DESCUBRIMIENTO DE CONOCIMIENTOS EN LA BASE DE DATOS ACADÉMICA DE LA UNIVERSIDAD AUTÓNOMA DE MANIZALES APLICANDO REDES NEURONALES

JAIRO ELIAS GUTIERREZ

Maestría en Gestión y Desarrollo de Proyectos de Software
Facultad de Ingenierías
Universidad Autónoma de Manizales
Manizales, Colombia
2012

Esta tesis es enviada a la Facultad de Ingenierías de la Universidad Autónoma de Manizales para optar al título de Magister en Gestión y Desarrollo de Proyectos de Software.

Autor:
Jairo Elías Gutierrez Londoño

Asesor Técnico:
Msc. Javier Hernández Cáceres
Docente Investigador UNAB

Universidad Autónoma de Manizales
Facultad de Ingenierías
Maestría en Gestión y Desarrollo de Proyectos de Software
Manizales, Colombia
2012

Resumen

Contexto: La educación superior en Colombia es un derecho de todos y es responsabilidad del Ministerio de Educación Nacional garantizarlo. Sin embargo, existen múltiples problemas que representan un reto a la hora de hacer efectivo este derecho. A los problemas propios del sistema educativo como son la baja calidad, la pertinencia y los bajos índices de cobertura, se suman otros problemas tales como la deserción y la poca vocación generados por factores propios del sistema de educación superior y factores externos relacionados con los estudiantes y su entorno social.

Objetivo: Este proyecto se propone generar conocimiento útil para encontrar posibles causas del problema de la deserción estudiantil de la Universidad Autónoma de Manizales a partir de las grandes cantidades de información académica generada por los sistemas transaccionales de la universidad.

Metodología: La primera fase de este proyecto propone verificar investigaciones previas acerca del problema de la deserción académica y otros problemas asociados a la educación superior a nivel nacional e internacional. Durante la segunda etapa se lleva a cabo el proceso de extracción de la información académica de los sistemas transaccionales de la Universidad Autónoma de Manizales; Y en la fase final se ejecuta el análisis de la información mediante técnicas de minería de datos las cuales son aplicadas de acuerdo al análisis realizado y las técnicas definidas después del proceso de extracción.

Resultados: Este proyecto pretende generar como resultados una fuente de datos consolidada y normalizada de información académica de la Universidad Autónoma de Manizales que sea utilizable durante la ejecución de este proyecto y en proyectos futuros de minería de datos e inteligencia de negocios, un framework de minería de datos con una implementación básica para este proyecto, pero extensible a gran variedad de nuevos problemas y técnicas, y por último un conjunto de conclusiones acerca del problema de la deserción a partir de la información académica y las técnicas de minería de datos aplicadas.

Palabras Clave: Minería de Datos, Almacén de Datos, Deserción Académica, Framework de Técnicas de Minería de Datos, Educación Superior, Cobertura, Pertinencia

Abstract

Context: In Colombia the educations is a right for all and must be guarantee by the National Minister of Education. However this right is complicated due to many problems. The educative system's problems such as Relevance and low index of coverage must be added to others problems such desertion and lack of vocation generated by external factors related to students and your social environment.

Objective: This project is oriented to generate a consolidate data source for find possible causes of student desertion problem in Universidad Autónoma de Manizales from the academic information generated by de transactional systems of the University.

Methodology: The first phase of project propose verify previous investigations about the education academic desertion problem and other problems on national and international institutions. By the second phase is executed the extraction of information process from the transactional systems on the University. On the last phase is executed the process of analysis of information through of the data mining techniques selected on previous phases.

Results: This project intended generate a consolidated and normalized data source useful by this project and futures projects about of data mining or business intelligence. Other result is a data mining framework with a basic implementation by this project, but extensible for variety of problems and needs. By last a set of conclusions about the academicals desertion problem on the Universidad Autónoma de Manizales.

Keywords: Data Mining, Datawarehouse, Academicals Desertion, Framework, Coverage, Relevance

Tabla de contenido

DESCUBRIMIENTO DE CONOCIMIENTOS EN LA BASE DE DATOS ACADÉMICA DE LA UNIVERSIDAD AUTÓNOMA DE MANIZALES APLICANDO REDES NEURONALES

Tabla de Ilustraciones

Índice de Tablas

Introducción

Educación Superior en Colombia

El Ministerio de Educación Nacional de Colombia define la educación como el proceso de formación permanente, personal, cultural y social y se fundamenta en la concepción integral de las personas. Las leyes colombianas establecen el marco normativo de la educación superior como un servicio público y encargan al estado la responsabilidad de velar por la calidad de dicho servicio y para que este cumpla con sus objetivos fundamentales de brindar una formación intelectual, cultural y física a las personas que acceden a este servicio.

Hasta antes de los primeros estudios de Schlutz (1961) y Becker (1961) acerca del capital humano la educación era considerada un gasto más que una inversión. Fue en este momento que se identificó la necesidad de las sociedades y los gobiernos de promover la educación como la herramienta encargada de llevar progreso a las naciones, permitiendo a quienes acceden a ella salir al mercado laboral con la oportunidad de aspirar a empleos mejor remunerados.

Problemas de la Educación Superior

Pese a que la Ley 30 de 1992 establece que la educación superior en Colombia debe ser de libre acceso a quienes demuestren cumplir las condiciones y tener la capacidad académica para acceder a ella, según las cifras presentadas por el Ministerio de Educación, en el año 2011 apenas se alcanzo una cobertura de 39% en educación superior en Colombia lo cual representa que al menos tres millones de jóvenes aptos para realizar estudios de pregrado no se encuentran estudiando debido en gran parte a los problemas que se describen durante este capitulo.

Cobertura

El Centro de Estudios sobre Desarrollo Económico de la Universidad de los Andes define la cobertura como el porcentaje de población entre los 18 y 25 años que se encuentra estudiando en un programa de pregrado en cualquiera de sus niveles. Según este análisis en el año 2010 esta cifra se situaba en el 31% mientras que en el año 2000 apenas alcanzaba el 23%. Pese a que la cifra parece alentadora, no lo es del todo ya que en este porcentaje están incluidos por ejemplo los estudiantes del SENA, universidades de carácter privado e instituciones de nivel intermedio, con lo cual la cifra de estudiantes matriculados en las universidades públicas del país apenas si pasa los 600mil, o cual corresponde apenas a un 12% del total de cobertura. Y si ver estas cifras es desalentador, es aún más compararlas con los números alcanzados en las

áreas rurales del país, donde la cobertura de la educación superior llega apenas a un 5%.

Pese a que los indicadores de educación superior en Colombia se encuentran en la actualidad cercanos al 40%, por encima de países como México (20%) y Brasil (30%), está muy por debajo de los países con mejores índices en Latinoamérica: Uruguay (64%) y Cuba (100%), pero el panorama es aún más desalentador si incluimos cifras económicas que afectan la cobertura educativa, por ejemplo en Colombia pese a que la constitución estipula que el principal responsable de la educación superior es el estado, este destina apenas un 4.8% del producto interno bruto (PIB) con este fin mientras que las familias proveen en promedio de un 28% de sus ingresos para los gastos de educación, esto sin contar con los gastos de sostenimiento; este 28% comparado con países como Francia o Suecia donde los gobiernos cubren la mayor parte de los gastos de educación superior incluyendo sostenimiento, y las familias apenas destinan para matriculas entre un 4 y un 6% de sus ingresos netos. Si a esto le agregamos que en Colombia apenas el gobierno destina un 4% de su PIB para créditos educativos comparado con el Reino Unido donde la cifra es cercana al 16%, se podría decir que es obligación del estado reforzar su participación en el proceso educativo cumpliendo no solo con lo que lo obliga la ley, sino viendo a la educación como una inversión rentable socialmente y con un alto índice de retribución.

Pertinencia

La pertinencia se refiere al grado de utilidad, oportunidad y eficacia con que los resultados de un programa educativo impactan a la sociedad en un sentido amplio o bien a sectores específicos [1]. En general cuando se habla de pertinencia se suele hablar del nivel de actualización que tienen los programas, contenidos, planes curriculares, estrategias de enseñanza y materiales, pero en realidad la pertinencia no se trata únicamente de actualidad, se refiere más al grado en el cual un egresado de determinado programa académico encaja en las necesidades que requiere una localidad, una región o un país.

El Ministerio de Educación Nacional de Colombia mide la pertinencia con base en la cantidad de egresados que se vinculan al mundo laboral, y los salarios promedio de los egresados, en el periodo comprendido entre el año 2001 y 2010, el porcentaje de egresados de programas académicos de pregrado (niveles técnicos y profesionales) que lograban vincularse a los sectores formales de la economía era cercano al 72% con salarios en promedio de 600 dólares en su primer empleo. Si bien esta cifra parece menor, se debe tener en cuenta que alcanzaba 2 salarios mínimos colombianos. Otros países latinoamericanos alcanzan cifras de empleo de 91% entre los recién egresados con salarios que alcanzan los 2600 dólares en promedio, y la diferencia es aún más notoria si se compara con países del primer mundo en Europa como Alemania donde un recién graduado alcanza ingresos promedio de 4500 dólares. [2].

Calidad

Para brindar garantías en los temas de evaluación, certificación y acreditación de la calidad de la educación superior en Colombia, se ha creado dentro del sistema educativo el denominado Sistema de Aseguramiento de la Calidad de la Educación Superior conformado por los organismos, las acciones y las estrategias que aplican desde el proceso mismo de creación y establecimiento de una institución de educación superior, hasta el desempeño del profesional que egresa del sistema (Ministerio de Educación Nacional Colombia, 2011).

Algunos de los criterios de calidad exigibles para un programa de pregrado, es preparar a los estudiantes para dar alternativas de solución a un problema dado y brindarles las herramientas suficientes para seleccionar la más indicada dependiendo del caso según los recursos y el entorno, a diferencia de los programas técnicos y tecnológicos que se enfocan exclusivamente en presentar las herramientas, procesos y procedimientos que están estandarizados y normalizados [3]. No necesariamente un criterio de calidad es la uniformidad, es posible encontrarse un mismo programa en diferentes universidades con contenidos y métodos diferentes. Sin embargo, lo importante es que siempre se encuentre dentro de los estándares de calidad impuestos por la entidad que los rige.

Dentro del proceso educativo, la calidad puede medirse en múltiples variables, calidad de la enseñanza, del aprendizaje o de las herramientas y metodologías. También se debe aclarar que el proceso de enseñanza está definido como un proceso colectivo, pero el proceso de aprendizaje depende de la actitud y las capacidades propias de cada individuo, es por estas razones que la calidad no debe medirse de forma objetiva con una calificación de x magnitud, sino que deben valorarse otros atributos que son más subjetivos [3].

Deserción

Se define la deserción como la comparación numérica entre la matrícula inicial menos el número de egresados de último año. Se atribuyen como causas de la deserción factores internos y externos de la universidad así como factores intrínsecos de cada estudiante. Sin embargo, gran cantidad de estudios sobre deserción coinciden en que las principales causas en Colombia son asociadas, las clases menos favorecidas tienen menores oportunidades de ingresar a la educación superior y cuando lo logran requieren mayor esfuerzo para culminar sus estudios de forma exitosa [4].

Históricamente se han realizado estudios desde múltiples perspectivas, y los estudios iniciales se enfocaban en factores propios de cada individuo, los cuales sugieren la deserción como el debilitamiento de las intenciones iniciales y del grado de persistencia de cada individuo. Un segundo grupo de estudios conciben la deserción desde un punto de vista social, toman en cuenta atributos como el grado de adaptación de cada estudiante al medio académico e institucional y concluyeron que el grado de adaptación del estudiante al medio ambiente influye directamente en su rendimiento académico. Un tercer grupo de investigaciones se enfoca en la

perspectiva institucional y estos asocian el fenómeno de la deserción a los beneficios estudiantiles, calidad de la educación, disponibilidad de recursos y actividades extracurriculares [5].

Existen varios periodos críticos durante la trayectoria académica de un estudiante en los cuales es más propenso a la deserción. El primer momento es el contacto inicial con la universidad y con el programa académico cuando se forman las primeras impresiones acerca del entorno académico y del entorno en general. El segundo periodo crítico ocurre más o menos al mismo tiempo, pero está relacionado con el cambio de un ambiente pequeño y controlado como lo es la educación media a un ambiente mucho más grande y rodeado de mayor diversidad. Esto ocurre sobre todo en grandes universidades donde los alumnos en muchos casos deben desplazarse a ciudades diferentes a la de origen [6].

Históricamente se ha señalado al bajo rendimiento académico como la principal causa de la deserción, pero en gran cantidad de los casos el bajo rendimiento académico se debe a otras causas tales como baja calidad de los docentes, falta de preparación en la educación media, problemas económicos, baja calidad de la educación y desmotivación de los estudiantes por problemas vocacionales a la hora de elegir el programa académico. [4].

1. Referente Teórico

Esta sección contiene un conjunto de términos que serán usados durante el desarrollo de este documento y que son importantes para un entendimiento correcto del mismo.

1.1. Minería de Datos (Data Mining)

Se trata de un proceso de descubrimiento de información a partir de una base de datos. Estos nuevos conocimientos deben ser válidos, útiles y sobre todo comprensibles. Se trata de una técnica de extracción de información para convertirla en conocimiento y apoyar el proceso de toma de decisiones en casi todas las áreas de la ciencia, las finanzas y la industria.

Algunos sistemas que son sólo parcialmente conocidos, producen una cantidad inmensa de datos que con frecuencia contienen valiosa información que puede resultar muy útil y ser vista como vetas de oro por los ojos de un ejecutivo de una corporación.

Las dimensiones de las base de datos grandes (montañas) y sus velocidades de crecimiento, hacen muy difícil para un humano su análisis y la extracción de alguna información importante; aún con el uso de herramientas estadísticas clásicas esta tarea es casi imposible.

El descubrimiento de conocimiento en base de datos (KDD) combina las técnicas tradicionales con numerosos recursos desarrollados en el área de la inteligencia artificial. En estas aplicaciones el término "Minería de Datos" (Data Mining) ha tenido más aceptación[7].

1.1.1. Técnicas de Minería de Datos

Las técnicas de minería de datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. La minería de datos toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva[8].

Existen gran variedad de técnicas de minería de datos cada una de ellas aplicable en diferentes contextos según las necesidades de cada problema. Entre las técnicas más utilizadas se encuentran las siguientes:

1.1.1.1. Clustering

Agrupan datos dentro de un conjunto de clases partiendo de un criterio de distancia o similitud dentro de cada clase. Su utilización ha proporcionado significativos resultados en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. Un problema relacionado con el análisis de clúster es la selección de factores en tareas de clasificación, debido a que no todas las variables tienen la misma importancia a la hora de agrupar los objetos.

1.1.1.2. Algoritmos Genéticos

Los algoritmos genéticos imitan la evolución de las especies mediante mutaciones, reproducciones y selectividad de los más fuertes. De esta forma los algoritmos genéticos son ampliamente utilizados en procesos de optimización debido a que su principal objetivo es la búsqueda de la efectividad y la eficacia.

1.1.1.3. Arboles de Decisión

Se trata de una técnica de aprendizaje supervisado, es decir, que necesita un conjunto de datos objetivo y la técnica se encarga de generar las reglas para llegar a dicho objetivo. Le deben su nombre a que representan rutas de decisiones en forma de árbol, en los cuales cada nodo es una decisión y cada rama es una ruta por la cual se llega a un objetivo.

1.1.1.4. Redes Neuronales

Las redes neuronales artificiales son sistemas de procesamiento de información, cuyo funcionamiento y estructura está basado en las redes neuronales biológicas. Se componen de un conjunto de elementos más simples denominados nodos o neuronas conectadas entre sí por un valor numérico modificable conocido como peso.
Debido a su fundamentación, las redes neuronales presentar grandes semejanzas con el cerebro; por ejemplo, ambos son capaces de aprender de la experiencia, generalizar a partir de casos anteriores y casos nuevos, y abstraer características relevantes a partir de un gran número de entradas que representan información irrelevante.
Esta técnica de inteligencia artificial en los últimos años se ha convertido en uno de los instrumentos de uso frecuente para detectar categorías comunes en los datos, debido a que son capaces de aprender patrones complejos y evaluar características de los datos

1.2. Inteligencia Artificial

Se trata de la ciencia que se encarga de diseñar y construir maquinas inteligentes, especialmente programas de computadora. Su principio fundamental es construir máquinas y programas que imiten las capacidades de aprendizaje y de realización de actividades que llevan a cabo entes inteligentes tales como seres humanos y animales usando técnicas de lenguajes matemáticos y lógicos.

Las primeras ideas de inteligencia artificial se remontan a Aristóteles[1], quien fue el primero que logró describir un conjunto limitado de reglas que describen el funcionamiento de la mente al momento de racionalizar algunos tipos de problemas. Sin embargo, fue solo hasta 1936 que Turing diseñó el primer dispositivo capaz de realizar cómputo sin asistencia y en 1943 Warren McCulloch y Walter Pitts presentaron su modelo de neuronas artificiales, el cual se considera el primer trabajo del campo, aun cuando todavía no existía el término el cual fue acuñado solo hasta 1956 por parte de John McCarty[2].

Las máquinas y programas basados en inteligencia artificial son ampliamente utilizados en múltiples ramas del conocimiento y la tecnología. Se destacan aplicaciones de reconocimiento de patrones, construcción de videojuegos, reconocimiento de texto, sistemas expertos, experimentación genética, etc. Las aplicaciones de inteligencia artificial son utilizadas en gran variedad de campos como la economía, salud, educación, manufactura, etc.[9].

[1] Filosofo, lógico y científico de la Antigua Grecia cuyas ideas han ejercido una enorme influencia sobre la historia intelectual de Occidente por más de dos milenios
[2] prominente informático que recibió el Premio Turing en 1971 por sus importantes contribuciones en el campo de la Inteligencia Artificial.

1.3. Framework de Software

Se trata de una porción de código fuente o programa construido con un fin genérico, pero lo más importante es que puede ampliarse su funcionalidad agregando código fuente o componentes adicionales fácilmente. Un framework puede ser diseñado y construido para dar solución a un problema específico, pero siempre pensando en el principio básico de la reutilización. Es por esto que en la mayoría de los casos un framework de software está constituido por un conjunto de librerías que contienen los componentes de bajo y medio nivel facilitando el trabajo a los desarrolladores de software que deben enfocarse en desarrollar la solución específica al problema en el que se encuentran trabajando.

En la actualidad, el término framework se ha extendido a otros conceptos además del software, incluyendo en su alcance frameworks de conceptos, de prácticas, de criterios, de reglas que hacen énfasis en su principio básico de reutilización. Es así como se habla de frameworks para gestión de calidad; por ejemplo, en los que éste se compone de un conjunto de plantillas y reglas que apoyan los procesos de calidad en las compañías.

2. Estrategia Metodológica

2.1. Preguntas de Investigación

En esta sección se plantean las preguntas que motivaron el desarrollo de esta tesis y que se planean responder durante el desarrollo de la misma.

2.1.1. Pregunta de Investigación 1 (PI1)

¿Es posible mediante el uso de tecnología informática consolidar los datos académicos y socioeconómicos de los estudiantes de la Universidad Autónoma de Manizales proveniente de fuentes de información heterogéneas y poco normalizadas?

2.1.2. Pregunta de Investigación 2 (PI2)

¿Se puede encontrar información pertinente que facilite la toma de decisiones acerca del fenómeno de la deserción estudiantil en la Universidad Autónoma de Manizales a partir de los datos académicos con los que cuenta en la actualidad la oficina de registro académico?

2.1.3. Pregunta de Investigación 3 (PI3)

Mediante el uso de técnicas de inteligencia artificial ¿es posible generar nuevo conocimiento que sea valioso al momento de realizar un análisis del problema de la deserción estudiantil en la Universidad Autónoma de Manizales?

2.2. Metodología

PI1, se realiza una investigación acerca de las técnicas de consolidación de información de fuentes heterogéneas disponibles en la industria y la educación. Durante la inspección se detecta que la técnica más usada para consolidación de datos se denomina ETL (Extracción, Transformación y Carga). Un Almacén de Datos provee una vista unificada y verificada de todos los datos operacionales de una compañía, a través de la integración de múltiples fuentes de datos, un Almacén de Datos utiliza como base fundamental un proceso de ETL que se encarga de cargar el Almacén de Datos a partir de las fuentes heterogéneas[10].

Teniendo en cuenta la utilidad de un proceso de ETL, se incluye como objetivo construir un Almacén de Datos que facilitará el posterior análisis de los datos ya consolidados. Para esto se debe diseñar el Almacén de Datos y luego se define el proceso de ETL a aplicar.

