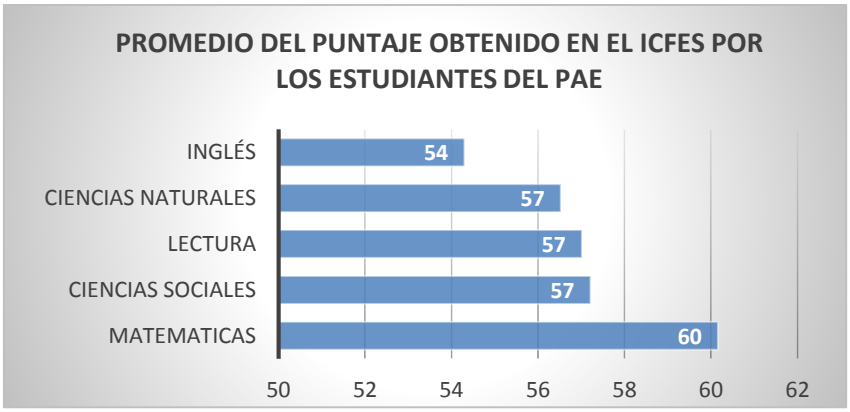
Grafica 6 Estudiantes según el tipo de colegio del que egresaron

- ***Estudiantes que pierden materias del componente de matemáticas según el colegio del que egresaron:*** En el siguiente gráfico, se muestra la división porcentual de estudiantes, entre quienes provienen de colegios oficiales (público) y quienes provienen de colegios no oficiales (privados). Es notable, cómo la línea roja (que representa los estudiantes que provienen de colegios no oficiales), no sobre pasa la línea azul (que representa a los estudiantes que provienen de colegios oficiales), lo que quiere decir, que la mayoría de estudiantes que pierden materias del componente de matemáticas provienen de colegios de carácter oficial.



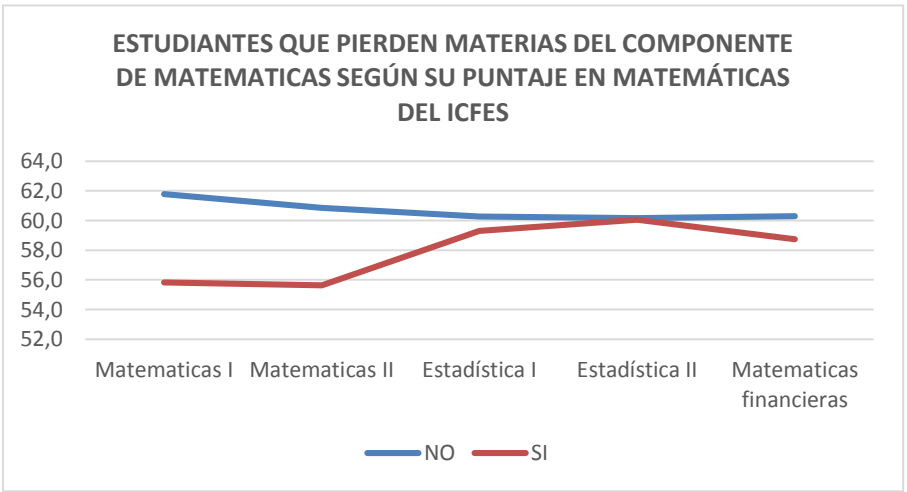
Grafica 7 Estudiantes que pierden materias del componente de matemáticas según el colegio del que egresaron

- ***Promedio del puntaje obtenido en el ICFES, por los estudiantes del PAE:*** El siguiente gráfico representa el porcentaje promedio que obtuvieron los estudiantes del Programa de Administración de Empresas en el ICFES. El área en la que los estudiantes que ingresan al PAE, tienen mayor puntaje, es en el área de matemáticas con 60 puntos obtenidos.



Grafica 8 Promedio del puntaje obtenido en el ICFES por los estudiantes del PAE

- ***Estudiantes que pierden materias del componente de matemáticas según su puntaje de matemáticas en el ICFES:*** Este gráfico plasma los datos de los estudiantes que pierden o no materias del componente de matemáticas. En la mayoría de los casos, los estudiantes que obtuvieron puntaje por debajo del promedio que reflejó el gráfico anterior en matemáticas, es decir, 60 puntos; perdieron alguna de las materias del componente de matemáticas. Este, es un patrón que se repite en todas las materias del componente.



Grafica9.Estudiantes que pierden materias del componente de matemáticas según su puntaje en matemáticas del ICFES

CAPÍTULO III: METODOLÓGIA

En este capítulo, se hará una descripción del proceso que se siguió para llegar a resolver el problema que se describió en un principio. Como ya se ha planteado, el método para hallar las causales de bajo rendimiento en los estudiantes activos a primer semestre de 2019 en el programa de Administración de Empresas, es la EDM (Educational Data Mining) y partiendo del punto de que esta, no es más que la adaptación del proceso de minería de datos al campo de la educación, se consideró los pasos para desarrollar este proceso.

Para realizar el proceso de minería de datos, Lara Torralbo en su libro "Minería de datos", plantea cada una de las etapas que se debe seguir y la forma de desarrollar cada una de ellas.

Paso 1. Recopilación de datos:

Muchas veces, los datos proceden no solo de la organización que ha solicitado el estudio, sino que también puede proceder de medios externos.

Lo importante es saber almacenar todos los datos en una sola estructura de ***data warehouse***

Para integrar las fuentes en un mismo almacén, es necesario orquestar un proceso que lea los datos de las diferentes fuentes, los limpie y los adecue a la estructura que tiene la data warehouse para su almacenamiento.

Paso 2. Selección y limpieza de datos:

Selección de datos:

Las técnicas de selección tienen como objetivo filtrar aquellos datos que no son relevantes para el análisis posterior.

El filtrado de esos datos puede ser:

- Filtrado de atributos: Es posible que algunos de los atributos de los datos a analizar no sean de interés.

  Por ejemplo, el número de cédula de ciudadanía, es más posible que dicho atributo debe ser filtrado ya que, de cara a un posible análisis de datos, este atributo no aporta nada.
- Filtrado de registros: En ocasiones, el objetivo puede ser eliminar algunos de los registros almacenados.

  Por ejemplo, es posible que la empresa sólo esté interesada, en este momento, en los clientes cuya edad es mayor que cierto umbral.

  Este tipo de filtrado, es común hacerlo, debido a una gran cantidad de registros existentes.

  Con un subconjunto menor de registros, denominado ***muestra***, se podrá hacer un análisis igual de efectivo, pero mucho más eficiente.

  Algunas técnicas de muestreo son:

  - Muestreo aleatorio simple: En este caso, todos los elementos del conjunto completo de datos tienen igual probabilidad de ser extraídos en la muestra.
  - Muestreo aleatorio estratificado: El objetivo de este tipo de muestreo es que todos los grupos o estratos que conforman el conjunto de datos completo estén representados de forma equilibrada en la muestra.
  - Muestreo por grupos: En este caso, solamente se seleccionan registros pertenecientes a un grupo determinado. Por ejemplo, aquellos clientes de una compañía que consuman más de 200 euros mensuales en sus productos.

