UNIVERSIDAD NACIONAL DE COLOMBIA

SEDE MANIZALES

MODELO DE DOMINIO ESPECÍFICO PARA ANÁLISIS Y MINERÍA DE DATOS EDUCATIVOS

Emilcy Juliana Hernández Leal

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura
Manizales, Colombia
2024

UNIVERSIDAD NACIONAL DE COLOMBIA

SEDE MANIZALES

MODELO DE DOMINIO ESPECÍFICO PARA ANÁLISIS Y MINERÍA DE DATOS EDUCATIVOS

Mg. Emilcy Juliana Hernández Leal

Tesis presentada como requisito para optar al título de:
Doctor en Ingeniería – Industria y Organizaciones –

Director
PhD. Néstor Darío Duque Méndez

Línea de Investigación:
Sistemas y Gestión de la Tecnología, la información, el conocimiento y la innovación tecnológica en la Industria y Organizaciones
Grupo de Investigación en Ambientes Inteligentes Adaptativos – GAIA –

Universidad Nacional de Colombia
Departamento de Ingeniería Industrial
Doctorado en Ingeniería - Industria y Organizaciones -
Manizales, Colombia
2024

Dedicatoria

En especial a mi familia y a todos los que aportaron paciencia y me dieron soporte. A los amigos y conocidos que sonrieron, me animaron y celebraron cada objetivo cumplido.

“Las cosas no tienen que cambiar el mundo para ser importantes”

Steve Jobs

Agradecimientos

Es momento de agradecer.

Agradezco a mi familia por ser siempre mi sustento y motivación para cumplir con las metas propuestas. A mis padres por sus oraciones, por el amor, por entender y aceptar mis ausencias y por apoyar todo este proceso. A mi hermanito por compartir los momentos de logros y de mayor esfuerzo. La familia es el lugar donde los grandes sueños comienzan y el amor nunca termina.

Gran gratitud al profesor Néstor Darío Duque Méndez, director de este trabajo doctoral, ha sido un bastón de apoyo y guía fundamental en los momentos donde parecía perder el norte. Agradezco sus contribuciones y acompañamiento académico en el desarrollo de esta tesis, también su amistad y confianza depositada en mí. Han sido años de apoyo a mi formación, crecimiento profesional, laboral y personal.

Mi agradecimiento a los integrantes del grupo de investigación GAIA, a quienes aún hacen parte y a quienes en algún momento pertenecieron a esta familia que nos acoge, nos ve crecer y nos motiva a vencer los límites cuando se trata de formarnos y aportar a la sociedad. A quienes pasaron de ser compañeros para convertirse en amigos y un poco más. Por los momentos de alegría, por hacerme sonreír, una sonrisa puede ser el mejor aporte en tiempos difíciles. Por los trabajos conjuntos, por las luchas ganadas y las derrotas, por los artículos aceptados y por los rechazados, por leerme y por dejarme leerlos…siempre hay un aprendizaje, con ustedes he aprendido mucho.

Agradezco a los integrantes del *“Laboratório de Criatividade e Inovação para Educação”* (CrIE) y del *“Laboratório Sistemas de Conhecimento”* (LSC) de la UFSC-Campus Araranguá, quienes me acogieron en mi pasantía. Su visión y aportes a la investigación fueron valiosos. En Brasil aprendí mucho, entre otras cosas, que el café es fundamental en la vida de un investigador y que los brigadeiros le hacen el mejor maridaje.

Agradecer es complejo porque no queremos dejar a nadie afuera, por ello agradezco a todos y cada uno de los que estuvieron presentes en este proceso, sin nombres, pero con la plena seguridad de que en mi corazón los llevo. A quienes se convirtieron en mi familia en Manizales, a los evaluadores que dedicaron su valioso tiempo y me dieron comentarios de mejora, a los que me tuvieron paciencia y siempre creyeron en mí, a los que aportaron granitos de arena y terrones de azúcar ;)

Resumen

El uso de técnicas de análisis de datos para el apoyo de procesos educativos, al igual que en otros dominios de datos, busca potencializar la toma de decisiones y la planeación de estos. Las tecnologías de información y comunicación contribuyen a dichos procesos de análisis. En particular, desde la minería de datos se tiene una opción para dar atención a las necesidades presentes en cuanto a gestión de datos académicos, datos que se producen desde el proceso de enseñanza-aprendizaje como tal, así como también desde procesos de carácter administrativo que están asociados. Dependiendo del nivel educativo, para el caso de Colombia estos niveles se distribuyen en educación pre-escolar, básica, media y superior, los sistemas de información donde son almacenados los datos educativos varían, influyendo también el carácter de la institución (pública o privada). Para el caso de la educación superior, estos sistemas de información o fuentes de datos suelen estar bastante estructurados, facilitando el acceso a los datos y por tanto la extracción de información y conocimiento. No obstante, a nivel de educación básica y media, las fuentes de datos resultan más difíciles de acceder y el tratamiento que requieren los datos antes de ser analizados puede ser considerable.

En este sentido, esta tesis doctoral propone un modelo conceptual con enfoque de dominio específico para minería de datos educativos, que ofrece mecanismos de solución a los problemas particulares de cada etapa del proceso de minería de datos educativos y en general de los modelos de dominio genérico, además, de atender la problemática asociada a los datos que provienen de múltiples fuentes y escalas para una aplicación puntual con datos de educación básica y media en Colombia, acotado también a técnicas de aprendizaje supervisado. De la mano del modelo conceptual, se presenta una estrategia de validación y aplicación de este.

El modelo propuesto puede ser aplicado a diferentes contextos educativos y para diferentes fuentes de datos, contando con el conocimiento de los expertos y con la información que puede ofrecer dicho contexto académico particular, teniendo como conclusión general que los procesos de análisis de datos educativos mediante minería de datos pueden ser abordados desde un enfoque de dominio específico, contribuyendo al logro de resultados satisfactorios en términos de los modelos de minería construidos y del apoyo al usuario por medio de la guía que puede ofrecer contar con el conocimiento del dominio particular. Además, se ofrecen modelos pre-entrenados y mecanismos de transferencia de aprendizaje que permiten aprovechar las ventajas de la minería de datos en ambientes con pocos datos y sin requerimientos de expertos en técnicas de análisis de datos.

Palabras clave: análisis de datos, datos educativos, dominio específico, educación básica y media, minería de datos.

SPECIFIC DOMAIN MODEL FOR EDUCATIONAL DATA MINING AND ANALYSIS

Abstract

The use of data analysis techniques to support educational processes, as in other data domains, seeks to enhance decision-making and planning. Information and communication technologies contribute to these analysis processes. From data mining there is an option to attend to the present needs in terms of academic data management, data that is produced from the teaching-learning process as such, as well as from administrative processes that are associated. Depending on the educational level, in the case of Colombia these levels are distributed in pre-school, basic, secondary, and higher education, the information systems where the educational data are stored vary, also influencing the nature of the institution (public or private). In the case of higher education, these information systems or data sources are usually quite structured, facilitating access to data and therefore the extraction of information and knowledge. However, at the basic and secondary education level, the data sources are more difficult to access and the treatment that the data requires before being analyzed can be considerable.

In this sense, this doctoral thesis proposes a conceptual model with a specific domain approach for educational data mining, which offers solution mechanisms to the problems of each stage of the educational data mining process and in general of generic domain models. In addition, to address the problems associated with data that come from multiple sources and scales for a specific application with data from basic and secondary education in Colombia, also limited to supervised learning techniques. Hand in hand with the conceptual model, a validation and application strategy of this model is presented.

The proposed model can be applied to different educational contexts and for different data sources, counting on the knowledge of the experts and with the information that this particular academic context can offer, having as a general conclusion that the educational data analysis processes through mining Data can be approached from a specific domain approach, contributing to the achievement of satisfactory results in terms of the mining models built and the support to the user through the guidance that having knowledge of the particular domain can offer. In addition, pre-trained models and transfer learning mechanisms are offered that allow taking advantage of data mining in environments with little data and without requiring experts in data analysis techniques.

Keywords: data analysis, educational data, specific domain, basic and secondary education, data mining.

CONTENIDO

Lista de figuras

Lista de tablas

Introducción

La minería de datos puede ser definida como el proceso de descubrimiento de patrones y tendencias en grandes conjuntos de datos (Larose & Larose, 2014). En otras palabras, el objetivo de la minería de datos es descubrir las relaciones que existen en el mundo real a través de los datos que se recolectan de él. Estas relaciones pueden ser del mundo físico, del mundo empresarial, del mundo científico o de algún otro dominio conceptual (Hand et al., 2001). Los datos que son recolectados por el mapeo de entidades en el dominio de interés simbolizan o representan medidas, asociaciones, variables o propiedades de dichas entidades y al ser almacenados se convierten en el sujeto de las actividades de minería de datos y en general de los procesos de análisis de datos.

En este orden de ideas, cada día la cantidad de datos que son recolectados y que requieren mecanismos para ser analizados tiene un crecimiento exponencial; algunos ejemplos de esto son los datos registrados en los supermercados a través de las cajas de pago, los datos recolectados por cámaras y sensores ubicados en las ciudades, los datos de miles de transacciones que se llevan a cabo por la web, los datos de visitas a páginas y sitios del ciberespacio, los datos de los comentarios e interacciones que se hacen en las redes sociales, entre otros.

En particular, en el campo educativo, los sistemas de información académicos pueden almacenar grandes conjuntos de datos provenientes de múltiples fuentes y que presentan diferentes formatos y niveles de granularidad. A la vez, cada problemática educativa particular tiene un objetivo específico con una serie de variables que requieren un tratamiento previo a las técnicas de minería de datos con las que se vaya a abordar (Romero & Ventura, 2013). Para este dominio de datos se ha constituido una línea de trabajo particular denominada Minería de Datos Educativos (EDM) (Scheuer & McLaren, 2012). Esta corriente de análisis de datos busca atender la diversidad de datos que se producen y almacenan en este campo y ofrecer posibilidades de análisis y apoyo en el entendimiento de los problemas y la generación de soluciones y estrategias no solo a nivel pedagógico y curricular, sino también en el nivel directivo e institucional.

La minería de datos educativos considera el tipo de datos que alimentan el proceso de descubrimiento de conocimiento; no obstante, por lo general hace uso de técnicas y algoritmos genéricos, lo que deja espacios de trabajo en el manejo e integración de diferentes fuentes de datos, que incluyen variedad de tipos, escalas y niveles de granularidad, así como también el carácter jerárquico y longitudinal de este tipo de datos, es decir, los datos deben ser convertidos a una forma adecuada para dar solución a cada problema específico. Otras dificultades se

presentan en la captura y representación de algunos eventos y en la selección de la técnica que más se ajuste al dominio y permita tratar el problema estudiado. Finalmente, cuando se llega a la interpretación y aplicación del nuevo conocimiento, es muy importante que los resultados puedan realmente soportar el proceso de toma de decisiones, esto lleva en ocasiones a preferir modelos de caja blanca y acompañar siempre de técnicas de visualización (Romero & Ventura, 2020).

En la búsqueda de solución a problemas y tópicos de interés como los enunciados, EDM hace parte de un grupo de enfoques relacionados y que incluye también las analíticas de aprendizaje, analíticas académicas, analíticas de enseñanza, Big Data en Educación (Big Data in Education BDE) e incluso la ciencia de datos educativos (Educational Data Science – EDS). Para ello, herramientas y frameworks de propósito general pueden ser usados; sin embargo, estas no siempre son asequibles para algunos de los stakeholders del ecosistema (profesores, directivos docentes) dado que requieren seleccionar la técnica y algoritmo adecuado y proporcionar los parámetros que permitan obtener buenos resultados. Ante esto, el planteamiento de herramientas de software específicas se ha tomado como una solución; no obstante, puede que estas iniciativas se limiten a casos o datos muy particulares, por ello se considera una buena alternativa la formulación de un modelo que pueda guiar el proceso de inclusión de los elementos del dominio específico en la solución de problemas de minería de datos educativos (Romero & Ventura, 2020).

En el proceso de formulación del problema de investigación, que busca ser atendido por medio de esta tesis, se realizó una revisión de literatura científica que permitió identificar trabajos relevantes relacionados con la temática que se expuso en los párrafos anteriores, además se revisaron otros trabajos relacionados que dieron fundamento a la problemática y orientaron el estudio hacia la propuesta de diseño de un modelo de dominio específico de minería de datos educativos. Al tratarse de un modelo de minería de datos se contemplan las diferentes etapas que hacen parte de ésta, como la identificación de fuentes de datos; la extracción, transformación y carga de los mismos; el almacenamiento; la construcción de los conjuntos de datos de prueba; la aplicación de técnicas y algoritmos y el análisis y reporte de resultados. En cuanto al dominio específico, se requiere lograr un entendimiento de los ambientes educativos y de los datos asociados a este, considerando simultáneamente la diversidad de escalas y niveles de granularidad de los registros que se puedan obtener.

El documento se estructura de la siguiente forma: se presenta una sección previa con la propuesta de investigación, donde se expone el problema y preguntas de investigación, las objetivos y metodología abordada. Posteriormente, en el capítulo

1 se realiza la revisión teórica de los conceptos asociados a los modelos de dominio específico. En el capítulo 2 se presenta una revisión de las técnicas de análisis de datos y en particular de minería de datos. Por su parte en el capítulo 3 se muestra la caracterización de las fuentes de datos educativos empleados para el desarrollo de esta tesis doctoral. En el capítulo 4 se da a conocer el proceso de definición, construcción y estructuración del modelo de domino específico para minería de datos educativos. Y en el capítulo 5 se describe las generalidades de la validación y aplicación del modelo de dominio específico. La tesis termina mostrando la producción académica, conclusiones, recomendaciones y trabajos futuros.

Propuesta investigativa

Problema de investigación

En la sociedad actual se producen cada vez un mayor número de datos en los diferentes sistemas de información, procesos y ambientes; esto, en gran medida, se debe al incremento en la capacidad de almacenamiento y potencia computacional. Lo anterior ha dado fundamento a la formulación de un conjunto de técnicas que combinan inteligencia artificial y estadística, y se conjugan bajo el apelativo de Minería de Datos (MD) (Pérez Marqués, 2014). Como lo indica su nombre, el fin de estas técnicas es la extracción de información relevante a partir de volúmenes de datos por medio de la utilización de algoritmos o técnicas que tratan de localizar información no trivial (diferencias, patrones, relaciones significativas, efectos de interacción, entre otros) (Castro & Lizasoain, 2012).

Un proceso de análisis de datos requiere de una serie de pasos que se pueden asociar a las fases que se siguen en un proceso de extracción de conocimiento en bases de datos (Knowledge Discovery in Databases – KDD) (Fayyad et al., 1996). La minería de datos es solo una etapa dentro del proceso de KDD. Sin embargo, no se puede desarrollar minería de datos aislada de las demás etapas, por lo cual, para efectos de este trabajo se toma el término de minería de datos como todo el proceso de KDD. Por tanto, las etapas que se requieren en la minería de datos van desde la identificación de diferentes fuentes de datos, pasando por los procesos de ETL (Extracción, Transformación y Carga) (Vaisman & Zimányi, 2014b), el almacenamiento bajo esquemas como Datawarehouse (Jaramillo Valvuena & Londoño, 2014; Vaisman & Zimányi, 2014a), la selección y adaptación de datasets para la aplicación de las técnicas de minería como tal y finalmente el análisis que permita la generación de conocimiento que apoye procesos posteriores como la toma de decisiones (Inmon & Linstedt, 2014).

La minería de datos tiene dos categorías principales, la predicción y la descripción; cada una maneja una serie de técnicas y algoritmos. La categoría predictiva corresponde a aquellas tareas que producen un modelo del sistema descrito por el conjunto de datos analizado; por su parte, la categoría descriptiva encierra las tareas que producen nueva información no trivial basada en el conjunto de datos disponible. Las principales tareas de la minería de datos son: clasificación, regresión, clustering, asociación, sumarización, modelado de dependencia, cambios y detección de la desviación (Kantardzic, 2011). No obstante que entre autores se presentan variaciones, lo que sí es claro es que se siguen dos paradigmas: el aprendizaje supervisado y el no supervisado. El primero se

caracteriza por la existencia de una clase en el conjunto de datos y el segundo tiene ausencia de esta clase, por lo cual las tareas y técnicas a aplicar varían (Cios et al., 2007).