El proceso de ETL se realizará apoyado en la herramienta rapidminer[3]. Se trata de una aplicación de código libre que incluye utilidades tales como conectores con diferentes fuentes de datos (Bases de datos, documentos, archivos de texto, hojas de cálculo, etc.), utilidades de conversión, transformación y limpieza basadas en funciones matemáticas, estadísticas, algorítmicas, etc.; y por último facilita la carga de datos a través de conectores que permiten generar la nueva información a múltiples destinos tales como documentos de texto, bases de datos y hojas de cálculo. El siguiente gráfico[11] describe de forma general los pasos que se tuvieron en cuenta a la hora del realizar el proceso de ETL y generar el Almacén de Datos.

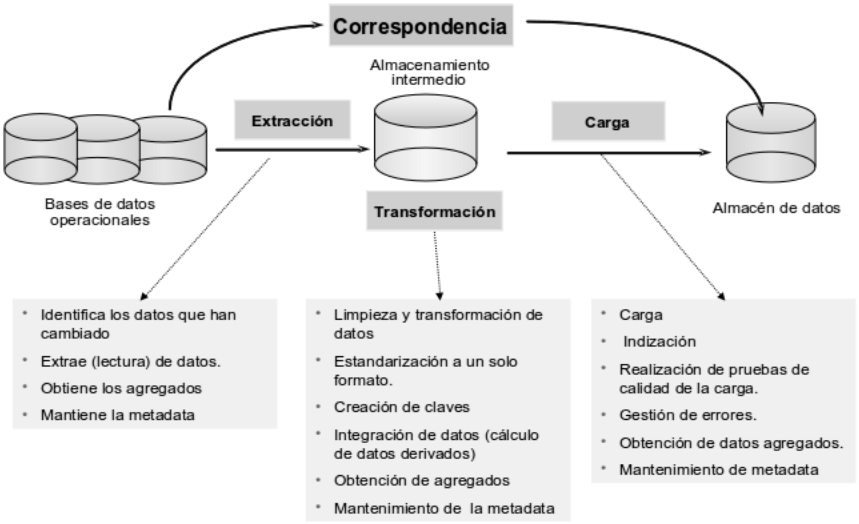


Ilustración 1. Proceso de ETL, tomado de www.carlosproal.com

[3] Herramienta informática usada para procesos de análisis de información y minería de datos desarrollada en la universidad de Dortmund en el año 2001.

PI2, Los datos académicos provenientes de los sistemas transaccionales al igual que en cualquier sistema OLTP [4]clásico no son analizables a simple vista. Se trata de sistemas construidos para soportar un objetivo de negocio y la operación del mismo. Sin embargo, si a dichos datos se les aplica un proceso de consolidación y limpieza (ETL), se puede generar nueva información a la cual se le pueden aplicar diferentes procesos de análisis con el fin de generar nuevo conocimiento.
Durante el desarrollo de este trabajo, la información generada a partir del proceso de ETL, se analiza mediante el uso de técnicas estadísticas, con el fin de identificar variables que pueden tener mayor impacto sobre el fenómeno de la deserción con el objetivo de encontrar algunas conclusiones acerca de la deserción, pero con el objetivo fundamental de crear una base pre-analizada sólida que facilite la búsqueda de nuevo conocimiento a través de técnicas avanzadas e tecnología informática.

PI3, La primera fase de análisis mediante inteligencia artificial será dedicada a analizar el estado del arte acerca de proyectos que usen inteligencia artificial, técnicas usadas, contextos en los cuales se utiliza cada una y casos de éxito aplicados a problemas similares a la deserción académica, con el fin de encontrar la técnica más adecuada a la problemática estudiada.
Como segundo punto se diseña e implementa un framework de software que permita aplicar la técnica de inteligencia artificial seleccionada al conjunto de datos generados como resultado del ETL. Por último se aplica la técnica de minería seleccionada y se evalúan los resultados del proceso.
Durante el proceso de minería de datos se aplica la metodología CRISP-DM, la cual está descrita en términos de un modelo de proceso jerárquico consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): tarea genérica, tarea especializada, e instancia de procesos. La metodología CRISP-DM esta descrita como un modelo estructurado en seis fases [11]algunas de las cuales son bidireccionales, lo cual indica que se puede dar una revisión parcial o total de fases anteriores.

[4] **O**n**l**ine **T**ransaction **P**rocessing, método de procesamiento de transacciones generalmente construido en arquitectura cliente servidor usados en las compañías para soportar las transacciones de en sus procesos y sistemas de negocio.

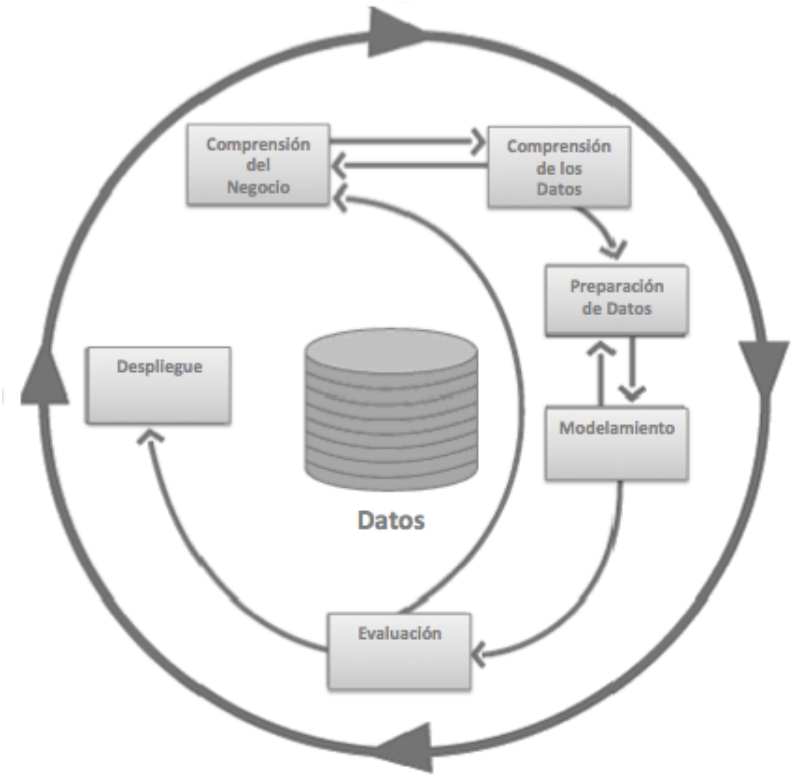


Ilustración 2. Fases metodología CRISP-DM

3. Antecedentes

Encontrar las causas de la deserción académica en la educación superior y por ende su solución han sido históricamente un reto de investigación que ha sido abordado desde múltiples perspectivas por gran cantidad de investigadores, llegando en cada uno de ellos a gran variedad de conclusiones. A partir de las conclusiones de dichos trabajos se ha llegado a un conjunto de modelos que plantean la deserción desde múltiples puntos de vista; cada uno con un grado de validez y que pueden ser complementarios entre sí:

3.1. Modelos Sicológicos

Estos modelos señalan que son algunos rasgos específicos de la personalidad de cada individuo los que lo hacen más o menos propenso a la deserción académica. Fishbein y Ajzen proponen la *Teoría de la Acción Razonada* que analiza el comportamiento como actitudes en respuesta a objetos específicos, considerando normas subjetivas que guían el comportamiento hacia esos objetivos y el control percibido sobre ese comportamiento. Estos autores señalan que la 'intención de tomar la acción' es determinada por dos factores: primero, 'actitud hacia tomar la acción', y segundo la 'norma subjetiva'. La norma subjetiva se refiere a cómo se espera que el individuo se comporte en la sociedad, la cual es determinada por una evaluación de la expectativa. En el caso de la decisión de desertar o permanecer se ve influida por: conductas previas, actitud acerca de la deserción o permanencia y normas subjetivas acerca de

estas acciones. En consecuencia, según estos autores [12]la deserción es el resultado del debilitamiento de las intenciones iniciales.

3.2. Modelos Sociológicos

En cuanto a los primeros estudios realizados desde la sociología[13], basado en la "teoría del suicidio" de Durkheim, sugiere que la deserción es el resultado de la falta de integración del estudiante a su entorno de educación superior. En este estudio se argumenta como la falta de integración del individuo trae como consecuencia un bajo rendimiento académico, descontento personal y falta de compromiso con la institución.

3.3. Modelos Económicos

Las primeras investigaciones desde el campo de la economía, hicieron énfasis en los modelos costo-beneficio, una aproximación poco profunda.
En el primer modelo, el individuo compara los beneficios económicos de estudiar con los beneficios de actividades alternativas como trabajar para decidir si abandona o no sus estudios. En el segundo modelo se intenta identificar grupos de estudiantes en riesgo de deserción por factores económicos, para evitar por medio de subsidios directos evitar que estos abandonen. En estas primeras investigaciones económicas, las variables económicas son consideradas de control y no factores determinantes o de riesgo.

Sin embargo con el tiempo, las investigaciones desde la economía han tenido una posición más fuerte frente a la deserción, estudiando la interacción entre los diferentes grupos de factores individuales, académicos, institucionales y socioeconómicos que inciden en el fenómeno de la deserción. Es así como empleando diferentes indicadores de rendimiento académico, junto con características individuales de la persona encontrando que una variable fundamental para explicar el nivel de deserción es el nivel académico de los padres: a mayor educación de los padres, menor el nivel de riesgo de deserción[14].

3.4. Estudios Sobre Deserción en Colombia

Colombia no ha estado exento del fenómeno de la deserción. A los grandes problemas de cobertura y calidad del sistema educativo nacional en los niveles superiores se ha sumado históricamente el problema de la deserción, y es por esta razón que se han realizado estudios acerca de dicho fenómeno. Sin embargo, la mayoría de estos han sido esfuerzos aislados de universidades o programas académicos específicos. A continuación se resaltan algunos de los más importantes.

En el estudio [15]realizado por la Universidad Cooperativa de Colombia seccional Santa Marta, que cuestiona el supuesto que dice que la deserción es la selección natural del proceso académico y trata de identificar los principales factores no

académicos asociados al fenómeno de la deserción. Como por ejemplo, falta de recursos económicos, estudiantes que laboran, o que tenga personas a su cargo. En este estudio se concluye que los factores socioculturales explican en gran parte el fenómeno de la deserción en esta universidad en específico.

En el artículo[16] se investiga la deserción en los programas de ingeniería de la Universidad Nacional de Colombia, haciendo énfasis en el hecho que esta Universidad es pública y la deserción implica un desperdicio de recursos. Encuentra una relación entre deserción y calidad de la educación, en contraste con los estudios anteriores que le dan más peso a los factores académicos propios del estudiante, y concluye que mejorar la calidad de la educación es un buen mecanismo para ahorrar recursos del estado.

Un estudio más general [17] sobre la equidad social en el acceso a la educación y los factores de la permanencia en la universidad pública en el país. El estudio tuvo en cuenta variables nunca antes estudiadas tales como habilidad académica, la edad, el sexo y otras variables relacionadas con el hogar. Los resultados obtenidos de este estudio muestran que el acceso a las universidades públicas colombianas está determinado por el resultado en las pruebas de estado "ICFES", que a su vez es determinado por condiciones estructurales como las características de la familia del estudiante.

Una investigación realizada en la universidad Militar Nueva Granada en Colombia [4] realizó un análisis del problema de la deserción que tenía como objetivo consolidar la información proveniente de estudios previos sobre el fenómeno y realizar un análisis de las conclusiones de cada uno de ellos. Este estudio logró dividir en tres grandes grupos los factores que hacen a un estudiante más propenso a la deserción: factores externos a la universidad, factores propios de la universidad y factores intrínsecos del estudiante. De esta forma se brinda a futuros investigadores una base de trabajo sobre variables y características propias del fenómeno de la deserción.

Un estudio realizado en la Universidad Pontificia Bolivariana de Medellín [18], indica que otro de los grandes factores que afecta la deserción académica en la Educación Superior en esta institución está relacionada con la preparación de los estudiantes en la educación media. Se logró identificar que los estudiantes cumplían con las competencias en diferentes niveles dependiendo del tipo de institución de la cual provenían. Esto generaba en los primeros semestres de pregrado grupos muy heterogéneos con altos índices de deserción.

En términos generales, los estudios realizados en Colombia sobre el fenómeno de la deserción contienen estimaciones empíricas de funciones de permanencia, y frecuentemente se focalizan en instituciones individuales lo que dificulta la generalización de los resultados y conclusiones.

3.5. Deserción y Tecnología Informática

3.5.1. Descubrimiento de Perfiles de Deserción en la Universidad de Nariño

Proyecto llevado a cabo por el grupo de investigación GRIAS de la Universidad de Nariño en Pasto, Colombia. El objetivo del proyecto era detectar perfiles de bajo rendimiento académico y posibilidades de deserción a través de técnicas de minería de datos aplicadas sobre la base de datos académica histórica de dicha universidad[19]. Durante la ejecución de la investigación usaron como base la herramienta TariyKDD desarrollada por el grupo de investigación en proyectos anteriores. En TariyKDD se encuentran implementados los algoritmos de minería de datos: Apriori, FPGrowth y EquipAsso, para la tarea de Asociación y los algoritmos C4.5 y Mate-tree para la tarea de clasificación. Sin embargo, durante este proyecto las conclusiones provenían en su mayoría de las reglas de clasificación generadas a partir de los algoritmos de asociación.

Durante la ejecución del proyecto se encontraron las principales dificultades a la hora de realizar el pre procesamiento (Limpieza y Transformación) de los datos, debido a que estos no eran de la calidad esperada debido a las múltiples fuentes de información que se registraron. Las conclusiones del proyecto arrojaron algunas reglas de asociación donde se identificaban algunos grupos de estudiantes más propensos que otros a la deserción.

3.5.2. Descubrimiento de Conocimiento en información académica para determinar factores de deserción y retención de estudiantes

El estudio realizado en la Universidad Politécnica de Valencia en España identificó la gran cantidad de datos académicos con los cuales contaba la universidad. Dichos datos eran generados a partir de los múltiples sistemas transaccionales encargados de soportar la operación académica de la institución. Sin embargo, la gran cantidad de datos dificultaban el proceso de análisis con respecto a fenómenos relacionados con el rendimiento académico, problemas de deserción y la retención de estudiantes.
El estudio se realizó en 5 fases que se describen a continuación: en la fase 1 denominada pre-procesamiento se realizó la recolección y limpieza de los datos con el fin de facilitar las fases posteriores del proceso. En la fase 2 denominada análisis estadístico, se identificaron a partir de técnicas de estadística las variables candidatas a un proceso de análisis más exhaustivo. En la fase 3 se utilizó la técnica de minería de datos denominada redes neuronales con el fin de encontrar subconjuntos de información con características comunes a partir de la información obtenida de las bases de datos transaccionales. Por último se aplicaron algoritmos de reglas de decisión, las cuales se encargaron de encontrar relaciones entre las variables identificadas en cada uno de los subgrupos generados a partir de la aplicación de redes neuronales.

Este proyecto generó como resultados un software genérico de minería de datos que puede seguir siendo aplicado a través del tiempo con el fin de generar nuevo conocimiento y adicionalmente identificó relaciones iniciales entre las variables que más impactan en la deserción o retención de estudiantes las cuales generan representaciones basadas en árboles de decisión.

3.5.3. Sistema Para la Prevención de la Deserción en Instituciones de Educación Superior SPADIES

El Ministerio de Educación Nacional como entidad encargada de velar por la calidad de la educación en Colombia, y en este caso puntual la educación superior, lanzó en el 2002 como parte del Sistema Nacional de Información de la Educación Superior (SNIES) el SPADIES (Sistema Para la Prevención de la Deserción en Instituciones de Educación Superior).

SPADIES tiene como objetivo principal consolidar y ordenar la información socio-académica de los estudiantes que ingresan a la educación superior en Colombia con el fin de facilitar el seguimiento del proceso educativo y de establecer los factores determinantes del fenómeno de la deserción. SPADIES pone a disposición del público en general la información consolidada, para apoyar el proceso de toma de decisiones acerca del problema de la deserción.

En la actualidad la plataforma está instalada en un 99% de las instituciones de educación superior del país y consolida la información de cerca de 3 millones de estudiantes. El sistema permite a cada institución realizar el seguimiento y monitoreo, además de generar alarmas ante la presencia de factores de riesgo de deserción dentro de la población estudiantil.

La plataforma SPADIES [20]fue diseñada basada en técnicas estadísticas que permiten identificar estudiantes en riesgo de deserción. Sin embargo, en la actualidad mediante el uso de técnicas avanzadas de tecnología informática es posible realizar caracterizaciones automáticas en grandes cantidades de datos e identificar relaciones entre las características y variables que facilitan la generación de nuevo conocimiento, ampliando aún más las capacidades de las técnicas estadísticas para facilitar el proceso de toma de decisiones acerca del fenómeno de la deserción.

4. Análisis del Problema

4.1. Dominio del Problema

Las cifras de desercion de los programas de pregrado en la ciudad de Manizales son cercanos al 50% (Fuente SPADIES) y este no es un fenomeno ajeno a la Universidad Autónoma de Manizales y sus programas académicos. El problema es mayor aun cuando observamos que no se han identificado las principales causas de la alta tasa de deserción.

La necesidad de la universidad autónoma de conocer las causas de la deserción académica no es una tarea fácil. Se trata de identificar razones académicas, sociales, económicas o personales que llevan a los alumnos a tomar la decisión de abandonar la universidad. Identificar las causas no es suficiente, se deben proponer soluciones que permitan a un porcentaje más alto de alumnos culminar sus estudios de pregrado y de la misma forma aumentar la demanda de los programas académicos que ofrece la universidad tanto a nivel de pregrado como posgrado.

4.2. Minería de Datos

Para la ejecución del proyecto se utilizó la metodología CRISP-DM [11]. Esta metodología ha sido ampliamente empleada en el diseño, implementación y despliegue de proyectos de minería de datos para múltiples sectores productivos y áreas del conocimiento.

Es fundamental llevar a cabo cada una de las fases de la metodología CRISP-DM dentro del proyecto, debido a que esta describe las actividades que se deben llevar a cabo durante la ejecución del mismo y los entregables que componen cada una de estas.

4.2.1. Fase 1, Entendimiento del Negocio

Se debe llevar a cabo un análisis del negocio que garantice un entendimiento completo del mismo lo cual llevara a cumplir los objetivos del proyecto en el cual nos embarcamos. Se debe comprender correctamente el negocio para garantizar una identificación correcta de las necesidades y cumplir con los objetivos del proyecto.

4.2.1.1. Objetivos del Negocio

Se debe realizar un análisis previo del objetivo del negocio de la Universidad Autónoma de Manizales, y para esto se debe tomar en cuenta la misión y la visión de la institución y de ser posible evaluar los objetivos de algunos programas de pregrado de alto impacto en la región.

En este paso se realiza la evaluación del problema que se desea estudias teniendo en cuenta el comportamiento de este en los últimos años, posibles causas, etc. Con el fin de realizar una evaluación de la situación actual que permita obtener resultados más precisos que permitan la toma de decisiones a partir de resultados reales.

4.2.2. Fase 2, Comprensión de los Datos

Es en esta fase del proyecto en la cual se recolectan, describen y exploran los datos disponibles para la realización del análisis mediante minería de datos. En esta fase se deben obtener los datos académicos de los programas de pregrado de la Universidad Autónoma de Manizales para realizar una exploración y análisis inicial de la calidad de dicha información que permita la selección de las variables a evaluar técnicas más recomendadas de acuerdo a las características identificadas en los datos, etc.

4.2.3. Fase 3, Preparación de Datos

A partir de la información académica recolectada de la Universidad Autónoma, se debe identificar el conjunto de datos que de acuerdo al análisis del problema y la exploración de los datos realizada en fases anteriores puede generar resultados más valiosos. También es importante a la hora de seleccionar los datos tener en cuenta las restricciones de las herramientas seleccionadas, calidad de los datos y la cantidad de información disponible.

Es en esta fase del proyecto donde se lleva a cabo el proceso de extracción, transformación y carga (ETL) de la información. El proceso de ETL se debe realizar en dos fases. Una inicial donde por simple exploración se eliminen los datos que pueden generar resultados inesperados en el proceso, y una segunda fase en la cual se debe utilizar una herramienta automática de limpieza, conversión y formateo de los datos.

La salida de esta fase es un conjunto de información que cumple con las características necesarias para ser analizados mediante el uso de técnicas de minería de datos que permitan encontrar información relevante acerca del problema de la deserción en la Universidad Autónoma de Manizales.

