

INSTITUCIÓN UNIVERSITARIA POLITÉCNICO GRANCOLOMBIANO
FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS
MAESTRÍA EN INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN BDA & ARS

Análisis de deserción en una Institución de Educación Superior
Colombiana mediante el uso de Minería de Datos para la
Educación (EDM)

Presenta:

Andrés Villanueva Manjarres

Cod: 1310020763

Asesor Temático

MSc. Luis Gabriel Moreno Sandoval

Bogotá, 2016

ÍNDICE GENERAL

1. INTRODUCCIÓN	7
1.1 Objetivo General	8
1.1.1 Objetivos Específicos	8
1.2 Justificación	8
1.3 Planteamiento del problema	9
2. MARCO TEÓRICO	10
2.1 Deserción	10
2.2 Modelo de Atención Integral al Estudiante MAIE	11
2.3 Educación tradicional	12
2.4 Tipos de ambientes educativos	12
2.5 Educación basada en Computador	13
2.6 Minería de Datos:	13
2.6.1 Implementación de un proyecto de Minería de Datos	14
2.7 Minería de Datos para la Educación	15
2.7.1 Antecedentes	16
2.7.2 Conferencias internacionales	17
2.7.3 Líneas de investigación futuras	19
3. MARCO METODOLÓGICO	20
3.1 Fase 1. Contextualización para el desarrollo del trabajo	20
3.2 Fase 2: Diseño del caso de estudio	21
3.3 Fase 3: Ejecución del caso de estudio	21
4. DESARROLLO E IMPLEMENTACIÓN	22
4.1 Documentación General del Proyecto	22
4.1.1 Clasificación de técnicas de minería de datos utilizadas en entornos educativos	22
4.1.2 Trabajos clasificados por técnicas de minerías de datos	22
4.1.3 Clasificación de trabajos por dominios educativos	24
4.1.4 Técnicas de minería de datos utilizadas por dominios educativos	26
4.2 Diseño del Caso de estudio	27
4.2.1 Institución educativa	28
4.2.2 Recolección de los datos	29
4.2.3 Selección de programas académicos	31
4.3 Ejecución Caso de Estudio	32
4.3.1 Preparación de los datos	32
4.3.2 Modelado	37

5. RESULTADOS	44
5.1 Resultados fase contextualización	44
5.2 Resultados análisis de datos.....	44
5.3 Resultados fase implementación Caso de Estudio	45
6. DISCUSIÓN Y CONCLUSIONES	47
7. REFERENCIAS.....	49

ÍNDICE DE TABLAS

Tabla 1. Abandono de estudios por semestre.	10
Tabla 2. Trabajos aceptados en las International Conference on Educational Data Mining	19
Tabla 3. Consolidado de trabajos revisados por técnicas de minería de datos	23
Tabla 4. Clasificación de trabajos por dominios presentes en la educación.....	25
Tabla 5. Datos relevantes para el desarrollo del caso de estudio.....	29
Tabla 6. Tabla archivos con datos utilizados en el caso de estudio	31
Tabla 7. Número de estudiantes por programas seleccionados para el análisis	32
Tabla 8. Datos utilizados para el caso de estudio	32
Tabla 9. Lista de características con valores nulos.....	34
Tabla 10. Análisis de atributos relevantes	36
Tabla 11. Atributos relevantes en los conjuntos de datos.....	37
Tabla 12. Distribución de datos para entrenamiento, validación y pruebas	37
Tabla 13. Precisión detallada IBk sede Cundinamarca.....	41
Tabla 14. Precisión detallada IBk facultad Ingeniería	41
Tabla 15. Precisión detallada OneR sede Cundinamarca.....	42
Tabla 16. Precisión detallada OneR facultad Ingeniería	42
Tabla 17. Clasificaciones por algoritmos sede Cundinamarca	46
Tabla 18. Clasificaciones por algoritmos facultad de Ingeniería.....	46

ÍNDICE DE FIGURAS

Figura 1. Modelo de Atención Integral al Estudiante MAIE.....	12
Figura 2. Tipos de entornos y sistemas educativos tradicionales y basados en computadoras..	13
Figura 3. Clasificación técnicas minería de datos.....	14
Figura 4. Principales áreas involucradas en la minería de datos para la educación	15
Figura 5. Publicaciones sobre EDM registradas en SCOPUS y ScienceDirect.....	17
Figura 6. Metodología de la investigación	20
Figura 7. Clasificación de trabajos por dominios presentes en la educación	26
Figura 8. Fases del modelo de referencia CRISP-DM	28
Figura 9. Deserción estudiantil por metodología de la sede de Cundinamarca	29
Figura 10. Datos deseados para el análisis de deserción y datos obtenidos	31
Figura 11. Árbol J48 estudiantes sede Cundinamarca	38
Figura 12. Árbol J48 estudiantes facultad Ingeniería 6 atributos relevantes	39
Figura 13. Árbol J48 estudiantes facultad Ingeniería 31 atributos	40

RESUMEN

Este trabajo se desarrolló con el fin de implementar algunas técnicas de minería de datos que nos permitieran identificar patrones de deserción en estudiantes de una institución educativa en Colombia. El fundamento teórico para el desarrollo del caso de estudio fue una revisión de literatura de técnicas de minería de datos aplicadas en entornos educativos, dicha revisión de literatura nos dio las pautas para el diseño y la implementación del caso de estudio a fin de obtener los patrones de deserción en la institución seleccionada, objetivo principal del trabajo.

ABSTRACT

This work was developed in order to implement some data mining techniques that allow us to identify patterns of dropout students in an educational institution in Colombia. The theoretical foundation for the development of the case study was a literature review of techniques of data mining applied in educational environments, the literature review gave us the guidelines for the design and implementation of the case study looking to obtain patterns desertion in the selected institution, the main objective of the work.

PALABRAS CLAVE

Minería de datos para la educación, Minería de datos, Sistemas educativos, Clasificación, Weka.

KEY WORDS

Educational Data Mining, Data mining, Educational Systems, Clasification, Weka.

1. INTRODUCCIÓN

El fenómeno de la deserción es una realidad que deben enfrentar todos los actores del sector educativo, principalmente aquellos vinculados al nivel de la educación superior pues este es, según el Ministerio de Educación Nacional de Colombia el que tiene más altos índices de deserción.

Son notables los múltiples trabajos que se han adelantado para analizar esta problemática desde diferentes perspectivas, una de ellas el análisis de los datos asociados a los procesos educativos, los cuales buscan identificar patrones que faciliten la detección de aquellos estudiantes en riesgo de deserción con el fin de establecer políticas que puedan llevar a las instituciones educativas a reducir los índices de deserción, situación que favorecería tanto a los estudiantes como a las propias instituciones.

El mundo actual es muy diferente de aquel en el que empezamos a vivir hace unas décadas, la cantidad de información generada, los diversos medios que producen datos, las disciplinas en las que nos podemos desempeñar, en fin, el mundo ha cambiado y de manera adaptativa nosotros lo hemos hecho con él. En este cambio la toma de decisiones es una actividad que cada vez se hace esperar menos, especialmente cuando estas decisiones son definitivas para el futuro de las organizaciones. En escenarios educativos, uno de los principales objetivos es conseguir que los estudiantes permanezcan en sus programas académicos hasta la finalización de estos y la reducción en los índices de deserción. Es aquí donde la Minería de Datos para la Educación o *Educational Data Mining (EDM)* por sus siglas en inglés aportan un valor fundamental en las instituciones educativas.

¿Cómo negar las múltiples ventajas y facilidades que ofrece la tecnología actualmente?, ¿cómo ocultar la importancia que Internet brinda de manera general?, ¿cómo ocultar el crecimiento e importante penetración de internet en todos los escenarios de la vida y de todos nosotros?[1]. La educación no es la excepción a los cambios que se están dando, los modelos educativos han cambiado, la manera de enseñar y de acceder al aprendizaje ha experimentado modificaciones significativas, los modelos de aprendizaje cambian día a día y sin duda la forma en la que los estudiantes interactúan con los recursos de aprendizaje y la influencia que tiene las redes sociales en los procesos de educación son cada vez más significativas [2].

La elaboración del presente trabajo busca integrar dos paradigmas, educación y minería de datos, para ello se adelantó una revisión de literatura de técnicas de minería de datos aplicadas a entornos educativos, y con base en la revisión realizada se seleccionarán algunas técnicas de minería de datos que se aplicarán con el fin de identificar riesgos de deserción en estudiantes de programas de educación superior en una institución colombiana. El desarrollo del presente

trabajo busca aportar elementos que apoyen la toma de decisiones respecto a las estrategias para reducir los niveles de deserción en la institución de educación superior seleccionada.

1.1 Objetivo General

Identificar riesgos de deserción en estudiantes de educación superior mediante la implementación de técnicas de minería de datos.

1.1.1 Objetivos Específicos

1. Clasificar las técnicas de Minería de datos usadas en escenarios educativos para definir las más apropiadas en la identificación de los riesgos de deserción.
2. Analizar los indicadores del modelo de prevención de la deserción (MAIE) para establecer los requerimientos técnicos necesarios que permitan aplicar las técnicas de Minería de datos seleccionadas.
3. Aplicar las técnicas de Minería de Datos seleccionadas para medir el riesgo de deserción en estudiantes de Educación Superior en la institución y los programas seleccionados.

1.2 Justificación

Los actores que toman decisiones y planifican día a día las actividades en las instituciones educativas deben disponer de manera constante de la información en la cual se puedan fundamentar sus decisiones, por otra parte, los actores que participan en los diferentes procesos educativos deben contar con esta información para poder evaluar el desempeño de las actividades realizadas. En general, todos los usuarios que tiene roles estratégicos en las organizaciones educativas necesitan disponer de información tanto consolidada como detallada de las actividades asociadas a los estudiantes, predecir tendencias y comportamientos de estos y adelantar actividades orientadas a los objetivos estratégicos de las instituciones.

Uno de los principales problemas que se presentan en la educación superior en la actualidad es la deserción, para no ir muy lejos, en Colombia, en 2013, la deserción en el nivel universitario alcanzó el 44.9%, en el nivel técnico alcanzó 62.4% y en el nivel y tecnológico alcanzó el 53.8%. Muchos son los estudios que se han desarrollado para entender este fenómeno. Por otra parte, a nivel mundial se han adelantado varios casos de estudio en los cuales se utilizaron técnicas de minería de datos para abordar problemáticas educativas.

Debido al interés que está generando la minería de datos educativa en diferentes entidades y al creciente número de investigadores resulta muy pertinente utilizar como base las investigaciones realizadas anteriormente para lograr identificar las técnicas de minería de datos que permitan identificar riesgos de deserción en estudiantes de programas virtuales de

educación superior y adelantar un caso de estudio en una institución de educación superior de Colombia.

1.3 Planteamiento del problema

En Colombia, en el año 2013, la deserción en el nivel universitario alcanzó el 44.9%, en el nivel técnico alcanzó 62.4% y en el nivel tecnológico alcanzó el 53.8% [3]. Estos niveles han motivado a diferentes entidades e instituciones a analizar el fenómeno de la deserción desde diferentes perspectivas una de ellas el análisis de los datos almacenados en sus sistemas de información. Hoy en día, las instituciones educativas tienen diversas fuentes de información (sistema de gestión académica, registros en hojas de cálculo, etc.) en las cuales se almacenan datos relacionados a los procesos educativos en todas sus fases, pero usualmente los datos almacenados no son analizados, tampoco se cuenta con un modelo o un sistema que permita hacer una exploración de los datos de carácter educativo para posteriormente generar estrategias pedagógicas, técnicas o administrativas y evaluar el resultado buscando oportunidades de mejora en todos los escenarios del proceso educativo.

Por otra parte, el crecimiento de la minería de datos en los últimos años es muy significativo [4], tanto que hoy en día existen diferentes sistemas que permiten hacer minería de datos, pero no existen herramientas desarrolladas o metodologías definidas específicamente para educación y estas son requeridas debido a la especificación de los actores y el objetivo que se busca. El desarrollo de un modelo para implementar EDM en los programas formación virtual estaría totalmente enfocado en un problema histórico de la enseñanza, ¿cómo conseguir un incremento importante en el aprendizaje de todos los estudiantes? y para ello se tendría como soporte las ventajas que la tecnología actual nos ofrece.

2. MARCO TEÓRICO

En este capítulo se busca contextualizar al lector en los conceptos relevantes en el desarrollo de este trabajo, se busca presentar de manera general los elementos en los que se fundamenta el desarrollo del trabajo. Se presentan de manera concreta algunas aclaraciones relacionadas a la deserción, las técnicas de minería de datos que se han implementado en educación y que ha aportado cada una de ellas; también se busca presentar los vacíos que las técnicas utilizadas han dejado y aportar los elementos que se utilizarán en el desarrollo de los experimentos.

2.1 Deserción

La presente sección no busca profundizar los aspectos relacionados con la deserción, tampoco ser un referente en cuando a la deserción de estudiantes de programas de educación superior, sin embargo, con el fin de delimitar el contexto en el que se elaborará el presente trabajo, en esta sección se presentarán de manera general los elementos que permiten entender el fenómeno de la deserción especialmente aquellos relacionados con los programas de educación superior.