La minería de datos tiene aplicación en una gran variedad de dominios de datos. Particularmente, se ha consolidado una corriente que se enfoca en la minería de datos sobre registros educativos, denominada EDM (por sus siglas en inglés Educational Data Mining) (Baker & Inventado, 2014a; Peña-Ayala, 2014). La EDM corresponde a la aplicación de técnicas de minería de datos a información que se ha originado en entornos educativos. Esta área surge debido al gran potencial que presentan los sistemas educativos para la extracción de información y conocimiento (Scheuer & McLaren, 2012).

Cabe resaltar nuevamente que los procesos educativos son generadores de una cantidad considerable de datos, lo cual ha propiciado un interés en cómo explotarlos en beneficio de la educación y las ciencias del aprendizaje (Baker & Inventado, 2014b). Se refleja lo anterior en el desarrollo de una serie de trabajos que se enmarcan en las tendencias de la EDM (Peña-Ayala, 2014) y que cubren espacios como el análisis de comportamiento de los estudiantes (Jindal & Borah, 2013), el diseño de sitios web usando patrones de acceso, las rutas de aprendizaje en educación a distancia (Nguyen & Yang, 2012), los modelos de transferencia de aprendizaje (Fernandez & Lujan-Mora, 2017), la predicción de rendimiento académico (Bhardwaj & Pal, 2012), el estudio de la deserción estudiantil (Yukselturk et al., 2014), la predicción en la elección de carreras profesionales (Zhu et al., 2016), entre otras.

De esta manera, se ha consolidado, desde la investigación, una contribución significativa a la educación apuntando a la mejora de los procesos académicos e institucionales desde diferentes niveles (Csikszentmihalyi & Wolfe, 2014; Probert & Ridgman, 2013; Winch et al., 2015). A pesar de los trabajos realizados, aún se encuentran una gran variedad de espacios de investigación abiertos (Jindal & Borah, 2013; Cristobal Romero & Ventura, 2013), para lograr desde la formulación de proyectos investigativos, la generación de plataformas académicas sólidas con un enfoque interdisciplinario y bases tecnológicas y aprovechar la historia de hechos para encontrar patrones y modelos que permitan entender los fenómenos relacionados.

Por otra parte, en las ciencias de la computación, se ha introducido el término "*dominio específico*", principalmente para referirse a un tipo particular de lenguajes de programación, los cuales están orientados a resolver un problema en particular, representar un problema específico y proveer una técnica para solucionar una

situación particular (Kosar et al., 2010); pero el término se ha acuñado en otros campos de la computación, específicamente en algunas técnicas de inteligencia artificial, como los planificadores inteligentes (Jiménez Celorrio & de la Rosa Turbides, 2009; Kautz & Selman, 1998; Winner & Veloso, 2003) y la minería de datos (Hübscher et al., 2007).

Los modelos de dominio específico se han utilizado en telecomunicaciones, finanzas y servicios gubernamentales, mostrando un potencial para fortalecer los procesos de KDD tradicionales con el logro de reglas realmente aplicables al negocio (Cao & Zhang, 2007). Los modelos de minería de datos que consideren el dominio pueden aprovechar el conocimiento específico del problema y los expertos humanos como descubridores de patrones. Cuando el usuario o investigador que busca conocimiento no sólo evalúa el resultado de un proceso automático de minería de datos, sino que participa activamente en el diseño de nuevas representaciones y tratamiento de los datos, es decir, se concentra en el dominio específico, se da lugar a una mayor comprensión de los datos (Hübscher et al., 2007).

En Cao & Zhang (2007) se muestra el KDD tradicional como un proceso de descubrimiento automatizado de prueba y error basado en los datos. También se presenta como objetivo de la minería, permitir, por medio de los datos, la demostración y verificación de resultados de investigación e impulsar nuevos algoritmos. En los escenarios del mundo real, donde se requiere minería de datos, los desafíos por lo general provienen de problemas específicos del dominio, involucrando esto, resolver necesidades propias de los usuarios, y para ello se tienen requerimientos que surgen primordialmente de problemas funcionales y no funcionales concretos y que involucran el dominio específico (Cao & Zhang, 2007).

No obstante, aún están por cubrir algunas expectativas en cuanto a problemas asociados a los dominios de datos específicos (Cao, 2010; Cao & Zhang, 2007; Che et al., 2013; Hübscher et al., 2007; Varde & Tatti, 2014). Ante esto, queda un espacio abierto para la formulación de propuestas que busquen mejorar las posibilidades que brinda la minería de datos para el descubrimiento de patrones y conocimiento. En Begum (2013) se resaltan como problemas o espacios actuales relacionados con las fuentes de datos: el manejo de datos heterogéneos que incluyen datos estructurados, semi-estructurados y no estructurados, y como problemas a futuro: el manejo de objetos de datos complejos con alta dimensionalidad, flujos de datos de alta velocidad y múltiples representaciones de objetos y datos temporales. Además, los datos adquiridos a diferente escala y nivel de granularidad hacen difícil la extracción, transformación, carga e

integración desde las fuentes (Dsilva et al., 2015), correspondiendo esto a la segunda etapa del proceso de minería de datos.

De acuerdo con lo anterior, una de las dificultades en la creación de modelos de minería de datos para dominios específicos ha sido la complejidad en el manejo de datos a diferentes escalas (Bermanis et al., 2013; Chen et al., 2012). Los datos educativos son producidos desde diferentes ambientes. Dentro del aprendizaje hibrído o Blended learning se encuentra el E-learning, que a su vez puede soportar tanto el aprendizaje tradicional como el aprendizaje basado en computador, este segundo, puede a su vez soportarse del aprendizaje en línea o del aprendizaje basado en la web. En el aprendizaje tradicional, cara a cara, suele tenerse una jerarquización en niveles educativos que parten desde la formación de la primera infancia hasta la educación superior, sin embargo, dichos niveles podrían también ser desarrollados desde ambientes educativos basados en computador (Romero & Ventura, 2020). Todos estos ambientes constituyen fuentes de datos educativos que aportan complejidad al proceso de desarrollo de un modelo de minería de datos y que dejan a disposición un espacio de investigación con alto potencial.

En la Figura 1 se muestra por medio de un árbol de causas y efectos la identificación del vacío de conocimiento. Se resaltan algunas de las principales causas que dieron lugar al planteamiento del problema principal de investigación, así como también los efectos generados por dicho problema.

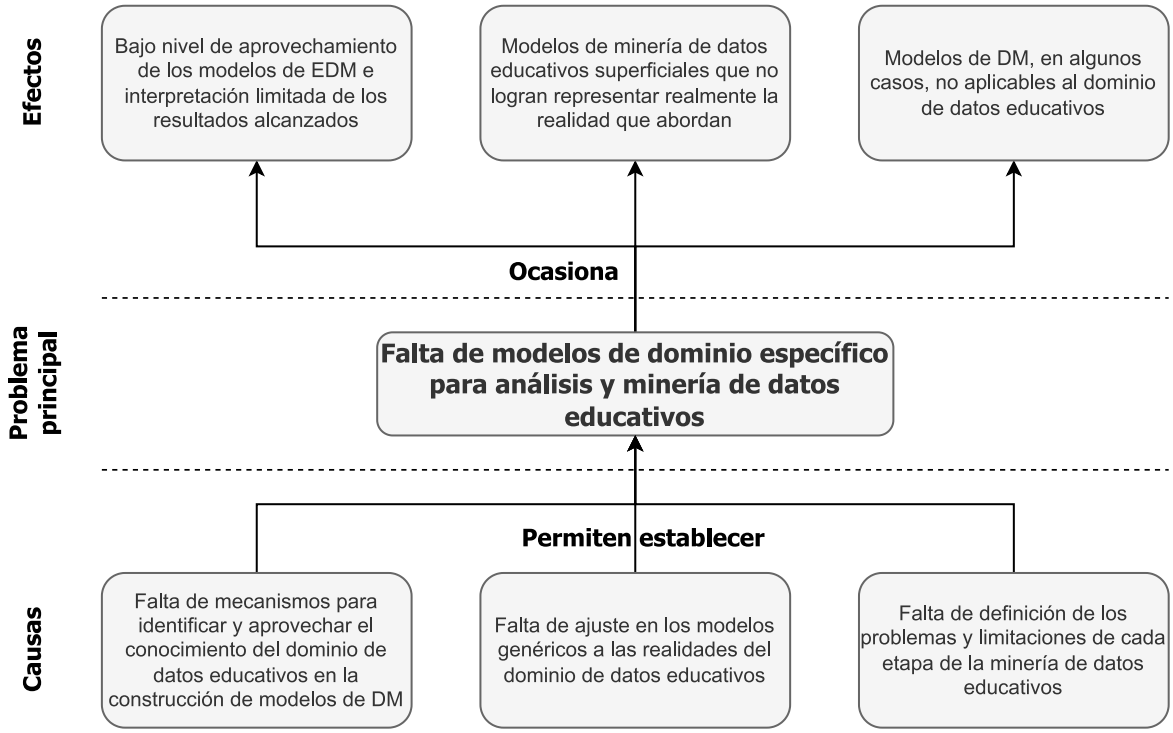


Figura 1. Declaración del vacío de conocimiento
Fuente: Elaboración propia

Preguntas de investigación

Como se aprecia en la declaración del vacío del conocimiento, se identificaron tres necesidades que motivaron el planteamiento de la pregunta de investigación. Se encontró, por un lado, que existen dificultades en la definición de los problemas asociados a cada etapa del proceso de minería de datos. Por otro, que los modelos genéricos de minería de datos presentan poco ajuste a las realidades de cada dominio de datos. Finalmente, se identificó la necesidad de construir modelos de análisis y minería de datos para dominios específicos. A partir de lo anterior, se formula la siguiente pregunta de investigación general:

¿En qué medida el desempeño del proceso de análisis y la aplicación de técnicas de minería de datos se afecta cuando se considera un modelo de dominio específico para datos educativos provenientes de diferentes fuentes y múltiples escalas?

Los procesos de análisis de datos son complejos e incluyen cierto número de pasos o etapas, cada una con su propia problemática, por lo cual se hace necesario que la pregunta de investigación general se fraccione en varias aproximaciones a partir de las sub-preguntas. La pregunta general empieza con

un “en qué medida” puesto que se deben realizar acciones para cada una de las etapas y de los resultados que se logren obtener en estas se podrá medir el desempeño general del proceso. Por lo anterior, desde la pregunta general se derivan las siguientes preguntas emergentes:

- ¿Qué elementos deben componer un modelo de dominio específico para análisis y minería de datos educativos?

- ¿Qué enfoques, técnicas y algoritmos de análisis y minería de datos educativos pueden ser aplicados en el modelo de dominio específico?

- ¿Cuáles fuentes de datos podrían alimentar el modelo de dominio específico?

- ¿Cómo integrar los elementos caracterizados para la construcción de un modelo de dominio específico para el análisis y minería de datos educativos?

- ¿Cómo se podría validar el modelo de dominio específico propuesto para el análisis y minería de datos educativos de diferentes fuentes y múltiples escalas?

Sistema de objetivos

En coherencia con la pregunta y sub-preguntas de investigación se plantea el siguiente sistema de objetivos, que además del objetivo general contempla cinco objetivos específicos.

Objetivo General

Diseñar, construir y validar un modelo de dominio específico para minería de datos educativos que contribuya de forma positiva en el desempeño del proceso de análisis de datos provenientes de diferentes fuentes y múltiples escalas.

Objetivos Específicos

1. Caracterizar los elementos que componen un modelo de dominio específico para minería de datos educativos.

2. Caracterizar los diferentes enfoques, técnicas y algoritmos de análisis y minería de datos educativos que pueden ser aplicados en el modelo.

3. Caracterizar las diferentes fuentes y múltiples escalas de datos que alimentarían el modelo de dominio específico.

4. Proponer un modelo de dominio específico para minería de datos educativos que soporte el análisis de datos provenientes de diferentes fuentes y múltiples escalas

5. Validar el modelo de dominio específico propuesto por medio del desarrollo de un prototipo y realizar pruebas con un caso de estudio particular.

A través de la ejecución de los objetivos específicos se construyó el puente para llegar a una serie de resultados que se conectan entre sí, como se muestra en la Figura 2.

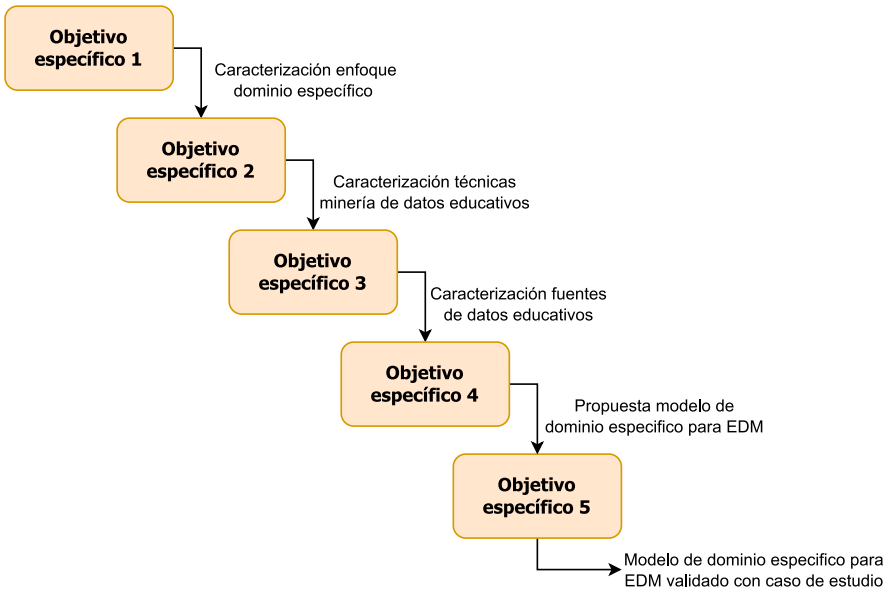


Figura 2. Sistema de resultados por objetivo
Fuente: elaboración propia

Metodología del estudio

Con el fin de dar cumplimiento a los objetivos planteados para esta investigación, se ejecutaron 6 etapas, las cuales se plantean y explican a continuación, en la Tabla 1.

Tabla 1. Etapas para abordar los objetivos

Etapa	Descripción	Objetivo asociado	Método / herramienta
1	Elaboración del marco teórico y revisión del estado del arte	Determinar y caracterizar los elementos que componen un modelo de dominio específico para minería de datos educativos	Revisión sistemática y narrativa de literatura Asesoría de expertos
2	Caracterización de las diferentes fuentes de datos a integrar	Determinar las diferentes fuentes de datos de educación básica y media presencial que alimentarían el modelo de dominio específico.	Trabajo con expertos del dominio de datos Revisión de literatura
3	Revisión y comparación de técnicas de análisis y minería de datos	Determinar y caracterizar los diferentes enfoques, técnicas y algoritmos de análisis y minería de datos educativos que pueden ser aplicados en el modelo.	Revisión de literatura Ejecución de algoritmos de MD Comparación de rendimiento
4	Definición del modelo de dominio específico para minería de datos educativa	Proponer un modelo de dominio específico para minería de datos educativos que soporte el análisis de datos de educación básica y media presencial provenientes de diferentes fuentes.	Plataforma de MD que integre las técnicas a evaluar según el modelo
5	Implementación del prototipo para la validación del modelo propuesto	Validar el modelo de dominio específico propuesto por medio del desarrollo de un prototipo y realizar pruebas con un caso de estudio particular.	Estrategia de integración y almacenamiento de los datos educativos del caso de estudio y técnicas aplicar.
6	Difusión de los resultados obtenidos	Para todos los objetivos	Estrategia: Identificación de revistas objetivo, preparación de artículos, revisión de idioma, someter a evaluación, correcciones y publicación

Fuente**: Elaboración propia**

Cada etapa estuvo formada por una serie de actividades con las que se planteó alcanzar el objetivo asociado y como resultado un producto o productos. Seguidamente se detalla el plan de actividades en la Tabla 2.