La fase de preparación de los datos es el punto de entrada del proceso de ETL descrito en capítulos posteriores y a partir del cual se genero el almacén de datos sobre el cual se aplico la técnica de minería de datos.

4.2.4. Fase 4, Modelado

Se trata de la fase en la cual se selecciona una metodología de minería de datos y se aplica. Dependiendo de la metodología o herramienta(s) seleccionada(s), es posible que sea necesario volver a la fase de preparación de datos para aplicar ajustes de acuerdo a los requerimientos de la(s) técnica(s) seleccionada(s).

Durante la ejecución del proyecto se planea aplicar dos técnicas de minería de datos, asegurando que los resultados obtenidos en una de ellas puedan ser verificados y validados con la segunda técnica.

El primer paso del proyecto es construir un prototipo software que permita aplicar el algoritmo de inteligencia artificial de Kohonen [21] sobre los datos con el fin de hallar un conjunto de resultados. Se llegó a la conclusión del uso de esta técnica teniendo en cuenta los antecedentes, el tipo de datos a estudiar y la base teórica/práctica del equipo de trabajo.

El segundo paso es aplicar una técnica de minería de datos diferente con cuyos resultados se pueda verificar los resultados obtenidos de aplicar la primera técnica mediante el prototipo. De esta forma es posible garantizar que las conclusiones del análisis de los datos pueden ser mucho más acertadas que aplicando una única técnica.

Para seleccionar una segunda técnica se debe construir un modelo de selección de técnica de minería de datos, teniendo en cuenta para la selección que la técnica se debe ajustar en la mayor medida posible a los datos obtenidos del proceso de preparación de datos realizado en la fase anterior. Esto garantizara que ambas técnicas se aplicaran sobre conjuntos de datos casi idénticos.

4.2.5. Fase 5, Evaluación

Durante esta fase se debe evaluar los resultados obtenidos de las diferentes técnicas, pero adicionalmente el modelo diseñado. Es importante validar que el modelo cubra la mayor cantidad de escenarios encontrados en la evaluación del negocio.
En la primera parte de la evaluación se debe validar el modelo obtenido contra los objetivos de negocio que se definieron como caso de estudio.
La primera fase de evaluación de resultados se realizará mediante el uso de técnicas de minería de datos e inteligencia artificial con el fin de hallar relaciones entre las diferentes variables e identificar como algunas características de las variables afectan directa o indirectamente otras, generando nuevo conocimiento que no podría encontrarse por simple inspección visual o matemática. Los datos generados serán analizados mediante la técnica de árboles de decisión que permiten clasificar la información y predecir relaciones entre las variables. Los árboles de decisión se utilizarán con el objetivo de generar nueva información que pueda complementar o negar los resultados encontrados en fases previas con el uso de técnicas de minería de datos. Los árboles de decisión permitirán además presentar de forma legible los resultados facilitando su análisis y la generación de conclusiones.

4.2.6. Fase 6, Despliegue

El despliegue o puesta en producción del modelo de minería de datos aplicado para este caso se trata de la realización de un informe que incluye los resultados, variables identificadas y el análisis de los resultados. El objetivo de dicho informe es que sea de utilidad para la toma de decisiones académicas en la universidad. En dicho informe deben plasmarse las conclusiones del modelo aplicado sobre el que variables impactan de forma directa o indirecta sobre el problema de la deserción académica.

4.3. Técnica de Minería de Datos

Desde sus inicios, la universidad Autónoma de Manizales ha recopilado gran cantidad de información académica. Esta ha sido recolectada a través de varios sistemas transaccionales que a su vez han ido evolucionando a través del tiempo soportando correctamente la operación académica de la universidad y las necesidades propias de cada momento. Sin embargo, esta gran cantidad de datos y su heterogeneidad ha dificultado la recolección de información que sea propicia para realizar un análisis de los múltiples problemas relacionados como la deserción y la calidad de la educación; en este punto aparece la Minería de Datos como una opción propicia para encontrar conocimiento pertinente en grandes cantidades de datos.

La minería de datos ofrece múltiples opciones y técnicas para diferentes tipos de proyectos, dependiendo de los objetivos del mismo y de la calidad, cantidad y características de los datos:

Clustering: Las técnicas de clustering permiten agrupar datos dentro de características preestablecidas de acuerdo a criterios de distancia o similitud, su utilización a generado resultados satisfactorios en reconocimiento de patrones o modelamiento de

Sistemas; no se considera pertinente el uso de Clustering dentro de este proyecto porque no se pretende determinar grupos o categorías a priori con el fin de no sesgar los resultados del mismo.

Algoritmos Genéticos: Los algoritmos genéticos imitan la evolución de los seres vivos mediante la mutación, reproducción y selección. Debido a sus características los algoritmos genéticos son ampliamente utilizados en problemas que requieren optimización de procesos o de algoritmos. Durante este proyecto no esta concebida la posibilidad de realizar optimización de procesos o actividades, se pretende generar un resultado que permita llegar a la optimización del proceso académico de la universidad, los algoritmos genéticos pueden ser utilizados como un complemento que permita analizar mejor los resultados obtenidos.

Arboles de Decisión: Se trata de una técnica de aprendizaje supervisado en la cual deben conocerse los resultados esperados con el fin de identificar las decisiones y reglas necesarias para llegar a ellos. Los arboles de decisión son fáciles de usar, se comportan relativamente bien con datos discretos y son de fácil interpretación sus resultados. No se consideran una buena opción de generación de conocimiento en este caso debido a que los datos no son lo suficientemente completos y en algunos casos se presentan dispersos.

Redes Neuronales: Técnica de minería de datos ampliamente utilizada en problemas de detección de patrones y detección de características comunes en los datos. La principal característica de las redes neuronales es la capacidad de aprender y generar conocimiento a partir de datos incompletos e incluso paradójicos. Debido a la naturaleza de los datos con los que se cuenta, se considera a las redes neuronales como la técnica mas adecuada para este proyecto. [22]

Existen tres tipos principales de redes neuronales de acuerdo al tipo de aprendizaje utilizado (Supervisado, No Supervisado y Por Corrección), cada una de ellas es aplicable en diferentes casos dependiendo de las necesidades y los objetivos.
Las redes neuronales supervisadas y por corrección requieren un grupo de datos de control contra los cuales verificar los resultados obtenidos con el fin de generar conocimiento, debido a las características del proyecto en el cual no se conocen los resultados esperados no se consideran a las técnicas supervisadas las mas adecuadas. En contraste las técnicas no supervisadas no requieren datos de control para verificar los resultados obtenidos debido a que considera todos los datos de entrada como variables aleatorias a partir de las cuales puede generar conocimiento. Debido al tipo de problemática analizada en este proyecto en la cual se requiere analizar grandes cantidades de información dispersa, posiblemente incompleta y la dificultad para identificar patrones objetivos se consideran mas adecuadas las técnicas no supervisadas.

Entre las técnicas no supervisadas más usadas que cuentan con más soporte teórico y en las cuales posee mayor experiencia el equipo del proyecto se encuentran los mapas auto-organizativos o red neuronal de Kohonen. Se trata de una arquitectura de

sistemas de clasificación no supervisada que permite encontrar individuos de una población que comparten características comunes. Esto los hace ideales para explorar espacios vectoriales en los que se desconoce la estructura de clasificación de los vectores[21].*T. Kohonen* presentó en 1982 un sistema con un comportamiento semejante al del cerebro. Se trataba de un modelo de red neuronal con capacidad para formar *mapas de características* de manera similar a como ocurre en el cerebro. En éste hay neuronas que se organizan en muchas zonas, de forma que las informaciones captadas del entorno a través de los órganos sensoriales se representan internamente en forma de *mapas bidimensionales.*

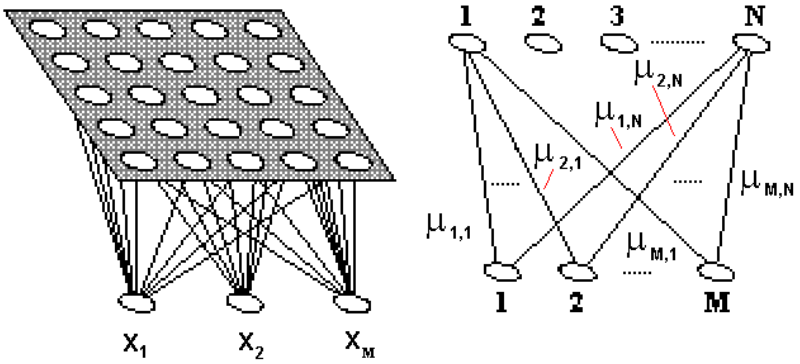


Ilustración 3. Mapa bidimensional de neuronas

La selección de los mapas auto-organizativos SOM de Kohonen se realiza debido a sus principales características. Aprendizaje no supervisado debido a que en este caso no se conoce a priori qué resultados se pueden obtener. Multicapa, lo cual le permite la generación de una mayor cantidad de conocimiento debido a que cada neurona es entrenada de forma independiente y a su capacidad de clasificar individuos de acuerdo a características comunes, dado que el principal objetivo del trabajo actual es encontrar características comunes a los estudiantes propensos a la deserción con el fin de poder predecir este comportamiento.

Además del uso de mapas auto-organizativos se utilizarán arboles de decisión con el fin de confirmar o refutar los resultados obtenidos de la aplicación de los mapas auto-organizativos. Los árboles de decisión permiten clasificar la información y predecir futuros comportamientos de las diferentes variables. Como valor agregado los árboles de decisión facilitan el análisis y presentación de resultados además de la generación de conclusiones.

4.4. Extracción, Transformación y Carga (ETL)

A la hora de aplicar una técnica de minería, los datos deben tener la menor cantidad de ruido posibles para generar resultados de alta confiabilidad que generen valor. En un análisis previo de los datos del proyecto actual se encontró que los datos tienen algunos problemas que es necesario solucionar antes de iniciar el proceso de minería. A esta parte del proceso se le denomina ETL, y durante este proyecto se realizará la

limpieza de los datos y se cargarán a una nueva fuente de datos denominada Almacén de Datos la cual contará con un diseño que se adapta al proceso de minería de datos que se aplicara posteriormente con el fin de encontrar información relevante que permita dar luces acerca del problema de la deserción. El proceso de ETL será preparado, diseñado y ejecutado de acuerdo a los pasos descritos dentro de la metodología CRISP-DM [11].

5. Diseño

En este capítulo se describe el diseño de los diferentes componentes que se definen y construyen durante la ejecución de este proyecto con el fin de resolver las preguntas de investigación.

5.1. Almacén de Datos y ETL

Esta fase tiene como objetivo llevar a cabo una definición del Almacén de Datos, para la cual se debe realizar un proceso de ETL que permita llevar los datos como se encuentran en la actualidad a un formato que permita realizar el proceso de minería de datos de una forma más efectiva. Durante esta etapa se definirá el Almacén de Datos de acuerdo a la metodología bottom-up propuesta por Ralph Kimball, en la cual se definen en detalle partes pequeñas del Almacén de Datos de acuerdo a necesidades puntuales del negocio hasta llegar a un diseño completo [23].

5.1.1. Comprensión de los Datos

Los datos se obtuvieron de la Oficina de Registro Académico de la Universidad Autónoma de Manizales. Los datos fueron entregados en formato Excel (XLS), pero fueron obtenidos de las bases de datos que soportan los sistemas de información de la oficina y que son utilizados para generar información acerca del registro de asignaturas, calificaciones y para consulta de los estudiantes.

Los datos incluyen información obtenida desde el semestre 1 del año 2006 hasta el semestre 1 del año 2011, ordenada semestre a semestre en hojas dentro del libro de Excel.

Además de la información académica enviada por parte de la universidad, se logró obtener también información de los colegios registrados en gran parte del eje cafetero y del país en general mediante la Universidad de Caldas, con el fin de completar los datos enviados desde el registro académico de la Universidad. La información de los colegios se encuentra en un archivo de texto plano.

A través del DANE, se obtuvo información acerca de la situación social, económica y demográfica del país distribuida por departamentos, el plan consiste en cruzar la información académica con la información del DANE mediante el nombre del departamento asociado al registro académico.

5.1.2. Problemas de los Datos

Esta sección describe algunos de los principales problemas encontrados en los datos, este análisis se realiza a partir de una primera inspección visual de los datos pero es

posible que durante la ejecución del proceso de ETL se encuentren problemas adicionales.
Dentro del proceso de exploración visual inicial de los datos se encontraron algunos problemas que pueden afectar de forma directa o indirecta el proceso de análisis mediante minería de datos:

- El código de cada estudiante identifica de forma única a cada uno de ellos dentro de un programa específico. El código contiene datos adicionales acerca del programa, el año/semestre de inicio, la modalidad y la jornada en la cual se encuentra matriculado. Este campo sin embargo por diversas razones ha cambiado en su formato a través del tiempo. Estos cambios se deben fundamentalmente al crecimiento de la universidad y de la oferta académica que brinda. Dentro de los datos enviados por la universidad, se identifican tres formas diferentes de codificación.

    - **XXXXYYZDDDD:** Es la codificación utilizada actualmente, se compone de caracteres distribuidos de la siguiente forma:
      XXXX = 4 caracteres que indican el código del programa académico en el cual se encuentra matriculado el estudiante.
      YY = Año de ingreso a la universidad.
      Z = Semestre del año en el que se matriculo (0 = enero, 1 = Julio).
      DDDD = Consecutivo numérico que identifica de forma única al estudiante.

    - **0XXYZZDD =** Codificación de 8 caracteres utilizada para soportar más programas y alumnos que la descrita anteriormente. Dónde:
      0 = Valor fijo de relleno
      XX = Año de ingreso a la universidad
      Y = Semestre de ingreso a la universidad (0=Enero, 1=Julio).
      ZZ = Consecutivo del estudiante.
      DD = Código del programa académico.

    - **XYYZDD =** Codificación de solo 6 caracteres, se dejó de usar debido a que no soportaba la cantidad de estudiantes y programas que tenía la universidad. Sin embargo, algunos estudiantes conservan esta codificación.
      X = Programa, cuando únicamente la universidad ofrecía 9 programas
      YY = Año de ingreso a la universidad, por ejemplo X5 = año 2005.
      Z = Semestre del año en el cual se matriculo (0 = enero, 1 = Julio).
      DD = Consecutivo del estudiante, solo soportaba hasta 99 estudiantes por cada programa en un semestre.

- Como parte de la información recibida, se incluyen datos como: Números de identificación, nombres de los estudiantes y nombres de los colegios que por sí solos no son muy útiles a la hora de aplicar una técnica de minería de datos.

- La modalidad en la cual se matriculó el estudiante (Diurno, Nocturno, Distancia) no se encuentra como un campo adicional, hace parte del campo nombre del programa, así:

  NOMBRE_DEL_PROGRAMA (XXXXXX), donde XXXXX es el nombre de la modalidad (Nulo = diurno, A DISTANCIA o NOCTURNO).

- El semestre en el cual terminó sus estudios, o se retiró de la Universidad no se encuentra como tal en un campo. Es necesario calcularlo, analizando el semestre actual, contra el semestre anterior.

- La Información académica solo incluye el nombre del colegio, debido a que este campo se debe ingresar por medio del sistema de información a la hora del registro del estudiante, es probable que debido a errores o diferencias de digitación, este nombre no corresponda en todos los casos. El campo nombre de colegio por sí solo no es muy diciente, pero es necesario tratar de cruzarlo con la información de los colegios para lograr encontrar información útil para el análisis del problema en cuestión.

Debido al problema encontrado en los nombres de los colegios, es complejo cruzar la información de los colegios con la información académica asegurando una confiabilidad de 100%. Se pretende realizar una inspección visual que permita ajustar algunos datos, y aplicar alguna técnica de ETL que permita al minimizar este riesgo.

5.1.3. Descripción de los Datos

En este capítulo se describen los datos obtenidos con el fin de brindar al lector y al investigador una visión más clara del estado actual de los mismos. A partir de esta descripción se pueden ir generando conclusiones acerca de que variables orientaran el proceso de minería, el cual a su vez orienta el proceso de ETL.

Cada registro de los datos académicos representa el registro de materias de un estudiante para el semestre en cuestión, la información se encuentra agrupada por semestres. Aproximadamente contiene 25300 registros en total.

DATOS ACADEMICOS		
VARIABLE	**DESCRIPCION**	**MINEABLE**
ID	Llave subrogada del registro del alumno.	NO
AÑO INGRESO	Año en el cual ingreso el estudiante a la universidad. Formato YYYY-MM	SI
PLAN	Código del programa en el cual se matriculó el estudiante.	SI
EXPEDIENTE	Junto con el PLAN, compone un único registro por cada estudiante en un programa.	NO
PROGRAMA	Nombre del programa al cual está registrado un estudiante. El nombre del programa contiene además la modalidad (Nocturno, Normal o A	SI

	Distancia)	
CODIGO	Código único del estudiante, contiene además información acerca del semestre de ingreso, el programa. Se debe unificar para identificar de manera única a un estudiante.	NO
DOCUMENTO	Numero de documento de identidad del estudiante. No sirve para diferenciar a un único estudiante, porque no tiene el tipo y adicionalmente es posible que el estudiante haya cambiado de documento por haber cumplido la mayoría de edad durante su época de estudios.	NO
NOMBRE	Nombre (Nombres y Apellidos) de un estudiante.	NO
SEXO	Sexo el estudiante (D=Femenino, H=Masculino).	SI
FECHA DE NACIMIENTO	Fecha exacta de nacimiento del estudiante.	SI
EDAD	Edad del estudiante al momento del registro del semestre del registro actual.	SI
PROMEDIO_SEM	Promedio alcanzado por el estudiante en el semestre en cuestión.	SI
PROMEDIO_ACUMULADO	Promedio acumulado por el estudiante en los semestres que lleva matriculado en el programa.	SI
CREDITOS_REGISTRADOS	Cantidad de créditos registrados por el estudiante en el semestre actual.	SI
ESTRATO	Estrato socioeconómico del estudiante.	SI
COLEGIO	Nombre del colegio del cual se graduó el estudiante. Se debe cruzar con la información de colegios para hacer parte del análisis.	NO POR SI SOLO
JORNADA	Jornada en la cual registra sus materias el estudiante.	SI

Tabla 1. Descripción de datos

La información obtenida contiene registros de gran cantidad de colegios de Caldas y algunos a nivel nacional. La información incluye en total 10000 registros. La información se pretende cruzar con el campo COLEGIO de la información académica para incluir en cada estudiante la información del colegio del cual proviene.

DATOS ACADEMICOS		
VARIABLE	**DESCRIPCION**	**MINEABLE**
COL_CODIGO	Código único que identifica cada colegio.	NO
COL_NOMBRE	Nombre del colegio, se utilizará para cruzar la información del colegio de cada estudiante registrado en la universidad.	NO
COL_JORNADA	Jornada en la cual realiza sus labores el colegio.	SI
COL_NATURALEZA	Indica si el colegio es privado o público. PRIVADO o PUBLICO	SI
COL_GENERO	Genero de los estudiantes del colegio. MIXTO, FEMENINO o MASCULINO.	SI
COL_CALENDARIO	Calendario académico utilizado por el colegio. A o B	SI
CIUD_ID	Código DANE de la ciudad donde realiza sus actividades el colegio.	SI
CIUD_NOMBRE	Nombre de la ciudad donde realiza sus actividades el	SI

	colegio.	
CIUD_DEPARTAMENTO	Departamento donde está ubicado el colegio.	SI

Tabla 2. Descripción de datos académicos

A continuación se relacionan algunas de las variables que se presumen de importancia para el proceso de minería de datos, que no se encuentran explícitas en la información recibida, pero que pueden llegar a ser calculadas.

VARIABLE	DESCRIPCION
DESEMPEÑO	Promedio de créditos aprobados/créditos tomados.
ULTIMO_SEMESTRE_CURSADO	Indica el último semestre en el cual el estudiante se matriculó en el programa.
GRADUADO	Indica si el estudiante logró aprobar con éxito todos los créditos necesarios para graduarse del programa.
INDICE_RETENCION	Indica cuántos de los estudiantes que ingresan en un semestre, continúan para el siguiente.
CATEGORIA	Indica dependiendo del promedio en qué categoría se encuentra un estudiante. Califica de forma cualitativa de acuerdo a la calificación cuantitativa. ALTO = Promedio acumulado superior a 4.5 MEDIO_ALTO=Promedio acumulado entre 3.5 y 4.5 MEDIO_BAJO=promedio acumulado entre 2.5 y 3.5 BAJO=Promedio acumulado inferior a 2.5

Tabla 3. Variables a calcular

5.1.4. Exploración de Datos

Durante esta fase se describirán agrupamientos y análisis estadísticos que permitan profundizar un poco más en la información y en cómo se encuentra distribuida la misma. Esto permitirá tener un contexto más claro del problema de la deserción académica y del camino a seguir durante el proceso de minería que permita llegar a conclusiones realmente valiosas.