Limpieza de datos:

Las tareas de limpieza de datos van normalmente encaminados a resolver dos problemas bastante habituales

- La ausencia de valores: Es muy común que, para muchos de los registros analizados, falte cierta información. A estos valores ausentes se les conoce con el nombre de ***missing values.*** Es importante que el ingeniero de minería de datos analice cautelosamente los valores faltantes ya que, en ocasiones, aportan información interesante.

  En general, ante un valor faltante, se pueden tomar las diferentes alternativas, siendo las siguientes, las más habituales:

  - Pasar por alto el valor faltante y continuar con el análisis.
  - Filtrar la columna asociada a dicho atributo.
  - Filtrar el registro que contiene el valor faltante.
  - Filtrar el registro que contiene el valor faltante.
  - Asignar un valor al atributo en cuestión.
- La existencia de valores erróneos: Habitualmente podemos encontrar valores que claramente son erróneos. Aunque existen diferentes técnicas para detectar este tipo de valores, la realidad es que la mayoría de veces, se realiza mediante procesos artesanales “ad – hoc”

  Una vez localizados, las opciones más comunes para su tratamiento suelen ser:

  - Pasar por alto el valor erróneo y continuar con el análisis.
  - Filtrar toda la columna asociada al valor erróneo.
  - Filtrar el registro que contiene el valor erróneo.
  - Reemplazar el valor erróneo por uno correcto.

Transformación de datos:

En la fase posterior de data mining se ejecutarán una serie de técnicas o algoritmos sobre los datos. Estos algoritmos, normalmente exigen que los datos se encuentren en un formato determinado, puede ser que sólo acepte datos cualitativos o sólo acepte datos cuantitativos.

Existen múltiples formas de transformación de datos, por su importancia y uso, a continuación, se encuentran algunas que son de las más usuales.

- Numerización: Transformar los atributos cualitativos en datos cuantitativos.
- Discretización: Transformar los atributos cuantitativos en datos cualitativos.

Lo importante es que el atributo al que se transforma, sea uno equivalente.

Creación de características:

Consiste en la creación de un nuevo atributo en los datos, normalmente calculado como función de otros atributos ya existentes.

Por ejemplo, se puede crear el atributo "sueldo bruto anual" en el caso que tengamos "sueldo bruto" y "número de pagas".

- Normalización: consiste en la transformación del rango de valores que toma un determinado atributo

  El caso más común de normalización es la ***normalización lineal uniforme***, que transforma los valores de un atributo a una escala en el intervalo (0,1) utilizando, para ello, la siguiente fórmula, que requiere conocer los valores mínimo y máximo que toma el atributo a normalizar.

$$valor\ normalizado = \frac{valor\ inicial - valor\ mínimo}{valor\ máximo - valor\ mínimo}$$

Reducción de dimensionalidad:

Las técnicas de reducción de dimensionalidad buscan reducir el número de atributos sobre los que realizan el análisis.

Existen múltiples técnicas, pero la más utilizada quizá es ***el análisis de componentes principales (PCA, Principal Components Analisys).*** Esta técnica proyecta los atributos recogen la mayor parte de la información relevante de los originales, pero con la ventaja adicional de que se eliminan las posibles redundancias y dependencias que había.

Paso 3. Data Mining:

En la etapa de data mining, se pueden aplicar diferentes técnicas para resolver diferentes tipos de problemas.

A los diferentes tipos de problemas que se pueden resolver mediante el uso de técnicas de data mining se les conoce como ***tareas.*** Estas tareas se pueden clasificar en:

- Tareas predictivas: Son aquellas tareas de data mining que se utilizan para predecir el valor desconocido de uno o varios atributos para uno o varios registros de la vista minable.
    - Clasificación: Consiste en encontrar un modelo que, aplicado a un nuevo ejemplo sin clasificar, lo catalogue dentro de un conjunto definido de clases.
    - Regresión: Esta tarea es similar a la de clasificación, con la diferencia de que, en este caso, el atributo a predecir no es cualitativo sino cuantitativo.
- Tareas descriptivas: Son aquellas tareas de data mining que generan modelos que, de alguna forma, describen los datos. El objetivo es, por lo tanto, describir los datos existentes.

- Clustering: La tarea de clustering pretende dividir una población heterogénea de objetos en grupos homogéneos, denominados ***clusters***, de forma que los objetos de cada grupo sean muy similares entre sí.

  También se le llama ***segmentación*** o ***agrupamiento***.
- Asociación: Pretende encontrar reglas que muestran la relación que existe entre los distintos atributos de los datos analizados.
- Detección de atípicos: la tarea de detección de atípicos consiste en encontrar objetos que, dentro de un conjunto, manifiesten características significativamente diferentes a las del resto de los objetos del conjunto.

Paso 4. Interpretación y evaluación:

Esta fase es importante dado que no todos los modelos obtenidos han de ser precisos y cumplirán sus objetivos.

Estos modelos deben ser precisos, comprensibles e interesantes. Normalmente, para evaluar un modelo, se usan un enfoque que consiste en reservar un pequeño subconjunto de los datos (conjunto de prueba) que se utilizará para validar el modelo construido con el resto de los datos (conjunto de entrenamiento).

Otra técnica, que es un poco más avanzada, es la técnica de validación cruzada ***n-fold cross validation.*** Para validar un modelo, se elige aleatoriamente **n%** de los datos como conjunto de prueba y con el **(100 – n) %** restante como conjunto de entrenamiento. Este proceso no se realiza una sola vez sino un ***n – número*** de veces. Un valor muy habitual para n, es 10. Una vez conocida la calidad de los modelos, es necesario expresarlos en términos del área de aplicación.

Es importante contar con técnicas de visualización de dichos modelos para que sean comprendidos e interpretados por los expertos de cada dominio.

CAPÍTULO IV CONTEXTUALIZACIÓN METODOLÓGICA

1. Construir un conjunto de datos de estudiantes, que contemplen información de tipo socioeconómico e historial académico.

El estudio está enfocado a encontrar patrones que nos ayuden a determinar causales de bajo rendimiento académico y sobre los estudiantes activos en el programa de Administración de Empresas que figuran para el periodo 2019-I.

Según el reglamento estudiantil de la Universidad del Cauca, capitulo VIII, articulo 6, dice que un estudiante se considera en situación de bajo rendimiento académico, cuando haya perdido alguna asignatura que curse en calidad de repitente por primera vez.

Para efectos de este trabajo, se adiciona al concepto de "bajo rendimiento académico", a los estudiantes que no se encuentran matriculados en las asignaturas que corresponde según el periodo en que ingresaron.

El primer paso de Minería de datos es tomar los datos y almacenarlos de forma conjunta en un repositorio llamado data warehouse.

La data warehouse para efectos de practicidad y como es mencionado en la teoría, se realizó en uno de los programas del paquete de office, Excel, en él se consignó la información socioeconómica, personal y académica de dichos estudiantes; tomada desde SIMCA y los resultados de las pruebas saber 11°.