Tabla 2. Plan de actividades detallado

Etapa 1. Elaboración del marco teórico y revisión del estado del arte	
Actividades	**Producto**
- Revisión bibliográfica de modelos de dominio específico y modelos de dominio genérico; comparación de los dos grupos. - Revisión bibliográfica de técnicas de análisis y minería de datos, teniendo en cuenta técnicas, herramientas y algoritmos; haciendo	● Marco teórico ● Estado del arte ● Estado de la práctica

énfasis en la minería de datos educativos. - Revisión bibliográfica de tratamiento de datos educativos y en general de datos multiescala, manejo e integración de diferentes fuentes. - Conceptualización de los elementos asociados al modelo.	
Etapa 2. Caracterización de las diferentes fuentes de datos a integrar	
Actividades	**Producto**
- Revisión y caracterización de las fuentes de datos educativos que alimentarán el modelo - Generación de estrategias para el tratamiento que requieren los datos antes de iniciar la integración, almacenamiento y análisis. - Diseño y construcción de una base de datos académicos para el almacenamiento de los datos del caso de estudio.	● Modelado de datos del dominio ● Base de datos académicos
Etapa 3. Revisión y comparación de técnicas de análisis y minería de datos	
Actividades	**Producto**
- Revisión de trabajos previos y de plataformas para el análisis y minería de datos, concentrándose en minería de datos educativos. - Comparación, desde la revisión literaria, de técnicas de análisis y minería de datos educativos para determinar las que reportan mayor ajuste. - Comparación de técnicas y algoritmos por medio de la aplicación sobre datos propios del caso de estudio.	● Conjunto de técnicas y algoritmos de análisis y minería de datos que harán parte del modelo.
Etapa 4. Definición del modelo de dominio específico para minería de datos educativos	
Actividades	**Producto**
- Diseño de los componentes del modelo y de la arquitectura que se asociará a este. - Estructuración y acoplamiento de los componentes del modelo. - Conformación de la comunicación entre los componentes del modelo por medio de la definición de las etapas desde la entrada de los datos, pasando al procesamiento y finalizando con el análisis de las salidas.	● Modelo de dominio específico para el análisis y minería de datos educativos.
Etapa 5. Implementación del prototipo para la validación del modelo propuesto	
Actividades	**Producto**
- Desarrollo, implementación e integración del prototipo del modelo. - Ejecución del prototipo para el caso de estudio, por medio de pruebas para una verificación preliminar. - Ejecución de pruebas y comparación del desempeño del modelo con un modelo genérico. - Evaluación de los resultados obtenidos.	● Documento de validación del modelo por medio de las pruebas preliminares y finales.
Etapa 6. Difusión de los resultados obtenidos	
Actividades	**Producto**
- Preparación de informes parciales. - Preparación de artículos académicos resultado del desarrollo de cada una de las etapas planteadas. - Preparación del documento de tesis de doctorado.	● Documento de tesis doctoral. ● Artículos sometidos a revistas indexadas. ● Ponencias para eventos internacionales.

Fuente: Elaboración propia

Resumen del capítulo

En este capítulo se realizó la presentación del problema de investigación, con un acercamiento a la motivación para el planteamiento de la investigación y de los objetivos que la orientaron. Así mismo, se mostró la metodología para el desarrollo de esta tesis doctoral.

CAPÍTULO 1 – Modelos de Dominio Específico

Los modelos de dominio específico, o en particular los enfoques de dominio específico son definidos como formas o estrategias de abordar una problemática en las cuales interviene conocimiento propio de un campo o dominio de datos y que a través de dicho conocimiento se puede llegar a tener un mayor entendimiento del problema y por lo tanto soluciones que se adaptan mejor y que pueden llevar a resultados más ajustados. Los modelos de dominio específico han sido abordados desde el punto de vista teórico o conceptual y también desde el práctico, generando propuestas en varios campos de estudio.

Teniendo en cuenta lo anterior, en este capítulo se realiza un acercamiento a la conceptualización de enfoques de dominio específico, para luego avanzar a analizar modelos de dominio específico, primero se hace una revisión conceptual y luego del estado del arte de modelos teóricos y prácticos que serán un punto de partida para el posterior planteamiento del modelo de dominio específico para minería de datos educativos.

1.1 Conceptualización de dominio específico

Desde el punto de vista más amplio, el enfoque de dominio específico es la oposición al enfoque de dominio general o genérico. A pesar de sonar bastante básico, es la primera aproximación que se tiene y que es lógicamente deducible. El término dominio específico ha sido usado en diferentes contextos, se puede decir que prácticamente en todos los campos de estudio, dado que cada campo como tal abarca un dominio específico. Por lo tanto, lo que se pretende con los enfoques de dominio específico es aprovechar ese conocimiento propio que se presenta en cada campo de estudio o problemática. En particular, en las ciencias de la computación, el concepto de dominio específico empieza a aparecer asociado a los lenguajes de modelado (Karagiannis et al., 2016).

Es así como el modelado de dominio específico ha ganado fuerza en los últimos años, no solo en la academia, sino también en la industria, surgiendo un número considerable, no solo de enfoques, sino también de herramientas. Sin embargo, algunas de estas herramientas son difíciles de integrar y se van a un extremo o aspecto demasiado específico del dominio modelado llevando a la tendencia del aislamiento. Por ejemplo, se tienen lenguajes de modelado de dominio específico que se centran únicamente en la parte grafica o de diagramas, mientras que otros se enfocan netamente en lenguajes textuales (Gerbig, 2017).

En concreto, un modelo de dominio específico (o Domain-Specific Model en inglés) es un modelo que se construye para un dominio de aplicación particular, es decir, para una tarea específica o para un tipo de problema concreto. Algunos ejemplos de dominios de aplicación podrían ser la simulación de procesos industriales, el diseño de sistemas de control de tráfico aéreo, o el análisis de datos médicos.

1.2 Estado del arte dominio específico

Se presenta a continuación, una revisión del estado del arte de modelos de dominio específico, la cual permitió identificar los aspectos, características, autores y enfoques que giran en torno a la temática y que fueron utilizados como insumos para la construcción y justificación de la propuesta y el desarrollo de esta Tesis. Los principales hallazgos se relatan seguidamente.

En diferentes ámbitos de la computación se ha introducido el concepto de dominio específico. Uno de ellos se relaciona con los modelos dirigidos por arquitectura (MDA por sus siglas en inglés de Model Driven Architecture), los cuales han alcanzado un mayor impacto al expandirse a modelos orientados por arquitectura de dominio específico (DSMDA), en estos se busca ayudar a los desarrolladores a realizar las representaciones de sus sistemas utilizando conceptos propios o familiares al dominio (Agrawal, 2003). Ahora bien, para los DSMA, se requiere de lenguajes de especificación de alto nivel, surgiendo otro enfoque asociado al dominio específico, los lenguajes de modelado específicos del dominio (DSML por sus siglas en inglés de Domain-Specific Modelling Languages), los cuales tienen el propósito de ayudar a definir y construir familias de metamodelos dentro de un dominio específico (De Lara & Guerra, 2012).

Igualmente, la computación integrada por modelos (MIC Model Integrated Computing), cobra atención como método para el desarrollo y mantenimiento de aplicaciones de dominio específico. Los MIC permiten modelar el entorno de modelado mediante metamodelos específicos del dominio deseado, los cuales permiten a su vez, que el entorno de diseño evolucione de forma eficiente frente a los requerimientos cambiantes del dominio (Nordstrom et al., 1999). Otro concepto asociado son los DSDE (Domain-Specific Design Enviroments) que ayudan a capturar especificaciones a través de modelos de dominio, son una herramienta que soporta el proceso de diseño mediante análisis automático y simulación de comportamientos de sistema, pero tienen un alto costo de desarrollo, por lo cual su penetración en el campo de la ingeniería es limitada (Leédeczi et al., 2001).

Como se ha evidenciado hasta el momento, son varios los conceptos o denominaciones que se van adhiriendo al enfoque de dominio específico; hacer una línea del tiempo para la evolución de las iniciativas asociadas puede ser difícil, dado que los autores van dando denominaciones particulares. En 2004, se acuña el término DDD (Domain-Driven Design), el cual se puede definir como un enfoque para el desarrollo de software donde se tiene o se crea una conexión fuerte entre los conceptos del modelo y el núcleo del dominio de estudio. DDD como tal, no es una metodología ni una tecnología, por lo que puede tornarse bastante abstracto, se basa en algunas premisas para proveer una estructura de prácticas que ayuden a tomar decisiones de diseño en proyectos de software que involucren dominios complejos. Las premisas incluyen: (1) enfocarse en el núcleo y la lógica del dominio; (2) tener un modelo como base de los diseños complejos; y (3) fortalecer la colaboración entre técnicos y expertos del dominio para conseguir estar lo más cercano posible a los conceptos fundamentales del dominio y de su problemática (Evans, 2004).

En el párrafo anterior se trajo a colación una premisa fundamental de los enfoques orientados al dominio, la participación de los expertos del dominio. El proceso de vinculación de los expertos en el dominio y de su comprensión de la problemática, es uno de los aspectos que se pueden tornar más difíciles en la computación. Se requiere hacer mapeos manuales de los conceptos presentes en el dominio del problema para que el diseño del objetivo se logre expresar en términos del dominio. Surge así, otro término asociado, los entornos de ingeniería de dominio específico, DSEE (Domain Specific Engineering Environment) (Patwari et al., 2016), este concepto más reciente, incluye la definición de un entorno de diseño que tenga en cuenta el dominio y se asocia a la Ingeniería Dirigida por Modelos (MDE) y a un Lenguaje de Modelado Específico de Dominio (DSML), término que se había mencionado anteriormente y que se ampliará a continuación.

Los lenguajes específicos de dominio (DSL Domain-Specific Language) son lenguajes especializados para un dominio de aplicación particular, contrastan con los lenguajes de propósito general (GPL General Purpose Language) (Iung et al., 2020). Existen varios tipos de DSL, algunos son lenguajes de modelado y otros son de programación como tal. Los lenguajes para fines especiales han existido en la computación desde hace mucho tiempo, pero se han empezado a tornar más populares con el aumento de la construcción de soluciones específicas para un dominio (Bettini, 2016). En concordancia, los DSL generalmente están restringidos, tanto en el dominio donde se usan como en su expresividad. Sin embargo, esta restricción da la ventaja de diseñar más fácilmente el lenguaje, incluida la sintaxis y la semántica de todos los elementos que lo vayan a componer. Volviendo al contraste entre un DSL y un GPL, que es ampliamente

aplicable sin ninguna característica para un dominio particular, los lenguajes de programación suelen ser GPL, y los lenguajes de modelado pueden estar más enfocados al dominio, pero existe el Lenguaje de Modelado Unificado (UML) que es de dominio general. Si se usa un DSL para propósitos de modelado, entonces se hace referencia a un lenguaje de modelado específico de dominio o DSML (Clark et al., 2015).

Entre las ventajas de los DSL se encuentra la capacidad de abstracción del dominio del problema, alimentando la idea de que los expertos del dominio puedan comprender, validar y modificar los programas de lenguaje específico, aunque pocas veces se consigue. Los DSL logran que en el proceso de desarrollo de una solución informática se termine involucrando más un grupo de expertos del dominio, que, a pesar de tener menos experiencia técnica, cuentan con el conocimiento profundo del dominio. En cuanto a desventajas, prevalece el costo de diseño, desarrollo y mantenimiento del lenguaje; dificultad para integrar con otros componentes o sistemas, poca presencia de expertos y personal que pueda dar soporte; dificultad para equilibrar la compensación entre la especificidad que brinda para el dominio frente a una solución construida con un lenguaje de propósito general (Bettini, 2016).

De acuerdo con esto, una de las ventajas a rescatar de los DSL es la integración a los equipos de desarrollo de personal experto del dominio que brinda su conocimiento especializado. Este empoderamiento del usuario o experto se debe dar en la parametrización de la solución, incluyendo el modelo, la composición del paradigma y la herramienta, todos propios del dominio (Desolda et al., 2017).

1.3 Modelos teóricos/prácticos con enfoque de dominio específico

Según Bellifemine et al., (2011) los modelos de dominio específico están entre el código de aplicación específica y los enfoques de middleware de propósito general, estos abordan específicamente y estandarizan los desafíos principales de los diseños dentro de un dominio de aplicación particular. Mientras se mantiene una alta eficiencia, los marcos de dominio específico permiten un desarrollo más efectivo de aplicaciones personalizadas y con la provisión de abstracciones de programación de alto nivel adaptadas para el dominio de aplicación de referencia. En este apartado se hace un recuento de algunas aplicaciones teóricas y prácticas de modelos donde se ha usado el enfoque de dominio específico, este recuento abarca una variedad de campos de estudio con el fin de ratificar la multiplicidad de usos que se pueden dar al enfoque.

En Desolda et al., (2017) se emplea el enfoque de dominio específico en el desarrollo de dispositivos IoT (Internet de las Cosas). Los autores afirman que los beneficios sociales y prácticos obtenidos por los dispositivos IoT aún no son los deseables, por lo cual identifican que las oportunidades que ofrece IoT pueden amplificarse si se conciben nuevos enfoques para permitir que los usuarios no técnicos se involucren directamente en "componer" sus objetos inteligentes mediante la sincronización de su comportamiento. Para ello, construyen un modelo que incluye nuevos operadores para definir reglas que combinan múltiples eventos y condiciones expuestas por los objetos inteligentes y para definir restricciones temporales y espaciales en la activación de reglas, lo anterior empleando herramientas y enfoque de dominio específico. Se involucraron en el desarrollo expertos en automatización del hogar para la definición de la arquitectura de la plataforma donde se admite la definición y ejecución de las reglas.

Por su parte, en Krahn, Rumpe, & Völkel (2014), se quiere demostrar cómo un proceso de desarrollo ágil, que usa código y modelos al mismo nivel de abstracción, puede ser usado para desarrollar de forma eficiente un sistema de software; lo anterior con el uso de los DSML para separar los aspectos tecnológicos y específicos de la aplicación. Los diferentes roles que juegan los desarrolladores en la realización del software se ven involucrados, a la vez, esta técnica simplifica la integración de los expertos del dominio con el equipo de desarrollo al proporcionarles herramientas específicas para expresar sus conocimientos sin la necesidad de profundizar en los problemas de software. En este trabajo se presenta el marco MontiCore que es utilizado para simplificar el desarrollo de DSML. Esta simplificación es asistida por herramientas fáciles de usar y ejecutadas rápidamente que permiten un proceso de desarrollo mucho más ágil. Además, se define la posibilidad de tener nuevos roles dentro un proyecto basado en DSML que serán llevados a cabo por desarrolladores y expertos en el dominio, respectivamente.

Los entornos de desarrollo específicos del dominio pueden ayudar a capturar especificaciones en forma de modelos de dominio. En Leédeczi et al., (2001) se explica cómo estas herramientas apoyan el proceso de diseño al automatizar el análisis y simular el comportamiento esencial del sistema. No obstante, el alto costo de desarrollar entornos de generación y modelado específicos del dominio impide su penetración en algunos campos de la ingeniería que tienen bases de usuarios limitadas. En este caso la MIC, un enfoque de la ingeniería basada en modelos ayuda a componer entornos de diseño específicos de dominio de manera rápida y tal vez más rentable (ver Figura 3). Los autores describen que la MIC proporciona una manera de crear dichos entornos mediante el uso de una

arquitectura de meta-nivel para especificar el lenguaje de modelado específico del dominio y las restricciones de integridad, conduciendo a reducciones significativas en los costos de desarrollo y mantenimiento.