5.1.4.1. Distribución de Estudiantes

La Universidad Autónoma de Manizales cuenta con 24 programas académicos de pregrado en los cuales se distribuye la cantidad de estudiantes matriculados de la siguiente forma:

Este gráfico representa la distribución de los estudiantes a través de los diferentes programas académicos contabilizando los registros académicos realizados desde el primero periodo del año 2006 hasta el semestre 1 del año 2011. En el gráfico se puede observar como el programa de odontología representa un 24% de la totalidad de la población duplicando a su seguidor inmediato Ingeniería Industrial. Debido a esta distribución, es también probable que el programa de odontología represente los índices más altos de deserción lo cual lo hace una fuente importante para la detección de posibles causas.

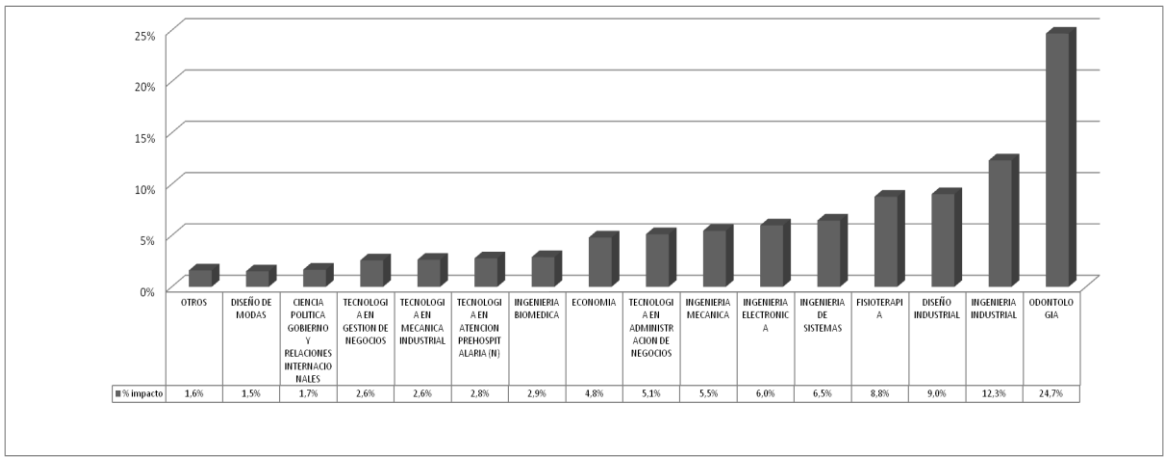


Ilustración 4. Estudiantes por Programa

Debido a su ubicación geográfica y su amplia oferta educativa, la ciudad de Manizales representa una buena oportunidad para jóvenes que desean realizar estudios en programas de pregrado técnicos y profesionales. Es por esta razón que gran cantidad de personas de diversas regiones del país deciden realizar sus estudios en la ciudad de Manizales, a continuación se presenta de forma gráfica como se distribuyen estos estudiantes dependiendo de la región de la cual provienen. Es importante distribuir los registros académicos dependiendo del departamento, porque este departamento nos permitirá cruzar la información académica con la información obtenida del DANE para hallar características sociales, económicas y educativas de la región de la cual proviene cada uno de los estudiantes.

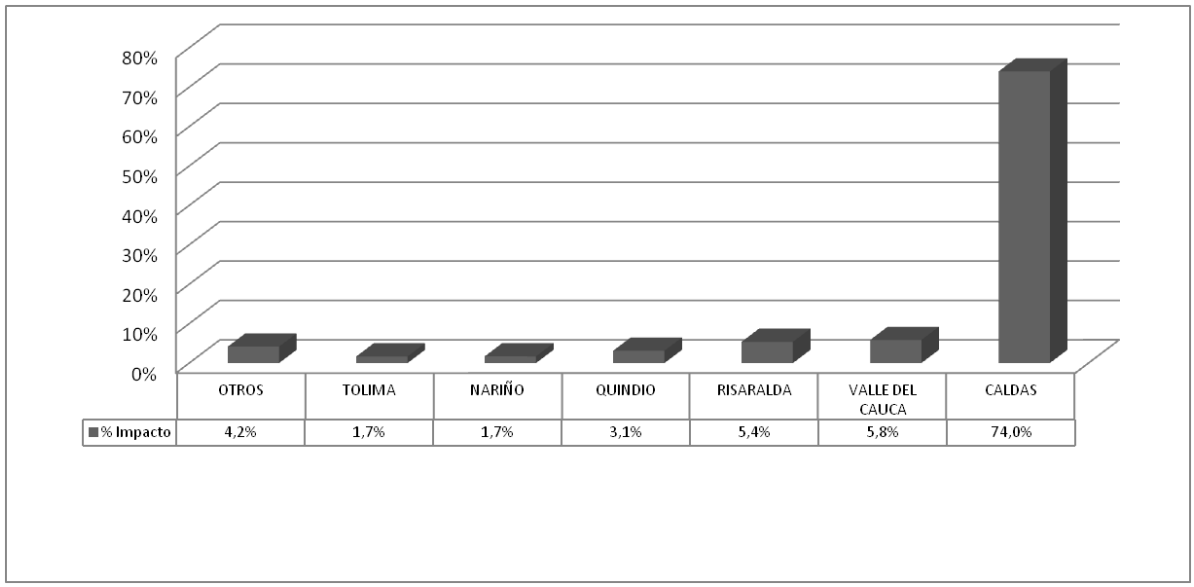


Ilustración 5. Estudiantes por Departamento

El estrato socioeconómico de un estudiante puede brindar bastante información acerca de la situación social y económica del medio que lo rodea. Además, permite identificar la población objetivo a la cual brinda sus servicios la Universidad Autónoma de Manizales. Se puede observar en el gráfico como la mayor parte de la población educativa de la universidad proviene del estrato 3.

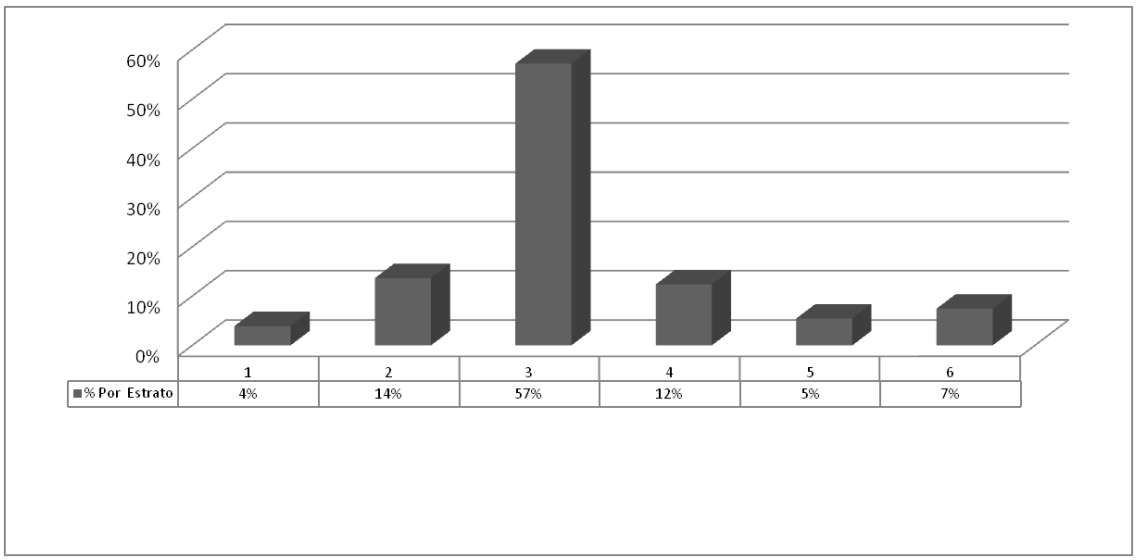


Ilustración 6. Estudiantes por estrato socioeconómico

5.1.4.2. Cifras Rendimiento Académico

De acuerdo a las reglamentaciones académicas, el promedio semestral se calcula sumando las notas obtenidas por el estudiante en cada una de las materias registradas (de 0.0 a 5.0) y dividido por el número de materias registradas en el semestre evaluado. Un análisis previamente realizado indicó que la mayoría de estudiantes que deciden desertar de los diferentes programas presentan bajo rendimiento académico en los semestres anteriores al de su deserción. A continuación se realiza un análisis gráfico del porcentaje de estudiantes que aprueban y los que no aprueban su promedio académico (promedio semestre >= 3); estos datos se calculan tomando en cuenta la información tomada al final de cada semestre desde el semestre 1 del 2006 hasta el semestre 1 de 2011.

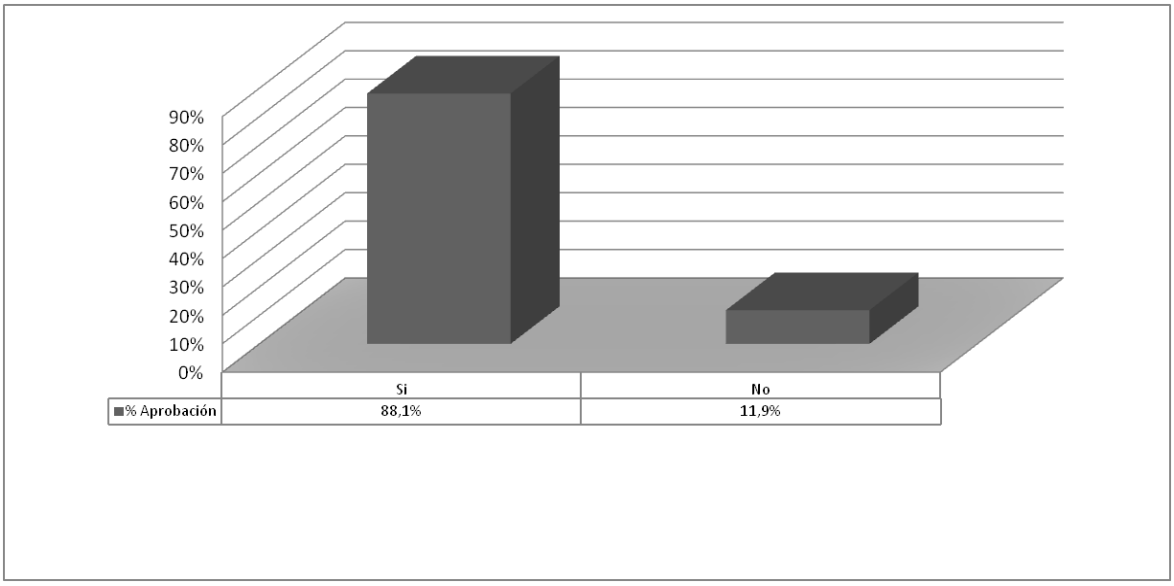


Ilustración 7. Estudiantes Con Promedio Aprobado

La reglamentación académica de la universidad exige registrar al inicio de cada semestre una cantidad determinada de créditos distribuidos entre materias tomadas, a continuación se presenta la distribución de los estudiantes que aprueban la totalidad de los créditos que inscriben al comienzo del semestre, con respecto a los estudiantes que reprueban al menos un crédito.

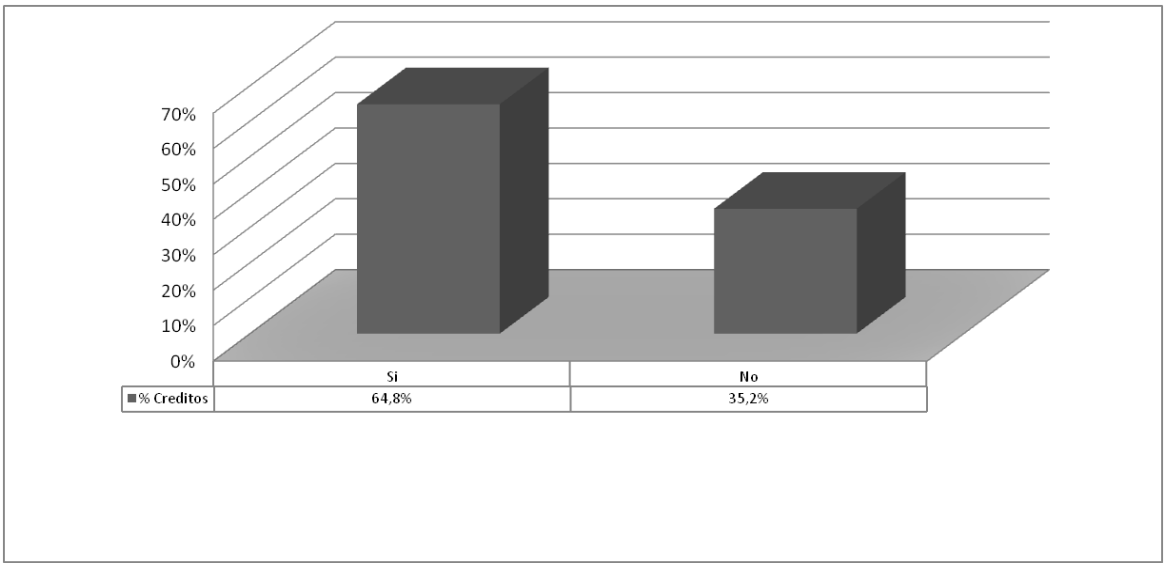


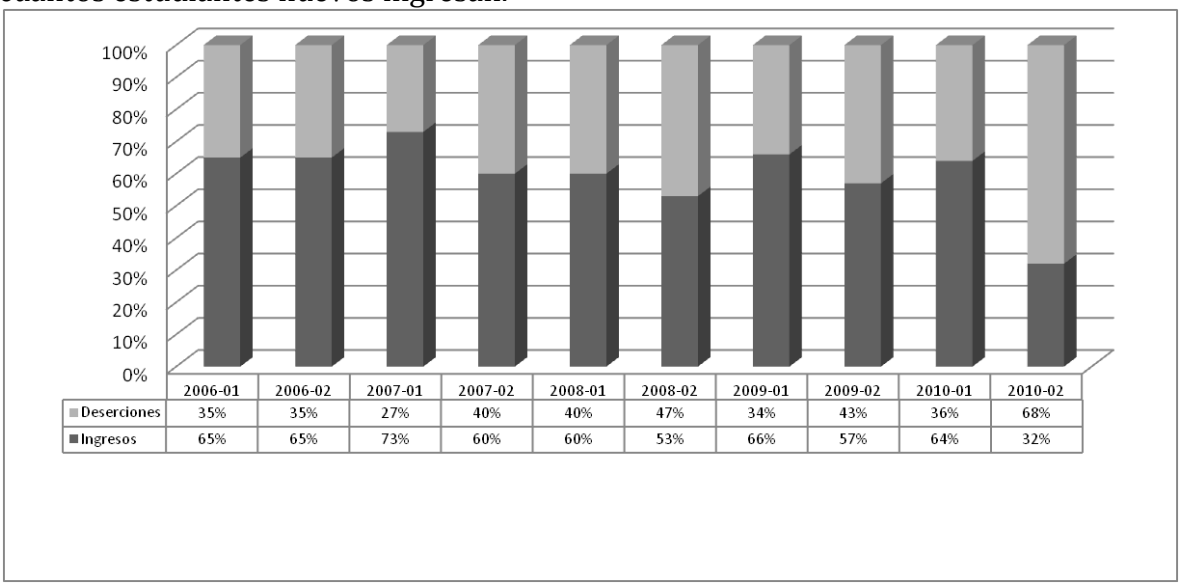
Ilustración 8. Estudiantes Créditos Aprobados

El gráfico muestra como un porcentaje representativo de estudiantes de al menos el 35% reprueba al menos un crédito de la totalidad de los inscritos al inicio del semestre. La cantidad de créditos reprobados es una variable que claramente afecta el rendimiento académico e influye en la deserción académica, por esta razón es importante relacionar la cantidad de créditos desaprobados con otras variables que

pueden afectar el rendimiento académico y por ende impactan directamente en los índices de deserción.

5.1.4.3. Cifras de Deserción

Se considera desertor a un estudiante del cual no se registra matricula académica por dos semestres consecutivos, de acuerdo a esta premisa a continuación se representa gráficamente cuantos estudiantes desertan cada semestre académico con respecto a cuántos estudiantes nuevos ingresan.



	2006-01	2006-02	2007-01	2007-02	2008-01	2008-02	2009-01	2009-02	2010-01	2010-02
Deserciones	35%	35%	27%	40%	40%	47%	34%	43%	36%	68%
Ingresos	65%	65%	73%	60%	60%	53%	66%	57%	64%	32%

Ilustración 9. Deserciones vs Ingresos

Aunque a partir del gráfico no es posible hacer ninguna conclusión acerca de las razones de la deserción, si es importante conocer previo al análisis de los datos cifras más precisas acerca de la cantidad de deserciones que se presentan en la Universidad Autónoma de Manizales. A continuación se agrupan la cantidad de deserciones que se presentan de acuerdo a la cantidad de estudiantes; se trata de cifras acumuladas de los últimos 4 años.

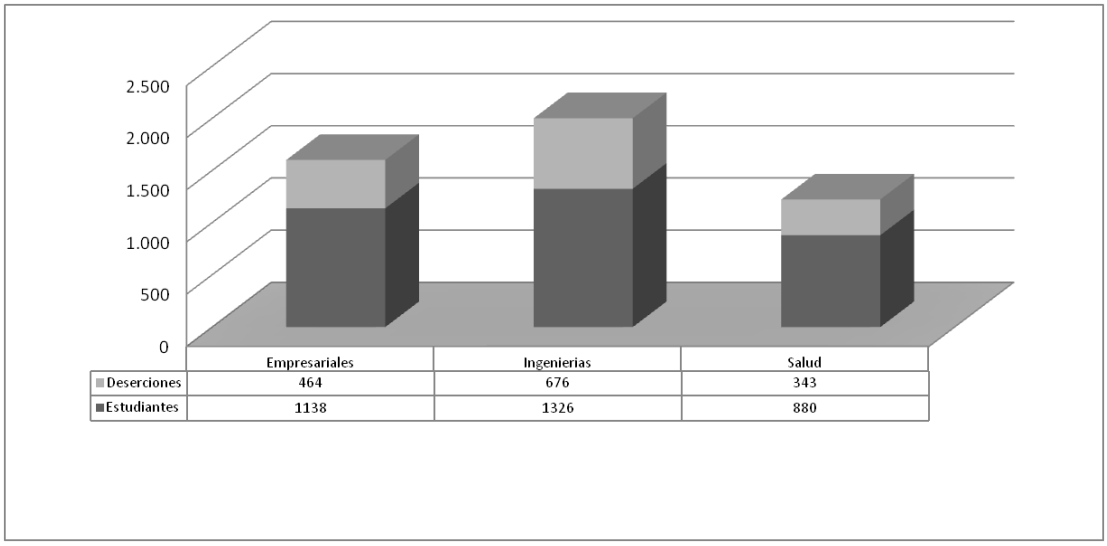


Ilustración 10. Deserciones por facultad

El siguiente gráfico representa como la cifra de deserción ha variado en los últimos semestres en la Universidad Autónoma de Manizales.