A tener en cuenta:

- Para el inicio del periodo de 2019-1, el número de estudiantes activos en el Programa de administración de empresas era de 496 estudiantes.
- Los estudiantes que ingresaron en el periodo 2018-1, apenas se encuentran cursando (en caso de que estén al día) tercer semestre, estos estudiantes no contribuyen de manera óptima al estudio, pues no han cursado más de dos semestres y por eso, no es posible que haya perdido una materia más de dos veces, que es una de las clases que se investigará en

este estudio. Algo parecido ocurre con los estudiantes que ingresaron al periodo 2018-2, que se encuentran en segundo semestre y 2019-1 que iniciaron primer semestre.

El número de estudiantes de ingreso 2018-1 es de 44

El número de estudiantes de ingreso 2018-2 es de 35

El número de estudiantes de ingreso 2019-1 es de 45

Para un total de 124 estudiantes

En ese sentido, el número de estudiantes sobre los cuáles se realizará el proceso de EDM será las 496 iniciales menos los 124 que no cumplen con las condiciones para el mismo, para un total de ***372 estudiantes***. Será sobre estos últimos sobre quienes se aplique la EDM, en ese sentido, sólo habrá que completar la data warehouse con sus datos.

Para poder dar inicio al almacenamiento de los datos que se utilizan en el estudio, primero se establece el valor de la muestra, ya que como se planteó en el marco teórico, una de las formas más eficientes de establecer los datos que son necesarios para realizar el estudio. Entonces a continuación, se muestra cómo se calculó la muestra de los estudiantes activos a primer semestre de 2019, sobre los cuales se aplicó minería de datos.

Para calcular la muestra se utilizó la siguiente fórmula:

$$n = \frac{N * Z_a^2 * p * q}{e^2 * (N-1) + Z_a^2 * p * q}$$

Donde:

n = Tamaño de la muestra (El valor que buscamos)
N = Tamaño de la población
Z = Parámetro estadístico que depende del nivel de confianza
e = Error de estimación máximo aceptado
p = Probabilidad de que ocurra el evento estudiado (éxito)
q = Probabilidad de que el evento estudiado no ocurra (fracaso)
Entonces:

N = 372 Z = 1,96 p = 0,5 q = 0,5 e = 0,04

$$n = \frac{(372) * (1{,}96^2) * (0{,}5) * (0{,}5)}{(0{,}04)^2 * (372 - 1) + (1{,}96)^2 * (0{,}5) * (0{,}5)}$$

$$n = 229{,}9 \ \ldots n = 230$$

El tamaño de la muestra es igual a 230.

Para obtener los datos del ICFES, se envió la solicitud de parte de la coordinación a la dependencia de archivo de la Universidad del Cauca, quienes comunicaron de forma verbal que el manejo de ese tipo de datos es exclusivo de DARCA. Es por eso, que para obtener los resultados de los estudiantes del programa de Administración de Empresas, se tuvo que buscar directamente en la página del ICFES (http://www2.icfesinteractivo.gov.co-resultados-saber2016-web/pages/publicacionResultados/autenticacion/autenticacion.jsf?id-1#No-back-button), desde este sitio, se obtuvo gran parte de los resultados que obtuvieron los estudiantes en este examen. Y fue con estos datos con los que se trabajó en este proyecto.

Los datos que se obtuvieron desde SIMCA, se obtuvieron accediendo al sistema desde la coordinación del Programa de Administración de Empresas y analizando una a una, la situación académica de los estudiantes activos en el programa en el primer periodo del año 2019.

Los datos que se utilizaron para este estudio fueron:

Socio económicos:

- Municipio de procedencia (donde realizó su bachillerato). Este dato se obtuvo de los resultados del ICFES.
- Tipo de colegio (público o privado). Este dato se obtuvo de los resultados del ICFES.

Datos personales:

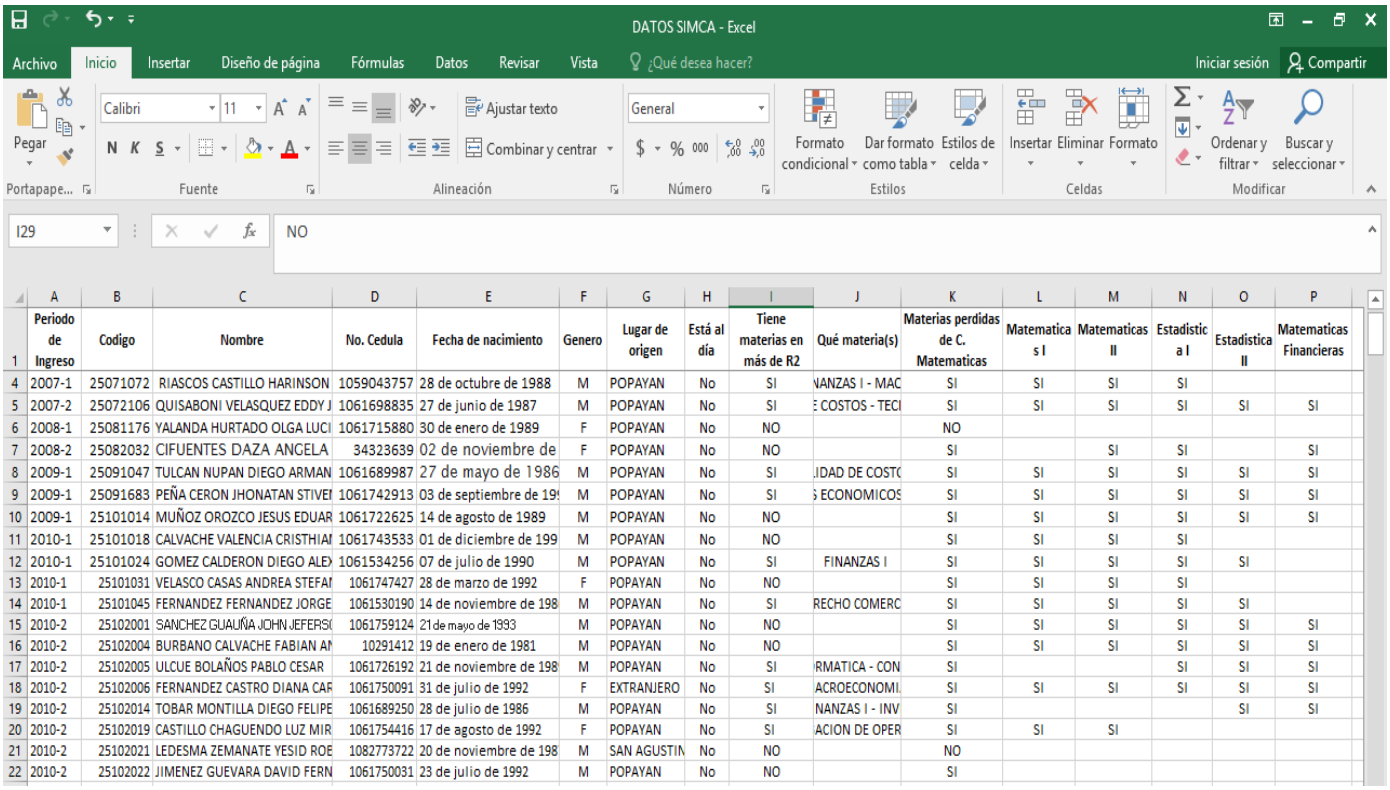
- Nombre completo (SIMCA).

- Numero de documento de identidad (SIMCA).
- Fecha de nacimiento (SIMCA).