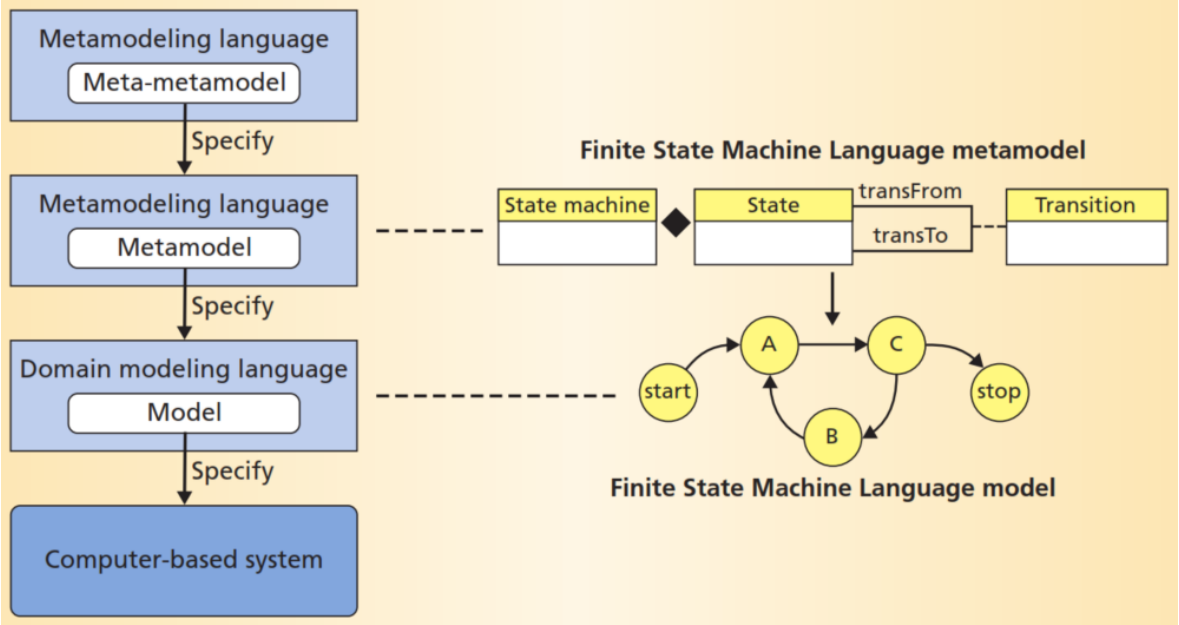


Figura 3. Arquitectura de metamodelado de cuatro capas de computación integrada (MIC)
Fuente: tomado de Leédeczi et al., (2001)

Las arquitecturas de dominio específico, las plataformas middleware y las técnicas de análisis apoyadas en conocimiento específico ayudan a los ingenieros a construir sistemas más eficientes; pero unir estos elementos puede no resultar fácil, para ello una forma de integrarlos son las DSDI (Domain-Specific Development Infrastructure). En Edwards & Medvidovic (2008) se propone, mediante dos estrategias, la incorporación de semánticas para simplificar y automatizar la integración de DSDI. Las estrategias son: (1) la creación de metamodelos de arquitecturas de referencia, plataformas de middleware y tecnologías de análisis; (2) la aplicación de transformaciones de modelos que realizan análisis y síntesis. Mediante esta metodología se mejora el soporte para la composición y validación automatizadas de metamodelos y la transformación automatizada de modelos al integrar y aprovechar la semántica tanto en la fase de meta-modelado como en la fase de interpretación (ver Figura 4).

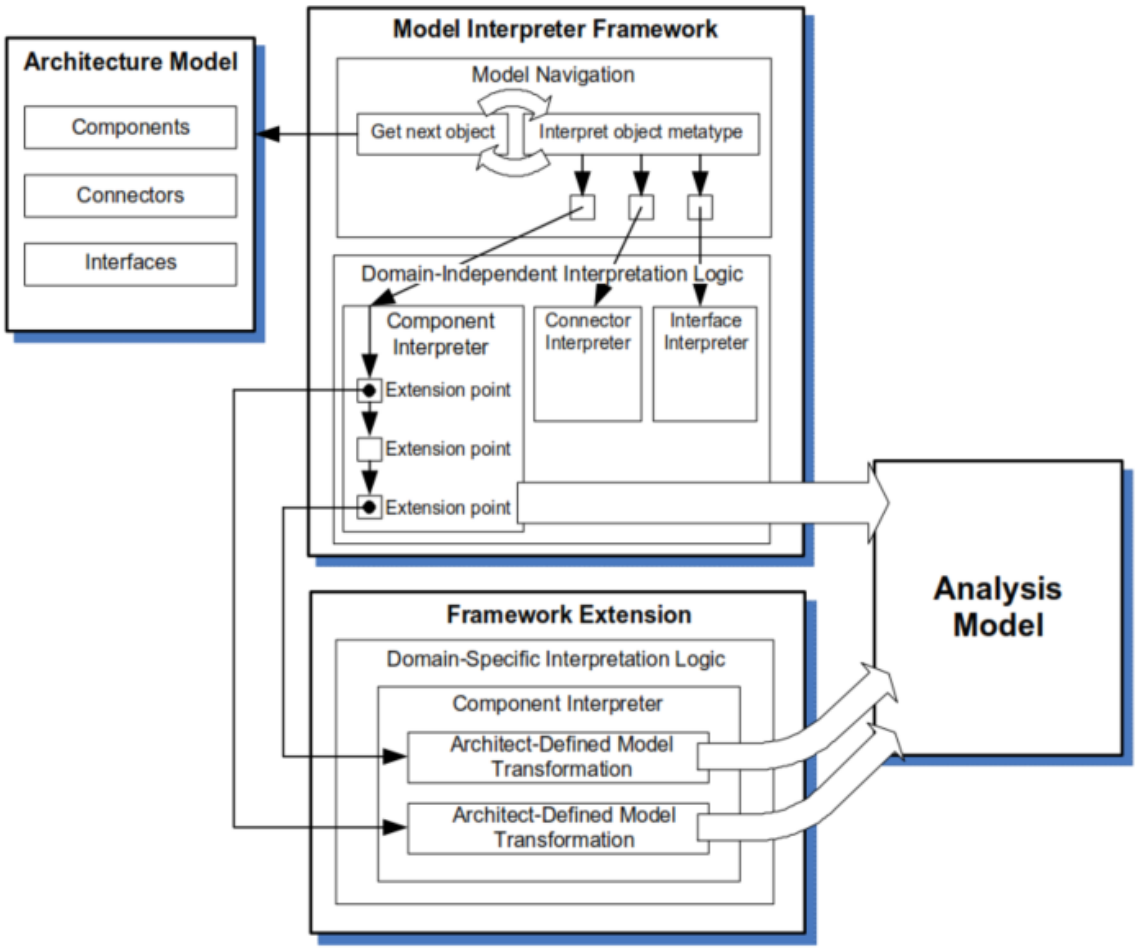


Figura 4. Diseño de alto nivel de un Framework Interpretador de modelos
Fuente: tomado de Edwards & Medvidovic (2008)

Uno de los campos de estudio donde se inició el uso de enfoque de dominio específico fue en la planificación o planificadores inteligentes. En Winner & Veloso (2003) se presenta una alternativa para la planificación independiente, usando el dominio para proporcionar planes de ejemplo que permitan demostrar cómo resolver problemas en un dominio particular y así mismo, usar esa información para aprender a planificar ese dominio específico. Los autores presentan los “dsPlanners”, o planificadores específicos de dominio. También el algoritmo DISTILL, usado para el aprendizaje de los dsPlanners automáticamente a partir de planes de ejemplo. Este algoritmo convierte un plan en un dsPlanner y luego lo combina con otros dsPlanners previamente aprendidos. Los resultados muestran que los dsPlanners aprendidos automáticamente por el algoritmo representan de manera compacta su experiencia de planificación específica. Además, los dsPlanners generalizan situacionalmente los planes de ejemplo dados, lo que les permite resolver eficientemente problemas que no se han encontrado previamente.

Continuando en el campo de la planificación, el primer enfoque de aprendizaje para los planificadores de dominio específico fue el aprendizaje inductivo supervisado, inicialmente se utilizó la programación genética y el aprendizaje de lista de decisiones, sin embargo, no se consiguieron resultados confiablemente buenos. Más adelante, se presentó un enfoque diferente basado en generalizar un plan de ejemplo en un programa de planificación y luego fusionar el código fuente resultante con los anteriores. Posteriormente surgió un enfoque alternativo para el aprendizaje controlado, consiste en programas de planificación con aprendizaje de dominio específico. Estos programas reciben como entrada un problema de planificación de un dominio fijo y devuelven un plan que resuelve dicho problema. En general, los planificadores de dominio específico tienen que enfrentar el problema de la generalización. Estas técnicas crean programas de planificación a partir de un conjunto dado de problemas resueltos, por lo que, en teoría, no pueden garantizar la resolución de problemas posteriores (Jiménez Celorrio & de la Rosa Turbides, 2009).

Las ontologías son otro campo de estudio en el cual se destaca el uso de enfoque de dominio específico. Para contextualizar, las ontologías son definidas como especificación explícita y formal de una conceptualización compartida de un dominio, estas proveen de un conocimiento común y compartido de un dominio de interés particular. Han sido ampliamente aceptadas como el modelo de representación del conocimiento más avanzado y se encuentran entre los bloques de construcción más importantes de la web semántica. En Parekh, Gwo, & Finin (2004) analizaron un método rápido para facilitar la evaluación y el enriquecimiento de las ontologías de dominio utilizando un enfoque de minería de texto (ver Figura 5). Los autores exploraron textos específicos y glosarios o diccionarios del dominio para generar automáticamente conjuntos de conceptos/términos que tienen relaciones taxonómicas o no taxonómicas entre ellos. Posteriormente un ingeniero experto en ontologías del dominio revisa estos conjuntos generados y los usa para evaluar y enriquecer la ontología de dominio. Los anterior fue validado en el campo de las ciencias ambientales, utilizando el enfoque en interacción con un experto en el dominio y se obtuvo, de forma empírica, que el enfoque puede ayudar a los ingenieros expertos a construir ontologías específicas de dominio de manera eficiente, además se ratificó la importancia de contar con la participación no solo de los expertos en el campo general de las ontologías, sino también en el caso particular de la ontología de dominio en proceso de construcción.

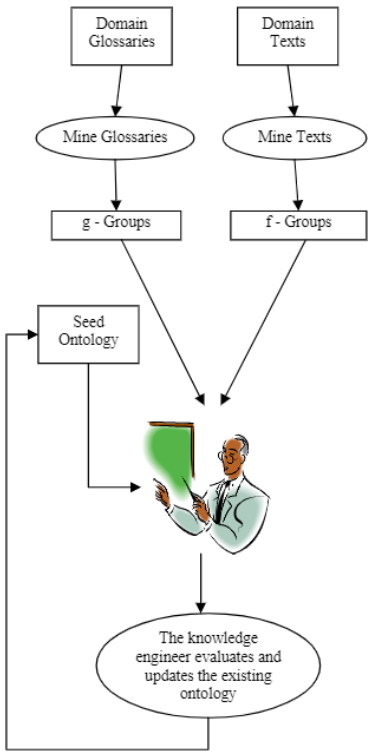


Figura 5. Construcción Ontología con uso de Dominio
Fuente: tomado de Parekh, Gwo, & Finin (2004)

Ejemplos de otras construcciones de ontologías con enfoque de dominio específico se encuentran en los siguientes trabajos. (El Ghosh et al., 2017), en este trabajo se construyó una ontología de dominio específico basada en conocimiento del sistema legal libanés, desde recursos textuales. Es de notar, que el aprendizaje en las ontologías no es una tarea trivial, debido a razones asociadas a la cantidad de trabajo de investigación previa al aprendizaje automático de una ontología específica de dominio a partir de los datos. En Xu, Rajpathak, Gibbs, & Klabjan (2019) proponen un sistema de clasificación en dos etapas para aprender automáticamente a partir de datos de texto no estructurados, primero hacen una recopilación de conceptos candidatos y en segundo lugar aplican un clasificador para obtener los diferentes tipos de concepto. En otro caso práctico, Selvaraj, Burugari, Sumathi, Nayak, & Tripathy (2019) hacen uso de las ontologías de dominio específico para apoyar un sistema de recomendación, buscando con esto resolver algunos problemas relacionados con las búsquedas. Los autores indican que la consulta es una de las funcionalidades básicas que se esperan de los sistemas de bases de datos y la eficiencia de la consulta se ve afectada negativamente por el aumento de las tablas; por lo tanto, los meta-motores de búsqueda combinan los resultados de diferentes motores y mejoran la efectividad de la búsqueda en la web gracias a una amplia cobertura de datos indexados; con su método utilizan múltiples

consultas en lugar de una sola consulta y se apoyan en una ontología de dominio específico para dar como resultado una búsqueda específica de la consulta.

En la revisión de literatura realizada hasta el momento, solo se identifica un trabajo que realiza un primer acercamiento a la inclusión del enfoque de dominio específico para un modelo de minería de datos educativos. No obstante, dicho trabajo se encuentra centrado en entender las rutas de navegación de los estudiantes en un sistema educativo hipermedia (CoMPASS) y no se encuentra una conceptualización y aplicación para un contexto educativo como el analizado en esta tesis doctoral, ni un modelo creado con base en la representación del conocimiento del dominio. Lo que se muestra en este documento es la propuesta de un enfoque para tomar las ventajas del conocimiento del dominio específico y de los expertos humanos para encontrar patrones en un problema asociado a un contexto educativo particular. Sin embargo, en el documento se resaltan las razones por las cuales usar las representaciones y visualizaciones del dominio específico presenta una ventaja respecto de enfoques generales, por lo cual se toma este trabajo como punto de referencia para el desarrollo del modelo propuesto en esta tesis doctoral.

Para dar cierre a la caracterización del uso del enfoque de dominio específico, se presenta en la Tabla 3 los resultados de la búsqueda de la relación entre diversas áreas temáticas con el enfoque de dominio específico. Para ello se usó como ecuación los términos “domain specific” OR “domain-specific” OR “specif* domain” y se restringió por áreas temáticas, para el período 2010 a 2022 y solo artículos de revista. Dicha búsqueda se realizó en la herramienta bibliográfica Web of Sciences. En la tabla se muestran las primeras 40 áreas temáticas, ordenadas a partir del área con mayor cantidad de trabajos que corresponde a Computer Science Software Engineering. En la Figura 6 se hace el compendio de las 10 áreas temáticas donde más trabajo se encuentran relacionados al enfoque de dominio específico.

Tabla 3. Trabajos con enfoque de dominio específico por áreas de estudio

Área temática	**Registros**	**% de 8343**	**Área temática**	**Registros**	**% de 8343**
Computer Science Software Engineering	1112	8,38	Psychology Applied	156	1,18
Computer Science Information Systems	856	6,45	Linguistics	153	1,15
Engineering Electrical Electronic	601	4,53	Biochemistry Molecular Biology	150	1,13
Computer Science Artificial Intelligence	576	4,34	Mathematical Computational Biology	149	1,12
Computer Science Theory Methods	526	3,96	Information Science Library Science	148	1,11

Área temática	Registros	% de 8343	Área temática	Registros	% de 8343
Neurosciences	512	3,86	Medical Informatics	136	1,02
Education Educational Research	460	3,46	Psychology Clinical	136	1,02
Computer Science Interdisciplinary Applications	447	3,37	Health Care Sciences Services	130	0,98
Psychology Experimental	388	2,92	Management	122	0,92
Psychology Multidisciplinary	377	2,84	Engineering Multidisciplinary	118	0,89
Psychology Educational	266	2,00	Operations Research Management Science	117	0,88
Multidisciplinary Sciences	265	2,00	Biochemical Research Methods	115	0,87
Clinical Neurology	241	1,82	Geriatrics Gerontology	113	0,85
Computer Science Hardware Architecture	241	1,82	Behavioral Sciences	111	0,84
Public Environmental Occupational Health	238	1,79	Environmental Sciences	100	0,75
Psychology Developmental	230	1,73	Gerontology	99	0,75
Psychology	224	1,69	Biotechnology Applied Microbiology	94	0,71
Psychiatry	221	1,66	Language Linguistics	87	0,66
Telecommunications	215	1,62	Business	85	0,64
Psychology Social	158	1,19	Radiology Nuclear Medicine Medical Imaging	85	0,64

Fuente: Elaboración a partir de resultados de Web of Science

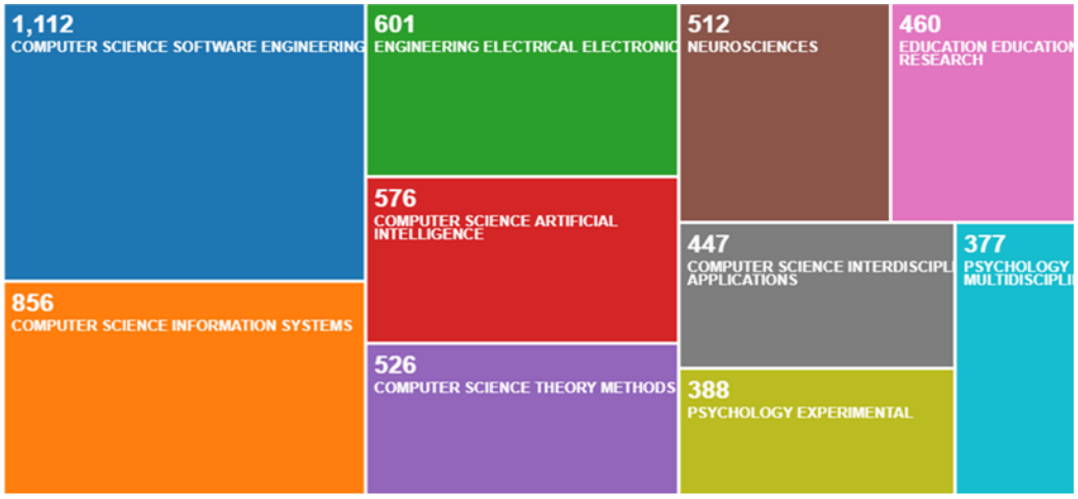


Figura 6. Áreas temáticas con mayor cantidad de trabajos asociados a Dominio Específico
Fuente: Web of Science

En concreto, se puede decir que el objetivo de los modelos de dominio específico es ser usados como medio para la definición de abstracciones de las entidades y

relaciones claves en un ambiente y a través de esto permitir la creación de especificaciones que puedan ser llevadas a desarrollos o construcciones de arquitecturas o sistemas determinados; en otras palabras, el modelo de dominio específico puede ser considerado como un lugar común donde varias implementaciones pueden ser formuladas y/o adaptadas teniendo en cuenta el conocimiento previo recolectado o a partir de la consolidación de este. Para esto se necesita tener un conjunto mínimo de los conceptos que describen los tipos de entidades involucradas en el dominio del problema, un conjunto de axiomas acerca de ese dominio, describir las relaciones entre los conceptos y ser independiente de las tecnologías o cualquier otro detalle más concreto de la implementación (De Almeida Neto & Castro, 2017).