De manera general, la deserción es considerada como el abandono temporal o definitivo del aula de clase por parte de los estudiantes, de los programas académicos para las cuales se matricularon[5], [6]. En Colombia, en 2013, la deserción en el nivel universitario alcanzó el 44.9%, en el nivel técnico alcanzó 62.4% y en el nivel y tecnológico alcanzó el 53.8% [3], niveles que superaron ampliamente los registrados en el año 2013. Cifras bastante alarmantes para todos los actores del sector educativo colombiano. En el caso de la educación superior, y analizando el “momento” en el que los estudiantes abandonan sus estudios se ha encontrado que estos abandonan principalmente durante los primeros semestres de sus carreras [6] como se puede ver a continuación:

Tabla 1. Abandono de estudios por semestre.

Semestre de abandono	Porcentaje
Primero	37%
Segundo	16%
Tercero	11%
Cuarto	8%
Quinto	6%
Sexto	6%
Séptimo	4%
Octavo	4%
Noveno	4%

Decimo	5%
--------	----

Fuente: Ministerio de Educación Nacional. [7]

En Colombia, la entidad encargada de analizar las cifras de deserción en educación superior es el Ministerio de Educación Nacional, para ello la entidad en el año 2008 creo la herramienta llamada SPADIES la cual se utiliza para hacer seguimiento sobre las cifras de deserción de estudiantes de la educación superior [8]. Este seguimiento se realiza analizando los datos suministrados por las instituciones de educación superior, se identifican y se ponderan los comportamientos, las causas, variables y riesgos determinantes para desertar. Los resultados presentados por el sistema SPADIES se presentan en el anexo 01 Cifras de deserción en Colombia según SPADIES – MEN.

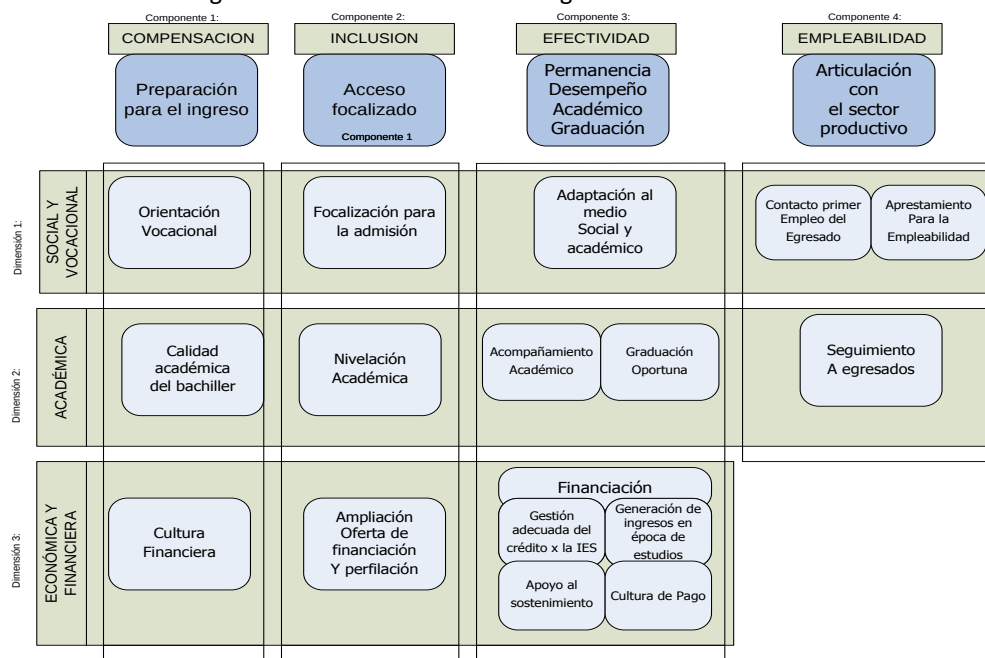
Según un artículo publicado por el Ministerio de Educación nacional en el año 2010, el fenómeno de la deserción se presenta por tres aspectos fundamentales, el rendimiento, el cual hace referencia al conocimiento adquirido por los estudiantes durante la educación básica y media; la capacidad económica; y la orientación vocacional de los estudiantes [7].

2.2 Modelo de Atención Integral al Estudiante MAIE

En el año 2011 el ICETEX, la entidad del estado colombiano que promueve la Educación Superior a través del otorgamiento de créditos educativos, publicó el modelo de atención integral al estudiante MAIE., el cual tiene como base fundamental la adopción de estrategias de acción afirmativa para que los usuarios de la financiación educativa permanezcan, se gradúen oportunamente y puedan crear empresas u obtener un empleo una vez culminen sus estudios.

El objetivo primordial del MAIE es en brindar oportunidades de acceso y permanencia en la educación superior. El modelo MAIE se compone de tres dimensiones las cuales son socio-vocacional, académica y económico-financiera, así mismo, cada una de estas dimensiones es analizada por medio de componentes como compensación, inclusión, efectividad y empleabilidad [9].

Figura 1. Modelo de Atención Integral al Estudiante MAIE.



Fuente: ICETEX

<https://www.icetex.gov.co/dnnpro5/es-co/iesycooperativas/modelodeatenci%C3%B3nintegral.aspx>

2.3 Educación tradicional

La Educación tradicional es el sistema educativo más utilizado y se refiere a las escuelas y universidades presenciales donde se ofrece la enseñanza a partir de cátedras en un salón de clases. En estos entornos, los sistemas recolectan información de todo lo que les rodea y el contexto en general, es decir: la asistencia, las notas, logros del plan de estudios a nivel general e individual. Estas instituciones educativas cuentan con fuentes diversas de información como: bases de datos tradicionales con información del estudiante y todo su proceso educativo e información en línea con información adicional de contenido educativo [10].

2.4 Tipos de ambientes educativos

Los tipos de ambientes educativos que existen hoy día y que pueden ser explorados y analizados por la Minería de Datos Educativa son la Educación tradicional y la Educación basada en Computador. Cada uno de estos dos ambientes proporciona datos de diferentes fuentes que deben ser pre-procesados de maneras particulares dependiendo de la naturaleza de cada uno de ellos, los problemas y las tareas específicas a ser resueltas [10].

2.5 Educación basada en Computador

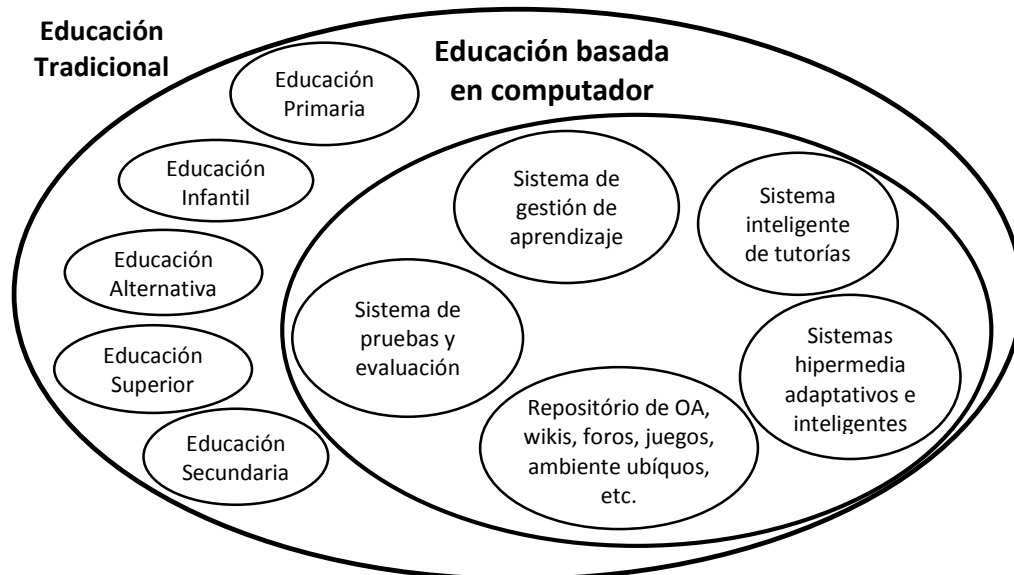
La Educación basada en Computador es el uso de Computadores para proporcionar dirección, instruir o gestionar las instrucciones dadas al estudiante.

En un inicio estas aplicaciones fueron de escritorio y se ejecutaban en computadores sin conexiones entre ellos y sin la utilización de técnicas de inteligencia artificial para el modelado, adaptación y personalización de los estudiantes hasta el momento en que Internet popularizó estos sistemas hasta el punto de ser denominados y conocidos masivamente como *e-learning*, sistemas de instrucción en línea, que hoy en día son de uso extensivo entre instituciones netamente virtuales e instituciones educativas tradicionales para extender sus actividades a más personas y lugares poco accesibles [11]. Entre ellos se encuentran los siguientes:

- LMS, ITS - Sistemas de gestión de aprendizaje
- AIHS - Sistemas Hipermedia Adaptativos
- Sistemas de pruebas y evaluación

Otros tipos de Sistemas de Educación basada por Computador (CBE)

Figura 2. Tipos de entornos y sistemas educativos tradicionales y basados en computadoras



Fuente: *Data mining in education*. [10].

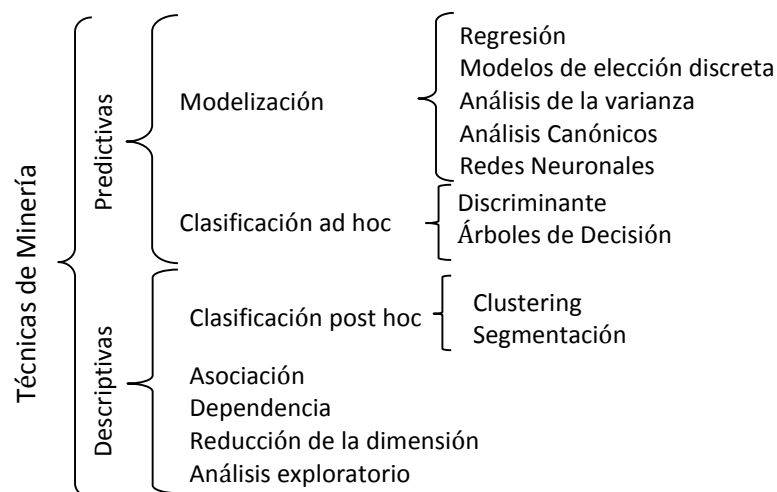
2.6 Minería de Datos:

La minería de datos es un conjunto de técnicas y procedimientos que puede ser desarrollado a partir de diversas fuentes de datos como bodegas de datos o bases de datos relacionales, hasta archivos planos sin formatos y realiza a partir de estos datos, análisis predictivos mediante técnicas de análisis estadístico para predecir o prever medidas de certeza basada en los hechos existentes. La minería de datos también puede definirse como un proceso no trivial de

identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos [12]. La minería de datos también es vista como la evolución de las tecnologías de la información [13], esto en gran medida apalancado por el crecimiento de internet que ha tomado gran relevancia en los últimos años [1][14].

La minería de datos es un campo multidisciplinario que permite obtener información relevante de grandes cantidades de datos en donde confluyen entre otras áreas inteligencia artificial, estadística, bases de datos, ciencias de la información, y otras. Los tipos de aplicaciones de la minería de datos se caracterizan por la respuesta que se busca responder, y de manera general, podríamos clasificar de la siguiente manera:

Figura 3. Clasificación técnicas minería de datos



Fuente: Minería de datos conceptos, técnicas y sistemas [15]

2.6.1 Implementación de un proyecto de Minería de Datos

Los proyectos de minería de datos se implementan con el objetivo de descubrir patrones de información relevante e interesante en grandes volúmenes de información. Esto se hace con el desarrollo de 4 fases [16][12], que usualmente son:

1. Filtrado de datos.
2. Selección de variables.
3. Extracción de conocimiento.
4. Interpretación y evaluación.

Estas fases se desarrollan indistintamente de la técnica de minería de datos que se presenta usar y en escenarios educativos no es la excepción. En general todas las técnicas han sido probadas en el ámbito educativo[17], y se han adelantado diferentes casos de estudio para evaluar el desempeño de las diversas técnica en escenarios educativos. Todos los casos de estudio han sido adelantados con el fin de satisfacer el principal objetivo de la Minería de datos para la educación, el cual busca identificar patrones de comportamiento de los alumnos en su entorno académico;

clasificar los tipos de alumnos de acuerdo al rendimiento registrado; clasificación de docentes de acuerdo a las actividades desarrolladas y el uso de las plataformas; identificación de patrones exitosos en el uso de las ambientes virtuales de aprendizaje entre muchos otros.