Académicos:

- Resultados por materia, en las pruebas saber 11° (ICFES).
- Materias no aprobadas (Historia académica SIMCA).

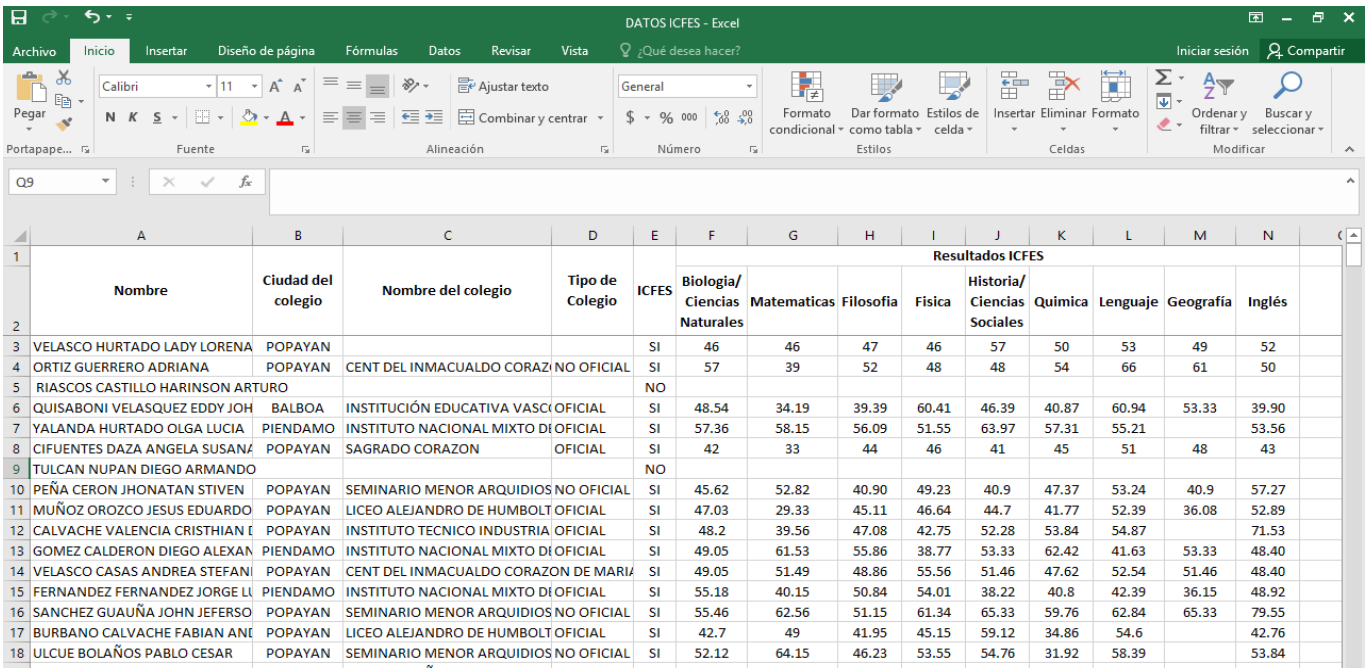
Los datos obtenidos de SIMCA se plasmaron en un libro de Excel, de la siguiente manera:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Periodo de Ingreso	Codigo	Nombre	No. Cedula	Fecha de nacimiento	Genero	Lugar de origen	Está al día	Tiene materias en más de R2	Qué materia(s)	Materias perdidas de C. Matematicas	Matematicas I	Matematicas II	Estadistica I	Estadistica II	Matematicas Financieras
4	2007-1	25071072	RIASCOS CASTILLO HARINSON	1059043757	28 de octubre de 1988	M	POPAYAN	No	SI	NANZAS I - MAC	SI	SI	SI	SI		
5	2007-2	25072106	QUISABONI VELASQUEZ EDDY J	1061698835	27 de junio de 1987	M	POPAYAN	No	SI	E COSTOS - TECI	SI	SI	SI	SI	SI	SI
6	2008-1	25081176	YALANDA HURTADO OLGA LUCI	1061715880	30 de enero de 1989	F	POPAYAN	No	NO		NO					
7	2008-2	25082032	CIFUENTES DAZA ANGELA	34323639	02 de noviembre de	F	POPAYAN	No	NO		SI		SI	SI		SI
8	2009-1	25091047	TULCAN NUPAN DIEGO ARMAN	1061689987	27 de mayo de 1986	M	POPAYAN	No	SI	IDAD DE COSTO	SI	SI	SI	SI	SI	SI
9	2009-1	25091683	PEÑA CERON JHONATAN STIVEI	1061742913	03 de septiembre de 19	M	POPAYAN	No	SI	S ECONOMICOS	SI	SI	SI	SI	SI	SI
10	2009-1	25101014	MUÑOZ OROZCO JESUS EDUAR	1061722625	14 de agosto de 1989	M	POPAYAN	No	NO		SI	SI	SI	SI	SI	SI
11	2010-1	25101018	CALVACHE VALENCIA CRISTHIAI	1061743533	01 de diciembre de 199	M	POPAYAN	No	NO		SI	SI	SI	SI		
12	2010-1	25101024	GOMEZ CALDERON DIEGO ALEX	1061534256	07 de julio de 1990	M	POPAYAN	No	SI	FINANZAS I	SI	SI	SI	SI	SI	
13	2010-1	25101031	VELASCO CASAS ANDREA STEFAI	1061747427	28 de marzo de 1992	F	POPAYAN	No	NO		SI	SI	SI	SI		
14	2010-1	25101045	FERNANDEZ FERNANDEZ JORGE	1061530190	14 de noviembre de 198	M	POPAYAN	No	SI	RECHO COMERC	SI	SI	SI	SI	SI	
15	2010-2	25102001	SANCHEZ GUAUÑA JOHN JEFERS(	1061759124	21 de mayo de 1993	M	POPAYAN	No	NO		SI	SI	SI	SI	SI	SI
16	2010-2	25102004	BURBANO CALVACHE FABIAN AI	10291412	19 de enero de 1981	M	POPAYAN	No	NO		SI	SI	SI	SI	SI	SI
17	2010-2	25102005	ULCUE BOLAÑOS PABLO CESAR	1061726192	21 de noviembre de 198	M	POPAYAN	No	SI	RMATICA - CON	SI			SI	SI	SI
18	2010-2	25102006	FERNANDEZ CASTRO DIANA CAI	1061750091	31 de julio de 1992	F	EXTRANJERO	No	SI	ACROECONOMI	SI	SI	SI	SI	SI	SI
19	2010-2	25102014	TOBAR MONTILLA DIEGO FELIPE	1061689250	28 de julio de 1986	M	POPAYAN	No	SI	NANZAS I - INV	SI				SI	SI
20	2010-2	25102019	CASTILLO CHAGUENDO LUZ MIR	1061754416	17 de agosto de 1992	F	POPAYAN	No	SI	ACION DE OPER	SI	SI	SI			
21	2010-2	25102021	LEDESMA ZEMANATE YESID ROE	1082773722	20 de noviembre de 198	M	SAN AGUSTIN	No	NO		NO					
22	2010-2	25102022	JIMENEZ GUEVARA DAVID FERN	1061750031	23 de julio de 1992	M	POPAYAN	No	NO		SI					