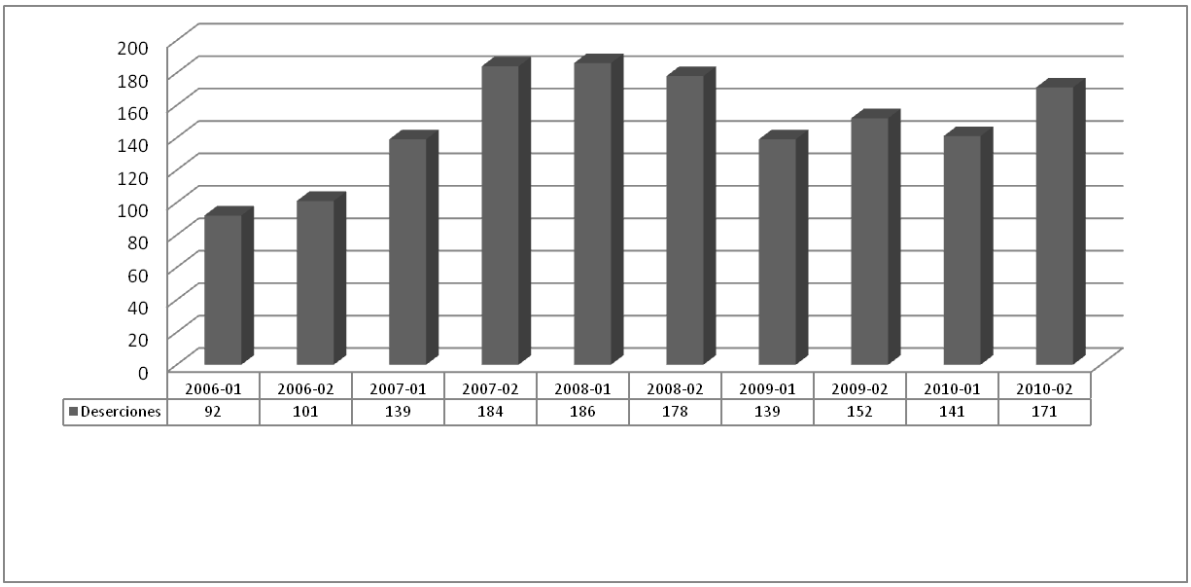


Ilustración 11. Deserciones por semestre académico

5.1.5. Diseño Almacén de Datos

Realizado el análisis de los datos, e identificadas las variables que se identificaron con mayor probabilidad de generar más información acerca del problema de la deserción, se procede a realizar un diseño de Almacén de Datos que lleve a un proceso de minería exitoso. Durante este diseño de acuerdo a la metodología bottom-up de Ralph Kimball [23], se identificarán necesidades puntuales y se diseñarán estrellas individuales que luego darán un diseño general con miras al proceso de minería de datos. El resultado del almacén de datos se incluye en el ANEXO A de este documento.

5.1.5.1. Rendimiento Académico

Para analizar el fenómeno de la deserción desde el punto de vista del rendimiento académico se hace fundamental separar las variables incluidas dentro de la información académica que permiten analizar el rendimiento. Estas variables son la cantidad de créditos aprobados y el promedio académico alcanzado por un estudiante. Adicional, se calcularán variables como el promedio acumulado y el porcentaje de créditos aprobados y no aprobados por un estudiante.

Es importante poder agrupar la información académica por diferentes características tales como el semestre académico y el programa al cual se encuentra adscrito el estudiante. De acuerdo a estas premisas se identifica la siguiente estructura de estrella para soportar esta información:

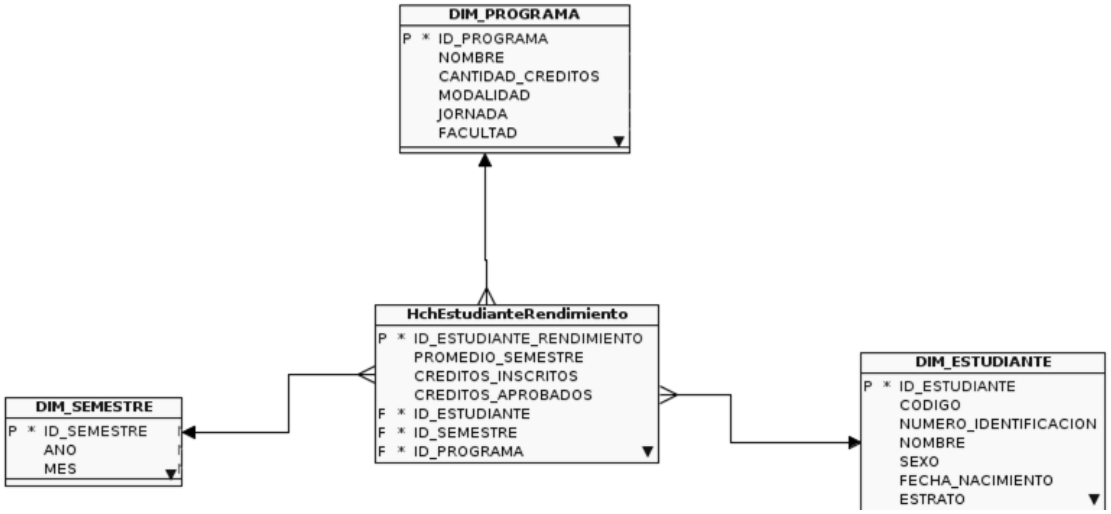


Ilustración 12. Estrella rendimiento académico

5.1.5.2. Estado académico estudiantes

El fenómeno de la deserción se debe analizar desde el punto de vista del estado académico en el cual se encuentra un estudiante, los posibles estados en los que se pueden encontrar un estudiante son 'ACTIVO', 'INACTIVO' y 'GRADUADO'. Un estudiante activo es uno que registra asignaturas en el semestre en curso, un estudiante inactivo se considera aquel que no registra actividad académica por dos semestres consecutivos, y un estudiante graduado es aquel que superó satisfactoriamente la totalidad de los créditos del programa académico en el cual se encontraba registrado. Adicional al estado actual del estudiante, se agrega información como el semestre en el cual el estudiante se registra por primera vez y en cual se generó su última matricula. Los estudiantes en estado INACTIVO serán considerados como desertados para este estudio debido a que no se tiene un dato exacto de los estudiantes considerados desertores.

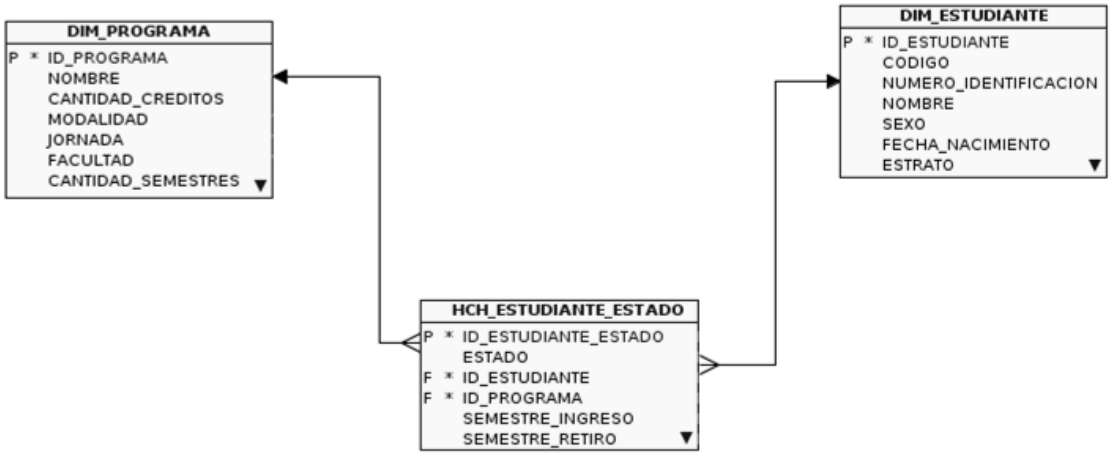


Ilustración 13. Hecho estado académico

5.1.5.3. Aspectos sociales y procedencia

En el estado del arte acerca del problema de la deserción se identificó en muchos casos que factores socioeconómicos que rodean a los estudiantes, o la formación previa que estos tienen y las instituciones de educación media de las cuales provienen pueden ser características que influyen de manera positiva o negativa en la decisión de desertar de un estudiante. Para este caso, se tomarán aspectos socio-económicos de cada uno de los estudiantes, debido a que no se proporcionó dicha información. Se analizará relacionando cada estudiante con las características de la institución en la cual realizó sus estudios de educación media y la región en la cual se encuentra la misma.

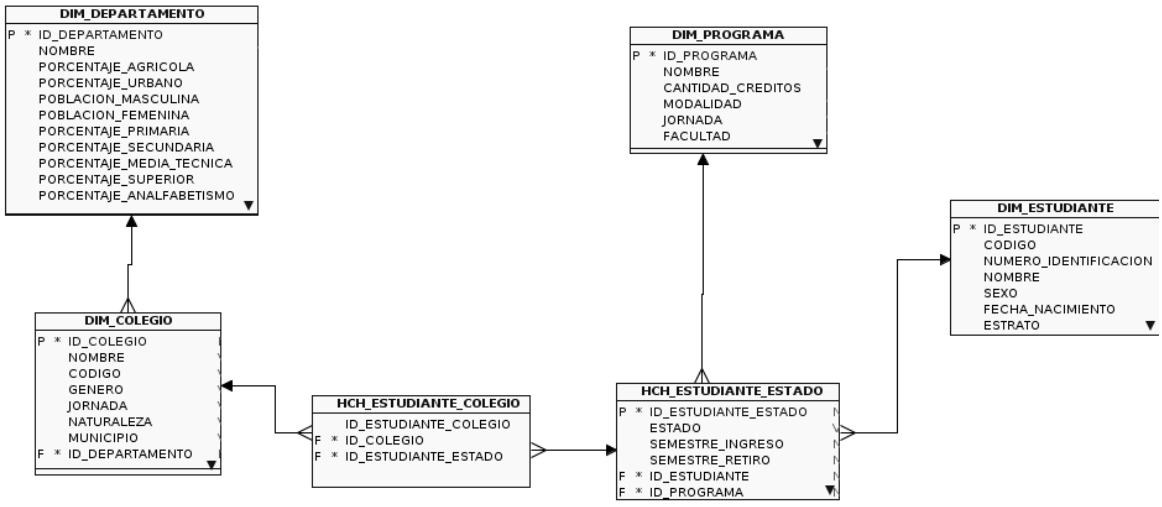


Ilustración 14. Hecho de aspectos socioeconómicos

5.1.6. Diseño proceso de ETL

El proceso de ETL se encarga de convertir los datos descritos en la fase de comprensión de datos en el numeral 4.2.2, a la estructura de Almacén de Datos diseñada de acuerdo a la necesidad puntual de encontrar información pertinente acerca del fenómeno de la deserción. Para cumplir su objetivo principal el proceso de ETL debe cumplir otros objetivos específicos que garanticen el resultado final:

- Corregir datos erróneos y nulos de las fuentes de datos.
- Garantizar la confiabilidad de los datos.
- Consolidar datos de múltiples fuentes en una única fuente.
- Generar una nueva fuente de datos que puede ser usada por otros sistemas o usuarios finales directamente.
- Remover y ajustar valores extremos con el fin de disminuir el ruido en la información en el destino.

El proceso de ETL utiliza múltiples herramientas para cumplir su objetivo: lógica difusa, técnicas estadísticas, SQL, transformaciones de tipos de datos, etc. Estas herramientas se encuentran en su mayoría implementadas dentro de algunas herramientas. Para este proyecto se utilizará como herramienta principal de ETL RapidMiner, una herramienta de código libre especializada en procesos de minería de datos[24].

A continuación se describen algunas de las actividades realizadas durante el proceso de ETL:

5.1.6.1. Cargue Rendimiento Académico

El proceso de cargue de rendimiento académico se encarga de generar la información que se almacenará en la estructura descrita en el numeral 5.1.5.1. El proceso de extracción y limpieza se realiza mediante Rapidminer [24] a partir de la información académica de la Universidad Autónoma de Manizales.
Se debe cruzar la información cargada previamente de programas con la información académica recibida de la oficina de registro; este cruce se realiza a través del campo PLAN. Sin embargo, debido a que el código del plan se puede repetir cambiando únicamente la jornada, se hace necesario usar como condición de cruce la jornada también.
Para cruzar con el semestre académico, se divide el campo SEMESTRE del Excel en dos campos diferentes “ANO” y “MES”, para lograr igualar con la dimensión de semestres.

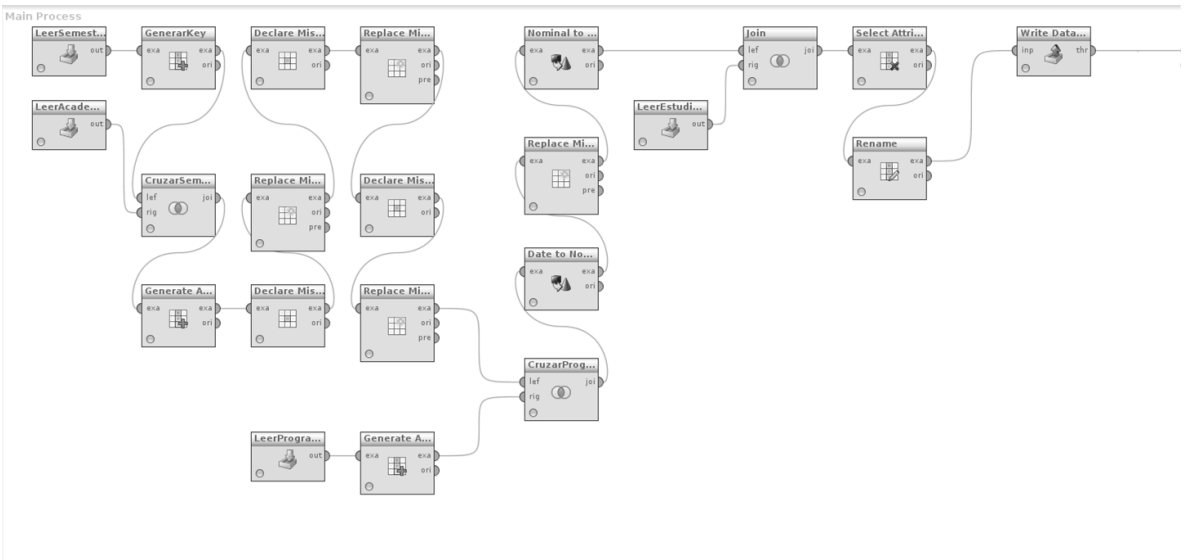


Ilustración 15. Modelo rapidminer diseñado para cargar la información del rendimiento académico

- *Paso1:* Se carga la información del Excel con la información académica de cada estudiante en todos los semestres analizados, previamente unificada en una única hoja de Excel.
- *Paso2:* Se lee la información del hecho de semestres.
- *Paso3:* Se genera la clave del Excel para cruzar con el hecho de semestres, se genera un campo "ANO" y un campo "MES".
- *Paso4:* Se cruza la información académica con los datos de cada semestre de acuerdo a los campos año y mes.
- *Paso5:* Se reemplazan los códigos de programa que no eran correctos en el Excel de datos académicos. Se realiza haciendo una inspección visual y luego operadores de reemplazo de valores de RapidMiner.
- *Paso6:* Se consultan los datos de la dimensión de programas.
- *Paso7:* Se cruza la información del Excel de datos académicos, con el programa respectivo a cada registro. Este cruce se realiza por el campo código del programa y la jornada.
- *Paso8:* Se consulta la información de los estudiantes de la dimensión respectiva.
- *Paso9:* Se realiza el cruce de la información de la dimensión de estudiantes con la información del Excel académico. Este cruce se realiza por el campo código previamente normalizado.
- *Paso10:* Se crean los registros con los atributos necesarios y se cargan en el hecho.

El resultado del proceso de cargue se encuentra en el ANEXO E.

5.1.6.2. Cargue estados académicos

El cargue de los estados académicos de los estudiantes no se encuentra explícitamente en los datos académicos. Es por esta razón que se debe calcular a partir de los datos con los que se cuenta apoyado en el uso de herramientas como SQL, programación en Java[5] y RapidMiner.

El campo ESTADO tiene los posibles valores "ACTIVO", "GRADUADO", o "INACTIVO".

Para llenar la tabla inicialmente se genera un archivo CSV que contiene el semestre en el cual ingreso el estudiante al programa, y el último semestre cursado. Para generar este archivo CSV se lee el archivo de datos académicos usando Java y se genera un nuevo CSV en el cual se incluye el código del estudiante, el primer semestre cursado y el último semestre cursado. Si el último semestre cursado es el actual, se considera que aún está activo, de lo contrario se considera inactivo o desertado.

```
cargarArchivoAcademico
mientras existe estudiantes entonces
      leerEstudiante
      si estudiante.ultimo_semestre=semestreActual entonces
            marcarEstudianteActivo
      sino si estudiante.cantidad_creditos=créditos_programa entonces
            marcarEstudianteGraduado
      sino
            marcarEstudianteInactivo
      fin si
      si estudianteInactivo entonces
            si semestresInactivo>2 entonces
                  marcarEstudianteDesertado
      fin si
fin mientras
escribirNuevoArchivoAcademico
```

Ilustración 16. Pseudocódigo, calcular estado académico estudiantes

Con el fin de completar la información referente al estado académico de un estudiante con la información del programa académico y demás datos de cada estudiante, se define el siguiente proceso de RapidMiner.

[5] Java es un lenguaje de programación Orientado a Objetos propiedad de Oracle y creado por Sun Microsystems durante la década de los 90. (http://www.java.com/es/about/).

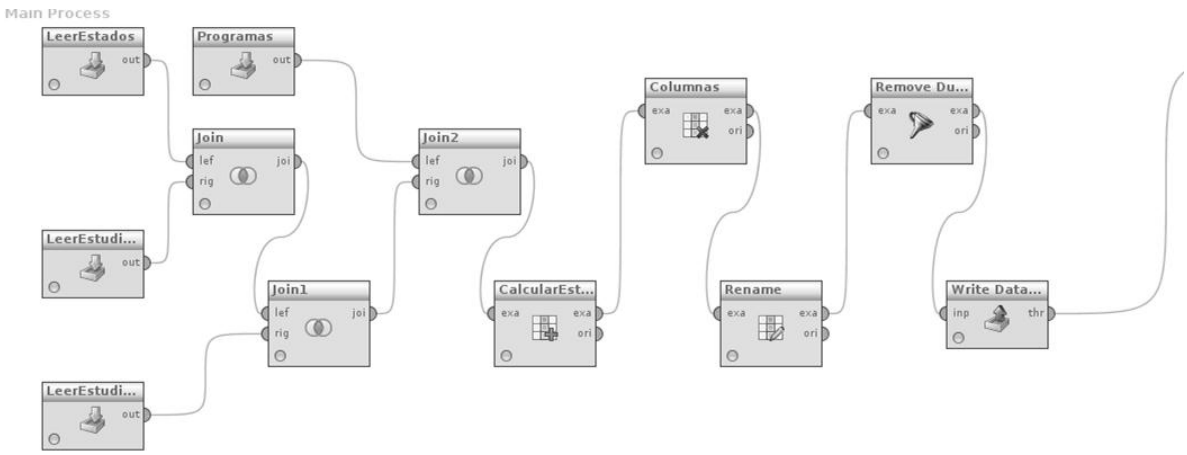


Ilustración 17. Modelo rapidminer cargue estados académicos

- *Paso1:* Leer el CSV de estados de los estudiantes generado en el paso anterior.
- *Paso2:* Leer la dimensión de estudiantes.
- *Paso3:* Cruzar la información en la dimensión de estudiantes con la calculada en el CSV a través del código de estudiante.
- *Paso4:* Se carga la información cargada en el hecho de estudiantes y programas.
- *Paso5:* Se realiza el cálculo del estado de los estudiantes calculando el número de semestres que lleva en la universidad y comparándolo con la cantidad de semestres que tiene normalmente el programa académico en el cual se encuentra inscrito. Si el estudiante se encuentra inactivo y la cantidad de semestres estudiados es mayor o igual a la cantidad de semestres del programa se asume que el estudiante se graduó, pero si se encuentra inactivo y la cantidad de semestres es menor se presume que abandono el programa académico.
- *Paso6:* Se seleccionan las variables a insertar.
- *Paso7:* Se insertan los registros calculados mediante el proceso de ETL.

5.1.6.3. Cargue aspectos socioeconómicos

Debido a que no se cuenta con información de los aspectos socioeconómicos de los estudiantes directamente en la información académica. Se cruza la información obtenida de los colegios de la región y de cada uno de los departamentos obtenidos con información del DANE a partir del colegio asociado a cada uno de los estudiantes.

- Se carga una tabla intermedia con los datos académicos que incluyen el código del estudiante, nombre del colegio y municipio del colegio. Esta información se toma del Excel de datos académicos y se carga en la base de datos usando RapidMiner.
- Se carga una tabla intermedia con la información de los colegios.

- Se carga una nueva tabla que cruza la información de los colegios con el colegio de cada uno de los estudiantes. Sin embargo, debido a que la información de los colegios incluida en los datos académicos fue digitada manualmente, no se encuentra normalizada por lo cual fue necesario utilizar funciones de lógica difusa de Oracle con el fin de lograr los mayores índices de coincidencia y cruzar una mayor cantidad de información con mayor certeza.
- En este punto se utiliza la información cargada para cargar la información al Almacén de Datos mediante RapidMiner.