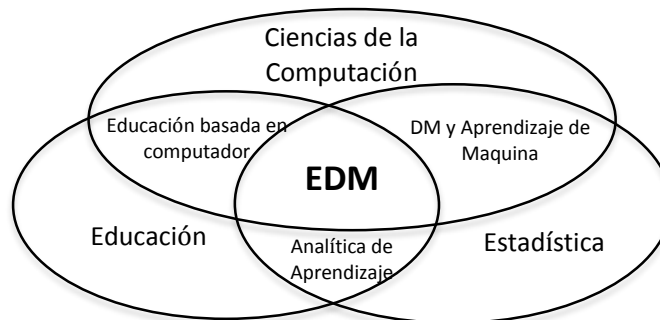
2.7 Minería de Datos para la Educación

La importancia que está adquiriendo en los últimos años la educación virtual o *e-learning*, ha dado paso al nacimiento de una disciplina que emerge como herramienta que permite dar solución a diferentes paradigmas que asociados a la educación virtual y la minería de datos que se pueda asociar a este tipo de educación. Esta disciplina se conoce como EDM [18] [19] [11] [20]. La Minería de datos para la Educación se centra en el desarrollo de métodos para la exploración de tipos específicos de datos que provienen del contexto educativo. Su objetivo global es lograr comprender cómo aprenden los estudiantes e identificar aquellos aspectos en los que es posible mejorar en aspectos educativos y de aprendizaje. La Minería de Datos para la educación es tan importante que existe en el año 2007 se fundó la *International Educational Data Mining Society*, organización sin ánimo de lucro dedicada a aumentar las oportunidades de participación en el área de la EDM. Tal es la importancia que ha suscitado esta área de investigación que se estima que para el año 2022 toda la investigación relacionada a educación involucrará análisis y minería de datos [21].

La información de los sistemas educativos en la actualidad almacena grandes cantidades de datos y su origen puede provenir de diversas fuentes, diversos formatos y diferentes niveles de granularidad [22]. Los problemas de minería de datos educativa, deben analizarse de manera particular debido a que su objetivo específico determina una singularidad al momento de resolverse mediante técnicas de minería de datos.

La minería de datos en educación puede analizar los datos generados por cualquier sistema de aprendizaje y enfocarse en diversos aspectos tanto individuales como grupales y tener en cuenta datos subyacentes; datos administrativos, datos demográficos, datos motivacionales que a su vez contienen múltiples niveles de jerarquía, contextos, niveles de granularidad y datos históricos. Se llama interdisciplinaria a la minería de datos en educación por que puede involucrar el análisis de redes sociales, psicopedagogía, psicología cognitiva, psicometría entre otras.

Figura 4. Principales áreas involucradas en la minería de datos para la educación



Fuente: *Data mining in education*. [10]

El campo más relacionado de las áreas del gráfico es *Learning Analytics* (Análisis del Aprendizaje) que se puede definir como la medición, recopilación, análisis y presentación de informes de datos sobre los estudiantes y sus contextos, para fines de la comprensión y el aprendizaje y la optimización de los entornos en los que se produce, por tanto, la Minería de Datos para Educación puede compartir muchos atributos entre todas y cada una de las áreas que la rodea [10].

Los tipos de ambientes educativos que existen hoy día y que pueden ser explorados y analizados por la Minería de Datos Educativa son la Educación tradicional y la Educación basada en Computador. Cada uno de estos dos ambientes proporciona datos de diferentes fuentes que deben ser pre-procesados de maneras particulares dependiendo de la naturaleza de cada uno de ellos, los problemas y las tareas específicas a ser resueltas [10].

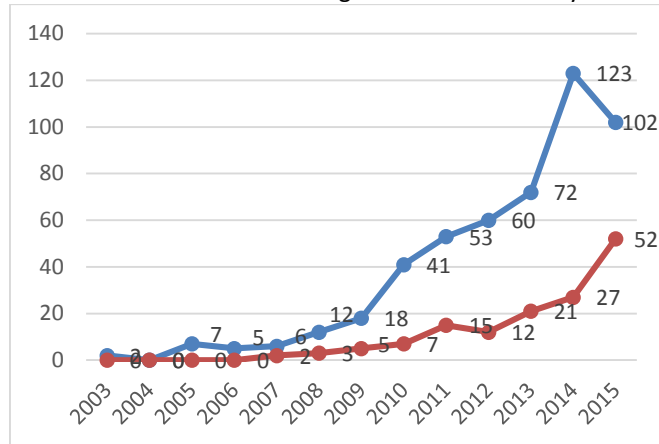
2.7.1 Antecedentes

Históricamente el proceso de enseñanza ha buscado incrementar las fortalezas de los profesores y principalmente maximizar el aprendizaje de los estudiantes. Hoy en día la educación se apoya de manera muy fuerte en la tecnología, especialmente en sistemas de enseñanza apoyados en WEB. La cantidad de datos que se almacenan en estos sistemas y en los sistemas tradicionales es cada vez mayor, así mismo la complejidad de analizar estos datos también se ha incrementado.

La minería de datos es un área que cada vez se aplica con mayor fuerza en más áreas como por ejemplo gobierno, empresa, educación, investigación espacial, deportes, análisis de textos, internet y muchas otras [16] y en los últimos años en el análisis de sistemas de e-learning con el fin de obtener mejores resultados en el proceso de aprendizaje de los estudiantes. En este proceso ha nacido una nueva área dedicada a realizar minería de datos en educación, es decir, implementación de técnicas de minería de datos a información registrada en entornos educativos. En relación a esta área que se ha denominado EDM (Educational Data Mining) a continuación se presentan los principales hitos.

Los primeros papers que se enfocaron en EDM se publicaron en 1995, año en el que se publicaron 3 papers, a partir de este año el número de papers publicados en conferencias y revistas ha crecido de manera importante. Para ello se han consultado algunas de las principales bases de datos bibliográficas de artículos de investigación o científicos.

Figura 5. Publicaciones sobre EDM registradas en SCOPUS y ScienceDirect.



Fuente: Elaboración propia tomado de <http://www.scopus.com> y <http://www.sciencedirect.com> propia el 10 de enero de 2016

- Publicaciones SCOPUS:
 - 1993: 3
 - 2015: 97
- Publicaciones ScienceDirect
 - 2007: 2
 - 2014: 27
 - 2015: 52

2.7.2 Conferencias internacionales

Debido a la importancia que ha tomado la minería de datos para la educación, la International Educational Data Mining Society (IEDMS) desde el año 2008 se está adelantando un encuentro internacional que se dedica exclusivamente a tratar temas relacionados con la Minería de Datos para la Educación [4], estos encuentros se han adelantado en diferentes países y los principales temas tratados han sido:

1. EDM'08: 1st International Conference on Educational Data Mining
 Montreal, Québec, Canadá
 Junio 20 al 21 2008
 Número de Papers Aceptados: 17
 Número de Posters Aceptados: 5
2. EDM'09: 2nd International Conference on Educational Data Mining
 Córdoba, España
 Julio 1 al 3 2009
 Número de Papers Aceptados: 20
 Número de Posters Aceptados: 14
3. EDM2010: 3th International Conference on Educational Data Mining
 Pittsburgh, PA, USA

- Junio 11 al 13 2010
 Número de Papers Aceptados: 24
 Número de Posters Aceptados: 17
4. EDM2011: 4th International Conference on Educational Data Mining
 Eindhoven, Holanda
 Julio 6 al 8 2011
 Número de Papers Aceptados: Completos 20, Cortos 17
 Número de Posters Aceptados: 30
 5. EDM2012: 5th International Conference on Educational Data Mining
 Chania, Grecia
 Junio 19 al 21 2012
 Número de Papers Aceptados: Completos 17, Cortos 16
 Número de Posters Aceptados: 17
 6. EDM2013: 6th International Conference on Educational Data Mining
 Memphis, Tennessee, USA
 Julio 6 al 9 2013
 Número de Papers Aceptados: Completos 21, Cortos 25
 Número de Posters Aceptados: 29
 7. EDM2014: 7th International Conference on Educational Data Mining
 Londres, Reino Unido
 Julio 4 al 7 2014
 Número de Papers Aceptados: Completos 24, Cortos 33
 Número de Posters Aceptados: 63
 8. EDM2015: 8th International Conference on Educational Data Mining
 Madrid, España
 Junio 26 al 29 2015
 Número de Papers Aceptados: Completos 41, Cortos 46
 Número de Posters Aceptados: 39
 9. EDM2016: 9th International Conference on Educational Data Mining
 Raleigh, NC, USA
 Junio 29 al 02 de julio de 2016
 Número de Papers Aceptados: Completos 30, Cortos 55
 Número de Posters Aceptados: 51

En la tabla No. 2 se presenta el detalle de los trabajos aceptados en la conferencia internacional de Minería de Datos para la Educación desde el año 2008 hasta el 2016. En la tabla se presenta el detalle de los papers y posters aceptados.

Tabla 2. Trabajos aceptados en las International Conference on Educational Data Mining

		Papers		Posters
		Costos	Largos	
International Conference on Educational Data Mining	1 – 2008	17		5
	2 – 2009	20		14
	3 -2010	24		17
	4 – 2011	20	17	30
	5 – 2012	17	16	17
	6 – 2013	21	25	29
	7 – 2014	24	33	63
	8 – 2015	41	46	39
	9 – 2016	30	55	51

Fuente: Elaboración propia totado de <http://www.educationaldatamining.org/IEDMS/events> el 13 de febrero de 2016.

2.7.3 Líneas de investigación futuras

Dentro de los principales asuntos que en la actualidad están despertando el interés de los investigadores de esta área se ha encontrado entre otras las siguientes¹:

1. Cerrando el ciclo entre investigación de datos en educación y resultados educativos.
2. Representación de conocimiento de dominio de datos
3. Detección y tratamiento de estados de ánimo y emocionales de los estudiantes
4. Minería de datos en entornos pedagógicos emergentes [23]
5. Minería de datos en aplicaciones Open Source [24]
6. Integración entre EDM y Análisis de aprendizaje
7. Aplicación de técnicas de minería de datos en dominios específicos [25]
8. Marcos genéricos y enfoques para EDM

De acuerdo a la lista de líneas de investigación futura se encuentran dos que le dan completa pertinencia al desarrollo de este trabajo, Minería de datos en aplicaciones Open Source y Aplicación de técnicas de minería de datos en dominios específicos.

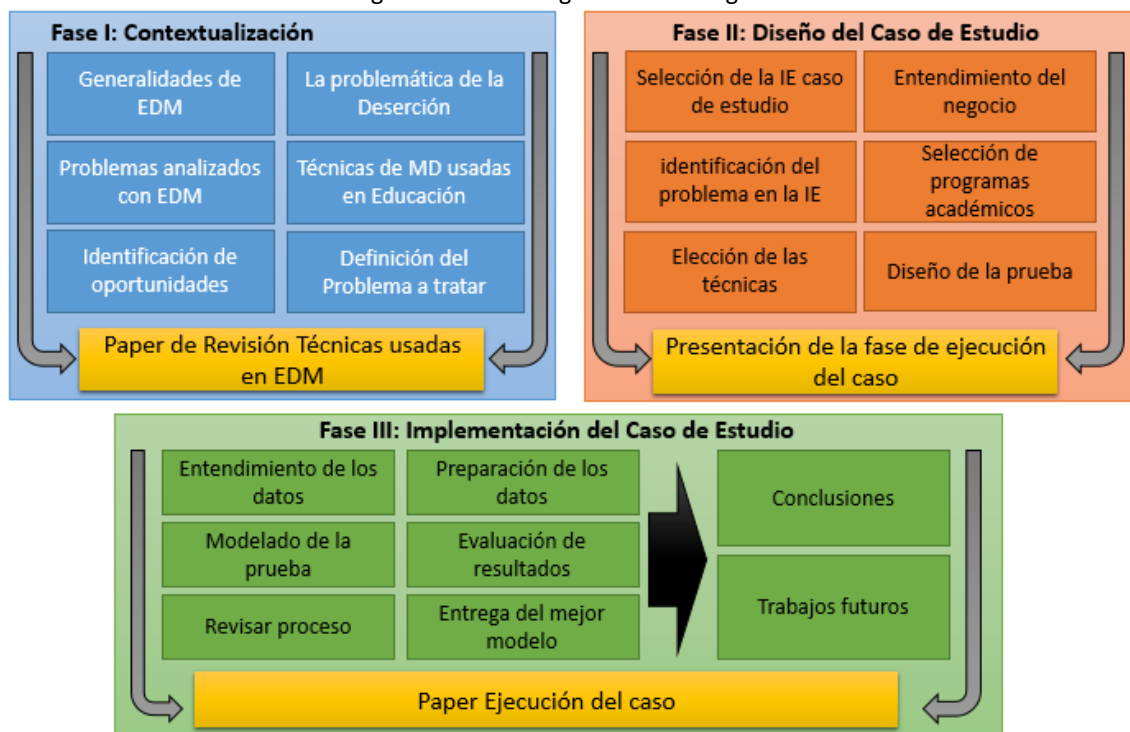
¹ <http://educationaldatamining.org/EDM2016/>

3. MARCO METODOLÓGICO

Mediante el desarrollo de este trabajo se pretende realizar una investigación que permita identificar las técnicas de minería de datos se han implementado en escenarios educativos y aplicar las técnicas que una vez evaluadas permitan identificar riesgos de deserción en estudiantes de programas de educación superior en una institución colombiana.