1.4 Elementos esenciales de los modelos de dominio específico

Los modelos de dominio específico en el campo de las ciencias de la computación han sido utilizados principalmente para la construcción de lenguajes de modelado, sin embargo, no se ha explorado a profundidad los modelos de dominio específico en la comunidad de dicha área. Al hacer referencia en particular a los elementos que se deben tomar en cuenta para la construcción de un modelo de dominio específico, se destaca que este paradigma o enfoque busca la construcción de múltiples metamodelos que representen conceptos y abstracciones particulares del dominio (Lin et al., 2017). De allí que las características esenciales en los modelos de dominio específico para los lenguajes de modelado se refieren a unidades atómicas, modelos como tal y conexiones. Los átomos tienen unos atributos que se utilizan para registrar información de las abstracciones y se definen por medio de su tipo, clase, nombre, y el conjunto de dichos atributos. Los modelos se conforman por una serie de estos átomos y otra información que estos contienen incluye sus relaciones con otros elementos, es decir, las conexiones con los vecinos. Estos elementos reciben también una representación gráfica que permite distinguirlos al hacer una estructura o flujo de funcionamiento (Lin et al., 2017).

Se rescata también, que los modelos de dominio específico en lenguajes de modelado son importantes en las primeras etapas de los proyectos de software, puesto que describen conceptos y relaciones de los respectivos campos de aplicación en términos específicos del dominio. De allí proviene una de las dificultades en la creación y uso de modelos de dominio específico, hace referencia a la experiencia y conocimiento detallado del dominio que debe tener quien está a cargo de la construcción del modelo, la recopilación de este conocimiento requiere de mucho tiempo y constituye un proceso manual ya que

rara vez se cuenta con una herramienta que brinde el soporte a esta tarea (Agt-Rickauer et al., 2019).

A pesar de la dificultad en la creación de los modelos de dominio específico, esto se puede ver luego reflejado en una reducción de los tiempos de implementación, puesto que brindan características como idoneidad funcional, compatibilidad y reutilización (Arslan & Kardas, 2020). Este es el caso del trabajo citado de Arslan y Kardas, en este se implementó un lenguaje de modelado de domino específico (DSML4DT) por medio de un metamodelo desarrollado de acuerdo con las especificaciones de los dispositivos DT usados en microprocesadores, este metamodelo inicial incluye las relaciones de componentes denotados como conceptos y que se asocian a una notación gráfica, dichos componentes incluyen también propiedades que definen, entre otras cosas, la capacidad de relación con otros conceptos.

En concreto, se puede decir que los elementos esenciales del modelado de lenguajes de dominio específico son los ***conceptos, relaciones y restricciones***. Siendo una ventaja importante de este tipo de modelado el mapeo de las particularidades del dominio que posteriormente permite generar soluciones aceleradas y optimizadas, mejorando la productividad y efectividad. En lenguajes de dominio específico se habla de la generación de metamodelos que son especificaciones del lenguaje. Y en cuanto a la representación de estos elementos, se pueden expresar en notación gráfica, textual o mixta según el contexto del dominio (Castellanos et al., 2021).

En cuanto a la aplicación del enfoque de dominio específico para problemas de minería de datos, es posible identificar que puede ser útil contar con elementos particulares del dominio representados en conocimiento para la intervención a nivel de fases como la preparación de los datos. Desde este punto de vista, el enfoque de dominio específico puede ser valioso para extraer información de los datos, incluso si estos son ruidosos, y esta información se puede representar bajo definiciones y teoremas o reglas que permitan preprocesar los datos, rectificando discrepancias y recuperando patrones útiles para tareas como la reconstrucción de series con faltantes o inconsistencias dentro de fuentes heterogéneas (Liu et al., 2019). En (Liu et al., 2019) usaron dominio específico para crear definiciones, lemas y teoremas con los que preprocesaron conjuntos de datos de viajes cortos en un sistema de transporte masivo, para luego aplicar técnicas de minería de datos y conseguir, por ejemplo, patrones de movilidad de los pasajeros. El principal logro para este caso fue usar el conocimiento del dominio para conseguir información enriquecida posterior a la fase de limpieza y preprocesamiento de los datos.

Como se aprecia, los elementos esenciales de un modelo de dominio específico pueden variar según el enfoque y la metodología utilizada, pero a grandes rasgos, los elementos fundamentales que deben incluirse son:

- ***Conceptos del dominio:*** Son las entidades y relaciones que forman parte del problema a resolver. Por ejemplo, si se está modelando un sistema de minería de datos educativos, los conceptos del dominio podrían incluir datos de instituciones educativas, profesores, estudiantes, directivos, matrículas, ingresos, calificaciones, interacciones en una plataforma de educación virtual, etc.

- ***Reglas del dominio:*** Son las reglas y restricciones que se aplican en el dominio. Por ejemplo, en el caso de un modelo para minería de datos educativos, podría haber reglas que establezcan las escalas de aprobación y reprobación de asignaturas, contexto sociodemográfico de una determinada institución educativa, o políticas que definan qué se enseña en cada nivel del sistema educativo.

- ***Comportamiento del sistema:*** Especifica cómo se lleva a cabo la generación, almacenamiento y procesamiento de los datos y cómo se realizan las acciones en el sistema o sistemas que involucran a los conceptos. Por ejemplo, cómo se matricula un nuevo estudiante, cómo se registra el cambio de institución educativa, cómo se reportan las calificaciones de los estudiantes, etc.

- ***Aspectos tecnológicos:*** Se refiere a los detalles técnicos relacionados con la implementación del modelo. Esto podría incluir detalles sobre el lenguaje de programación utilizado, la plataforma de ejecución, los patrones de diseño utilizados, etc.

A partir de lo anterior, se puede aportar que, para la construcción de un modelo de dominio específico para minería de datos educativos, los elementos esenciales pueden incluir:

- ***Datos educativos:*** Son los datos que se utilizan como fuente para el análisis y la toma de decisiones. En este caso, los datos educativos incluirían información sobre los estudiantes, sus notas, asistencia, historial o trayectoria académica, comportamiento, interacciones, entre otras.

- ***Problemáticas de interés:*** Son las variables que se quieren analizar en el contexto educativo. Por ejemplo, la tasa de deserción estudiantil, el rendimiento académico, la eficacia de las políticas de intervención, entre otras.

- ***Métodos de análisis de datos:*** En este caso se refiere a las técnicas de minería de datos que se utilizarán para analizar los datos educativos. Estas técnicas pueden incluir la regresión, el clustering o agrupación, la clasificación y visualización.

- ***Indicadores y métricas:*** Son las medidas que se utilizan para evaluar la eficacia de las intervenciones educativas o de los modelos de minería logrados.

- ***Contexto educativo:*** Es importante considerar el contexto en el que se desarrolla el modelo de dominio específico. Por ejemplo, el nivel educativo de los estudiantes, el tipo de escuela, el entorno socioeconómico de los estudiantes, entre otros.

- ***Tecnologías y herramientas:*** Tecnologías y herramientas de software que se utilizarán para implementar y aplicar el modelo de dominio específico para minería de datos educativos. Esto puede incluir herramientas de visualización de datos, plataformas de análisis de datos, lenguajes de programación y librerías, entre otros.

Resumen del capítulo

En el presente capítulo se relató el proceso de revisión teórico-práctica de enfoques de dominio específico, los principales aspectos asociados, conceptos, ventajas, desventajas y aplicaciones a diferentes campos de estudio. La representación del dominio es importante en un modelo de dominio específico. La representación adecuada del dominio permite definir los elementos esenciales, conceptos, objetos y relaciones relevantes del dominio de aplicación de manera precisa y clara, lo que a su vez facilita la comprensión y el análisis de los datos y el conocimiento que se extrae de ellos.

En la minería de datos educativos, la representación del dominio puede ayudar a identificar las variables e indicadores de interés, a seleccionar las técnicas de minería de datos adecuadas para el análisis, a interpretar los resultados del análisis y a evaluar la eficacia de las intervenciones educativas. Además, la

representación del dominio puede ayudar a los expertos en el dominio, como los profesores y directivos educativos, a comprender mejor los datos y a tomar decisiones informadas basadas en los resultados del análisis.

En resumen, la representación del dominio es esencial para un modelo de dominio específico efectivo en la minería de datos educativos, ya que permite definir y comprender los conceptos y relaciones relevantes del dominio, lo que a su vez facilita la interpretación y aplicación de los resultados del análisis. Este capítulo corresponde al cumplimiento del primer objetivo específico, la conceptualización de los elementos esenciales que se toman en cuenta para el desarrollo del modelo de dominio específico para minería de datos educativos.

CAPÍTULO 2 – Minería de datos

La minería de datos se encuentra dentro de las técnicas de análisis de datos más utilizadas en las últimas décadas. En esta tesis doctoral se aborda la construcción de un modelo de dominio específico para minería de datos educativos, por lo cual, en este capítulo se busca dar una contextualización sobre técnicas de análisis de datos y en particular sobre minería de datos.

2.1 Generalidades

Con el aumento en la generación de datos y la capacidad de almacenamiento y procesamiento de los sistemas de cómputo, la comunidad científica se ha visto en la necesidad y en la tarea de desarrollar técnicas de análisis de datos, las cuales son una combinación de métodos matemáticos, estadísticos y computacionales para abordar el procesamiento de datos con diferentes fines.

De acuerdo con lo anterior, se encuentra una oportunidad incalculable en estos grandes cúmulos de datos que pueden ser usados como punto de partida para una variedad de estudios en diferentes campos y áreas de trabajo investigativo. Las técnicas de análisis de datos acompañan a los investigadores en la tarea de entender e interpretar lo que está sucediendo en el mundo real, siendo una labor difícil, pues los datos acompañan la complejidad de los fenómenos físicos y humanos que se analizan. Así mismo, las técnicas de análisis comprenden una serie de pasos que implican no solo la aplicación de un algoritmo a través de alguna herramienta computacional, sino que también incluyen el procesamiento previo o preparación de los datos, los diferentes objetivos que atienden los estudios exigen diferentes técnicas y cada técnica implica una parametrización de acuerdo con los datos que va a recibir.

Las técnicas y algoritmos de análisis permiten extraer conclusiones e información útil, principalmente con el ánimo de acompañar la toma de decisiones, probar/validar hipótesis e indagar en general de algún aspecto. Dependiendo del dominio de datos, las técnicas y herramientas pueden variar y ser adaptadas. A continuación, se parte de los modelos para minería de datos, para luego introducir el concepto de minería de datos como tal y finalmente particularizar en EDM y algunas herramientas asociadas a esta.

2.2 Sobre modelos para minería de datos

Tres de las metodologías más populares para llevar a cabo minería de datos son KDD (Knowledge Discovery in Databases), CRISP-DM (Cross-Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model y Access). KDD fue un término acuñado en 1989 y hace referencia al proceso de encontrar conocimiento en los datos, en 1996, Fayyad consideró la minería de datos como parte de las fases del proceso de KDD tomando en cuenta que ésta principalmente se centra en los patrones y la extracción de estos a partir de los datos (U. M. Fayyad et al., 1996). Con lo anterior se puede definir KDD como un modelo iterativo e interactivo en el cual se rescata conocimiento en los datos y se enfatiza en el alto nivel de especificidad de los métodos de minería de datos.

En la literatura (Shafique & Qaiser, 2014) se puede encontrar que en los procesos de KDD se siguen nueve pasos principales: (1) desarrollo y entendimiento del dominio de aplicación, (2) creación/selección de un conjunto de datos objetivo, (3) limpieza y preprocesamiento, (4) transformación, (5) selección de la técnica de minería de datos adecuada, (6) selección del algoritmo de minería de datos adecuado, (7) aplicación del algoritmo, (8) interpretación de los patrones/modelo y finalmente, (9) uso del conocimiento descubierto. En algunos casos estos pasos se resumen en cinco, los (5), (6) y (7) se resumen en minería de datos y el (1) y (9) se omiten como obvios.

Por su parte, CRISP-DM fue concebido en 1999 por Daimler Chrysler, SPSS y NCR (Chapman et al., 1999). Ofrece un marco de trabajo y una guía para minería de datos que consiste en seis fases o estados bien estructurados: (1) Entendimiento del negocio, (2) entendimiento de los datos, (3) preparación de datos, (4) modelamiento, (5) evaluación y (6) despliegue. El modelo de proceso SEMMA fue desarrollado por el Instituto SAS, tiene el objetivo de permitir que los proyectos de minería de datos sean entendidos, organizados, desarrollados y mantenidos (Matignon, 2007). Enfocado a soluciones empresariales, está conformado por un ciclo de cinco estados o pasos: (1) muestra, (2) exploración, (3) modificación, (4) modelo y (5) acceso.

Para crear puntos de referencia entre estos tres modelos para procesos de minería de datos, en la Tabla 4 se presenta un resumen y comparativo.

Tabla 4. Comparativo KDD, CRISP-DM y SEMMA

Modelo de proceso DM	**KDD**	**CRISP-DM**	**SEMMA**
No. Pasos	9	6	5
Nombre de los pasos/fases/estados	Desarrollo y entendimiento del dominio de aplicación	Entendimiento del negocio	----------
	Creación/selección de un conjunto de datos objetivo	Entendimiento de los datos	Muestra
	Limpieza y preprocesamiento	Preparación de datos	Exploración
	Transformación		Modificación
	Selección de la técnica de minería de datos adecuada	Modelado	Modelo
	Selección del algoritmo de minería de datos adecuado		
	Aplicación del algoritmo		
	Interpretación de los patrones/modelo y finalmente	Evaluación	Evaluación
	Uso del conocimiento descubierto	Despliegue	----------

Fuente: Adaptado de (Shafique & Qaiser, 2014).

2.3 Sobre minería de datos

Según (López, 2007), "*la minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos*". La minería de datos es un enfoque de análisis de datos que surge para dar respuesta a la disponibilidad, cada vez mayor, de datos almacenados bajo conjuntos estructurados en bases de datos o datos semiestructurados o sin estructura, pero que crecen cada día más.

Como se ha indicado, la extracción de patrones, tendencias y modelos para entender y describir mejor los datos es la función de la minería de datos. Sin embargo, para llegar a ella, se llevó a cabo un proceso de evolución de los sistemas de información y de la forma de utilización de los datos. Es decir, antes de llegar a la minería de datos se fueron adoptando algunos enfoques y tecnologías que permitieron llegar a los análisis actuales, esto de la mano del crecimiento constante en la cantidad de datos producidos y almacenados. Teniendo en cuenta, desde un punto de vista organizacional, la evolución de los enfoques se dio como se aprecia en la Figura 7.