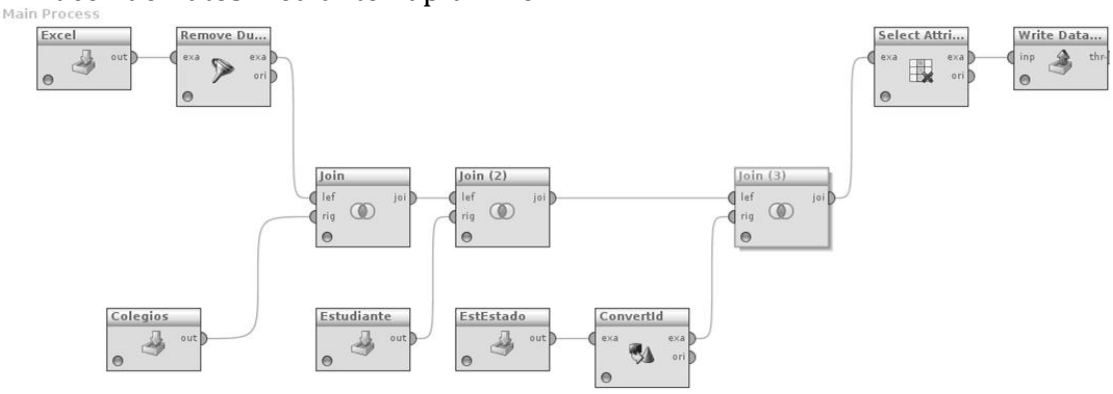


Ilustración 18. Proceso cargue variables socioeconómicas rapidminer

- *Paso1:* Se carga la información generada de la sentencia Oracle que aplica lógica difusa.
- *Paso2:* Se cargan los datos de la dimensión de colegios.
- *Paso3:* Se cargan los datos de la dimensión de Estudiantes.
- *Paso4:* Se cargan los datos de la tabla de hechos de Estudiantes Por Estado.
- *Paso5:* Se cruza la información de colegios cargada de Excel con la información de la dimensión de colegios a través del código del mismo.
- *Paso6:* Se cruza la información del Excel con los datos de la dimensión de estudiantes a través del código del estudiante.
- *Paso7:* Se cruza la información de la dimensión de estudiantes con los datos de la tabla de hechos de estudiantes por estado a través del id de cada registro en la dimensión de estudiantes.
- *Paso8:* Se seleccionan solamente las variables necesarias para generar los datos de la nueva tabla de hechos.
- *Paso9:* Se escriben los datos cargados a la nueva tabla de hechos.

Los componentes usados para el proceso de ETL se incluyen en el ANEXO B de este documento.
El resultado del proceso de ETL es el almacén de datos incluido como adjunto a este documento ANEXO E.

5.1.7. Diseño Framework para aplicación de Redes Neuronales

El diseño del framework de redes neuronales pretende dar una solución a la problemática puntual de este proyecto, pero en su diseño contempla la posibilidad de convertirse en un componente genérico con características de extensibilidad que le permitan ser usado o adaptarse a cualquier otro proyecto que use redes neuronales. El framework se dispone en tres componentes fundamentales, cada uno de ellos independiente de los demás, lo cual le brinda la capacidad de adaptarse a las necesidades de cada tipo de problema con poco esfuerzo.

El framework se diseñó utilizando patrones de diseño de programación orientada a objetos. Los patrones de diseño fueron creados para definir la forma en la que se deben solucionar problemas típicos de programación, con el fin de promover la reutilización y la mantenibilidad del código, garantizando que un tipo dado de problema se solucione siempre usando la misma plantilla o patrón[25].

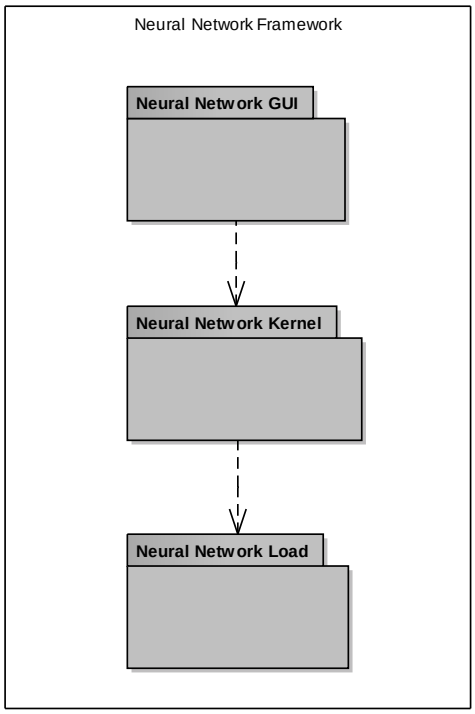


Ilustración 19. Componentes básicos framework de redes neuronales

A continuación se describe en detalle cada uno de los componentes básicos del framework con el fin de dar más claridad acerca del diseño de cada uno de ellos:

5.1.7.1. Componente Neural Network Load

Componente que tiene como objetivo principal cargar la información desde una fuente de datos indicada al formato requerido por la red neuronal para realizar su entrenamiento y posterior proceso de generación de conocimiento. El componente se diseña usando un patrón de fábricas abstractas en el cual se permite utilizar diferentes tipos de fuentes de datos (Archivos planos, archivos CSV, Bases de datos

relacionales, Almacén de Datos, etc.), realizando la implementación específica según el caso. Sin embargo, en la versión desarrollada para este proyecto solo se implementa la utilidad de cargue a partir de archivos posicionales CSV. El patrón Abstract Factory (Fabricas abstractas) permite soportar varias familias de productos o componentes sin modificar los clientes de los mismos.

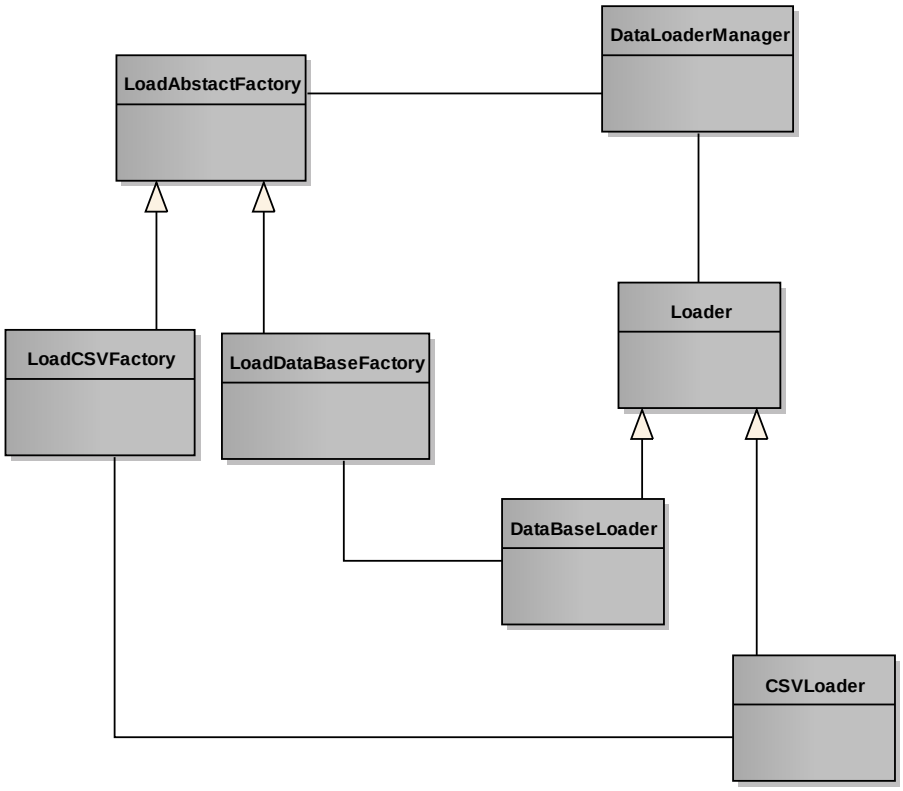


Ilustración 20. Diagrama de clases componente cargue de datos del framework de redes neuronales

El componente de carga tambien incluye las funcionalidades de validación y conversión de datos. Estas funcionalidades le permiten validar los datos que provienen de la fuente seleccionada de forma que cumplan con los requisitos de la red neuronal que se pretende cargar y con la estructura definida para la fuente de datos seleccionada. La validación de los datos incluye también la generación de errores encontrados en los datos.

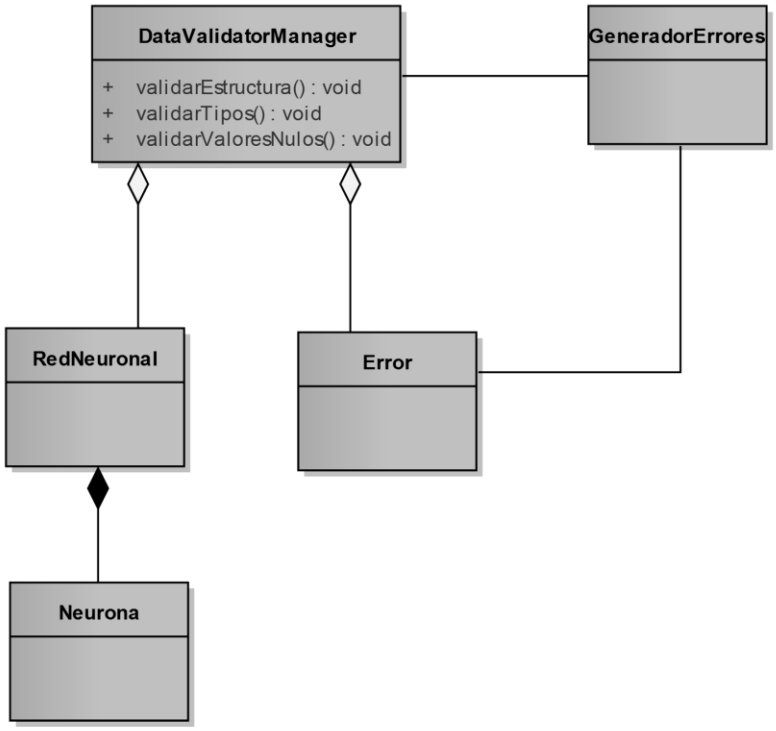


Ilustración 21. Diagrama de clases componente de validación de datos del framework de redes neuronales

5.1.7.2. Componente Neural Network Kernel

Este componente es considerado el núcleo del framework, es el encargado de aplicar la lógica propia de cada red neuronal implementada, aunque la versión inicial solo incluye soporte para la técnica seleccionada para este proyecto; el diseño quedará preparado para soportar nuevas técnicas.

El diseño encargado de extender la posibilidad de nuevas técnicas se implementa utilizando el patrón Abstract Factory, debido a que el componente de entrenamiento y aprendizaje tarda un tiempo prudencial en ejecutarse, se utilizó el patrón Observer con el fin de controlar el estado del proceso. Según [25] el patrón Observer, es aplicable en casos en los cuales un objeto debe notificar sus cambios de estado a otros objetos tal como el caso actual en el cual el objeto encargado de implementar el entrenamiento o el aprendizaje debe notificar a sus clientes su estado actual.

La técnica implementada en este proyecto utiliza el algoritmo de Kohonen, algoritmo que realiza su entrenamiento de forma no asistida. En el proceso de entrenamiento forma una red utilizando como entrada una muestra de los datos, a partir de ellos forma mapas con cada una de las variables agrupando en zonas específicas de la red las neuronas con características similares utilizando el cálculo de la distancia euclidiana entre ellas. La red considera como finalizado su entrenamiento cuando se cumple el número de iteraciones configurado para la misma o cuando se logra un nivel de error igual o menor al configurado como entrada. A mayor cantidad de iteraciones de entrenamiento o menor valor de error soportado configurados para el entrenamiento, se formará una red que genera un nuevo conocimiento más efectivo.

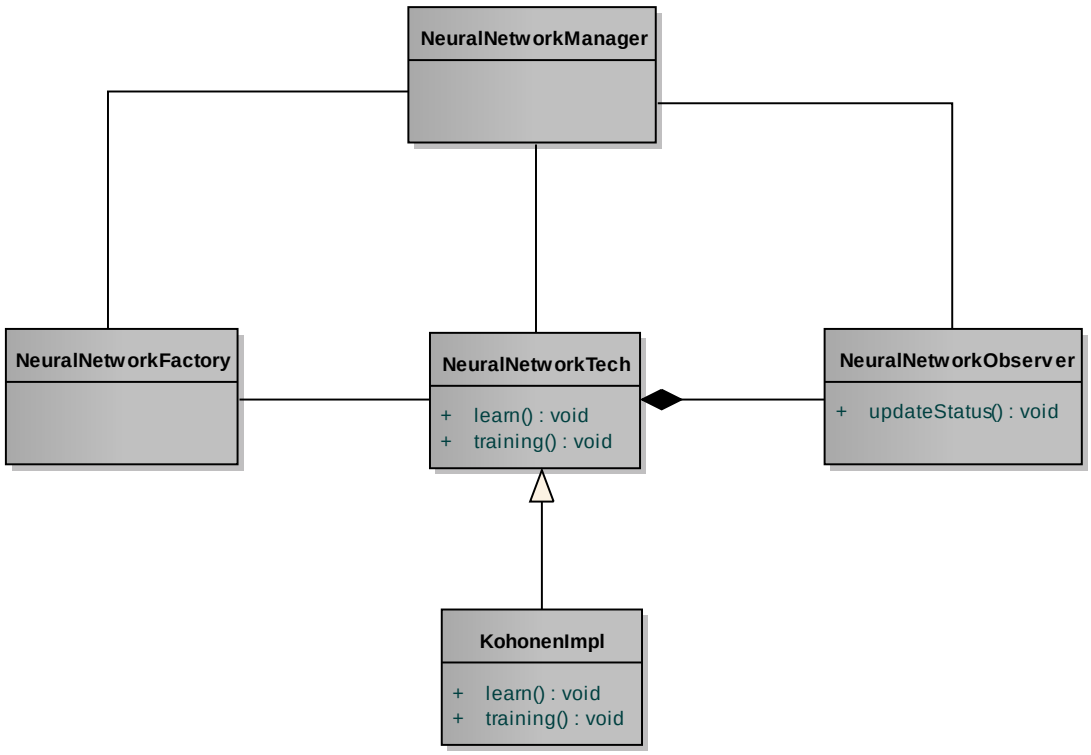


Ilustración 22. Modelo de clases del kernel de framework de redes neuronales

5.1.7.3. Diseño Componente GUI

Los componentes de cargue de datos y kernel fueron diseñados para funcionar independientemente del cliente utilizado. En este caso se diseñó un cliente pensando en la problemática tratada durante este proyecto, en el cual se permite al usuario a través de una interfaz gráfica, interactuar con los datos y con la técnica de red neuronal construida.

Funcionalmente la interfaz se diseñó como un asistente (wizard) de tres pasos. En el primero de ellos el usuario debe cargar los datos en este caso provenientes del Almacén de Datos generado con el diseño realizado durante la ejecución del proyecto. En el segundo paso se seleccionan las variables con las cuales se realizará el entrenamiento y los parámetros de configuración del mismo. Por último en el tercer paso, se brinda al usuario la posibilidad de generar nuevo conocimiento a partir de las variables con las cuales se realizó el entrenamiento utilizando la red neuronal.

Para la construcción de la GUI, se utilizó programación orientada a objetos teniendo en cuenta el patrón arquitectónico Modelo Vista Controlador (MVC). El MVC es ampliamente utilizado en la construcción de interfaces de usuario avanzadas y es muy conveniente debido a que separa la lógica de negocio de la lógica de visualización e interacción del usuario con las formas, lo cual facilita la mantenibilidad de la capa de presentación en la cual ocurren la mayor cantidad de errores o cambios. En el MVC el modelo es el encargado de almacenar los datos que se presentan al usuario en las formas, el controlador es el encargado de aplicar la lógica de interacción del usuario y

manejar los eventos ocurridos en los formularios y por último la vista es la encargada de presentar los datos y las acciones al usuario a través de formularios [26].
El componente de GUI cuenta con tres formularios, cada uno de ellos implementado siguiendo el patrón arquitectónico Modelo Vista Controlador descrito en el punto anterior.

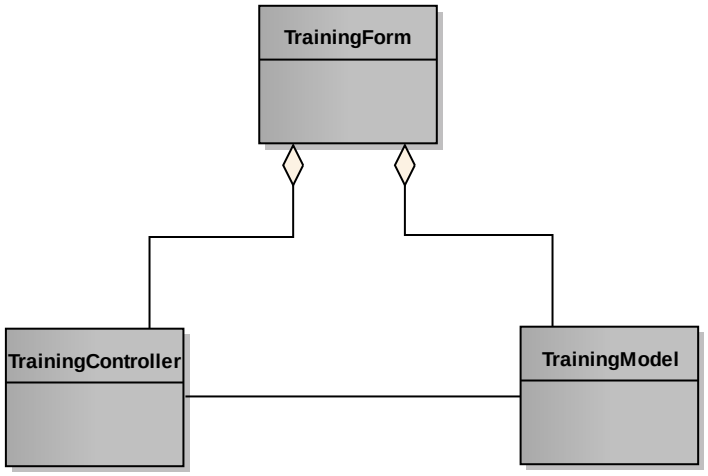


Ilustración 23. Modelo de clases GUI, formulario de entrenamiento

El código fuente mediante el cual se implemento el framework de minería de datos se adjunta a este documento y se describe en el ANEXO C. El componente ejecutable del framework de minería de datos se encuentra anexo a este documento en el ANEXO D.

6. Experimento

En este capítulo se aplica la metodología de trabajo propuesta con el fin de responder a las preguntas planteadas en el capítulo 2. Primero se aplicará el framework de redes neuronales construido para analizar la información generada a partir del proceso de ETL.

6.1. Análisis información Rendimiento Académico

La información de rendimiento académico obtenida del proceso de ETL descrito en el capítulo 5.1.6.1 es analizada a través del framework de rendimiento académico construido durante este proyecto y descrita en el numeral 5.1.7.

6.1.1. Fase 1, Entrenamiento

La primera fase del análisis corresponde al entrenamiento de la red neuronal. A continuación se describe la configuración con la cual se realizó este entrenamiento.