En desarrollo del presente trabajo está dividido en tres fases principales, las cuales son: 1. Documentación general del proyecto, 2. Diseño del caso de estudio y 3. Implementación del caso de estudio.

Figura 6. Metodología de la investigación



Fuente: Elaboración propia

3.1 Fase 1. Contextualización para el desarrollo del trabajo

La primera fase de la tesis consiste en realizar una investigación detallada de los diferentes temas que son relevantes en el desarrollo del presente trabajo. En esta fase se busca adelantar las siguientes tareas.

El entregable que se espera producir como resultado de las actividades adelantadas en la fase 1 es un paper de revisión de las técnicas de minería de datos utilizadas para el análisis de datos de entornos o dominios educativos.

3.2 Fase 2: Diseño del caso de estudio

La segunda fase se concentra en realizar el diseño del caso de estudio, para ellos se integrará de manera activa la Institución educativa que estará involucrada en el desarrollo del trabajo de tesis. En esta fase se busca adelantar las siguientes tareas.

El entregable que se espera producir como resultado de las actividades adelantadas en la fase 2 es la presentación del diseño del caso de estudio de los involucrados en la institución educativa.

3.3 Fase 3: Ejecución del caso de estudio

La tercera fase se concentra en la ejecución del caso de estudio.

El entregable que se espera producir como resultado de las actividades adelantadas en la fase 3 es el paper resultado de la ejecución del caso de estudio.

4. DESARROLLO E IMPLEMENTACIÓN

4.1 Documentación General del Proyecto

Como producto de la contextualización general para el desarrollo del trabajo, se elaboró un paper titulado “Data mining techniques applied in educational environments: Literature Review” el cual se pretende enviar a dos Journals buscando ser publicado, por esta razón, en el documento de tesis no se detallará cada uno de los puntos presentados en el paper, pero se presentan las conclusiones de la revisión adelantada.

4.1.1 Clasificación de técnicas de minería de datos utilizadas en entornos educativos

Son múltiples los casos de estudio en los que se aplican técnicas de minería de datos en entornos educativos, cada uno de ellos buscando una respuesta a una situación particular que aporte al objetivo buscado por la EDM. Para el desarrollo del presente trabajo se tuvieron en cuenta trabajos publicados sobre EDM entre el año 1993 y el año 2015, de los trabajos revisados se tuvieron en cuenta 130 trabajos en los que se aplicaron al menos una técnica de minería de datos con el fin de analizar una situación asociada al proceso de enseñanza-aprendizaje. La revisión de literatura adelantada se organizó en cuatro apartados así:

1. Revisión de trabajos en los cuales se han utilizado técnicas de minerías de datos en entornos educativos.
2. Trabajos clasificados por dominios educativos.
3. Trabajos clasificados por técnicas de minerías de datos
4. Listado de autores relevantes en EDM.

4.1.2 Trabajos clasificados por técnicas de minerías de datos

En los trabajos analizados se encontró que en total se utilizan 13 técnicas de minerías de datos, estas son:

- Análisis de Correlación
- Árboles de decisión
- Árboles de Regresión
- Cadenas de Markov
- Clasificación
- Clustering
- Differential Sequence Mining
- Patrones Secuenciales
- Redes Bayesianas
- Redes Neuronales
- Reglas de asociación

- Regresión Lineal

Los trabajos analizados en la revisión de literatura están asociados a las diferentes técnicas de minería de datos, en algunos casos el mismo paper utilizó diferentes técnicas, el paper es clasificado en tantas técnicas como haya utilizado para considerar todas las perspectivas utilizadas por los diferentes actores. En la tabla No. 3 se presentan todos los trabajos asociados a las técnicas enunciadas anteriormente.

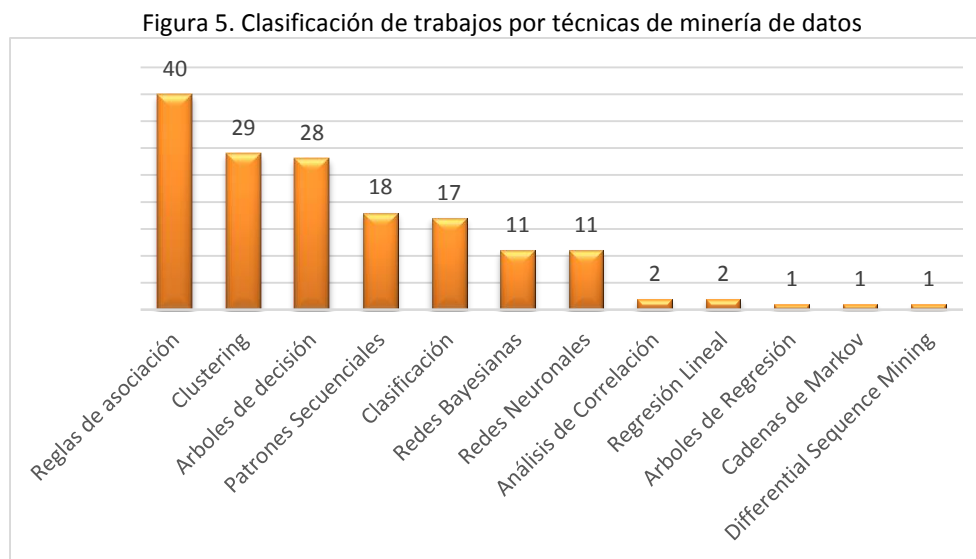
Tabla 3. Consolidado de trabajos revisados por técnicas de minería de datos

Técnica Implementada	No. de Publicaciones	Autores Relevantes	Años con publicaciones
Análisis de Correlación	2	Chen C., Chen M., Patarapichayatham, C. Kamata, A. Kanjanawasee, S.	2009, 2012
Arboles de decisión	27	Baradwaj B., Pal S., Sheard J., Ceddia J., Hurst J., Tuovinen J., Guo Q., Zhang M., Kovačić, Z. J	Entre el 2003 y 2015
Arboles de Regresión	1	Hu Y., Lo C., Shih S.	2014
Cadenas de Markov	1	Stamper J., Barnes T.	2009
Clasificación	15	Bhardwaj B., Pal S., Marquez-Vera C., Romero C., Ventura S., Chen C., Chen M., Bayer J., Bydzovská H.,	Entre el 2009 y 2015
Clustering	27	Wang F.H., Shao H.M., Baradwaj B., Pal S., Kumar V., Chadha A., Chen C., Chen M.	Entre el 2003 y 2015
Differential Sequence Mining	1	Martinez-Maldonado R., Yacef K. Kay J.	2013
Patrones Secuenciales	17	Karampiperis P., Sampson D., Romero C., Ventura S., Zafra A., De bra P., Kay J., Maisonneuve N., Yacef K., Zaiane O.R.	Entre el 2004 y 2015
Redes Bayesianas	10	Pardos Z. A., Heffernan N. T., Anderson B., Heffernan C. L., Schools W., Pandey U., Pal, S.	Entre el 2006 y 2013
Redes Neuronales	10	Oladokun V.O., Adebajo A.T., Charles-owaba O.E. , Härmäläinen W, Vinni M	Entre el 1993 y 2015
Reglas de asociación	35	Ha S., Bae S., Park S., Zaiane O., Romero C., Ventura S., De Bra P., Merceron A., Yacef K., Baradwaj B., Pal S.	Entre el 1995 y 2014
Regresión Lineal	2	Härmäläinen W, Vinni M	2010

Fuente: Elaboración Propia

De acuerdo a los trabajos analizados en la revisión de literatura se pudo identificar que las técnicas más utilizadas para analizar situaciones relacionadas a entornos educativos son Reglas de asociación con un 24,8%, seguido de Clustering con 18% y Arboles de decisión con 17%, de otra parte, las técnicas menos utilizadas son Differential Sequence Mining, cadenas de Markov y Arboles de Regresión cada una con un solo trabajo lo cual corresponde a un 0,62%.

En la siguiente figura se presenta la gráfica del número de trabajos analizados por técnicas de minería de datos



Fuente: Elaboración Propia

El análisis de técnicas utilizadas es importante para el desarrollo de las fases 2 y 3 del trabajo ya que la revisión de la literatura son los soportes para el desarrollo de las siguientes actividades. Es importante precisar que para las fases 2 y 3 serán tenidas en cuenta especialmente las técnicas que se han utilizado en el dominio de la deserción, independientemente de que sean las más utilizadas de manera general en EDM.

4.1.3 Clasificación de trabajos por dominios educativos.

Cada uno de los trabajos revisados fue clasificado en uno de los 7 dominios de tipo educativo que se tomaron para el desarrollo de la tesis. Es importante precisar que estos dominios surgen a partir de la revisión elaborada, cada trabajo revisado se asoció a un dominio de la educación. Los dominios de clasificación se seleccionaron a partir de los temas que tratan los trabajos, estos dominios son:

- *Análisis de Deserción o Retención:* trabajos en los cuales se analizan factores relacionados a deserción estudiantil o estudios que buscan favorecer la retención estudiantil.

- *Análisis de OVA o EVA*: aquellos trabajos en los cuales se han analizado objetos virtuales de aprendizaje OVA o entornos virtuales de Aprendizaje EVA
- *Análisis de Rendimiento o Evaluación de estudiantes*: Trabajos en los cuales el uso de técnicas de minería de datos se orientó a analizar el rendimiento de los estudiantes o las evaluaciones realizados a estos durante los cursos presenciales o virtuales.
- *Generación de Recomendaciones*: dominio en el cual se asociaron los trabajos que generaron recomendaciones a ser usadas durante el proceso educativo.
- *Identificación de patrones de aprendizaje*: análisis de datos de entornos educativos que permitieron identificar patrones de aprendizaje a ser usados por las instituciones educativas o instructores.
- *Identificación de patrones en estudiantes*: análisis de datos de entornos educativos que permitieron identificar patrones en los estudiantes.
- *Predicción relacionada a estudiantes*: en este dominio se encuentran los trabajos que permitieron realizar predicciones relacionadas a los estudiantes, predicciones en notas finales, desempeño, comportamiento en cursos específico, etc.

En la siguiente tabla se presentan los trabajos relacionados a cada uno de los dominios presentados anteriormente.

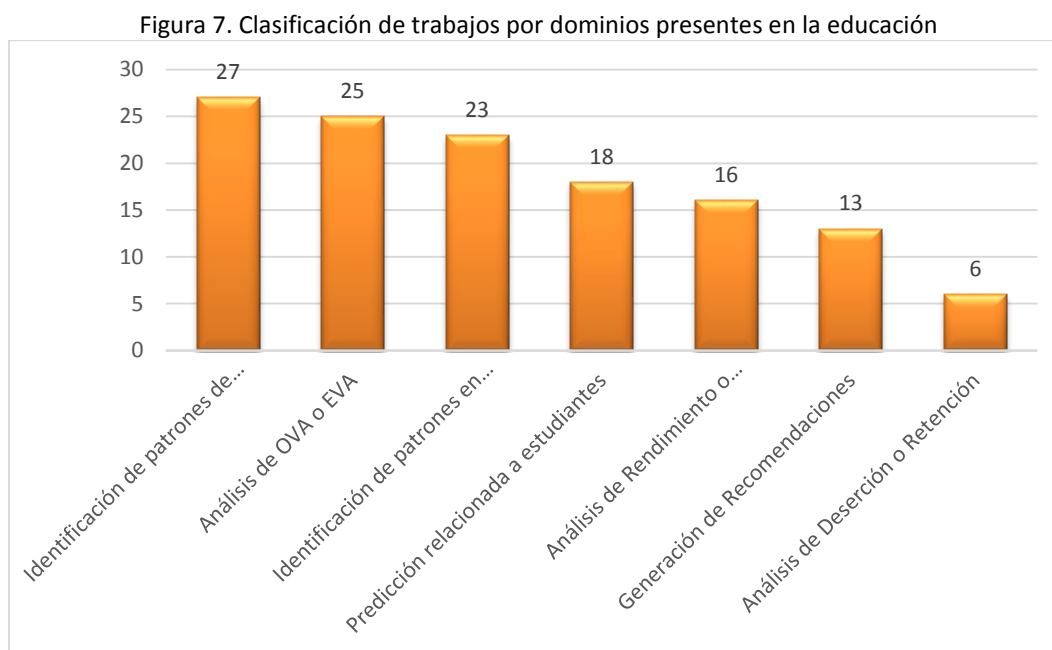
Tabla 4. Clasificación de trabajos por dominios presentes en la educación.