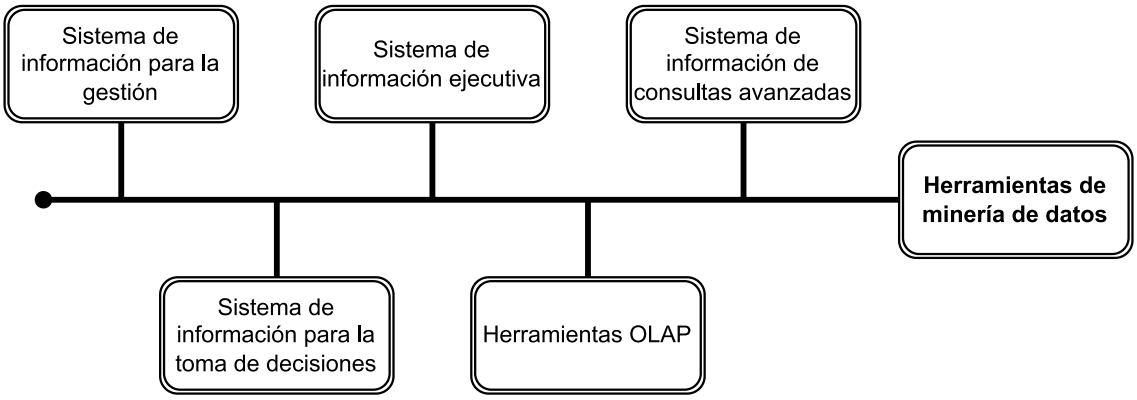


Figura 7. Evolución de los sistemas de información a la minería de datos
Fuente: elaboración propia a partir de (López, 2007)

Se inició con la conformación de sistemas de información para la gestión, los cuales aparecieron en la medida que se realizaba la informatización de las organizaciones y con el fin de atender a las necesidades básicas de análisis. Posteriormente se empieza a requerir un apoyo mayor, y surgen los sistemas de información para la toma de decisiones, los cuales aún coexisten con las demás tecnologías. Como una especialización dirigida a los ejecutivos de la organización surgen los sistemas de información ejecutiva que presentan información del estado de las actividades de gestión. Las herramientas OLAP son más genéricas, no solo están dirigidas a las organizaciones, y permiten hacer agregaciones y combinaciones de datos más complejas, funcionan sobre sistemas de información y son de gran ayuda porque permiten visualizaciones avanzadas de los datos. Por su parte, los sistemas de información de consultas avanzadas suelen estar asociados a sistemas relacionales y dan resultados en representación tabular. Finalmente llega la minería de datos, que permite analizar los datos y los otros sistemas que se han citado permiten el acceso a la información, convirtiéndose en elementos de apoyo a esta (López, 2007).

Es así como el término minería de datos, que viene del inglés ***data mining***, engloba una serie de técnicas de análisis de datos que combinan principios de estadística e inteligencia artificial. En la Figura 8 se muestra la ubicación de la minería de datos dentro de un proceso de KDD y a la vez, se presentan algunas de las bases a partir de las cuales se crean y constituyen estas técnicas.

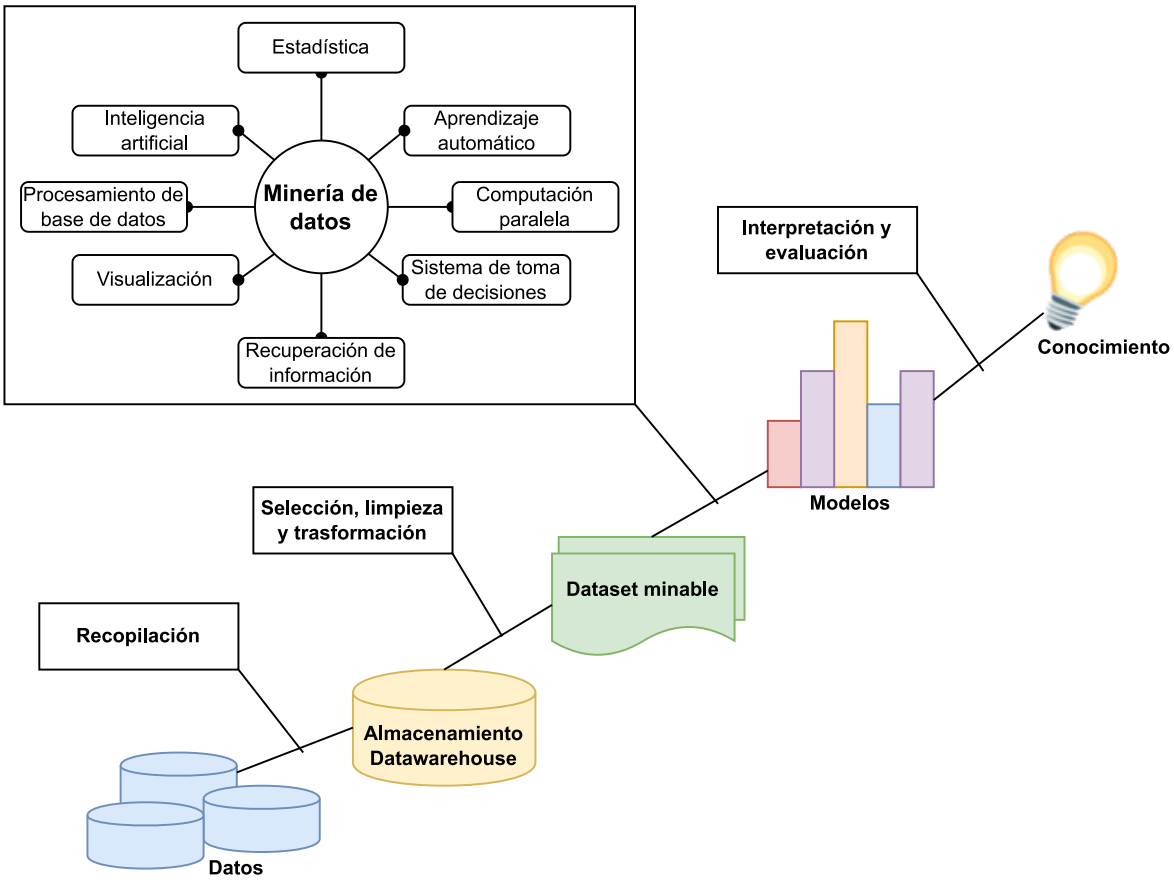


Figura 8. Minería de datos dentro del proceso de KDD
Fuente: elaboración propia

La minería de datos ostenta amplia aplicabilidad en campos de estudio, en general se podría decir que desde que se cuente con datos, la minería puede ser aplicada; sin embargo, no se debe olvidar que cada objetivo exige diferentes técnicas. A continuación, según Oliveira & Da Silva (2009) y Riquelme, Ruiz, & Gilbert (2006) se presentan algunas áreas en las cuales la minería de datos ha sido aplicada de forma satisfactoria:

- Retención de clientes: se identifican los perfiles para determinados productos
- Bancos: identificación de patrones para apoyar la gestión de relaciones con los clientes
- Tarjetas de crédito: identificación de segmentos de mercado, patrones de rotación
- Cobranza: detección de fraudes
- Telemarketing: facilidad de acceso a información del cliente
- Electoral: identificación de perfiles de posibles votantes
- Medicina: refinamiento de diagnósticos más precisos
- Seguridad: detección de actividades terroristas y criminales

- Investigaciones biométricas
- Toma de decisiones: filtro de información relevante para fortalecer bases como indicadores de probabilidad
- Educación: estudio de los perfiles estudiantiles y comportamientos en los procesos académicos
- Agencias de viaje: aumento en el volumen de ventas direccionando paquetes a clientes según perfiles
- Farmacia: efectividad de los tratamientos
- Astronomía: identificación de nuevas estrellas y galaxias
- Agricultura: identificación de áreas para uso en diferentes cultivos
- Ciencias ambientales: identificación de modelos de funcionamiento de ecosistemas naturales
- Ciencias sociales: estudio de flujos de la opinión pública, planificación urbana

Cabe resaltar que no son estas todas las áreas de aplicabilidad, es un ejemplo de algunas entre muchas otras. Esta amplia aplicabilidad puede verse explicada en que la minería de datos combina o acoge un conjunto de técnicas y herramientas que dan soporte al análisis de diferentes tipos de datos. Existen varias clasificaciones para las técnicas de minería de datos, pero en general los autores reconocen dos tipologías principales, las técnicas de minería de datos descriptivas y las técnicas de minería de datos predictivas. Para efectos de este trabajo se va a traer a colación la clasificación presentada por Pérez Marqués (2014), quien indica: "*Inicialmente las técnicas de minería de datos pueden clasificarse en técnicas de modelado originado por la teoría (en las que las variables pueden clasificarse en dependientes e independientes), técnicas de modelado originado por los datos (en las que todas las variables tienen inicialmente el mismo valor) y técnicas auxiliares*".

Ampliando la clasificación, las técnicas de modelado originado para la teoría se usan para construir modelos cuando se cuenta con un conocimiento teórico previo, estas técnicas también son asociadas al aprendizaje supervisado y conocidas como técnicas predictivas. En este grupo se pueden encontrar los tipos de clasificación, regresión y asociación, el análisis de varianza y covarianza, el análisis discriminante y las series temporales. Para la construcción de estos modelos, se deben llevar a cabo una serie de fases: la identificación objetiva (reglas para identificar el modelo que mejor se ajuste a los datos), estimación (cálculo de los parámetros del modelo con ajuste a los datos estudiados), diagnóstico (contraste del modelo estimado frente a la realidad) y predicción (utilización del modelo que ha sido identificado, estimado y validado para predecir valores futuros en las variables dependientes) (Pérez Marqués, 2014).

En el caso de las técnicas de modelado originadas por los datos, asociadas al aprendizaje no supervisado y en las cuales no se asigna ningún valor predeterminado en los atributos o variables, por lo cual no se distingue entre variables dependientes e independientes, ni se supone una existencia previa de un modelo para los datos. Por lo tanto, los modelos se deben crear de forma automática a partir del reconocimiento de patrones. Para la construcción del modelo se debe utilizar el conocimiento obtenido antes y después de la minería de datos, estos modelos deben permitir el descubrimiento de relaciones complejas entre variables sin tener que hacer intervención externa (Pérez Marqués, 2014).

Y finalmente las técnicas auxiliares son herramientas más superficiales y limitadas, se incluyen en este grupo las técnicas de estadísticas descriptivas e informes, también de visualización. En la Figura 9 se presenta de forma gráfica la clasificación de las técnicas.

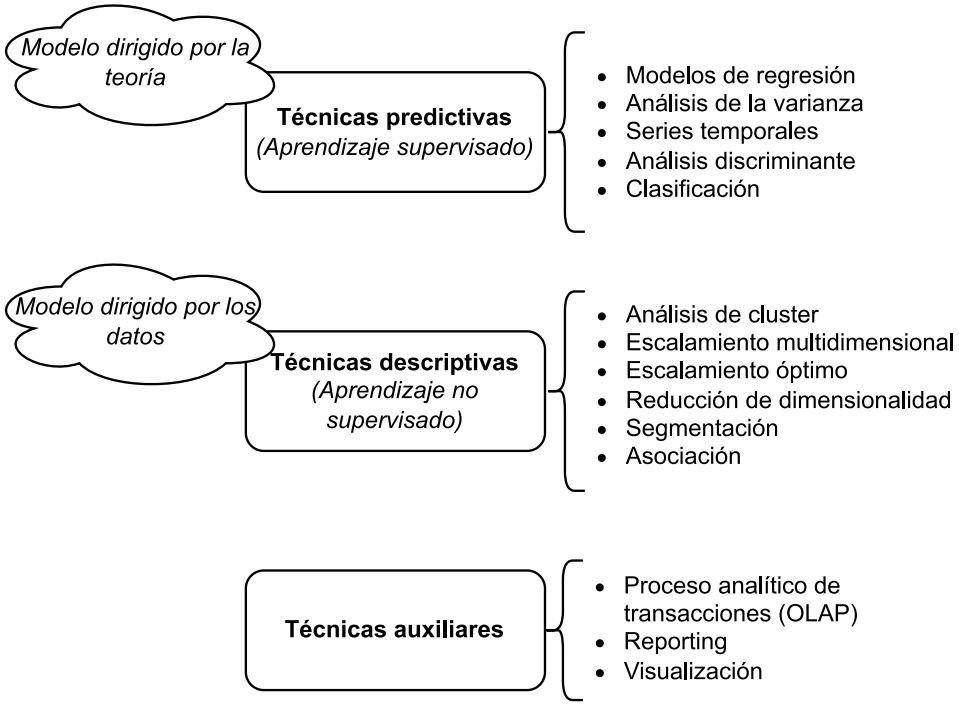


Figura 9. Clasificación de técnicas de minería de datos
Fuente: elaboración propia

Otra posibilidad de clasificación para la minería de datos es considerar las tareas que pueden ser llevadas a cabo mediante esta. De esta forma, según (Oliveira & Da Silva, 2009), se tienen las siguientes:

- Descripción: esta tarea es utilizada para realizar una exploración inicial de los datos y comprobar la influencia de algunas variables en los resultados; permite describir patrones y tendencias en los datos.

- Clasificación: es una de las más comunes, por medio de esta se puede encontrar a cuál clase pertenece un registro determinado. Para llevar a cabo esta tarea los datos deben tener la indicación de la clase a la que pertenece cada uno, en la tarea, el modelo analiza esto con el fin de aprender cómo se comportan los datos y poder clasificar un registro nuevo, por esto es por lo que pertenece a las tareas donde se utiliza el aprendizaje supervisado.

- Estimación: o regresión, es similar a una clasificación, pero esta es más usada cuando el registro es identificado como un valor numérico y no como una clase categórica. Con esta tarea se puede estimar el valor de una variable haciendo el análisis de los valores de las demás.

- Predicción: esta tarea es similar a las tareas de clasificación y estimación, sin embargo, esta es usada para descubrir el valor futuro de un atributo específico. Algunos métodos considerados como de clasificación o de regresión pueden ser usados para hacer predicción con las adaptaciones requeridas.

- Agrupamiento: también conocida como clustering, es una tarea usada para identificar y aproximar los registros similares. Un agrupamiento o clúster es una colección de registros similares entre sí pero que se diferencian de los otros registros agrupados en los demás clústeres. Esta tarea se diferencia de la clasificación por la ausencia de un atributo clase, es decir, no se requiere de registros previamente categorizados, por lo cual se reconoce como aprendizaje no supervisado. Normalmente esta tarea es combinada con otras tareas y son usadas en la fase de preparación de los datos.

- Asociación: la tarea de asociación consiste en identificar cuáles atributos están relacionados, se presentan en forma de atributo *X* entonces atributo *Y*, los buenos resultados que presenta esta tarea la hacen ser una de las más conocidas en la minería de datos para casos como el análisis de la cesta de compras.

2.3 Sobre minería de datos educativos

Uno de los campos de aplicación de la minería de datos es la educación, de allí surge la minería de datos educativos o más conocida como EDM por sus siglas en inglés de Educational Data Mining. La EDM emerge como un paradigma que se orienta al diseño de modelos, tareas, métodos y algoritmos que irán a ser utilizados en la exploración de datos educativos (Peña-Ayala, 2014). A continuación, se profundizará en el concepto y objetivos de EDM.

2.3.1 Qué es la EDM?

Como se indicó, la EDM se caracteriza por ser una aplicación de la minería de datos para un tipo de datos particular, los datos que provienen de ambientes educativos. El fin principal de la EDM es entender cómo se da el proceso de aprendizaje, cómo los estudiantes aprenden mejor y en general explicar los fenómenos educativos (Romero & Ventura, 2013).

Se puede decir que la EDM aparece de la combinación de tres áreas principales: las ciencias de la computación, la educación y la estadística; así mismo, al interceptar estas áreas sobresalen tres enfoques estrechamente relacionados con la EDM: la educación basada en computador, la minería de datos y aprendizaje máquina y las analíticas de aprendizaje (ver Figura 10). Siendo de estos últimos tres enfoques, las analíticas de aprendizaje, el más cercano a la EDM (Romero & Ventura, 2020).