Imagen 1 Muestra del libro de EXCEL para los datos de SIMCA

Los datos obtenidos de ICFES se plasmaron en un libro de Excel, de la siguiente manera:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1						Resultados ICFES								
2	Nombre	Ciudad del colegio	Nombre del colegio	Tipo de Colegio	ICFES	Biologia/ Ciencias Naturales	Matematicas	Filosofia	Fisica	Historia/ Ciencias Sociales	Quimica	Lenguaje	Geografía	Inglés
3	VELASCO HURTADO LADY LORENA	POPAYAN			SI	46	46	47	46	57	50	53	49	52
4	ORTIZ GUERRERO ADRIANA	POPAYAN	CENT DEL INMACUALDO CORAZ	NO OFICIAL	SI	57	39	52	48	48	54	66	61	50
5	RIASCOS CASTILLO HARINSON ARTURO				NO									
6	QUISABONI VELASQUEZ EDDY JOH	BALBOA	INSTITUCIÓN EDUCATIVA VASC	OFICIAL	SI	48.54	34.19	39.39	60.41	46.39	40.87	60.94	53.33	39.90
7	YALANDA HURTADO OLGA LUCIA	PIENDAMO	INSTITUTO NACIONAL MIXTO DI	OFICIAL	SI	57.36	58.15	56.09	51.55	63.97	57.31	55.21		53.56
8	CIFUENTES DAZA ANGELA SUSANA	POPAYAN	SAGRADO CORAZON	OFICIAL	SI	42	33	44	46	41	45	51	48	43
9	TULCAN NUPAN DIEGO ARMANDO				NO									
10	PEÑA CERON JHONATAN STIVEN	POPAYAN	SEMINARIO MENOR ARQUIDIOS	NO OFICIAL	SI	45.62	52.82	40.90	49.23	40.9	47.37	53.24	40.9	57.27
11	MUÑOZ OROZCO JESUS EDUARDO	POPAYAN	LICEO ALEJANDRO DE HUMBOLT	OFICIAL	SI	47.03	29.33	45.11	46.64	44.7	41.77	52.39	36.08	52.89
12	CALVACHE VALENCIA CRISTHIAN D	POPAYAN	INSTITUTO TECNICO INDUSTRIA	OFICIAL	SI	48.2	39.56	47.08	42.75	52.28	53.84	54.87		71.53
13	GOMEZ CALDERON DIEGO ALEXAN	PIENDAMO	INSTITUTO NACIONAL MIXTO DI	OFICIAL	SI	49.05	61.53	55.86	38.77	53.33	62.42	41.63	53.33	48.40
14	VELASCO CASAS ANDREA STEFANI	POPAYAN	CENT DEL INMACUALDO CORAZON DE MARIA		SI	49.05	51.49	48.86	55.56	51.46	47.62	52.54	51.46	48.40
15	FERNANDEZ FERNANDEZ JORGE LU	PIENDAMO	INSTITUTO NACIONAL MIXTO DI	OFICIAL	SI	55.18	40.15	50.84	54.01	38.22	40.8	42.39	36.15	48.92
16	SANCHEZ GUAUÑA JOHN JEFERSO	POPAYAN	SEMINARIO MENOR ARQUIDIOS	NO OFICIAL	SI	55.46	62.56	51.15	61.34	65.33	59.76	62.84	65.33	79.55
17	BURBANO CALVACHE FABIAN AND	POPAYAN	LICEO ALEJANDRO DE HUMBOLT	OFICIAL	SI	42.7	49	41.95	45.15	59.12	34.86	54.6		42.76
18	ULCUE BOLAÑOS PABLO CESAR	POPAYAN	SEMINARIO MENOR ARQUIDIOS	NO OFICIAL	SI	52.12	64.15	46.23	53.55	54.76	31.92	58.39		53.84

Imagen 2 Muestra del libro de EXCEL para los datos de SIMCA

Al final, al unir todos los datos disponibles, se obtuvo el Data Warehouse.

2. Identificar los diferentes factores que afectan el rendimiento académico del estudiante en el pregrado.

Siguiendo el proceso de EDM o Minería de Datos, para establecer los factores que afectan el rendimiento académico del estudiante de pregrado se utilizaron los siguientes pasos.

El segundo paso en el desarrollo de Data Mining, es la ***selección y limpieza de datos***. Para esto, se utilizó el filtrado de registros. El fin de este paso es sólo dejar los registros que cuenten con el total de sus atributos para poder realizar una minería de datos eficiente.

Primero se hizo una reducción de aquellos registros que no tenían la información completa de su examen del ICFES, paso seguido, de los estudiantes a quienes no les apareciera registro de la institución en dónde estudiaron, pues sin esa información, no podríamos saber si provienen de un colegio de tipo OFICIAL o NO OFICIAL.

Aunque se tenían los registros de las personas sobre los cuales se va a trabajar, aún existían demasiados datos innecesarios. El atributo “Nombre” o “Código” no aportaba nada para el estudio. Es verdad que estos atributos ayudaron a mantener un orden durante la recolección de datos, sin embargo, en este paso es necesario reducir los atributos, a los que sean realmente útiles para el estudio.

Los atributos Código, Nombre, Número de Cedula, Lugar de origen, ¿Qué materias en más de R2?, Matemáticas I, Matemáticas II, Estadística I, Estadística II, Matemáticas Financieras, Nombre del Colegio, ICFES, Biología/Ciencias Naturales, Filosofía, Física, Historia/Ciencias Sociales, Química, Geografía e inglés, fueron eliminados de la data warehouse debido a que no eran necesarios en este estudio.

Otros datos fueron usados para hallar atributos que si fueran útiles para el estudio como la fecha de nacimiento y la fecha de ingreso (para saber la edad a la que iniciaron el programa de pregrado).

Por otra parte, los datos de la procedencia del colegio, también son cambiados o reducidos a los departamentos.

El siguiente paso, fue realizar la limpieza de los datos.

Dado que el objetivo del estudio es descubrir qué atributos determinan que un atributo de clase sea SI o NO, en este sentido, la respuesta es de tipo Nominal por lo tanto cada uno de los registros deben encontrarse de la misma manera, se procede entonces a la discretización de los datos.

Según el reglamento estudiantil de la Universidad del Cauca, capitulo VIII, articulo 6, dice que un estudiante se considera en situación de bajo rendimiento académico, cuando haya perdido alguna asignatura que curse en calidad de repitente por primera vez.

Para efectos de este trabajo, se adicionará al concepto de “bajo rendimiento académico”, a los estudiantes que no se encuentran matriculados en las asignaturas que corresponde según el periodo en que ingresaron.

El objetivo del estudio era descubrir qué atributos determinan que un atributo de clase sea “SI” o “NO”, en este sentido, la respuesta es de tipo Nominal por lo tanto cada uno de los registros deben encontrarse de la misma manera, es por eso que se realizó la discretización de todos los datos.