Dominio	No. de Publicaciones	Autores Relevantes	Años con publicaciones
Análisis de Deserción o Retención	6	Marquez-Vera C., Romero C. , Ventura S. , Bayer J., Kovačić, Z. J	Entre el 2010 y 2015
Análisis de OVA o EVA	26	Romero C. , Ventura S. , De Bra P., Karampiperis P., Sampson D.	Entre el 2001 y 2015
Análisis de Rendimiento o Evaluación de estudiantes	17	Bhardwaj B., Pal S., Romero C. , Espejo, P., Zafra A., Romero R., Ventura S.	Entre el 2007 y 2015
Generación de Recomendaciones	14	Wang F.H., Shao H.M., Vialardi C., Brav J. Shafti L., Ortigosa A.	Entre el 2004 y 2014
Identificación de patrones de aprendizaje	25	Zaïane O., Garcia E., Romero C. , Ventura S. , Castro C., Ha S., Bae S., Park S.	Entre el 2000 y 2015
Identificación de patrones en estudiantes	23	Baradwaj B., Pal S., Mor E., Minguillón J.	Entre el 1995 y 2015
Predicción relacionada a estudiantes	20	Oladokun V.O., Adebajo A.T., Charles-owaba O.E,	Entre el 1993 y 2015

Fuente: Elaboración Propia

De acuerdo a la clasificación realizada, pudimos identificar que la mayor parte de los trabajos de EDM publicados entre el año 1993 y 2015 se orientaron a la identificación de patrones de aprendizaje con un 21%; seguido de análisis de OVA y EVA con un 19% e identificación de patrones en estudiantes con un 17,8%. Los dominios con menor número de trabajos publicados son, análisis de rendimiento de estudiantes con un 12%, generación de recomendaciones con un 10% y por último, el dominio que tiene menor porcentaje de trabajos realizados es el de Análisis de Deserción o Retención que tan solo se analizaron 6 trabajos los que corresponden al 4,6%.

En la siguiente figura se presenta la gráfica del número de trabajos de EDM por trabajos por dominios.



Fuente: Elaboración propia

En el resultado obtenido de los trabajos asociados a los dominios de datos se puede evidenciar que el dominio con el menor número de trabajos es el de Análisis de deserción o retención, lo cual llama especialmente la atención porque el fenómeno de la deserción es una de las situaciones más analizadas en los entornos educativos, sin embargo, de acuerdo a la revisión realizada, no han sido muchos los análisis al fenómeno de la deserción a partir de la implementación de técnicas de minería de datos.

4.1.4 Técnicas de minería de datos utilizadas por dominios educativos.

Como parte del análisis realizado en la revisión de literatura, se identificaron las técnicas utilizadas en cada uno de los dominios EDM identificados. En la siguiente tabla se pueden ver las técnicas de minería de datos utilizadas en los dominios educativos mencionados en el apartado 4.1.3.

Tabla 3. Técnicas de minería de datos utilizadas por dominios educativos

Técnica \ Dominio	Análisis de Deserción o Retención	Análisis de OVA o EVA	Análisis de Rendimiento o Evaluación de estudiantes	Generación de Recomendaciones Educativas	Identificación de patrones de aprendizaje	Identificación de patrones en estudiantes	Predicción relacionada a estudiantes
Análisis de Correlación		1				1	
Arboles de decisión	5	3	8	2	2	6	2
Arboles de Regresión			1				
Cadenas de Markov				1			
Clasificación	4	2	4		3	1	3
Clustering		7	3	5	3	9	2
DSM						1	
Patrones Secuenciales		4	1	1	7	3	2
Redes Bayesianas		2	1		1	1	6
Redes Neuronales	1	2	2		1		5
Reglas de asociación		8	1	7	14	9	1
Regresión Lineal					1		1

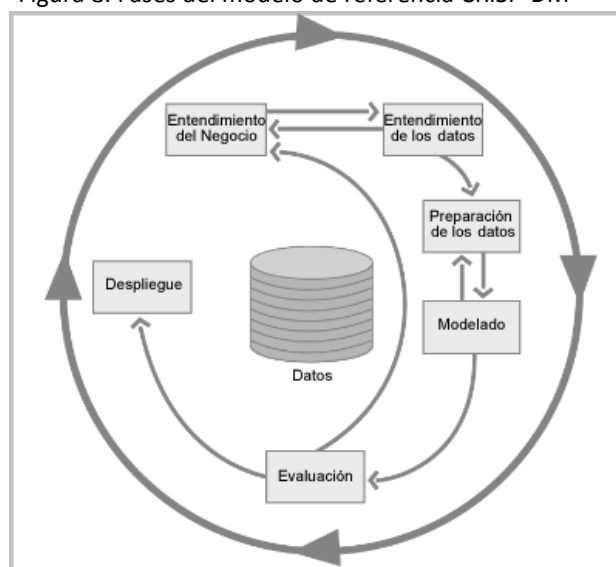
Fuente: Elaboración propia

Esta tabla busca ofrecer al lector un panorama general de las técnicas usadas en los dominios EDM identificados previamente. Si bien en el desarrollo de las fases 2 y 3 solo se tendrán en cuenta las técnicas aplicadas en análisis de deserción esta información es útil para aquello que deseen aplicar técnicas en entornos educativos.

4.2 Diseño del Caso de estudio

La segunda fase del trabajo se estructuró en actividades que permitieron hacer el diseño del caso de estudio, para el desarrollo de las fases dos y tres se utilizó la metodología CRISP-DM la cual define las fases que se deben implementar en cualquier proyecto de minería de datos, siendo estas entendimiento de negocios, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue [26]. En la siguiente figura se muestran las fases de CRISP-DM.

Figura 8. Fases del modelo de referencia CRISP-DM



Fuente: Crisp-Dm 1.0 [26]

4.2.1 Institución educativa

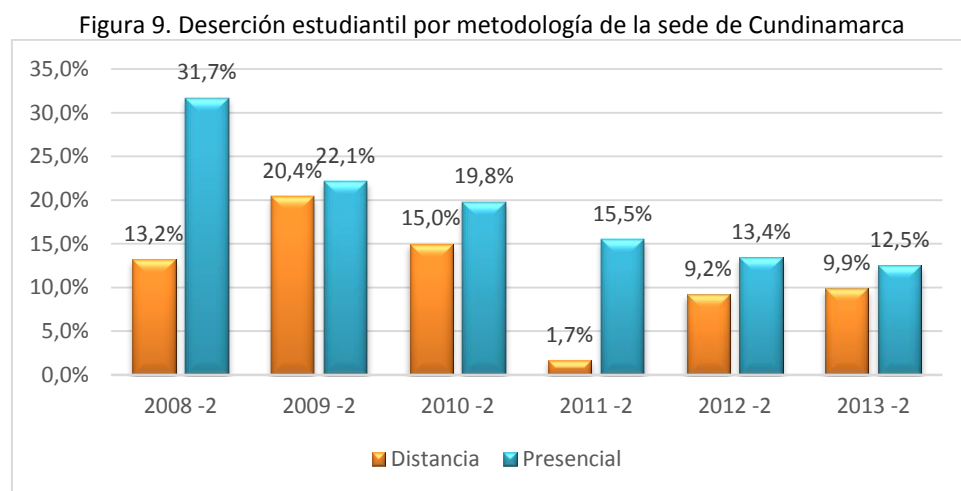
Para el desarrollo del trabajo se contó con el apoyo de la Corporación Universitaria Minuto de Dios – UNIMINUTO, particular mente de la Sede Cundinamarca, sede que facilitó los datos y el apoyo para el desarrollo de caso de estudio. Esta institución se constituyó en el año 1990 y empezó a operar en 1992, inició sus trabajos con 226 estudiantes y en la actualidad cuenta con más de 79.000 estudiantes distribuidos en 6 sedes, las cuales cubren 14 departamentos del país [27] .

La sede Cundinamarca comenzó su operación en el año 2000 con dos centros regionales, Girardot (2000) y Soacha (2002). En abril de 2014 la sede de Cundinamarca era la que tenía el mayor número de estudiantes con 15.447 lo cual correspondía al 20% de toda la población estudiantil [27].

En la sede Cundinamarca, el nivel de formación con mayor número de estudiantes es Profesional Universitario con 78%, seguido de nivel Tecnólogo con 11% y Articulación con 10% [27], es por ello que el caso de estudio estará enfocado a estudiantes de programas de pregrado, tecnologías y licenciaturas [27]. Así mismo, el 68% de la población estudiantil son mujeres, el 67% corresponde a estudiantes tiene una edad entre 16 y 26 años, por otra parte, el 69% de los estudiantes son estrato 2, el 14% son estrato 3 y el 11% estrato 1 [27].

La sede ofrece en total 34 programas académicos, 8 técnicos profesionales, 12 tecnológicos, 11 universitarios y 3 en posgrado, para estos programas se puede evidenciar que los niveles de deserción han disminuido en más de 19% por debajo de las cifras nacionales desde que se implementaron estrategias dirigidas a brindar oportunidades de acceso y permanencia en la educación superior. Estas estrategias se canalizan a través del área de bienestar estudiantil, y se

fundamentan en el acompañamiento a los estudiantes de primer año con estrategias de aprendizaje, actividades extracurriculares, promoción de hábitos saludables y actividades de pastoral, entre otras [27]. Sin embargo, estas estrategias se implementan para todos los estudiantes de primer año independientemente del programa o situación particular.



Fuente: UNIMINUTO Plan de Desarrollo 2013 – 2019 [27]

4.2.2 Recolección de los datos

Tal como se ha indicado previamente, para el desarrollo del caso de estudio se tomó como eje fundamental el Modelo de Atención Integral al Estudiante - MAIE. Sin embargo al analizar el modelo se puede encontrar que este está estructurado en cuatro componentes que abordan la historia de los estudiantes desde la preparación para el ingreso hasta la articulación con el sector educativo, lo cual es mucho más amplio que el objetivo general del desarrollo de este proyecto, razón por la cual a partir de este modelo se consideraron aquellos datos de los 4 componentes que podrían ser analizados para identificar patrones de deserción.

Una vez seleccionada la institución académica en la que se desarrollará el caso de estudio, se realiza la búsqueda de los datos requeridos para el desarrollo del proyecto; estos datos están enmarcados en las tres dimensiones en las cuales es estudiado el fenómeno de la deserción. En la siguiente tabla se presentan los dos datos que se consideran relevantes para el desarrollo del proyecto.

Tabla 5. Datos relevantes para el desarrollo del caso de estudio

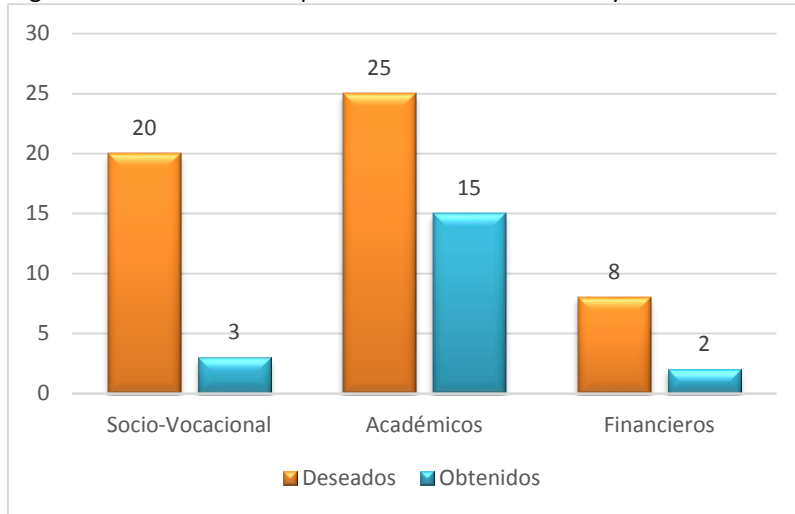
Dimensiones		
Socio-vocacional	Académica	Económico-financiera
Orientación para seleccionar el programa académico	Sede, Facultad, Nivel, Programa, Jornada.	Empleado
Tipo de estudiante	Fecha de Matricula	Estrato

Estado civil al inicio del programa	Fecha Inscripción de Materias	Beca
Estado Civil actual	Estudios de pregrado Previos	Apoyo económico de la familia
Nivel Educativo del Padre	Estudios de pregrado Previos culminados	Crédito Educativo
Nivel educativo de la madre	Puntaje de ICFES	Crédito Educativo ICETEX
Padre empleado	Icfes Saber 11 Lenguaje	Subsidio económico
Madre empleada	Icfes Saber 11 Matemática	Crédito Cooperativa
Número de hermanos	Icfes Saber 11 Ciencias Sociales	SISBEN
Orden de nacimiento	Icfes Saber 11 Filosofía	
Número de hijos al iniciar el programa académico	Icfes Saber 11 Biología	
Estimulo de la familia	Icfes Saber 11 Química	
Modo de transporte a la universidad	Icfes Saber 11 Física	
Tiempo de transporte a la universidad	Icfes Saber 11 Inglés	
Etnia	Icfes Saber 11 Colegio	
Tiempo dedicado el trabajo	Modalidad de bachillerato	
Lugar de nacimiento	Sanción académica registrada	
Lugar de residencia	Número de materias perdidas en el programa actual	
Número de años en ciudad de residencia	Créditos del programa académico	
	Créditos aprobados del programa académico.	
	Promedio académico	

Fuente: Elaboración propia

Es importante precisar que los datos listados en la tabla anterior son aquellos datos relevantes para el análisis de la deserción, sin embargo, contar con todos estos datos implica un levantamiento de información que requiere una preparación logística y tiempo significativo por parte de la institución académica. Debido a que muchos datos no los tenía la institución académica, se diseñó una encuesta para obtener la totalidad de los datos pero el levantamiento de esta información no se ha finalizado. A continuación se presenta una gráfica que muestra el número de datos deseados versus los datos obtenidos para el desarrollo del caso de estudio.