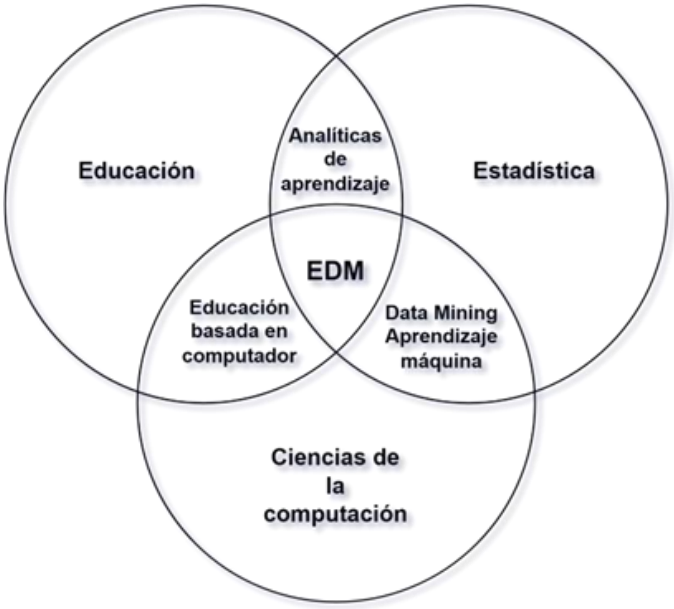


Figura 10. Principales enfoques involucrados en la EDM
Fuente: Adaptado de (Romero & Ventura, 2020).

A pesar de las múltiples similitudes entre las analíticas de aprendizaje (LA por sus siglas en inglés Learning Analytics) y la minería de datos educativos, se pueden

reconocer diferencias en algunos aspectos clave. Estas diferencias son presentadas en la Tabla 5.

Tabla 5. Diferencias entre EDM y LA

Aspecto	EDM	LA
Técnicas	Clasificación, clustering, modelos bayesianos, minería de relaciones y modelos de descubrimiento	Estadísticas, visualización, análisis de sentimientos, análisis del discurso, análisis de influencias, análisis de concepto, modelos de toma de sentido
Orígenes	Software educativo, modelado del estudiante, predicción de los resultados de curso	Semántica web, currículo inteligente, intervenciones sistémicas.
Énfasis	Descripción y comparación de las técnicas de DM usadas	Descripción de los datos y resultados
Tipo de descubrimiento	Aprovechar el conocimiento humano es clave, se usa descubrimiento automatizado como acompañamiento.	Aprovechar el descubrimiento automatizado es clave, se usa el conocimiento humano como acompañamiento.

Fuente: Construido a partir de (Romero & Ventura, 2013)

La comunidad científica en EDM se fue constituyendo en los últimos 20 años. En el año 2008 se llevó a cabo la primera conferencia de EDM en Canadá y se ha seguido desarrollando año a año en diferentes países. La conferencia EDM surge después de haber tenido Workshop en el tema de 2005 a 2008, esta conferencia es organizada por el grupo de trabajo internacional en EDM (International Working Group on Educational Data Mining) (Romero & Ventura, 2020).

El proceso que se sigue para la minería de datos educativos es semejante al desarrollado en la minería de datos general. En la Figura 11 se muestra un esquema general, se destaca el inicio en un ambiente educativo y la importancia de finalizar con la interpretación y evaluación, puesto que de ella se puede desprender la necesidad de volver en la formación de hipótesis o cualquier otra de las etapas previas.

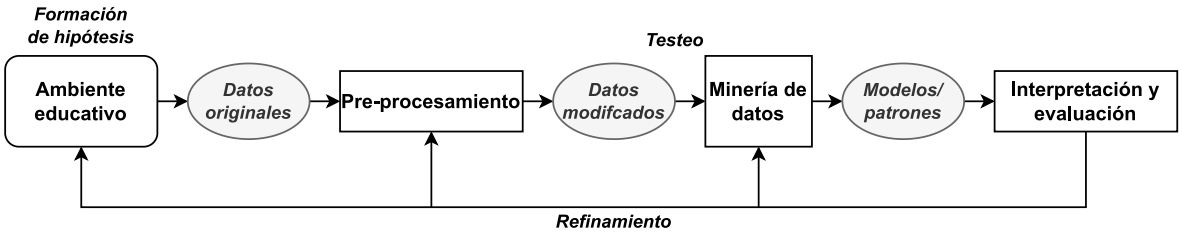


Figura 11. Proceso general de la EDM
Fuente: Adaptado de (Romero & Ventura, 2013)

2.3.2 Objetivos de la EDM

Se destacan entre los objetivos de la minería de datos educativos los siguientes:

- Hacer el análisis de rendimiento y comportamiento de cada estudiante.
- Analizar procesos de aprendizaje en varios ambientes y con diferentes métodos de minería de datos.
- Evaluar la calidad de las instituciones educativas.
- Generar patrones de acceso y permanencia en el sistema educativo.
- Proveer diferentes medios para hacer el análisis de logs en ambientes educativos virtuales.

Los objetivos de la EDM pueden estar, además, orientados por los diferentes actores (Romero & Ventura, 2007), de esta forma:

- Orientado a los estudiantes. Para recomendar actividades, recursos, tareas, para mejorar su proceso de aprendizaje.
- Orientado a los educadores. Para darles retroalimentación de su trabajo, evaluar la estructura de los cursos, contenidos y efectividad en el proceso de aprendizaje.
- Orientada a los administrativos y responsables académicos. Para dar parámetros para mejorar la efectividad de los servicios, mejorar la organización institucional y de los programas.

2.4 Herramientas de Minería de Datos Educativos

Hablar de las herramientas de minería de datos educativos no es una tarea sencilla, puesto que existe una cantidad considerable y el dilema inicia en la definición de qué puede ser considerado una herramienta propia de EDM. Si se acude a las revisiones de literatura sobre EDM, se encuentran diversas clasificaciones adoptadas por los autores de acuerdo con sus criterios. En este apartado se muestran algunas de las herramientas de EDM teniendo en cuenta tres revisiones de literatura, que se han considerado relevantes para el objeto de esta tesis.

2.4.1 Herramientas EDM según Jindal & Borah (2013)

En Jindal & Borah (2013) hacen una revisión de las herramientas de EDM de acuerdo con un criterio: las más usadas. Los autores indican que el rápido crecimiento de la EDM ha impulsado iniciativas por crear herramientas para el campo particular, sin embargo, siguen siendo más usadas las herramientas genéricas como Weka o RapidMiner. En la Tabla 6 se describen.

Tabla 6. Herramientas para EDM según Jindal & Borah

Nombre/ desarrollador	Recurso (abierta/libre /Comercial)	Función/ Características	Técnicas/ Algoritmos	Ambiente/ SO
Intelligent Miner (IBM)	Comercial	Proporciona una estrecha integración con el sistema de base de datos relacional DB2 de IBB, Algoritmo de escalabilidad de minería	Asociación, Clasificación, Regresión, Predicción, Modelado, Desviación, Clustering, Análisis de Patrón secuencial	Windows, Solaris, Linux
MSSQL Server 2005 (Microsoft)	Comercial	Proporciona funciones de DM tanto en el sistema de base de datos relacional como en el entorno de Data Warehouse (DWH)	Integra los algoritmos desarrollados por terceros proveedores y usuarios de aplicaciones	Windows, Linux
MineSet (SGI)	Comercial	Proporciona herramientas gráficas robustas como visualización de reglas, visualización de árboles, visualización de mapas y visualización de dispersión	Asociación, Clasificación, Estadísticas avanzadas y herramientas de visualización	Windows, Linux
Oracle Data Mining (Oracle Corporation)	Comercial	Proporciona una infraestructura DWH integrada para el análisis de datos multidimensionales.	Asociación, Clasificación, Predicción, Regresión, Clustering, Análisis secuencial Búsqueda de similitud	Windows, Mac, Linux
SPSS Clementine (IBM)	Comercial	Proporciona un entorno de desarrollo de minería de datos integrado para usuarios finales y desarrolladores.	Asociación, Clustering, Clasificación, Predicción, Herramientas de visualización	Windows, Solaris, Linux
Enterprise Miner (SAS Institute)	Comercial	Proporciona una variedad de herramientas de análisis estadístico.	Asociación, Clasificación, Regresión, Análisis de series de tiempo Análisis estadístico Clustering	Windows, Solaris, Linux
Insightful Miner (Insightful Incorporation)	Comercial	Proporciona una interfaz visual, que permite a los usuarios conectar componentes para crear programas de auto-documentación	Limpieza de datos, Clustering, Clasificación, Predicción, Análisis estadístico	Windows, Solaris, Linux
CART (Salford Systems)	Comercial	Proporciona división binaria y poda posterior para clasificación (árbol de decisión) y predicción (árboles de regresión).	Clasificación, Arboles de decisión y regresión	Windows, Linux
TreeNet(R) (Salford Systems)	Comercial	Proporciona una selección automática de predictores candidatos, capacidad para manejar datos sin preprocesamiento	Clasificación, regresión	Windows, Linux
Random Forests (Salford Systems)	Comercial	Proporciona altos niveles de precisión predictiva y un innovador conjunto de pantallas gráficas para revelar patrones	Clustering	Windows, Linux

Nombre/ desarrollador	Recurso (abierta/libre /Comercial)	Función/ Características	Técnicas/ Algoritmos	Ambiente/ SO
		inesperados en los datos.		
GeneSight (Inc. of El Segundo,CA)	Comercial	Proporciona al investigador la posibilidad de explorar grandes conjuntos de datos de múltiples grupos experimentales utilizando herramientas avanzadas de normalización, visualización y soporte de decisión estadística.	Visualización. K-means, Redes Neuronales, Clusteriing, Análisis de series de tiempo	Windows, Mac, Linux
PolyAnalyst (Megaputer Intelligence)	Comercial	Proporciona al tomador de decisiones el conocimiento derivado de grandes volúmenes de texto y datos estructurados.	Clustering, Clasificación, Predicción, Asociación	Windows
iData Analyzer (Microsoft)	Open /Free	Proporciona una plataforma para el entorno de aprendizaje visual.	Pre-procesamiento, ESX, Redes Neuronales, Generador de reglas, Generador de reportes	Windows, Linux, Solaris
See5 and C5.0 (RuleQuest)	Open/free	Proporciona análisis de árbol de decisión, versión comercial del algoritmo C4.5 DT	Árbol de decisión	Windows, Unix
TANAGRA (SPAD)	Open/free	Proporciona análisis de datos a principios de los 90. Software gratuito de minería de datos con fines académicos. Capacidad para diseñar la GUI y de adición de nuevo algoritmo	Análisis factorial de datos mixtos, Máquinas de soporte vectorial, Análisis factor principal	Windows, Linux, Mac OS, Solaris
SIPINA (Ricco Rakotomalala Lyon, France)	Open/free	Proporciona un entorno para algoritmos de aprendizaje supervisados, maneja datos continuos y discretos	C4.5, ID3	Windows, Linux
ORANGE (University of Ljubljana, Slovenia.)	Open/free	Proporciona una herramienta de visualización y análisis de datos de código abierto para principiantes y expertos.	Minería de texto y complementos de bioinformática	Windows, Linux
ALPHA MINER (E-Business Technology Institute)	Open/free	Proporciona la mejor relación costo-rendimiento para aplicaciones de minería de datos.	Funciones de minería de datos versátiles	Windows, Linux, Mac
WEKA (University of Waikato, New Zealand)	Open/free	Proporciona algoritmos de aprendizaje automático para tareas de minería de datos. Muy adecuado para desarrollar nuevos esquemas de aprendizaje automático.	Pre-procesamiento, Clasificación, Regresión, Clustering, Reglas de asociación, Visualización	Windows, Linux
Carrot	Open/free	Proporciona componentes listos para usar para obtener resultados de búsqueda desde varias fuentes	Clustering	Windows, Linux

Fuente: Jindal & Borah (2013)

Se destaca en esta revisión de herramientas que muchas son de carácter comercial, lo cual hace más difícil su acceso, sobre todo para fines investigativos donde no siempre se cuenta con un presupuesto para adquirir licencias de uso. También se resalta que todas pueden ser usadas en ambientes Windows y la mayoría en Linux. Finalmente, es claro que estas herramientas son de carácter

genérico para minería de datos, pueden ser usadas en EDM, pero en particular no van a presentar funcionalidades enfocadas en este tipo de dominio específico, esto a pesar de que los autores indican que se trata de una revisión de herramientas para EDM.

2.4.2 Herramientas EDM según Peña-Ayala (2014)

Peña-Ayala (2014) se concentra en la presentación de herramientas de EDM que surgen de un ámbito investigativo, en su mayoría trabajos menos conocidos o comerciales, pero más especializados o enfocados en aspectos específicos del dominio educativo. Es de resaltar que la mayoría de los trabajos se concentran en aplicación para EDM en LMS o en registros producidos en educación virtual o educación soportada en computador, que tienen que ver, por ejemplo, con las interacciones de los estudiantes en algún tipo de plataforma virtual. En la Tabla 7 se presenta un resumen de estas herramientas.

Tabla 7. Herramientas de EDM según Peña-Ayala (2014)

Autor	Tipo	Nombre	Propósito
Krüger, Merceron, and Wolf (2010)	Extracción	ExtractAndMap	Representa y despliega funcionalidades relacionadas con la extracción de datos de LMS
Pedraza-Pérez, Romero, and Ventura (2011)	Extracción	Java desktop Moodle mining	Ofrece un asistente para facilitar la extracción de datos de registro y la ejecución de procesos DM
Mostafavi, Barnes, and Croy (2011)	Soporte del aprendizaje	Logic Question Generator	Genera problemas de prueba que respaldan y satisfacen los requisitos conceptuales del curso de lógica deductiva
Rodrigo, Baker, McLaren, Jayme, and Dy (2012)	Ingeniería de características	Workbench	Busca y sugiere características apropiadas de entornos educativos como ITS
Johnson and Barnes (2010)	Visualización	InfoVis	Supervisa a los estudiantes en el aprendizaje para facilitar la supervisión del tutor.
Macfadyen and Sorenson (2010)	Visualización	Learner Interaction Monitoring System	Captura datos que demuestran la participación del alumno en línea con los materiales del curso
Maull, Saldivar, and Sumner (2010a)	Visualización	Curriculum Customization Service	Apoya la planificación curricular en línea y observa el comportamiento de los docentes provocado durante la planificación curricular
Rabbany, K., M., and O. R. (2011)	Visualización	Meerkat-ED	Adapta y visualiza instantáneas de los participantes en los foros de discusión, sus interacciones y el seguimiento del líder / estudiantes
García-Saiz and Zorrilla (2011)	Visualización	e-Learning Web Miner	Descubre los perfiles y modelos de comportamiento de los estudiantes sobre cómo navegan y trabajan en LMS

Autor	Tipo	Nombre	Propósito
Johnson, Eagle, Joseph, and Barnes (2011)	Visualización	EDM Vis	Facilita la visualización de información para explorar, navegar y comprender los registros de datos del alumno
Cohen and Nachmias (2011)	Soporte al análisis	Web-log based	Evalúa los procesos pedagógicos que ocurren en entornos LMS y las actitudes de los estudiantes
García, Romero, Ventura, and de Castro (2011)	Soporte al análisis	Continuous Improvement of e-Learning Courses Framework	Descubre las relaciones descubiertas en los datos de uso de los estudiantes a través de las reglas de recomendación If – Then; así como también comparte y califica las reglas obtenidas previamente por instructores en cursos similares con otros instructores y expertos en educación
Fritz (2011)	Soporte al análisis	Check My Activity	Ayuda a los estudiantes a comparar su propia actividad en Blackboard versus un resumen anónimo de sus compañeros del curso
Anjewierden, Gijlers, Saab, and DeHoog (2011)	Soporte al análisis	Brick	Explora patrones de secuencias de acción derivadas de un entorno de aprendizaje basado en simulaciones en el que los alumnos colaboran en parejas.
Moreno, González, Estévez, and Popescu (2011)	Soporte al análisis	SIENA	Logra una evaluación inteligente de la construcción social del conocimiento utilizando mapas conceptuales con nodos de aprendizaje multimedia.
Dyckhoff, Zielke, Chatti, and Schroeder (2011)	Soporte al análisis	eLAT	Permite a los maestros explorar y correlacionar el uso del contenido, las propiedades del usuario, el comportamiento del usuario y los resultados de la evaluación a través de indicadores gráficos
Devine, Hossain, Harvey, and Baur (2011)	Soporte al análisis	Data Miner for Outcomes based Education	Apoya el análisis de los tutores de los resultados de aprendizaje y los registros de rendimiento de sus estudiantes. Utiliza la selección de funciones supervisadas para producir patrones de aprendizaje e interpreta los resultados para proporcionar información para la optimización del curso.
Pechenizkiy, Trcka, Bra, and Toledo (2012)	Soporte al análisis	CurriM	Analiza los estudiantes y las perspectivas responsables de la educación sobre la minería del currículo y muestra los logros de un proyecto interesado en desarrollar el currículo

Fuente: Peña-Ayala (2014)

El autor agrupa las herramientas bajo tres grupos, de acuerdo con el tipo; en el primer grupo se describen las herramientas de extracción, las cuales facilitan

procesos como la búsqueda, representación y almacenamiento de datos desde sistemas educativos en un formato minable. El segundo grupo corresponde a herramientas de visualización, soportan el proceso de minería, el análisis e interpretación de resultados. El tercer grupo son las herramientas que dan soporte al análisis, tienen funciones orientadas a aspectos como: la evaluación del comportamiento y rendimiento de los estudiantes, apoyo a la solución de problemas, desarrollo de capacidades cognitivas, el monitoreo de la actividad de los estudiantes, entre otras. El inconveniente con algunas de estas herramientas y en general con las herramientas que son desarrolladas con un fin más investigativo que comercial, es que no siempre quedan disponibles para su uso o lo están por un tiempo y luego dejan de tener soporte y tienden a desaparecer.