De la siguiente manera para cada atributo:

- **Edad a la que ingresó:**
  - Mayor de edad
  - Menor de edad
- **Género:**
  - Femenino
  - Masculino
- **Lugar de origen del Colegio:**
  - Cauca
  - Costa pacífica caucana
  - Huila
  - Nariño
  - Popayán
  - Putumayo
  - Valle del Cauca
- **Tipo de Colegio (ICFES):**
  - Oficial
  - No oficial

- **Matemáticas:** Este atributo hace referencia al resultado que se obtuvo en el examen ICFES. Para efectos de practicidad en el estudio, se dividió los atributos en rangos, según los resultados obtenidos de la siguiente manera.

Tabla 2 Discretización de datos para atributos "lenguaje" y "Matemáticas"

Rango 1	DE 0 A 30
Rango 2	DE 31 A 40
Rango 3	DE 41 A 50
Rango 4	DE 51 A 60
Rango 5	DE 71 A 80
Rango 6	DE 81 A 100

- **Lenguaje:** De la misma manera que en el atributo “Ciencias sociales”, este atributo hace referencia a los resultados que el estudiante obtuvo en el examen del ICFES y se dividió los resultados en rangos, igual que con dicho atributo.
- **Atributo de clase:** el atributo de clase es en el cual está basado el estudio. En este caso, queremos conocer las causales de bajo rendimiento académico, es por eso que se debe tener en cuenta las características utilizadas para determinar que un estudiante está en bajo rendimiento que para efectos de este estudio son dos. La primera, es si el estudiante se encuentra al día o no, de acuerdo a su periodo de ingreso en contraste con el semestre que debe estar cursando de acuerdo a su plan de estudios; la segunda, es si el estudiante ha perdido una materia más de dos veces.

  Debido a que no se puede correr la información en el programa con dos atributos de clase, se realizó un consenso con ambos atributos para generar un atributo final de la siguiente manera.

Tabla 3 Unificación de atributos de clase

Tiene materias en más de R2	Está al día	Estudiante en bajo rendimiento
NO	SI	NO
NO	NO	SI
SI	NO	SI

Es así como se logró unir los dos atributos de clase, para formar uno sólo don de la respuesta fuera "SI" o "NO". Seguido a esto, pasamos los datos al programa obtener el modelo.

3. Diseñar el modelo predictivo utilizando técnicas de minería de datos.

Continuando con el proceso, ***el paso número tres*** es el que lleva su nombre ***data mining***. Para llevar a cabo este paso se pueden aplicar diferentes técnicas que sirven para resolver diferentes tipos de problemas

A los diferentes tipos de problemas que se pueden resolver mediante el uso de técnicas de data mining se les conoce como ***tareas***. Estas tareas se pueden clasificar en:

- **Tareas predictivas:** que son aquellas tareas de data mining que se utilizan para predecir el valor desconocido de uno o varios atributos para uno o varios registros de la vista minable.
    - **Clasificación:** la tarea de clasificación consiste en encontrar un modelo que, aplicado a un nuevo ejemplo sin clasificar, lo clasifique dentro de un conjunto predefinido de clases.
    - **Regresión:** la tarea de regresión es similar a la de clasificación, con la diferencia de que, en este caso, el atributo a predecir no es cualitativo, sino cuantitativo

- **Tareas descriptivas:** Son aquellas tareas de data mining que generan modelos que, de alguna forma, describen los datos.
    - **Clustering:** La tarea del Clustering pretende dividir una población heterogénea de objetos en grupos homogéneos, denominados clusters, de forma que los objetos de cada grupo sean muy similares entre sí.
    - **Asociación:** Pretende encontrar reglas que muestran la relación que existe entre los distintos atributos de los datos analizados.
    - **Detección de atípicos:** la tarea de detección de atípicos consiste en encontrar objetos que, dentro de un conjunto, manifiesten características significativamente diferentes a las del resto de los objetos del conjunto.

Según el objetivo general de esta práctica, podemos decir que la tarea que se debe desarrollar, es de tipo ***descriptiva***, pues por medio de ella podríamos generar un modelo que nos describa la información que le brindemos; Sin embargo, el último de los objetivos específicos, dice "Diseñar el modelo predictivo utilizando técnicas de minería de datos" es decir, que necesitamos un modelo que nos asocie las causales de bajo rendimiento en los estudiantes activos en el programa de administración de empresas para el primer semestre de 2019 y que a la vez nos sirva para predecir, cuándo los estudiantes estén en riesgo potencial de pasar a tener bajo rendimiento.

En ese sentido, el modelo a usar, es de ***tipo predictivo*** y no descriptivo.

El modelo de tipo predictivo a utilizar será, ***la clasificación***, pues dentro de los datos que encontramos en la vista minable, se encuentra el lugar de procedencia del estudiante, factor que queremos saber si es clave o no en los estudiantes que ingresan a estudiar al programa de administración de empresas.

Una vez seleccionada la tarea de minería de datos que se va a ejecutar, solo resta pasar los datos a WEKA.

WEKA es un software de uso libre desarrollado por la Universidad de WAIKATO, en Australia, que se utiliza para ejecutar algoritmos de minería de datos en un tiempo muy corto y eficiente.