Figura 10. Datos deseados para el análisis de deserción y datos obtenidos



Fuente: Elaboración propia

Por esta razón el caso de estudio se diseñó con los datos a los que se pudo acceder, los cuales fueron entregados por la institución educativa en 26 archivos distribuidos de la siguiente manera:

Tabla 6. Tabla archivos con datos utilizados en el caso de estudio

Archivo	Descripción	No. de Estudiantes Relacionados
créditos aprobados botón 14 280320161123.xls	Contiene el detalle de los créditos inscritos y créditos aprobados por estudiantes matriculados en los diferentes programas.	50.455
Listado de Deserción	5 archivos con la información de los estudiantes que desertaron en 5 periodos académicos	9.877
Archivos ICFES (19 Archivos)	19 archivos con información de resultados ICFES de los estudiantes inscritos por periodos académicos desde el 2002 hasta el primer periodo del 2016.	321.334
DATOS FALTANTES 13-04-2016.xls	Hoja de cálculo con información general de los estudiantes inscritos en programas de la Sede Cundinamarca.	61.685

Fuente: Elaboración propia

4.2.3 Selección de programas académicos

Debido al enfoque de los programas ofertados en la sede Cundinamarca y a los datos disponibles para el análisis se decidió trabajar con programas de larga duración, es decir, programas de pregrado dejando por fuera del análisis cursos libres, diplomados y tecnologías. Tampoco se analizaron estudiantes matriculados en programas de nivel técnico laboral y especialización toda vez que no se tenía un número de estudiantes significativos para analizar.

El análisis de los datos y la aplicación de algoritmos de minería de datos se aplicaron en dos niveles, un nivel superior en el que se analizaron todos los estudiantes de los que se tenían datos para el análisis, en total 18.587 estudiantes de los cuales 4.502 desertaron y otro nivel en el que se analizaron solo los estudiantes de la facultad de ingeniería, en total 2.726 de los cuales 820 desertaron.

Tabla 7. Número de estudiantes por programas seleccionados para el análisis

Programas	Estudiantes desertores	Estudiantes no desertores	Estudiantes analizados
Todos los programas	4.502 (24.22%)	14.085 (75,78%)	18.587
Programas Ingeniería	820 (30.08%)	1.906 (69,92%)	2.726

Fuente: Elaboración Propia

Los valores que se toman para la característica “Desertó” son asignados a partir de la definición y clasificación manejada por la entidad la cual es: *“Desertor es todo aquel estudiante que no presenta matricula durante dos periodos consecutivos o más al momento de generar el consolidado”*.

4.3 Ejecución Caso de Estudio

La tercera fase del trabajo consistió en la ejecución del caso de estudio. En esta fase se buscó adelantar las tareas que permitieron realizar el entendimiento de los datos, la preparación de los mismos, realizar el modelado y la evaluación y el despliegue.

4.3.1 Preparación de los datos

El proceso de preparación de datos fue uno de los que más tiempo tomo debido al desconocimiento de la estructura de los datos y las diversas fuentes de datos que tiene la entidad. Por supuesto en esta fase fue fundamental el apoyo de la entidad que en repetidas ocasiones aclaró las dudas que se tenían respecto a los datos que se recibieron. Una vez aclaradas las dudas de los datos se realizó la integración de datos generar un *datawarehouse* el cual tenía un total de 53.681 estudiantes. En este almacén de datos se encontraron problemas como datos vacíos, valores incorrectos, y estudiantes inscritos en programas repetidos. Una vez realizada la limpieza de datos, quedó un total de 18.634 estudiantes, de los cuales 4.528 desertaron.

Para todos los estudiantes se estructuró un *datawarehouse* con las siguientes características

Tabla 8. Datos utilizados para el caso de estudio

Dato	Descripción
Tipo Estudiante	Nominal: B, C, D, E, G, H, I, J, L, N, P, R, S, T, V, W, X, Y, Z, 0
Sede	Nominal: GUA, SOA, MAD, GIR, ZIP, APU, VEG, VPI, CHC, UBT, PAN, RLP, GCH, MED, SRS
Facultad	Nominal: IE, CE, IN, CS, ED, CH, UC, PC, IF, DU

Nivel	Nominal UG, TC, CL, LI, TP, ES, DP, TL
Programa académico	Nominal: ADFU, ADM, AEMD, ASOD, C, CCPI, CCPU, CFIF, CIIF, CIIN, CLGQ, CLIC, CLIP, CLIU, CLUC, CLVF, CONP, COPD, CSGP, CSOC, CSOD, CSRN, DEGO, DGFN, DPLI, EGEU, EGPR, IAGR, ICIV, LBA, LBH, LIN, LPID, LPIN, PSIC, PSID, TAIN, TCG, TCGE, TCOS, TGME, TINF, TLEC, TLOG, TPCE, TPDG, TPFI, TPMM, TRED, TRSO
Jornada	Nominal: D, M, N, T
Periodo admisión programa	Ordinal
Fecha Admisión	Ordinal
Periodo inicio materias	Ordinal
Fecha inicio materias	Ordinal
Admisión programa Vs Inicio Materias	Nominal: Si, No
Periodos diferencia entre admisión programa e inicio materias	Escala
Tipo de Documento	Nominal: CC, TI, CE
Estrato	Nominal: 1, 2, 3, 4, 5
SISBEN	Nominal: Si, No
Genero	Nominal: M, F
Edad admisión programa	Ordinal
Edad inscripción materias	Ordinal
Estado civil	Nominal: L, S, C, P, V, D, R
Etnia	Nominal: N, 69, A, 19, A3, 46, 20, 43, 49, 56, 61, 66, 73, 77, 80, 88, 92, A2, A1, A4
Departamento Nacimiento	Ordinal
Ciudad Nacimiento	Ordinal
Departamento actual	Ordinal
Ciudad actual	Ordinal
Departamento de nacimiento Vs Departamento actual	Nominal: Si, No
Municipio nacimiento Vs Lugar actual	Nominal: Si, No
Créditos inscritos	Ordinal
Créditos aprobados	Ordinal
Porcentaje Créditos Aprobados	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Promedio acumulado	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Promedio Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Matemáticas Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Biología Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Filosofía Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Física Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Lenguaje Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Inglés Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Química Icfes	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Ciencias Sociales ICFES	Nominal: Bajo, MedioBajo, MedioAlto, Alto
Desertor	Nominal (Si, No)

Fuente: Elaboración propia

Las características marcadas con “*” se crearon a partir de las características iniciales. Así mismo, en la fase de desratización, todos los resultados Icfes se cambiaron de los valores numéricos (de 0 a 100) a un formato nominal con el fin de facilitar el análisis de los datos, este procedimiento también se realizó para la característica Porcentaje créditos aprobados y Promedio Acumulado. Los valores utilizados en la etapa de desratización fueron:

Icfes:

- Menor o igual a 30: Bajo
- Mayor a 30 y menor o igual a 45: Medio Bajo
- Mayor a 45 y menor o igual a 70: Medio Alto
- Mayor a 70: Alto

Promedio Acumulado

- Menor o igual a 2: Bajo
- Mayor a 32 y menor o igual a 3: Medio Bajo
- Mayor a 3 y menor o igual a 4: Medio Alto
- Mayor a 4: Alto

Por otra parte, se validó la calidad de los datos eliminando aquello en los que se tenían valores que superaban los límites admitidos o clasificados de manera incorrecta. Con los datos restantes se hizo un análisis de valores obteniendo los siguientes resultados.

Tabla 9. Lista de características con valores nulos

Dato	No. de valores vacíos o Nulos	% de valores vacíos	Valores Distintos
Tipo Estudiante	0	0%	13
Sede	0	0%	15
Facultad	0	0%	6
Nivel	0	0%	4
Programa académico	0	0%	28
Jornada	0	0%	4
Periodo admisión programa	0	0%	45
* Fecha Admisión	0	0%	25
Periodo inicio materias	0	0%	8
* Fecha inicio materias	0	0%	5
* Admisión programa Vs Inicio Materias	0	0%	2
* Periodos diferencia entre admisión programa e inicio materias	0	0%	61
Tipo de Documento	0	0%	2
Estrato	27	0%	6
SISBEN	0	0%	2
Genero	0	0%	2
Fecha Nacimiento	0	0%	5905

* Edad admisión programa	0	0%	317
* Edad inscripción materias	0	0%	322
Estado civil	7	0%	7
Etnia	18	0%	21
Departamento Nacimiento	0	0%	32
Ciudad Nacimiento	0	0%	619
Departamento actual	0	0%	19
Ciudad actual	0	0%	272
* Departamento de nacimiento Vs Departamento actual	0	0%	2
* Municipio nacimiento Vs Lugar actual	0	0%	2
Créditos inscritos	408	2%	254
Créditos aprobados	408	2%	195
* Porcentaje Créditos Aprobados	415	2%	4
Promedio acumulado	12	0%	4
Promedio Icfes	0	0%	3
Matemáticas Icfes	0	0%	4
Biología Icfes	755	4%	4
Filosofía Icfes	894	5%	4
Física Icfes	778	4%	4
Lenguaje Icfes	39	0%	4
Inglés Icfes	3331	18%	4
Química Icfes	770	4%	4
Ciencias Sociales ICFES	4765	26%	4
Desertor	0	0%	2

Fuente: Elaboración propia

Para el caso en el que los valores nulos o vacíos de las variables eran bajos (menos del 20%) se procedió a aplicar la técnica imputación por moda, esta técnica sugiere reemplazar los datos faltantes faltante de acuerdo a la moda de un grupo apropiadamente definido de valores disponibles [28]. Para la variable que tiene más del 20% de valores perdido se decidió eliminar la variable pues aplicar alguna técnica de imputación puede incrementar la probabilidad de errores. En consecuencia después de aplicar la técnica de imputación de moda y la eliminación de variable los valores vacíos se deja en cero.

Es importante precisar que la generación de nuevos atributos se realizó una vez integrador todos los datos en el datawarehouse. En total se crearon seis nuevos atributos considerados relevantes para analizar el fenómeno de la deserción, estos fueron:

- Inicia materias cuando se inscribe al programa
- Número de periodos entre la admisión al programa y la inscripción de materias
- Edad admisión
- Edad inscripción de materias
- Departamento de nacimiento Vs departamento actual

- Municipio de nacimiento Vs municipio actual

En el Anexo 01 se presenta el detalle de los datos con los cuales se adelantó el caso de estudio, en dicho documento se presenta el tipo de dato, el número de datos perdidos, y el detalle de cada uno según el tipo.

Debido al número de atributos que se tiene (33) y con el fin de identificar aquellos que son más relevantes en el análisis de la deserción de los dos conjuntos de datos disponibles se realizó una selección de atributos, los métodos de análisis de atributos se presenta a continuación

Tabla 10. Análisis de atributos relevantes

Evaluador	Sede Cundinamarca	Facultad Ingeniería
CfsSubsetEval [29]	Créditos inscritos, Porcentaje créditos aprobados, Promedio programa	Tipo de Estudiante, Créditos aprobados, Porcentaje créditos aprobados, Promedio
GainRatioAttributeEval [29]	Porcentaje créditos aprobados, Promedio, Inicia materias Vs. Inscribe al programa, Créditos Inscritos, Créditos aprobados	Porcentaje créditos Aprobados, Promedio, Inicia Materias Vs. Inscribe al Programa, Créditos aprobados, tipo de estudiante
InfoGainAttributeEval [29]	Créditos aprobados, Promedio, Porcentaje créditos aprobados, créditos inscritos, Periodos de ingreso después de matrícula	Créditos aprobados, Porcentaje créditos aprobados, Promedio, Créditos inscritos, Tipo de estudiante,
OneRAttributeEval [29]	Promedio, Porcentaje Créditos Aprobados, Créditos aprobados, Créditos Inscritos, Estrato	Promedio, Porcentaje Créditos Aprobados, Créditos aprobados, Créditos Inscritos, Periodos de ingreso Vs. Matricula en el programa
PrincipalComponents [29]	Facultad, Créditos aprobados, Créditos Inscritos, Tipo de Estudiante, Inicia Materias cuando se inscribe al programa	Promedio Icfes, Créditos Aprobados, Química Icfes, Tipo de estudiante, Créditos Inscritos
ReliefAttributeEval [29]	Tipo de estudiante, Sede, Programa, Porcentaje créditos aprobados, Promedio	Sede, Programa, Jornada, Porcentaje Créditos Aprobados, Promedio
SymmetricalUncertAttributeEval [29]	Porcentaje Créditos Aprobados, Tipo de estudiante, Inicia Materias cuando se inscribe al programa, Porcentaje créditos aprobados, Créditos aprobados, Promedio	Porcentaje créditos aprobados, Créditos aprobados, Promedio, Inicia Materias cuando se inscribe al programa, Tipo estudiante

Fuente: Elaboración propia

El resultado del análisis de atributos relevantes permitió identificar aquellos significativos para el análisis de la deserción. El mismo análisis se realizó al datawarehouse con los valores vacíos

obteniendo resultados muy similares. En la siguiente tabla se presenta el resultado de este análisis.