FASE 1, Entrenamiento	
PARAMETRO	**VALOR**

Cantidad de registros	2500 (10%)
Error Deseado	0.01
Iteraciones de Entrenamiento	100000
Factor de Reducción	0.9
Tasa de Aprendizaje	0.2
Método de Aprendizaje	Aditivo

Tabla 4. Entrenamiento información académica

Las siguientes son las variables con las cuales se realizó el entrenamiento de la red neuronal:

VARIABLE	POSIBLES VALORES
PROMEDIO_SEMESTRE	Valores entre 0.0 y 5.0
CREDITOS_INSCRITOS	Máximo 28, Mínimo 0
CREDITOS_APROBADOS	Máximo 28, Mínimo 0
CANTIDAD_CREDITOS	Cantidad de créditos del programa académico, Valores entre 92 y 185
SEMESTRE_INGRESO	Valores entre 20011 y 20112
COD_MODALIDAD	1 = Presencial, 2 = Distancia
COD_JORNADA	1 = Diurna, 2 = Nocturna, 3 = Virtual, 4 = Distancia
COD_PROGRAMA	Valores entre 1 y 21 con el código del programa correspondiente.
COD_FACULTAD	1 = Empresariales, 2 = Salud, 3 = Ingenierías
COD_GENERO	1 = Masculino, 2=Femenino

Tabla 5. Variables entrenamiento/aprendizaje información académica

6.1.2. Fase 2, Aprendizaje

El proceso de aprendizaje dentro de la información académica tiene como objetivo identificar cuáles variables afectan con mayor intensidad el rendimiento académico de los estudiantes. A continuación se presentan los resultados conseguidos al aplicar la red neuronal a la información académica del Almacén de Datos. Los siguientes resultados gráficos representan las variables que tienen mayor influencia sobre la variable seleccionada y en qué porcentaje la impactan:

Variable Analizada: PROMEDIO_ACADEMICO

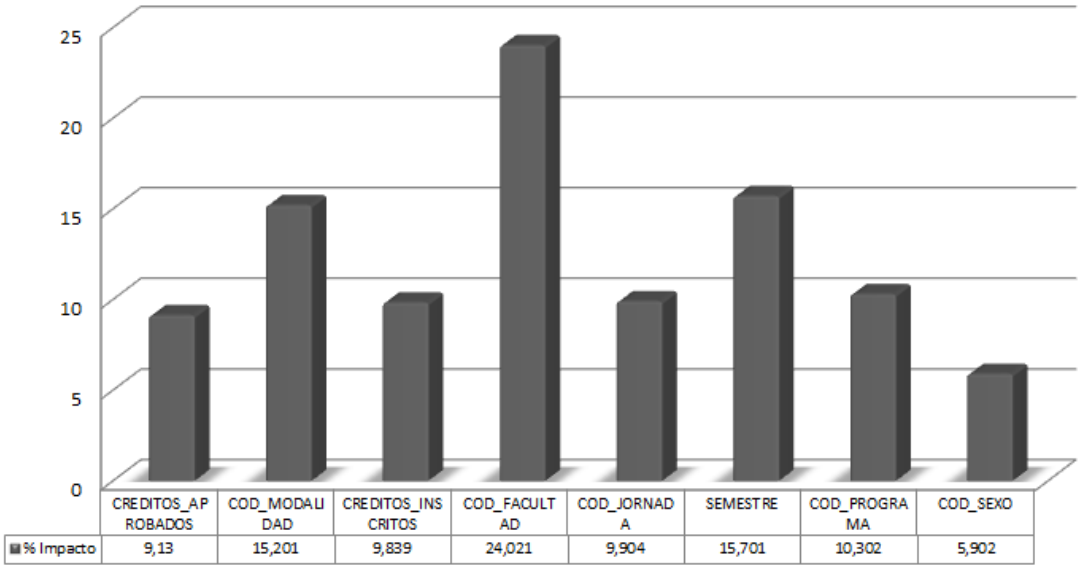


Ilustración 24. Gráfico resultado análisis con redes neuronales de la variable PROMEDIO_ACADEMICO

Variable Analizada: CREDITOS_APROBADOS

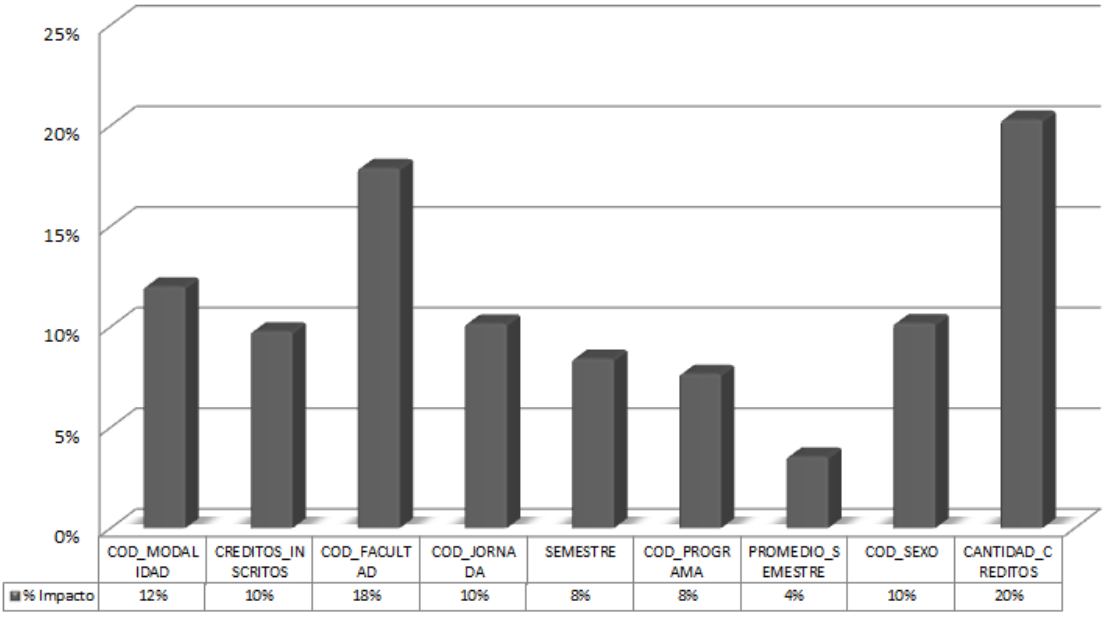


Ilustración 25. Resultado análisis mediante redes neuronales de la variable CREDITOS_APROBADOS

6.1.3. Fase 3, Análisis de Resultados

El promedio académico es impactado en gran porcentaje por la facultad a la cual se encuentra adscrito el estudiante. Esto indica que los promedios varían principalmente entre una facultad y otra, es decir, que algunas facultades presentan promedios mucho más bajos, lo cual según estudios previos es una de las principales causas de deserción[4]. Desde el punto de vista de los factores institucionales, este resultado puede ser un reflejo de problemas de uniformidad en los procesos académicos, criterios de selección de personal o estudiantes que ingresan.

La cantidad de créditos aprobados se ve impactada fundamentalmente por la cantidad de créditos inscritos. Sin embargo, este resultado no se puede analizar como un descubrimiento del proceso de minería por tratase de una conclusión manifiesta. Lo que sí se puede concluir al igual que el promedio académico, es que la cantidad de créditos aprobados se ve altamente impactada por la facultad a la cual se encuentra adscrito el estudiante, confirmando con esto los resultados obtenidos con la variable del promedio académico.

6.2. Análisis estado académico

En esta sección se realiza el proceso de análisis de la información de los estados académicos de los estudiantes de la Universidad Autónoma de Manizales que se encuentran en el Almacén de Datos y que fueron generados por el ETL descrito en el numeral 5.1.6.2.

6.2.1. Fase 1, Entrenamiento

El entrenamiento de la red neuronal se realiza usando los datos del Almacén de Datos y se realiza con los parámetros descritos a continuación:

FASE 1, Entrenamiento	
PARAMETRO	**VALOR**
Cantidad de registros	2000 (30%)
Error Deseado	0.01
Iteraciones de Entrenamiento	100000
Factor de Reducción	0.9
Tasa de Aprendizaje	0.2
Método de Aprendizaje	Aditivo

Tabla 6. Parámetros entrenamiento información estado académico

Las siguientes son las variables con las cuales se realizó el entrenamiento de la red neuronal:

VARIABLE	POSIBLES VALORES
SEMESTRES_CURSADOS	Valores numéricos enteros entre 1 y 15
CREDITOS_PROGRAMA	Cantidad de créditos totales del programa académico
ESTRATO_SOCIOECONOMICO	Valores entre 0 y 5
GENERO	1 = Masculino, 2 = Femenino
COD_MODALIDAD	1 = Presencial, 2 = Distancia
COD_JORNADA	1 = Diurna, 2 = Nocturna, 3 = Virtual, 4 = Distancia
COD_FACULTAD	1 = Empresariales, 2 = Salud, 3 = Ingenierías
ESTADO	Activo = 1, Desertado = 2, Graduado = 3

Tabla 7. Variables entrenamiento/aprendizaje información estados académicos

6.2.2. Fase 2, Aprendizaje

El proceso de aprendizaje en este caso tiene como objetivo utilizar la red neuronal construida y entrenada para identificar las variables que más afectan los cambios de estados académicos de los estudiantes. Los resultados alcanzados son los siguientes:

Variable Analizada: ESTADO_ACADEMICO

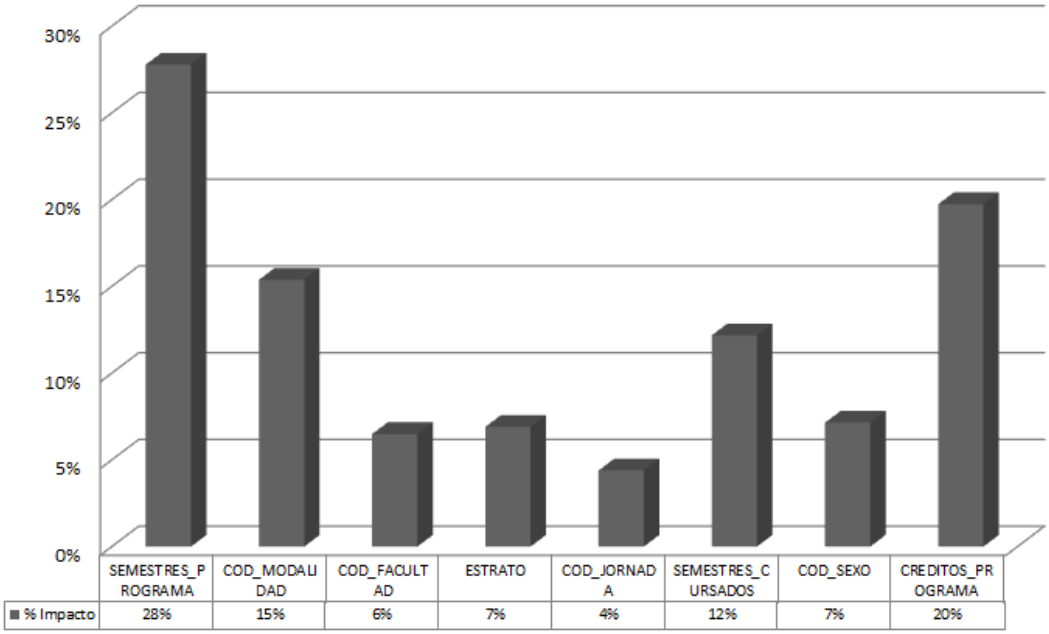


Ilustración 26. Resultado del análisis mediante redes neuronales del estado académico

6.2.3. Fase 3, Análisis de Resultados

Los resultados muestran que la mayor cantidad de cambios de estado de los estudiantes se producen en los programas académicos de mayor duración, es decir, en los programas académicos con mayor cantidad de semestres y mayor cantidad de créditos, el porcentaje con el cual las variables de duración afectan el estado académico es de un 48% acumulado entre ambas. En este punto estos resultados aún no indican si los estudiantes de programas de larga duración son más o menos propensos a la deserción, pero si son útiles para dirigir un análisis que se realizará posteriormente con el objetivo de complementar este análisis.

Este resultado puede ser analizado desde el punto de vista social si se toma como referencia que los programas académicos más largos presentan índices de deserción más altos. Esto es el reflejo de problemas tales como la falta de empleo, el costo de vida, problemas en el núcleo familiar, problemas de vocación y en general a los problemas sociológicos y sicológicos estudiados en los capítulos 3.1 y 3.2.

6.3. Análisis información de factores socioeconómicos vs académicos

En esta sección se analizará la información correspondiente a los factores socioeconómicos de la región y los datos correspondientes a la institución de educación media de cada estudiante, con respecto a su desempeño académico. Para este análisis, se aplicará la red neuronal construida a la información del Almacén de Datos obtenida en el numeral 5.1.6.3 de este documento.

6.3.1. Fase 1, Entrenamiento

El entrenamiento de la red neuronal para analizar la información socioeconómica se realizará con los siguientes parámetros:

FASE 1, Entrenamiento	
PARAMETRO	**VALOR**
Cantidad de registros	1000 (43%)
Error Deseado	0.01
Iteraciones de Entrenamiento	50000
Factor de Reducción	0.9
Tasa de Aprendizaje	0.2
Método de Aprendizaje	Aditivo

Tabla 8. Parámetros entrenamiento información socioeconómica

Las siguientes son las variables con las cuales se realizó el entrenamiento de la red neuronal:

VARIABLE	POSIBLES VALORES
PORCENTAJE_AGRICOLA	Porcentaje de agricultura en el departamento origen del estudiante
PORCENTAJE_URBANO	Porcentaje de urbanismo del departamento de origen del estudiante
POBLACION_MASCULINA	Porcentaje de población masculina del departamento de origen del estudiante
POBLACION_FEMENINA	Porcentaje de población femenina del departamento de origen del estudiante
PORCENTAJE_PRIMARIA	Porcentaje de personas que alcanzan solo educación primaria en el departamento de origen del estudiante.
PORCENTAJE_SECUNDARIA	Porcentaje de personas que logran terminar la educación secundaria en el departamento de origen del estudiante.
PORCENTAJE_MEDIA	Porcentaje de personas que terminan la educación media en el departamento de origen del estudiante.
PORCENTAJE_SUPERIOR	Porcentaje de personas que culminan estudios de educación superior en el departamento de origen del estudiante.
PORCENTAJE_ANALFABETISMO	Porcentaje de analfabetismo en el departamento origen del estudiante.
COD_ESTADO	Estado académico del estudiante. Activo = 1, Desertado = 2, Graduado = 3
NATURALEZA_COLEGIO	Indica si proviene de un colegio privado u oficial. 1 = Privado, 2 = Oficial

Tabla 9. Variables aprendizaje información socioeconómica

6.3.2. Fase 2, Aprendizaje

En esta fase se analizará mediante la red neuronal implementada cuáles variables relacionadas con las características socioeconómicas de la región y del colegio del cual proviene cada estudiante impactan en mayor grado sobre el estado académico del mismo.

Variable Analizada: ESTADO_ACADEMICO

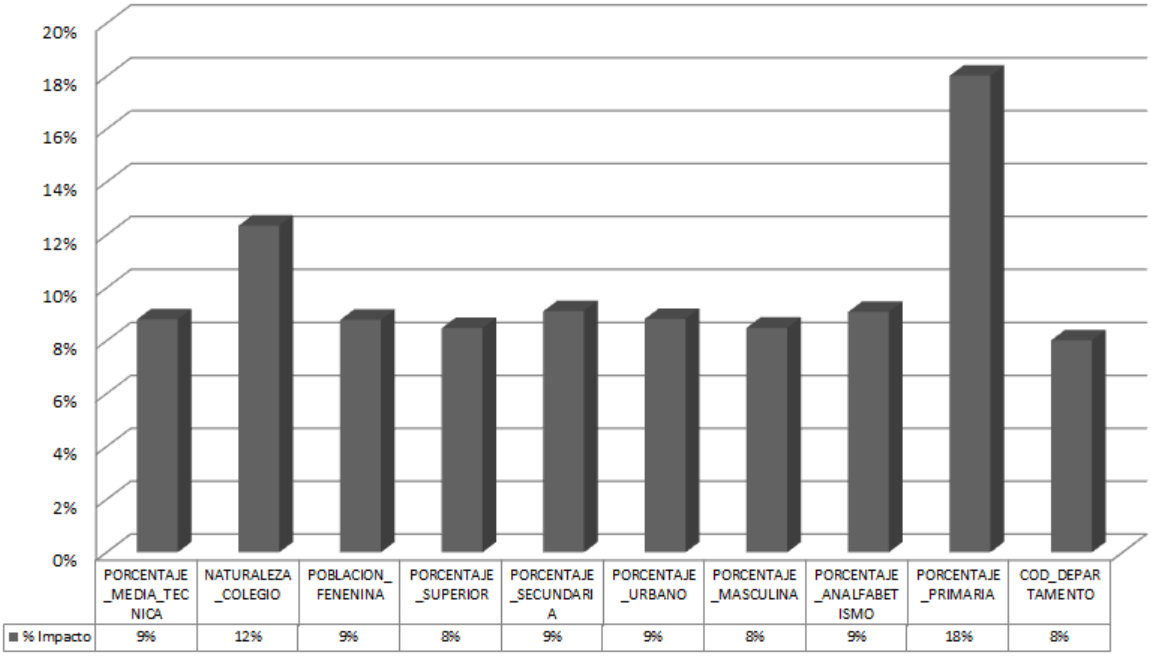


Ilustración 27. Análisis mediante redes neuronales de variables académicas con respecto a aspectos socioeconómicos

6.3.3. Fase 3, Análisis de Resultados

El resultado obtenido del análisis mediante la red neuronal, arroja como resultado que las variables que más afectan el estado académico de los estudiantes de la Universidad Autónoma de Manizales son el porcentaje de personas que alcanzan la primaria en sus respectivas regiones y la naturaleza del colegio del cual proviene cada uno de los estudiantes (Oficial o Privado). En un análisis posterior se pretende evaluar si este impacto es negativo o positivo en cada uno de los casos.

Visto desde el impacto de los factores socioeconómicos propios de la región podemos ver cómo las costumbres propias de cada región y los niveles de educación de cada una de ellas afecta la decisión de seguir o no en la universidad de un estudiante. También se puede analizar desde el punto de vista solamente económico teniendo en cuenta que las regiones con niveles académicos más bajos son también las regiones más pobres del país; presumiblemente los estudiantes de estas regiones realizan sus estudios rodeados de dificultades que los hacen propensos a la deserción.

6.4. Análisis de Resultados con Arboles de Decisión

Los árboles de decisión son una herramienta ampliamente utilizada en la inteligencia artificial. Se constituye por diagramas lógicos que permiten representar las condiciones de un problema a partir de los datos de entrada ingresados.

Durante la primera parte de este experimento, se realizó el análisis de la información del Almacén de Datos utilizando el framework de redes neuronales construido durante este proyecto. Los resultados generados de este experimento permitieron identificar las variables que más afectan al rendimiento académico y las que más influyen en el fenómeno de la deserción en la Universidad Autónoma de Manizales. Sin embargo, no es posible saber si este impacto es positivo o negativo. Con el fin de complementar estos resultados y determinar el tipo de impacto de estas variables en el problema de la deserción, se aplicarán arboles de decisión sobre la información aislando los resultados del experimento de redes neuronales.

6.4.1. Arboles de Decisión, rendimiento académico

En los capítulos 6.2 y 6.1 se realizó el análisis del rendimiento académico mediante redes neuronales dando como resultado que las variables que más afectan el rendimiento son la modalidad, la facultad y las variables relacionadas con la duración del programa. Con el fin de analizar si las variables afectan negativa o positivamente el rendimiento académico, se construye un árbol de decisión que permitirá observar de forma gráfica dicho comportamiento.

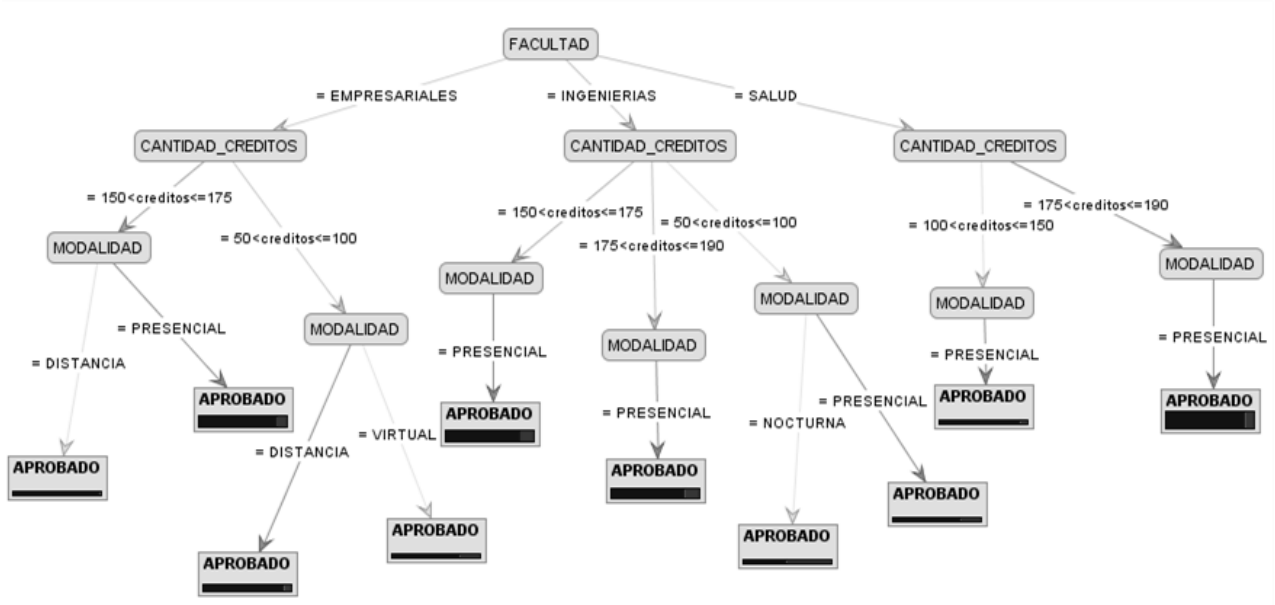


Ilustración 28. Árbol de decisión de variables relacionadas con el rendimiento académico

Según los resultados obtenidos con la técnica de árboles de decisión, la mayor cantidad de reprobaciones se presentan en los programas cortos (menos de 100 créditos) de la facultad de Ingenierías, en la modalidad nocturna. Mientras que los mejores rendimientos académicos se presentan en los programas largos (más de 150 créditos académicos) de la facultad de Empresariales en la modalidad a distancia. Los programas largos de la facultad de Salud (más de 175 créditos) también presentan un alto índice de reprobaciones. Sin embargo, este fenómeno se presenta porque también son los programas de mayor demanda y mayor cantidad de estudiantes registrados.
Analizando el fenómeno de las reprobaciones en los programas cortos de la facultad de ingenierías en el horario nocturno al nivel institucional, puede detectarse que la carga de trabajo académico extra no puede ser cubierta por personas que en su mayoría realizan actividades laborales adicionales a las académicas propias del programa. Es posible que se requiera de una reestructuración de los programas o de los métodos académicos utilizados.