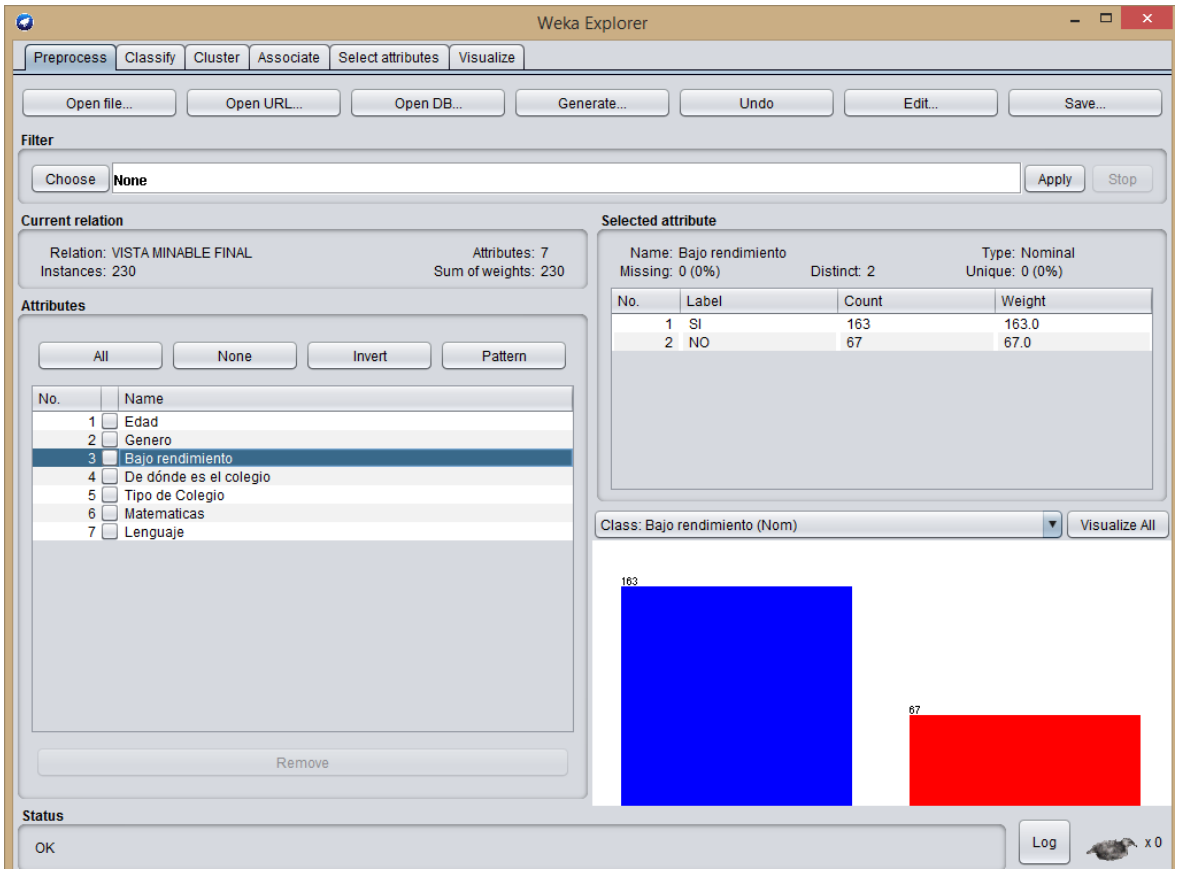


Imagen 3 Datos en WEKA

Paso cuatro del proceso de minería de ***datos Interpretación y evaluación***. La evaluación de resultados obtenidos en la etapa de data mining, es una tarea muy importante. No siempre se cumplirá con los objetivos esperados para que resulte en conocimiento.

El método de evaluación depende del tipo de tarea desarrollada durante la etapa de data mining; sin embargo, el método más comúnmente utilizado es el de n-fold cross

La evaluación de los modelos obtenidos en la fase anterior es una tarea crucial. O todos los modelos obtenidos cumplirán con las características esperadas en ellos para poder obtener conocimiento. En particular, los modelos han de ser precisos, comprensibles e interesantes.

Aunque depende de la tarea de data mining en cuestión, en general, para evaluar un modelo se suele utilizar un enfoque consistente en reservar un pequeño subconjunto de los datos (conjunto de prueba) que se utilizará para validar el modelo construido con el resto de los datos (conjunto de entrenamiento). Este enfoque se conoce como validación simple.

Una técnica algo más avanzada, a la vez más utilizada, es la técnica de validación cruzada n-fold cross validation. En este caso, para validar un modelo, se elige aleatoriamente el n% de los datos como conjunto de prueba y con el (100-n) % restante, como conjunto de entrenamiento, se construye el modelo. Este proceso no se realiza una única vez, sino que se repite n veces, variando cada vez, los conjuntos de prueba y entrenamiento. Un valor muy habitual para n es 10 (10- fold cross validation).

Este método de validación, está incluido en el programa WEKA por lo que él mismo se encarga de hacerlo todo y no hay que hacerlo de forma manual.

Para encontrar el modelo utilizamos el algoritmo de árboles de decisión J48 en WEKA y el resultado obtenido es el siguiente:

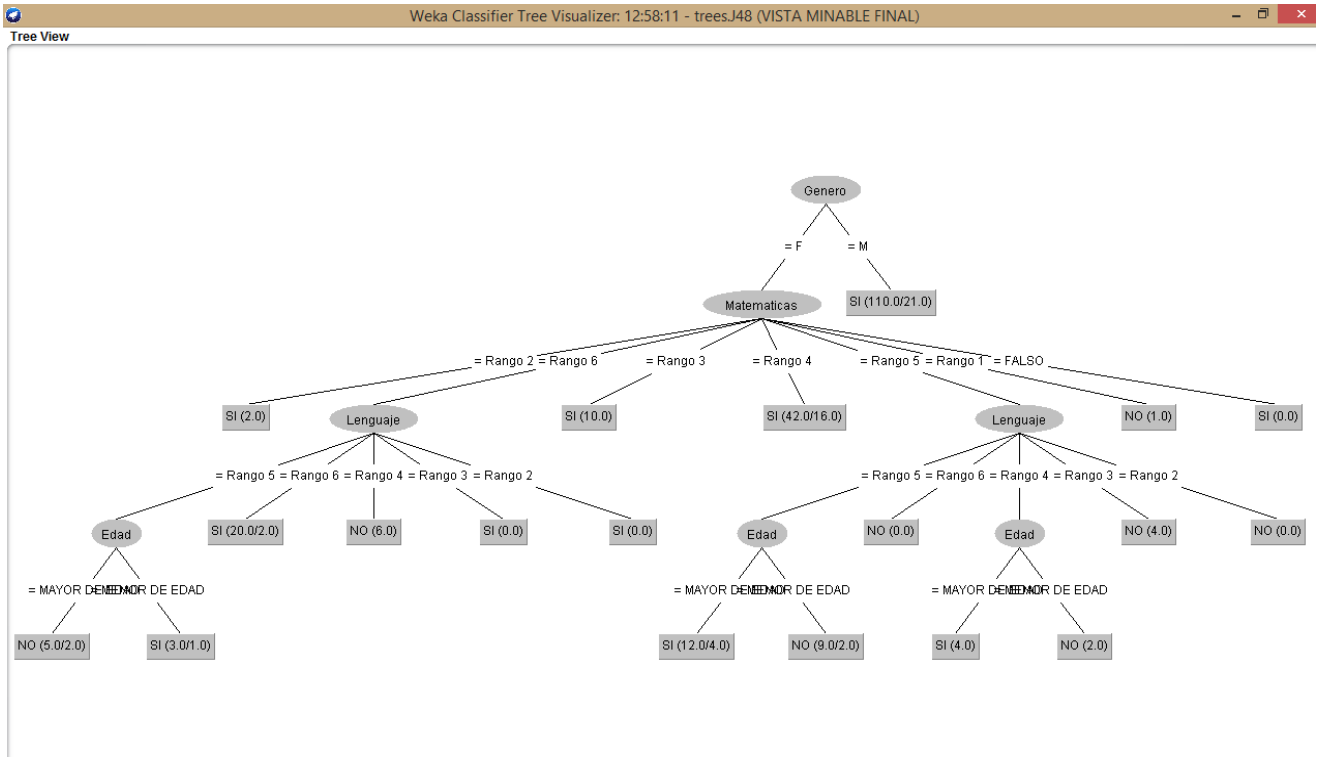


Imagen 4 Árbol de decisión obtenido en WEKA

1. Genero = F y Matemáticas = Rango 2
2. Genero = F y Matemáticas = Rango 6 y Lenguaje = Rango 5
3. Genero = F y Matemáticas = Rango 6 y Lenguaje = Rango 5 y Edad = MENOR DE EDAD
4. Genero = F y Matemáticas = Rango 6 y Lenguaje = Rango 6
5. Genero = F y Matemáticas = Rango 6 y Lenguaje = Rango 3
6. Genero = F y Matemáticas = Rango 6 y Lenguaje = Rango 2
7. Genero = F y Matemáticas = Rango 3: SI
8. Genero = F y Matemáticas = Rango 4: SI
9. Genero = F y Matemáticas = Rango 5
10. Genero = F y Matemáticas = Rango 5 y Lenguaje = Rango 5
11. Genero = F y Matemáticas = Rango 5 y Lenguaje = Rango 5 y Edad = MAYOR DE EDAD
12. Genero = F y Matemáticas = Rango 5 y Lenguaje = Rango 4
13. Genero = F y Matemáticas = Rango 5 y Lenguaje = Rango 4 y Edad = MAYOR DE EDAD
14. Genero = M.