Tabla 11. Atributos relevantes en los conjuntos de datos

Evaluador	Sede Cundinamarca	Facultad Ingeniería
Créditos Inscritos	6	3
Porcentaje de Créditos Aprobados	6	6
Promedio	6	6
Créditos Aprobados	5	6
Inicia Materias Cuando Se Inscribe al Programa	4	4
Tipo de Estudiante	2	5

Fuente: Elaboración propia

4.3.2 Modelado

En la fase de modelado se realizaron las actividades que permitieron aplicar las técnicas de minería de datos seleccionadas a los datos preprocesados.

La primera actividad realizada en la fase de modelado es la selección de técnicas de minería de datos a aplicar. En este caso, las técnicas que se aplicarán se han tomado de la fase 1 del presente trabajo, en donde se realizó un análisis de las técnicas de minería de datos aplicadas al dominio de la deserción el cual es el eje central del desarrollo del presente trabajo. Así mismo se validó la disponibilidad de implementar dichas técnicas en la herramienta seleccionada para el desarrollo del proyecto como resultado de dicho análisis de concluyo que las técnicas de minería de datos a aplicar son:

- KNN
- Árbol de Decisión

A partir de los datos disponibles se realizó una distribución de estos con el fin de generar los grupos de entrenamiento, validación y pruebas a continuación se presenta la distribución de datos en estos grupos:

Tabla 12. Distribución de datos para entrenamiento, validación y pruebas

	Desertó	Entrenamiento (40%)	Validación (40%)	Prueba (40%)
Sede Cundinamarca	Si	1800	1800	900
	No	1809	9459	2817
Facultad Ingeniería	Si	328	328	164
	No	328	1197	381

Fuente: Elaboración Propia

De acuerdo a lo anterior, los datos se han distribuido en seis archivos, tres para el análisis de los estudiantes de la sede Cundinamarca y tres para el análisis de los estudiantes de la facultad de

ingeniería. Una vez definidos las técnicas de minería de datos, preprocesado, seleccionadas los atributos más relevantes y preparados los conjuntos de datos se ejecutaron las técnicas de minería de datos, esta actividad implicó múltiples ciclos y ajustes en los datos con el fin de conseguir un modelo que ofreciere altos niveles de confianza.

Los resultados obtenidos aplicando las técnicas seleccionadas para los 6 atributos seleccionados se presentan a continuación, es de aclarar que los resultados obtenidos se aplicaron a los datos de entrenamiento respectivamente y posteriormente se validaron estos resultados con los datos de pruebas.

Resultado árbol de decisión sede Cundinamarca

Para el caso de estudio se trabajó con el algoritmo J48, el cual es una implementación del algoritmo C4.5, uno de los más utilizados en minería de datos [30]. Después de realizar varios ajustes el resultado es el siguiente:

Figura 11. Árbol J48 estudiantes sede Cundinamarca



Fuente: Visualizador WEKA

Algoritmo J48

No. total de instancias: 3607

No. de atributos: 7 (Tipo Est., Inicia Materias Cuando Se Inscribe al Programa, Créditos Inscritos, Créditos Aprobados, % Créditos Aprobados, Promedio, Desertó)

Correctamente Clasificados: 2820 (78,18%)

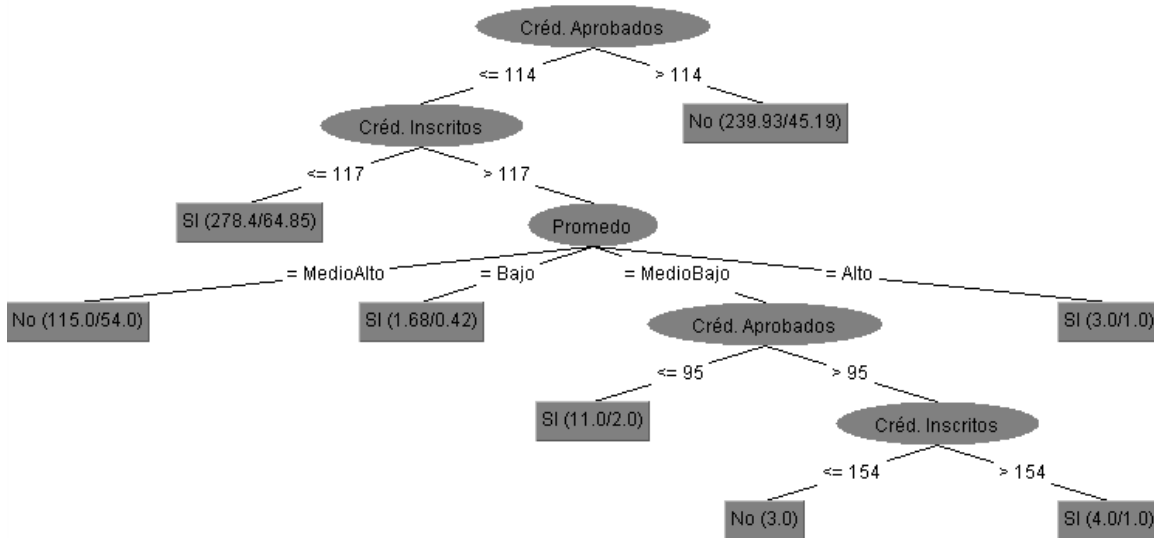
Incorrectamente Clasificado: 787 (21.81%)

Tiempo construcción del Modelo: 0,03 segundos

Resultado árbol de decisión Facultad Ingeniería

Uno de los análisis realizados fue el análisis de los estudiantes inscritos en programas largos de la facultad de ingeniería el árbol resultado de aplicar el algoritmo J48 a los 6 atributos relevantes se presenta a continuación:

Figura 12. Árbol J48 estudiantes facultad Ingeniería 6 atributos relevantes



Fuente: Visualizador WEKA

Algoritmo J48

No. total de instancias: 656

No. de atributos: 7 (Tipo Est., Inicia Materias Cuando Se Inscribe al Programa, Créditos Inscritos, Créditos Aprobados, % Créditos Aprobados, Promedio, Desertó)

Correctamente Clasificados: 471 (71,79%)

Incorrectamente Clasificado: 185(28.20%)

Tiempo construcción del Modelo: 0,07 segundos

Así mismo y con el fin de verificar la pertinencia de los atributos seleccionados como relevantes se corrió el modelo del árbol para los datos de la facultad de ingeniería, pero aplicado a los 29 atributos obtenidos. Lo cual nos dio como resultado lo siguiente:

Figura 13. Árbol J48 estudiantes facultad Ingeniería 31 atributos



Fuente: Visualizador WEKA

Algoritmo J48

No. total de instancias: 656

No. de atributos: 29 (Tipo Est., Sede, Nivel, Programa, Jornada, Inicia Materias Cuando Se Inscribe al Programa, Periodos de ingreso después de Matricula en la U, Estrato, SISBEN, Genero, Edad Admisión, Edad Inscripción Materias, Estado Civil, Etnia, Mismo Dpto Nacimiento Ubicación, Nacimiento Vs Ubicación - Municipio, Créd. Inscritos, Créd. Aprobados, % Créditos Aprobados, Promedio, Promedio Icfes, Matemáticas, Biología, Filosofía, Física, Lenguaje, Inglés, Química, Desertó)

Correctamente Clasificados: 471 (71,79%)

Incorrectamente Clasificado: 185(28.20%)

Tiempo construcción del Modelo: 0,07 segundos

Resultado vecinos más cercanos IBk sede Cundinamarca

Para el caso de estudio se trabajó con el algoritmo IBk de Weka el cual es la implementación de KNN [29]. Después de realizar varios ajustes a la ejecución de IBk para los 6 atributos relevantes sobre los datos de la sede Cundinamarca se obtuvo el siguiente resultado:

Algoritmo: IBK

No. total de instancias: 3607

No. de atributos: 7

Correctamente Clasificados: 3389 (93,95%)

Incorrectamente Clasificado: 218 (6.04%)

Tiempo construcción del Modelo: 1,37 segundos

Tabla 13. Precisión detallada IBk sede Cundinamarca

	Verdaderos Positivos	Falsos Positivos	Precisión
Si	0,875	0,244	0,783
No	0,756	0,125	0,857
Promedio Ponderado	0,815	0,185	0,820

Fuente: WEKA

Resultado vecinos más cercanos IBk facultad Ingeniería

Para el caso de estudio se trabajó con el algoritmo IBk de Weka el cual es la implementación de KNN [29]. Después de realizar varios ajustes a la ejecución de IBk para los 6 atributos relevantes sobre los datos de la facultad de ingeniería se obtuvo el siguiente resultado:

Algoritmo: IBK

No. total de instancias: 656

No. de atributos: 7

Correctamente Clasificados: 633 (96,49%)

Incorrectamente Clasificado: 23 (3.51%)

Tiempo construcción del Modelo: 0,19 segundos

Tabla 14. Precisión detallada IBk facultad Ingeniería

	Verdaderos Positivos	Falsos Positivos	Precisión
Si	0,994	0,064	0,939
No	0,936	0,006	0,994
Promedio Ponderado	0,965	0,035	0,967

Fuente: WEKA

Resultado Clasificador OneR sede Cundinamarca

Para el caso de estudio se trabajó con el algoritmo de clasificación disponible en WEKA OneR. Este clasificador permite seleccionar el atributo que mejor “explica” el atributo dependiente [30], obteniendo el siguiente resultado:

Cred. Aprobados:

< 43.5 -> SI

< 44.5 -> No

< 84.5 -> SI

< 85.5 -> No

< 98.5 -> SI

< 99.5 -> No

< 105.5 -> SI

< 106.5 -> No

< 108.5 -> SI

< 110.5 -> No

< 111.5 -> SI

< 112.5 -> No
 < 113.5 -> SI
 >= 113.5 -> No

Algoritmo: OneR
 No. total de instancias: 3607
 No. de atributos: 7
 Correctamente Clasificados: 2786 (77.23%)
 Incorrectamente Clasificado: 821 (22.76%)
 Tiempo construcción del Modelo: 0,04 segundos

Tabla 15. Precisión detallada OneR sede Cundinamarca

	Verdaderos Positivos	Falsos Positivos	Precisión
Si	0,724	0,179	0,803
No	0,821	0,276	0,747
Promedio Ponderado	0,772	0,227	0,775

Fuente: WEKA

Resultado Clasificador OneR facultad Ingeniería

La ejecución del algoritmo OneR sobre los datos de la facultad de ingeniería nos arrojó el siguiente resultado:

Créd. Inscritos:
 < 117.5 -> SI
 < 118.5 -> No
 < 123.5 -> SI
 < 125.5 -> No
 < 131.5 -> SI
 < 135.5 -> No
 < 138.5 -> SI
 >= 138.5 -> No
 ? -> SI

Algoritmo: OneR
 No. total de instancias: 656
 No. de atributos: 7
 Correctamente Clasificados: 499 (76,06%)
 Incorrectamente Clasificado: 157 (23.93%)
 Tiempo construcción del Modelo: 0,01 segundos

Tabla 16. Precisión detallada OneR facultad Ingeniería

	Verdaderos Positivos	Falsos Positivos	Precisión
Si	0,790	0,268	0,746
No	0,732	0,201	0,777

Promedio Ponderado	0,761	0,229	0,762
-------------------------------	-------	-------	-------

Fuente: WEKA

5. RESULTADOS

5.1 Resultados fase contextualización

Como resultado de la fase uno del proyecto se obtuvo una revisión detallada de técnicas de minería de datos aplicadas a entornos educativos desde 1993 hasta el año 2015, adicionalmente se elaboró una lista de dominios que permite clasificar los casos de estudio e investigaciones de EDM de acuerdo al objetivo perseguido en los trabajos revisados.

Por otra parte se elaboró una matriz que permite relacionar las técnicas de minería de datos asociadas a los diferentes dominios identificados en el proceso de revisión. Esta matriz les permite a los lectores conocer el tipo de técnicas utilizadas para estudiar las diferentes situaciones que se pueden analizar sobre datos almacenados en sistemas de educación académica. Dicha matriz es un buen punto de partida y resulta de gran utilidad para aquellas entidades que deseen analizar situaciones educativas particulares apoyadas en el uso de técnicas de minería de datos.