6.4.2. Arboles de Decisión, Estado Académico

El análisis realizado con redes neuronales en el capítulo 6.2 acerca del estado académico de los estudiantes de la Universidad Autónoma de Manizales, arrojó como resultado que las variables que más generan cambios en el estado académico de los estudiantes son los relacionados con la duración del programa (Cantidad de Créditos, y Semestres del programa), la modalidad del programa y la cantidad de semestres cursados por el estudiante. Mediante el análisis de árboles de decisión, se pretende determinar cómo estas variables afectan el estado de los estudiantes.

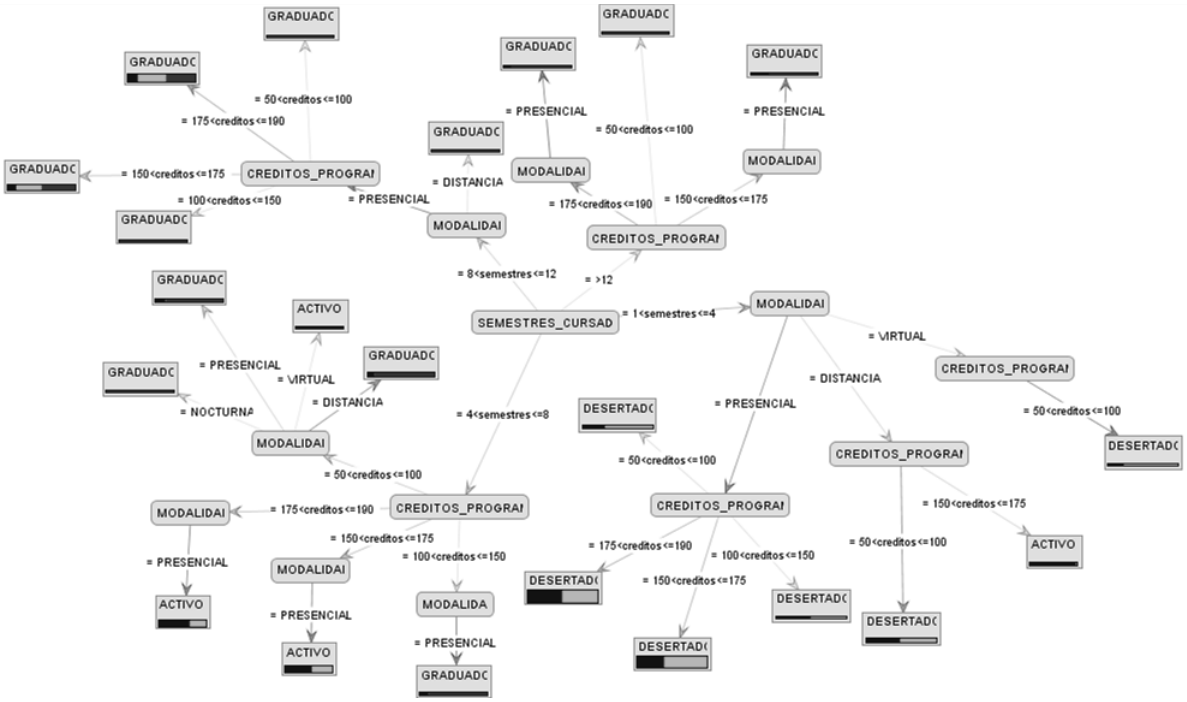


Ilustración 29. Análisis de árboles de decisión de estados académicos

De acuerdo al análisis realizado por medio de árboles de decisión de los datos del estado académico, se puede observar que los índices más altos de deserción se presentan en los programas de la modalidad presencial, en los programas de duración media (100 – 150 créditos) y en mayor cantidad en estudiantes que están cursando los primeros 4 semestres de sus programas académicos. El estudio [13] hace referencia a los problemas de adaptación de los estudiantes al entorno como una causal de deserción. Esto se manifiesta en este análisis debido a que la etapa considerada de adaptación se cumple en los primeros semestres de universidad y es el momento en el cual los estudiantes tienen sus primeros contactos con un nuevo entorno social, lo cual puede traer buenas o malas experiencias. Otro posible inconveniente que puede manifestarse en estos primeros estudiantes son problemas vocacionales asociados a los estudios previos y la preparación de los estudiantes para la educación superior.

6.4.3. Árboles de Decisión, información socioeconómica

El análisis de los datos académicos de la Universidad Autónoma de Manizales mediante redes neuronales descrito en el capítulo 6.3, arrojó como resultado que las variables socioeconómicas que más afectan el rendimiento académico y que más impactan en los cambios de estado de los estudiantes son la naturaleza del colegio del cual provienen, y los estudiantes que provienen de regiones donde los índices de población que solo alcanzan formación básica primaria son altos. A continuación se realiza el análisis de la información mediante árboles de decisión con el fin de determinar el impacto de estas variables sobre el fenómeno de la deserción.

Ilustración 30. Arboles de decisión información socioeconómica

Los resultados obtenidos con respecto a la región de la cual proviene el estudiante y el porcentaje de personas que terminaron la primaria en la región posiblemente no son muy contundentes teniendo en cuenta que la mayor concentración de estudiantes es

del departamento de Caldas y esto afecta los resultados. Sin embargo, si puede verse una diferencia en la cantidad de estudiantes desertados que provienen de colegios de naturaleza privada con respecto a la cantidad de estudiantes graduados provenientes de colegios oficiales la cual es ligeramente mayor.

7. Conclusiones

PI1. ¿Es posible mediante el uso de tecnología informática consolidar los datos académicos y socioeconómicos de los estudiantes de la Universidad Autónoma de Manizales proveniente de fuentes de información heterogéneas y poco normalizadas?

En la sección 5.1 se diseña y construye un Almacén de Datos encargado de mantener la información académica de la Universidad Autónoma de Manizales consolidada con la información socioeconómica obtenida del DANE y el ICFES. La carga se realiza mediante un proceso ETL (Extracción, Transformación y Carga) en el cual se presentaron algunos inconvenientes debido a problemas en los datos los cuales fueron resueltos y finalmente se logró el objetivo de cargar la información en una fuente de datos única usable en este y otros procesos de generación de conocimiento.

El almacén de datos es de suma importancia porque se trata de una fuente de información académica que surge a partir de gran cantidad de datos dispersos que por si solos no constituyen una fuente fiable y que no son analizables; el almacén de datos al mismo tiempo que el proceso realizado y descrito puede ser de vital importancia a la hora de emprender nuevos proyectos que tengan como finalidad analizar la información académica de la Universidad Autónoma de Manizales y pueden apoyar procesos de toma de decisiones por parte de las directivas de la misma.

PI2. ¿Se puede encontrar información pertinente que facilite la toma de decisiones acerca del fenómeno de la deserción estudiantil en la Universidad Autónoma de Manizales a partir de los datos académicos con los que cuenta en la actualidad la oficina de registro académico?

En la actualidad la información académica de la Universidad Autónoma de Manizales soporta los sistemas transaccionales de la universidad de forma adecuada. Estos sistemas generan grandes cantidades de información lo cual complica su análisis mediante técnicas tradicionales como la estadística o las inspecciones visuales. La gran cantidad de información sumada a los múltiples problemas de normalización a causa de la evolución propia de las aplicaciones hace complejos los procesos de análisis de los datos y generación de conocimiento que pueda ser útil a la hora de tomar decisiones acerca del problema de la deserción.

Pese a algunos inconvenientes identificados con los datos académicos durante el proceso de ETL, se puede determinar que si es posible realizar procesos de ETL incluso cruzar información de otras fuentes académicas y no académicas con el objetivo de contar con un almacén de datos más completo que permitan realizar análisis mas profundos y efectivos de diferentes problemas académicos.

Cabe resaltar también que a partir de algunos de los problemas encontrados durante el proceso de ETL pueden surgir recomendaciones y sugerencias para realizar

mejoras en la forma como los sistemas transaccionales académicos recolectan los datos y en como estos apoyan a los proyectos de inteligencia de negocios y minería de datos.

PI3. Mediante el uso de técnicas de inteligencia artificial ¿es posible generar nuevo conocimiento que sea valioso al momento de realizar un análisis del problema de la deserción estudiantil en la Universidad Autónoma de Manizales? Durante la ejecución de este proyecto se realizó un proceso cuyo objetivo era la identificación y predicción de relaciones entre las variables relacionadas en la información académica y socioeconómica conseguida a través del proceso de ETL. Este proceso se realizó en dos fases: en una fase inicial se analizó la información mediante técnicas de minería de datos 6.1, y los resultados de este análisis permitieron identificar las variables que más impactaban en el rendimiento académico y en el estado de los estudiantes. La segunda etapa del experimento 6.4 se realizó el análisis de dichas variables mediante árboles de decisión que permitieron analizar el tipo de impacto que tenían las variables identificadas en la primera fase sobre el problema de la deserción académica en la Universidad Autónoma de Manizales. En cumplimiento del objetivo planteado se detectaron algunas de las variables que afectan directa o indirectamente el rendimiento académico e impactan en el problema de la deserción. Sin embargo, no es posible identificar en el desarrollo de este trabajo si esta información es o no útil a la hora de apoyar la toma de decisiones; esto es una medida que puede ser validada exclusivamente por las entidades encargadas en la Universidad Autónoma de Manizales.

8. Trabajo Futuro

Los resultados obtenidos indican que los estudiantes mas propensos a la deserción académica provienen en gran proporción de regiones alejadas de la Universidad Autónoma de Manizales 6.4.3, este resultado obtenido a través del análisis mediante minería de datos puede ser una entrada importante en futuros proyectos de definición de estrategias que promuevan como objetivo la permanencia de estudiantes de esta población altamente vulnerable a la deserción.

Durante el proyecto se identifico que los programas que presentaban mayor cantidad de deserciones son aquellos de más larga duración 6.4.2, por esta razón se hace necesario dentro de los proyectos de diseño de los programas académicos tener en cuenta la el riesgo latente de deserción a mayor duración del programa, incluyendo en los nuevos programas (y en los existentes) actividades académicas y extra-académicas que promuevan la permanencia de los estudiantes.

El trabajo realizado puede ser considerado línea base para futuros proyectos desde diferentes perspectivas, algunas de ellas y como pueden ser útiles:

8.1. Almacén de Datos

El almacén de datos diseñado y construido para como fuente de información para este proyecto contiene información académica de la Universidad Autónoma de Manizales relacionada en gran parte con la información socioeconómica de los estudiantes de la misma, dicha información puede ser de utilidad para proyectos que requieran aplicar técnicas de minería de datos o inteligencia de negocios con el objetivo de analizar el fenómeno de la deserción u otros problemas similares de la Universidad Autónoma de Manizales.

El proceso de ETL definido y aplicado sobre los datos académicos y socioeconómicos puede ser de utilidad al momento de actualizar el almacén de datos generado durante este proyecto o complementarlo con nuevas variables que permitan analizar desde otras perspectivas problemas como el de la deserción académica o con el fin de analizar los resultados de las decisiones tomadas a partir de los resultados de este proyecto.
El proceso de ETL generado y aplicado durante este proyecto puede ser usado en futuros procesos de minería de datos o inteligencia de negocios como apoyo al momento de diseñar sus propios procesos de ETL debido a que en este se describe la forma como se enfrentaron los diferentes inconvenientes que se presentaban dentro de los datos transaccionales usados como fuente de información académica los cuales no son ajenos a ningún proceso que toma como base información de sistemas transaccionales.

8.2. Framework de Minería de Datos

El framework diseñado, construido y utilizado durante la ejecución de este proyecto permite el análisis de información mediante la técnica de redes neuronales seleccionada lo cual puede ser de utilidad para algunos procesos de minería de datos, sin embargo el diseño del framework permite que nuevas técnicas sean incluidas con un esfuerzo relativamente bajo lo cual permitiría a futuros proyectos invertir mas recursos en otras fases del proceso de minería de datos.

Referencias

[1] L. M. Caraveo, “La Flexibilidad en la Educación Superior,” 2002.

[2] Y. Navarro, “¿ Profesionistas del futuro o futuros taxistas ? Los egresados universitarios y el mercado laboral en México,” *Doctor*, 2009.

[3] G. M. Arango, *La educación superior en Colombia Análisis y estrategias para su desarrollo*. 2004.

[4] A. S. Escarria, “Deserción Universitaria en Colombia,” pp. 50–60, 2010.

[5] F. Zuluaga, A. Jaramillo, and U. Eafit, “Revista de Economía del Rosario Determinantes de la deimanda por educación superior en Colombia,” vol. 11, no. 1, pp. 121–148, 2008.

[6] J. Tinto, Vincent; Cullen, “desercion.vincent.tinto.pdf.” 1973.

[7] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[8] S. Vallejos, “Minería de Datos,” 2006.

[9] J. McCarthy, “What is artificial intelligence?,” 2007.

[10] Z. E. Akkaoui, E. Zimányi, J. Mazón, and J. Trujillo, “A Model-Driven Framework for ETL Process Development,” *Computing Systems*, pp. 45–52, 2011.

[11] Data Prix, “La metodología CRISP-DM,” 2009. [Online]. Available: http://www.dataprix.com/la-metodologí-crisp-dm.

[12] I. Fishbein, M.; Ajzen, “Understanding attitudes and predicting social behavior.” 1980.

[13] W. Spady, “Dropouts from higher education: An interdisciplinary review and synthesis,” *Interchange*, vol. 1, no. 1, pp. 64–85.

[14] L. Grecia, Luciano; Porto, Alberto; Ripani, “Rendimiento de los Estudiantes de las Universidades Públicas Argentinas,” 2002.

[15] O. Londoño, “Estudio del fenómeno de la deserción voluntaria estudiantil de la jornada nocturna del programa de Administración de Empresas de la Universidad Cooperativa, Seccional Santa Marta en el periodo 1986 - 1996,” 2000.

[16] E. Cárdenas, “Estudio de la deserción estudiantil en programas de ingeniería de la Universidad Nacional de Colombia,” 1996.

[17] F. Sánchez, M. Quirós, C. Reverón, and A. Rodríguez, “Equidad Social en el acceso y permanencia en la universidad pública determinantes y factores asociados,” vol. 7191, pp. 1–48, 2002.

[18] G. Lopez, C. Cardozo, M. Posada, and D. J. Cuartas, “Specific Actions for Desertion Reduction , Competence Identification and Guidance for New Students of an Engineering Program , a Case Study,” *Therapy*, pp. 165–170, 2010.

[19] R. Timarán, “La Minería de Datos en el Descubrimiento de Perfiles de Deserción Estudiantil en la Universidad de Nariño,” 2009.

[20] M. de E. Nacional, “SPADIES - Sistema Parao la Prevencióon de la Deserción en la Educación Superior,” 2002. .

[21] L. Longo and S. Barrett, “A Context-Aware Approach Based on Self-organizing maps to study Web-Users ’ tendencies from their behaviour,” *Training*, pp. 12–17, 2000.

[22] G. Moreno, “Técnicas más usadas en la Mineria de Datos,” 2010. [Online]. Available: http://gamoreno.wordpress.com/2007/10/03/tecnicas-mas-usadas-en-la-mineria-de-datos/.

[23] J. Wiley, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. 2004, p. 416.

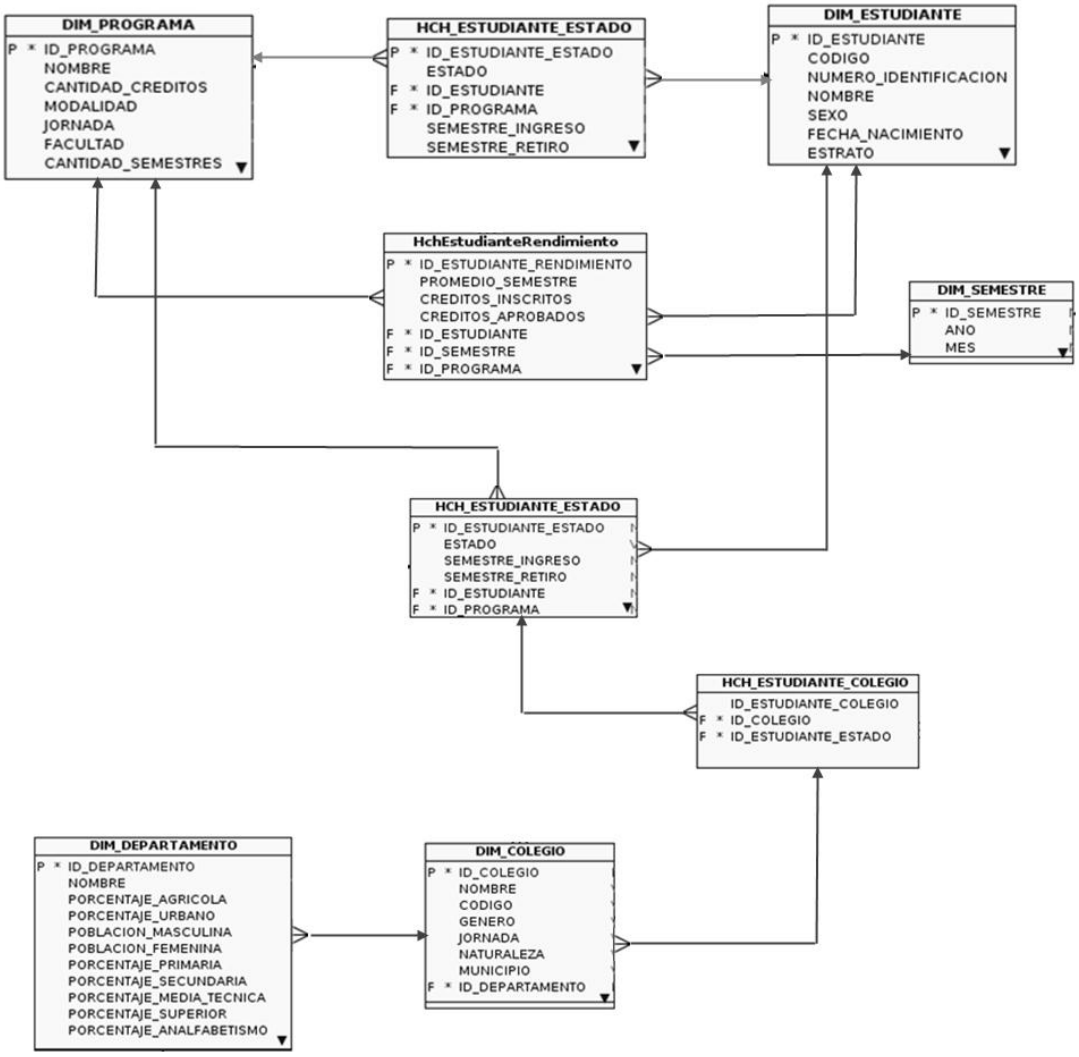
[24] Rapid-I, “RapidMiner,” *http://rapid-i.com/content/view/181/190/*, 2011. .

[25] J. Gamma, Erich; Helm, Richard; Johnson, Ralph; Vlissides, *Design Patterns, Elements of REusable Object-Oriented Software*. 1999.

[26] R. Lago, “Patrón: Modelo Vista Controlador,” *http://www.proactiva-calidad.com/java/patrones/mvc.html*, 2007. .

Anexos

ANEXO A:
Diseño Almacén de Datos
Imagen del modelo Entidad/Relación resultante del proceso de diseño del almacén de datos.



ANEXO B:
Proceso ETL
Recursos usados durante el proceso de ETL, entre ellos se pueden encontrar ejecutables del software RapidMiner, clases Java y hojas de calculo (xlsx). Se adjuntan en el directorio *ETL*.

ANEXO C:
Código Fuente Framework Minería de Datos
Código fuente java usado para la construcción del framework de minería de datos. Adjunto en el directorio *framework.src.*

ANEXO D:
Ejecutable Framework Minería de Datos
Archivo ejecutable *datamining.framework.jar* del framework de minería de datos construido durante el desarrollo de este proyecto. Archivo ejecutable de la tecnología Java. Adjunto en el directorio *framework.exe.*

ANEXO E:
Almacén de Datos
Se anexa almacén de datos generado durante el proyecto y utilizado para el proceso de análisis, se anexa en formato legible (XLS) el mismo usado por el framework de minería de datos. El almacén de datos generado se encuentra adjunto en el directorio *datawarehouse.*