CAPÍTULO V. CONCLUSIONES Y SUGERENCIAS

El estudio buscaba obtener un listado de causas que afectan el rendimiento académico de manera negativa.

Existe una fuerte relación entre los estudiantes que egresaron de colegios públicos y los que en la universidad pierden materias del componente de matemáticas.

El programa de Administración de Empresas tiene una fuerte influencia en el departamento del Cauca, con estudiantes provenientes del 45% de los municipios del Cauca.

Para el caso de las pruebas ICFES, el resultado en matemáticas, muestra una fuerte relación entre las personas que pierden materias del componente de matemáticas y las personas que obtienen un puntaje de ICFES en este componente por debajo de 60 puntos.

Los estudiantes que provienen del municipio del Bordo, suman un número cercano a estudiantes que provienen del municipio de Timbio y mayor que los estudiantes que provienen de Piendamó, esto es interesante en la medida que los dos últimos municipios quedan a minutos de distancia de Popayán, mientras que el Bordo queda a 3 horas. Esto configura un gran cambio para estos estudiantes y una variable interesante a analizar en próximos estudios

Con la cantidad de datos disponibles, se obtuvo que las dos causales que influyen de manera mayoritaria en el bajo rendimiento académico son, el género, el puntaje obtenido en las pruebas ICFES y la edad del estudiante en el momento de ingreso.

Además, se planteó un modelo predictivo con 15 reglas de decisión que marcan, los casos en los que un estudiante es propenso a tener bajo rendimiento.

El modelo es un 70,4% exacto, lo que quiere decir que hubo una gran influencia sobre el estudio, del hecho de no obtener el 100% de los datos del estudiante. Si bien el modelo muestra que quienes son más propensos a tener bajo rendimiento académico son los hombres y las mujeres que cumplan ciertas características a la vez en cuanto a las variables, resultados del ICFES en matemáticas y lenguaje, además de si es o no mayor de edad. Esto es sólo una parte de la información. Si se tuvieran datos como el estrato, si tiene hijos, la cantidad de dinero de la que dispone mensualmente, si vive en calidad de arrendatario, si utiliza transporte público, etc. Podríamos tener una mayor exactitud en el modelo.

Durante el desarrollo de este estudio, el factor humano tuvo una gran participación dado que afectó la calidad de los datos y la pronta disponibilidad de ellos, es por eso que fue necesario obtener los datos directamente de la plataforma del ICFES, proceso que finalmente retardó enormemente la etapa de recolección de los mismos.

Muchos estudiantes presentaron el examen de estado como menores de edad, por lo que para el momento tenían tarjeta de identidad que en la mayoría de casos para los nacidos antes del año 2.000, era completamente diferente al número de cédula.

Como se mencionó en el párrafo anterior, la calidad de los datos personales de los estudiantes en SIMCA, no es la mejor. En muchos casos, el número de documento de identidad, el lugar de origen y de más datos, no eran correctas. Es por eso que se tuvo que buscar desde otras fuentes estos datos que se suponía estaban disponibles en la plataforma de la universidad.

En 230 casos, estas limitaciones fueron superadas con éxito y se pudo recoger información completa de este número de estudiantes; sin embargo, el número total de estudiantes sobre los cuáles se debía realizar el estudio era de 372. Es por eso que se contó

sólo con una muestra de la cantidad de estudiantes y no con el total como hubiera sido óptimo.

Para nuevos proyectos de minería de datos del programa de administración de empresas, sería recomendable contar con una base de datos propia del programa, que contenga datos mucho más amplios para poder realizar un estudio de minería de datos mucho más exacto y asertivo para la generación de conocimiento que es el objetivo final de todo proceso de minería de datos.

REFERENCIAS BIBLIOGRAFICAS

- ARTUNDUAGA, M. (2005). Factores asociados al rendimiento académico y a la deserción en la Universidad. [Tesis doctoral] Universidad Complutense de Madrid
- ASTIN, A. (1993). Assessment for Excellence: The philosophy and practice of assessment and evaluation in higher education. Nueva York, , Oryx
- CROMBACH, L. (1968). Psicología Educativa. Paidós, México, D.F
- ESCUDERO MUÑOZ, J. M. (1990). "Tendencias actuales en la investigación educativa: los desafíos de la investigación crítica". En: Qurrículum. No. 2. pp. 3-25.
- ESCUDERO, E. (1981). Selectividad y rendimiento académico de los universitarios: condicionantes psicológicos y educacionales. Universidad de Zaragoza. Aragón.
- FUEYO, B. (1990). El fracaso escolar: entre la ideología y la impotencia. Revista Educadores. Nº 153. Madrid p 25 – 40
- GONZÁLEZ M, PILAR. (1989). Aplicación del LISREL al análisis del rendimiento estudiantil. Facultad de Ciencias Universidad de Los Andes, Lima, Perú.
- LARA TORRALBO, Juan (2014) Minería de Datos. Centro de Estudios Financieros Universidad UDIMA. P. 9 – 115.
- MARADONA & CALDERÓN. (2004). Una aplicación del enfoque de la función de producción en educación. Revista de Economía y Estadística, Universidad Nacional de Córdoba, XLII. P. 36 – 37
- NATEK, S., & ZWILLING, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 41, 6400–6407.
- PHRIDVIRAJ, M. & GURURAO, C. (2014). Data mining – past, present and future – a typical survey on data streams. Procedia Technology (12), 255 – 263.

- QUEZADA, R. (1991). Guía para evaluar el aprendizaje teórico y práctico. Editorial Limusa, México.
- RODRIGUEZ RODRIGUEZ, Jorge (2010) Fundamentos de Minería de Datos. Universidad Distrital Francisco José de Caldas. P 23 – 91.
- SANTALLA, (2003) Guía para la elaboración formal de reportes de investigación. P 26
- SHAPIRO K. (2011). Bajo rendimiento escolar: Una perspectiva desde el desarrollo del sistema nervioso.
- TIMARÁN PEREIRA, Ricardo; CALDERÓN ROMERO, Andrés & JIMÉNEZ TOLEDO, Javier (2013). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. En: Ventana Informática. No.28 (ener. – jun.). Manizales (Colombia): Facultad de Ciencias e Ingeniería, Universidad de Manizales. P. 31- 47, ISS: 0123-9678.
- TSAI, H. (2013). Knowledge management vs. data mining: Research trend, forecast and citation approach. Expert Systems with Applications, 40, 3160-3173.

WEBGRAFÍA

- http://portal.unicauca.edu.co/versionP/documentos/acuerdos/acuerdo-no-002-de-1988
- https://consultasrc.registraduria.gov.co:28080/ProyectoSCCRC/
- http://www2.icfesinteractivo.gov.co/resultados-saber2016-web/pages/publicacionResultados/autenticacion/autenticacion.jsf?id=1#No-back-button
- https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908
- https://jedm.educationaldatamining.org/index.php/JEDM/article/view/6
- http://educationaldatamining.org/