5.2 Resultados análisis de datos

En el anexo 2, se presenta el resultado del análisis de los datos, este análisis permite identificar por cada una atributo y sus posibles valores los estudiantes que han desertado o no. Dentro de este análisis se puede resaltar que:

- Los estudiantes de tipo reingreso y reintegro tienen mayor probabilidad de deserción con 0,55, y 0,50 respectivamente, por su parte, los estudiantes de tipo continuo son los que menor probabilidad tiene de desertar con 0,17.
- Aquellos estudiantes registrados con estrato 6 tiene una probabilidad de deserción mucho más alta que los estudiantes de estratos más bajos. En el caso de los estudiantes estrato 6 matriculados en programas en la sede Cundinamarca tiene una probabilidad de deserción de 0,66, mientras que los estudiantes de otros estratos tiene una probabilidad de deserción que oscila entre 0,23 y el 0,27.
- Respecto al atributo SISBEN no se identifican diferencias marcadas entre los que tiene SISBEN o no ya que en cualquier caso la probabilidad de deserción es de 0,24.
- En relación al atributo género se identificó que los hombres tiene una probabilidad de deserción de 0,28 la cual es mayor a la de las mujeres quienes registran una probabilidad de 0,22.
- Respecto al atributo estado civil se identificó que aquellos estudiantes con estado civil divorciado tiene mayor probabilidad de deserción con 0,28 mientras que los estudiantes con estado civil viudo tienen menos probabilidad de deserción con 0,14.
- Aquellos estudiantes que están matriculados en el mismo departamento en el que nacieron tiene una probabilidad de deserción de 0,23 un poco menor que aquellos que estudian en un departamento diferente al departamento de nacimiento quienes tienen una probabilidad de deserción de 0,27.

- Respecto al atributo promedio de programa se identificaron diferencias muy significativas respecto a los que tienen promedio altos y bajos mostrando una probabilidad de deserción muy alta para aquellos estudiantes con promedios bajos. Las probabilidades de deserción según el promedio de la carrera son las siguientes:
 - Promedio alto: probabilidad de deserción 0,08
 - Promedio medio alto: probabilidad de deserción 0,21
 - Promedio medio bajo: probabilidad de deserción 0,64
 - Promedio bajo: probabilidad de deserción 0,73
- Respecto al atributo resultado ICFES pudo identificar que los estudiantes con un resultado bajo tiene una mayor probabilidad de deserción 0,36 que aquellos con ICFES medio bajo 0,26 o medio alto con probabilidad de 0, 21. Este comportamiento es muy similar a los obtenidos con el atributo Matemáticas Icfes Biología Icfes, Química Icfes y Física Icfes en donde se evidencia q los estudiantes con resultados altos en estos atributos tiene probabilidades menores de desertar que aquellos con resultados bajos en estas áreas.

5.3 Resultados fase implementación Caso de Estudio

Durante la fase de implementación del caso de estudio se ejecutaron las técnicas de minería de datos seleccionadas sobre los conjuntos seleccionados y consolidados. La ejecución de las técnicas se realizó sobre los 29 atributos y sobre los 6 atributos seleccionados después de ejecutar los 7 algoritmos de identificación de atributos relevantes con el fin de contrastar los resultados.

Todos los resultados presentados se obtuvieron con el uso de la herramienta WEKA, la cual tiene un licenciamiento GPL desarrollada por la universidad de Waikato [31]. Esta herramienta fue seleccionada toda vez que es una de las que más se ha utilizados en proyectos de minería de datos y aprendizaje de máquina [32]. Otra de las razones que nos llevaron a seleccionar la herramienta WEKA para el desarrollo de caso de estudio es lo indicado por la IEDMS respecto a las líneas de investigación futura pues siguen adelantando casos de estudio apoyados en el uso de herramientas de código abierto.

Al analizar los resultados obtenidos los cuales no solo se aplicaron a los 6 atributos seleccionados como relevantes si no a los conjuntos de datos con los 29 atributos se identificó que existen dos atributos muy significativos en el análisis de la deserción, estos son “Créditos Aprobados” y “Promedio Programa”. Respecto a los créditos aprobados cuando los estudiantes aprueban más del 50% de los créditos del programa tiene menor riesgo de desertar y cuando el promedio académico del programa es medio bajo o bajo las probabilidades de desertar son muy altas respecto a los otros estudiantes. Este supuesto que se obtuvo en el análisis de los árboles se corroboró al ejecutar el clasificador OneR sobre los conjuntos de datos confirmando el resultado obtenido y el análisis realizado.

Por otra parte se identificó que el algoritmo con mayor nivel de aciertos para la sede Cundinamarca fue el IBk.

Tabla 17. Clasificaciones por algoritmos sede Cundinamarca

	Correctamente Clasificados	Incorrectamente Clasificados	Tiempo construcción del modelo
J48	78,18%	21,81%	0,03 seg.
lbk	96,49%	3,51%	0,19 seg.
OneR	77,23%	22,76%	0,04 seg.

Fuente: Elaboración propia

Así mismo para los datos de la facultad de ingeniería se consiguió un mejor resultado con el algoritmo lbk.

Tabla 18. Clasificaciones por algoritmos facultad de Ingeniería

	Correctamente Clasificados	Falsos Positivos	Tiempo construcción del modelo
J48	71,79%	28.20%	0,07 seg.
lbk	93,95%	6.04%	1,37 seg.
OneR	76,06%	23.93%	0,01 seg.

Fuente: Elaboración propia

Los resultados obtenidos permitieron corroborar que los atributos seleccionados como relevantes para el análisis de la deserción resultaron muy apropiados pues en las diferentes pruebas realizadas se pudo validar la pertinencia de dichos atributos.

6. DISCUSIÓN Y CONCLUSIONES

Debido al auge que tiene los sistemas de información en la actualidad, los cuales son parte vital de todo negocio, y a la gran cantidad de datos que se almacenan en dichos sistemas de información, emprender un proyecto de minería de datos resulta relativamente sencillo, por supuesto partiendo del supuesto de que los datos disponibles son de buena calidad. El almacenamiento ordenado de estos datos no es una situación que se presente muy a menudo en las organizaciones lo cual implica una gran cantidad de actividades de preprocesado para llegar a un nivel de madurez que permita adelantar proyecto de minería de datos que resulten exitosos.

Trabajar el proyecto tomando como uno de los ejes fundamentales el modelo MAIE nos permitió identificar las necesidades en cuanto a los datos que se requerían para adelantar el caso de estudio. Este modelo separa el análisis de la deserción en 3 dimensiones, las cuales están muy alineadas con los autores que han estudiado el análisis de la deserción.

Para el desarrollo del caso de estudio se planteó el desarrollo de una encuesta que permitiera obtener algunos datos de estudiantes tanto desertores como no desertores, sin embargo el levantamiento de dicha información implicó un tiempo mayor, por ello el caso de estudio se adelantó con los datos disponibles. La falta de algunos datos fue la principal dificultad para adelantar el análisis de la deserción en las tres dimensiones indicadas en el modelo MAIE, es por ello que los resultados obtenidos al ejecutar el análisis de los datos y las diferentes técnicas de minería de datos arrojaron resultados orientados a la dimensión académica, sin embargo el planteamiento del desarrollo del proyecto y específicamente las actividades adelantadas con los datos de la dimensión académica y la ejecución de las técnicas de minería de datos podrían ampliarse a los datos de las dos dimensiones para llegar a la identificación de patrones más completos en el análisis de la deserción.

Basados en los resultados obtenidos y al análisis de estos, llama la atención que los resultados ICFES para las diferentes áreas no son muy relevante cuando se refiere al análisis de deserción con el uso de las técnicas utilizadas, en general se identificó que los atributos créditos aprobados, créditos inscritos y promedio de la carrera son muy significativo en el análisis de la deserción. Por otra parte, en el análisis de los datos se identificó que existe una tendencia respecto a la probabilidad de deserción y el resultado Icfes, siendo inversamente proporcionar el resultado Icfes con la probabilidad de deserción.

Basados en los resultados obtenidos también se pudo identificar que la ejecución de algoritmos de selección de atributos relevantes son muy importantes para simplificar el análisis de datos cuando se tiene una gran cantidad de atributos. En nuestro caso particular, todas las pruebas adelantadas tanto para el análisis de la sede Cundinamarca como para la facultad de ingeniería nos dieron resultados en donde los atributos relevantes aparecían con cierta importancia.

Disponer de los datos identificados y catalogados por dimensiones permitiría realizar el análisis que se había planeado, dicha disponibilidad de datos facilitaría la identificación de patrones por cada una de las tres dimensiones analizadas que en conjunto ofrecerían patrones mucho más

completos orientados a la disminución de los niveles de deserción enfocadas a situaciones particulares (por dimensiones) y no general, tal como se hace en la actualidad.

7. REFERENCIAS

- [1] I. Society, "Informe Global de Internet 2014," p. 12, 2014.
- [2] R. Mason and F. Rennie, *E-learning and social networking handbook: Resources for higher education*. 2013.
- [3] C. A. Nacional, Ministerio de Educación, "Cifras generales de deserción 2014 SPADIES," *Minist. Educ. Nac.*, vol. 57, no. 1, pp. 10–15, 2014.
- [4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man, Cybern. C Appl. Rev.*, vol. 40, no. X, pp. 601–618, 2010.
- [5] C. Arturo and C. Isaza, "Estudio de causas de deserción de los estudiantes de la universidad la técnica de análisis de correspondencias simples," no. 30, pp. 261–266, 2006.
- [6] G. J. Paramo and C. A. C. Maya, "Deserción estudiantil universitaria. Conceptualización," *Revista Universidad EAFIT*, vol. 35, no. 114. pp. 65–78, 2012.
- [7] Ministerio de Educación Nacional, "Capital Humano para el avance colombiano," 2010.
- [8] U. de los Andes, "Determinantes de la Deserción," vol. 25, no. 49, pp. 122 – 164, 2014.
- [9] ICETEX, "Modelo de Atención Integral al Estudiante - MAIE," 2011.
- [10] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, Jan. 2013.
- [11] O. Scheuer and B. M. McLaren, "Educational Data Mining," 2011.
- [12] B. R, Sánchez-Guzmán A, and R. García, "Minería de datos educativa: Una herramienta para la investigación de patroness de aprendizaje sobre un contexto educativo. (Spanish)," *Latin-American J. Phys. Educ.*, vol. 7, no. 4, pp. p662–668, 2013.
- [13] H. Jiawei and M. Kamber, "Data mining: concepts and techniques," *San Fr. CA, itd Morgan Kaufmann*, pp. 377–385, 2001.
- [14] MINTIC, "Panorama TIC | Publicado en marzo de 2015," *PanoramaTIC*, 2015.
- [15] S. D. Pérez César, "Minería de datos conceptos, técnicas y sistemas," p. 789, 2007.
- [16] F. Virseda Benito and J. Carrillo, "Minería de datos y aplicaciones," 2008, p. 8.

- [17] G. Siemens and R. S. J. D. Baker, "Learning analytics and educational data mining," *Proc. 2nd Int. Conf. Learn. Anal. Knowl. - LAK '12*, p. 252, 2012.
- [18] R. S. J. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.
- [19] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.
- [20] R. S. j. d. B. G. Siemens and R. S. j. d. B. G. Siemens, "Learning analytics and educational data mining: towards communication and collaboration," *Proc. 2nd Int. Conf. Learn. Anal. Knowl.*, 2012.
- [21] P. S. Baker, Ryan S J, "Educational data mining and learning analytics," 2014.
- [22] D. G. Reina, S. L. Toral, and F. Barrero, "Metodologías de Análisis de los Big Data en las Plataformas Educativas."
- [23] A. Rodríguez and J. González, "Prospección de la colaboración utilizando herramientas de minería de datos en ambiente abiertos de aprendizaje colaborativo con el objetivo de mejorar la gestión del proceso de colaboración," 2009.
- [24] E. Hochsztain, "MoDaWeEd : un framework que integra Moodle , Data Mining y Web Usage Mining en el ámbito de la Educación," *Moodle moot*, no. Minian 1999, pp. 1–5, 2011.
- [25] E. García Salcines, "Usando minería de datos para la continua mejora de cursos E-Learning," UNIVERSIDAD DE CÓRDOBA, 2012.
- [26] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. R. Daimlerchrysler, C. S. Spss, and R. W. Daimlerchrysler, "Crisp-Dm 1.0," 2000.
- [27] UNIMINUTO, "UNIMINUTO Plan de Desarrollo 2013 - 2019," p. 320, 2013.
- [28] B. Ezequiel, "Un proceso de transformación de datos para proyectos de explotación de información," *Un proceso Transform. datos para Proy. Explot. Inf.*, vol. 1, no. 2, p. 90, 2013.
- [29] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.
- [30] M. G. Jiménez and a. Á. Sierra, "Análisis de Datos en WEKA–Pruebas de Selectividad," *It.Uc3M.Es*, pp. 1–9, 2003.

- [31] M. Hall, E. Frank, and I. H. Witten, "Practical Data Mining Tutorial 1 : Introduction to the WEKA Explorer," 2012.
- [32] D. García Morate, "MANUAL DE WEKA," 2012.