2.4.3 Herramientas EDM según Slater et. al. (2016)

Los autores realizan la clasificación de las herramientas de EDM de acuerdo con la tarea a la que apuntan o apoyan dentro del proceso de minería de datos. Es decir, las herramientas se muestran con relación a la fase del proceso de minería en el que puedan ser usadas. Los autores indican que las fases fundamentales del proceso de EDM son manipulación y limpieza de datos; aplicación de algoritmos y análisis y visualizaciones. Sin embargo, además de las anteriores categorías, incluyen una categoría para aplicaciones especializadas para EDM y LA y otra para herramientas que integran todas las fases (Slater et al., 2016). En la Figura 12 se presenta un resumen de las herramientas.

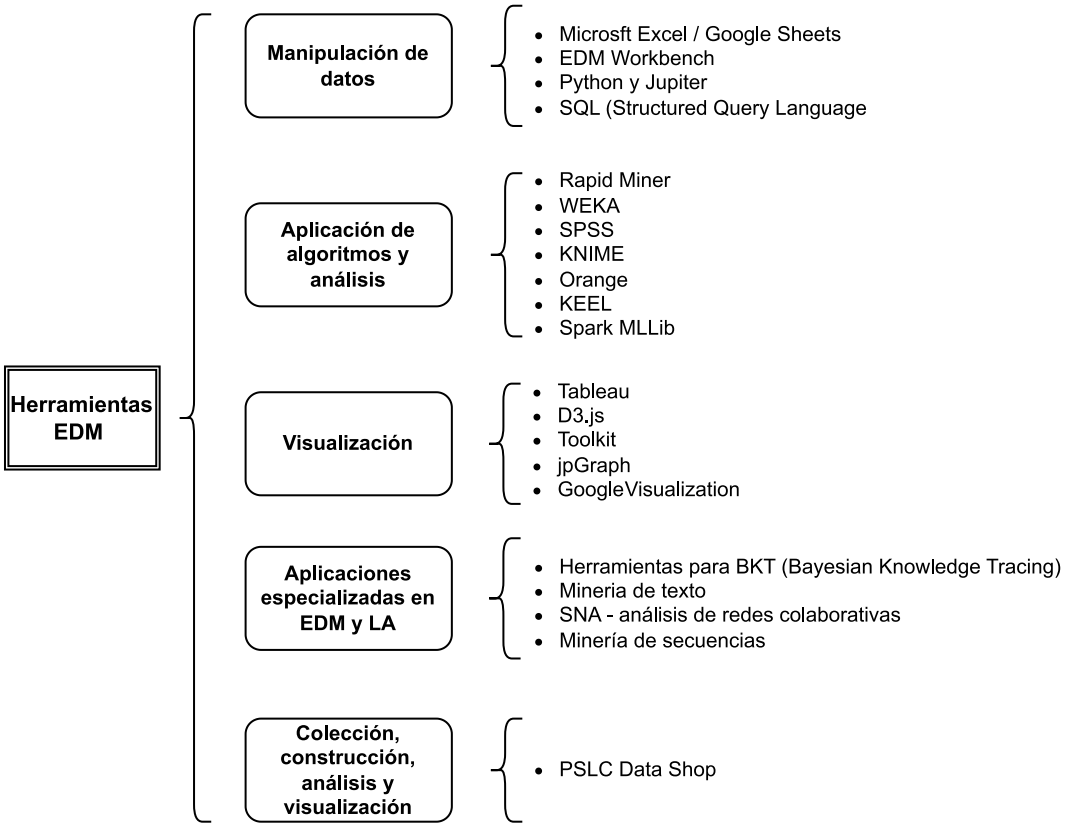


Figura 12. Herramientas de EDM según Slater et. al. (2016)
Fuente: construido a partir de (Slater et al., 2016)

Los autores abordan en su revisión tanto herramientas comerciales como herramientas que surgen de trabajos de investigación, buscando seleccionar las herramientas más usadas, accesibles y potentes.

2.4.4 Herramientas EDM según Romero & Ventura (2020)

Romero & Ventura (2020) hacen la actualización de su primera revisión respecto a EDM y presentan una serie de herramientas que combinan en una tabla junto con algunos datasets o fuentes de datos. Para efectos de este apartado, se adapta esta tabla dejando solo las herramientas, las fuentes de datos serán tratadas en el capítulo siguiente. De acuerdo con esto, en la Tabla 8 son presentadas dichas herramientas.

Tabla 8. Herramientas para EDM según Romero & Ventura (2020)

Nombre	Enlace	Descripción
DataShop	https://pslcdatashop.web.cmu.edu/	Proporciona un repositorio central para proteger y almacenar datos de investigación, y un conjunto de herramientas de análisis y generación de informes.

Nombre	Enlace	Descripción
GISMO	http://gismo.sourceforge.net/	Herramienta gráfica de monitoreo interactivo que proporciona a los instructores una visualización útil de las actividades de los estudiantes en los cursos en línea.
Inspire	https://moodle.org/plugins/tool_inspire	API de Moodle Analytics que proporciona un motor de análisis descriptivo y predictivo, implementando backends de aprendizaje automático.
LOCO-Analyst	http://jelenajovanovic.net/LOCO-Analyst/	Herramienta destinada a proporcionar a los profesores retroalimentación sobre los aspectos relevantes del proceso de aprendizaje que tiene lugar en un entorno de aprendizaje basado en la web.
Meerkat-ED	http://www.reirab.com/MeerkatED/index.html	Herramienta para analizar la actividad de los estudiantes en un curso ofrecido sobre herramientas de aprendizaje colaborativo asistidas por computadora.
MDM Tool	http://www.uco.es/kdis/research/software/	Marco para aplicar algunas técnicas de minería de datos en la versión Moodle 2.7.
Performance Plus	https://www.d2l.com/higher-education/products/performance/	Paquete para brindar potentes herramientas de análisis para ayudar a los administradores, educadores y estudiantes a ahorrar tiempo de calidad mientras maximizan el impacto y conducen al éxito.
SNAPP	https://web.archive.org/web/20120321212021/http://research.uow.edu.au/learningnetworks/seeing/snapp/index.html	Herramienta que permite a los usuarios visualizar la red de interacciones resultantes de publicaciones y respuestas en foros de discusión.
Solutionpath StREAM	https://www.solutionpath.co.uk/	Sistema en tiempo real que aprovecha los modelos predictivos para determinar todas las facetas de la participación de los estudiantes.

Fuente: Adaptado de Romero & Ventura (2020)

Si se realiza un contraste de las herramientas destacadas por las cuatro revisiones, se puede ver que algunas herramientas coinciden, por ejemplo, Weka, RapidMiner, Workbench, InfoVis, Orange y Datashop. En Mitrofanova et al., (2019) se considera que uno de los factores que contribuyeron al desarrollo de la EDM fue la mejora de las herramientas de software de análisis de datos, puesto que esto permitió involucrar a los especialistas y expertos en el campo educativo pero que no tienen mucha experiencia en programación. Sin embargo, de lo presentado en las revisiones y de la exploración de las herramientas mencionadas por los autores, no se evidencia ningún trabajo en el cual se permita una intervención directa de los expertos del campo educativo sobre la herramienta o los algoritmos.

De acuerdo con la revisión de los autores, se puede establecer que la mayoría de los artículos publicados en las revistas, así como en conferencias sobre EDM y Learning Analytics utilizan herramientas gratuitas (o que tienen una versión de uso académico) como RapidMiner, R, Weka, KEEL y SNAPP. Dichas herramientas tienen algoritmos de clasificación, agrupamiento, predicción, regresión, visualización, entre otros. Además, proporcionan soporte para el preprocesamiento de los datos que se utilizarán dentro del marco de estas técnicas y soporte para pruebas estadísticas de adecuación del modelo y visualización de datos (Mitrofanova et al., 2019). No obstante, y como se destaca en la última revisión presentada (apartado 2.4.4), en la mayoría de los casos, las herramientas que son realmente específicas se han diseñado para atender casos relacionados con datos provenientes de plataformas virtuales, principalmente LMS Moodle.

2.5 Comparativo de herramientas EDM

En este apartado se presenta en la Tabla 9 la comparación de algunos trabajos relacionados con el desarrollo de frameworks, modelos o herramientas para procesos de minería de datos educativos. Esta comparación fue realizada señalando aspectos de interés para el desarrollo del modelo propuesto en esta tesis. Para facilitar el manejo del espacio y comprensión de la tabla, se crearon las siguientes convenciones para designar los aspectos a comparar.

Técnicas de minería de datos que incluye/soporta:

1. Predicción/Clasificación
2. Agrupamiento/Clustering
3. Visualización
4. Otras

Tipos de datos educativos que incluye-soporta:

5. Resultados académicos (calificaciones)
6. Socioeconómicos/familiares
7. Interacción con Plataformas Virtuales (LMS)
8. Otros tipos de datos

Tabla 9. Comparativo herramientas/modelos de EDM

Modelo/ Herramientas	Autor(es)- Año/Propietario	Técnicas				Datos				Disponible al público			Múltiples fuentes		
		1	2	3	4	5	6	7	8	Si	No	NR	Si	No	NR
FlexDM https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.710.5148&rep=rep1&type=pdf	Kyle DeFreitas & Margaret Bernard, 2014	X	X					X				X	X		
TADA-Ed (**T**ool for **A**dvanced **D**ata **A**nalysis for **Ed**ucation) https://www.researchgate.net/publication/297863840_TADA-Ed_for_educational_data_mining	Agathe Merceron & Kalina Yacef, 2005	X	X					X			X			X	
GISMO https://www.researchgate.net/publication/297863840_TADA-Ed_for_educational_data_mining	RiccardoMazza & ChristianMilani, 2004			X				X		X				X	
eLAT http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1032.3892&rep=rep1&type=pdf#page=367	A.L. Dyckhoff, D. Zielke, M.A. Chatti, U. Schroeder Rwth, 2011			X				X		X			X		
PSLC Data Shop https://pslcdatashop.web.cmu.edu/	Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J., 2010			X	X	X		X	X	X			X		
MDM Tool https://dl.acm.org/doi/10.1002/cae.21782	Luna, J. M., Castro, C., & Romero, C., 2017	X	X					X		X				X	
Loco-Analisty http://jelenajovanovic.net/LOCO-Analyst/	Jovanović, J., Gašević, D., Brooks, C. A., Eap, T., Devedžić, V., Hatala, M., Richards, G., 2008				X			X		X					X
Meerkat-ED http://www.reirab.com/MeerkatED/index.html	Reihaneh Rabbany, Mansoureh Takaffoli, Osmar R. Zaiane, 2011			X	X			X		X					X
Moodle Learning Analytics API: IntelliBoard and LearnerScript https://moodle.com/functionality-with-moodle/learning-analytics-for-moodle/	Moodle, 2024			X	X	X		X		X					X
Open edX Insights https://github.com/openedx/edx-analytics-dashboard	Open edX community / Open edX Consortium			X				X	X	X				X	
SNAPP (Social Networks Adapting Pedagogical Practice) https://dl.acm.org/doi/abs/10.1145/2090116.2090144	Aneesha Bakharia, E. A. Heathcote, S. Dawson., 2009				X			X				X			X
KEEL 3.0 https://sci2s.ugr.es/keel/pdf/keel/articulo/triguero-et-al-KEEL3.pdf	Triguero I., González, S., et. al. (2017)	X	X	X					X	X					X
DM Toolkit http://sersc.org/journals/index.php/IJGDC/article/view/35181	Hembade, S., More, A., Mahajan, N., & Yelikar, B. (2020).			X	X			X				X			X

*NR - No Reporta

Fuente: Elaboración propia**.**

2.6 Comparativo de técnicas y algoritmos de DM para datos educativos

Así mismo, se realizó un comparativo de técnicas y algoritmos de minería de datos aplicados a datos educativos, para identificar las posibles ventajas y desventajas de estos para el dominio particular y con el fin, también, de establecer cuáles sería adecuado tomar para la validación del modelo propuesto de acuerdo con el alcance y datos del caso de estudio seleccionado.

Cabe aclarar, que a pesar de que se habla del enfoque de EDM como la particularización de la minería de datos aplicada a datos educativos, las herramientas analizadas y la literatura reporta el uso de las técnicas y algoritmos tradicionales de DM. Por ello, el objetivo de este comparativo fue establecer, de acuerdo con las principales problemáticas de interés en el campo y al alcance definido para la validación del modelo, las técnicas y algoritmos que, según lo reportado en la literatura, mejores resultados suelen obtener para el dominio.

A continuación, en la Tabla 10, se presenta una comparación de las técnicas de DM a partir de sus ventajas y desventajas y casos en los que funciona bien para problemáticas abordadas en la EDM.

Tabla 10. Técnicas de DM aplicadas a problemáticas contexto educativo

Técnica	Ventajas	Desventajas	Casos de éxito en EDM
Árboles de decisión	Fácil de entender e interpretar, capaz de manejar variables categóricas y numéricas, puede manejar grandes cantidades de datos, no requiere normalización de datos	Puede sobreajustar los datos, puede ser sensible a ruido y valores atípicos	Predecir el rendimiento académico de los estudiantes e identificar los factores que influyen en él
Regresión logística	Fácil de entender e interpretar, capaz de manejar variables categóricas y numéricas, no requiere normalización de datos	Puede no ser adecuado para datos no lineales, puede ser sensible a ruido y valores atípicos	Predecir el abandono escolar y la satisfacción del estudiante con el curso
Redes neuronales	Capaces de modelar relaciones complejas y no lineales entre variables, pueden manejar grandes cantidades de datos, pueden generalizar bien	Requieren normalización de datos, pueden ser difíciles de interpretar, pueden ser sensibles al sobreajuste	Identificación de estudiantes en riesgo de abandono, predicción del rendimiento académico

Técnica	Ventajas	Desventajas	Casos de éxito en EDM
Análisis de componentes principales	Capaz de reducir la dimensionalidad de los datos, puede identificar patrones y tendencias en los datos, puede manejar grandes cantidades de datos	Puede no ser adecuado para datos no lineales, no es bueno para identificar relaciones entre variables categóricas	Análisis de las interacciones entre los estudiantes y el contenido del curso en un entorno de aprendizaje en línea.
Clustering	Capaz de identificar patrones en los datos, puede manejar grandes cantidades de datos	Puede no ser adecuado para datos no lineales, la interpretación de los grupos puede ser subjetiva	Identificación de grupos de estudiantes con necesidades educativas similares.
XGBoost	Buen rendimiento en grandes conjuntos de datos, buena capacidad para manejar variables categóricas, fácil de paralelizar y escalar, buen manejo de datos desbalanceados	Requiere ajuste de hiperparámetros cuidadoso, requiere más recursos de computación que otros algoritmos, puede sufrir de sobreajuste	Predecir el rendimiento académico

Fuente: Elaboración propia

Ahora bien, en el caso concreto de las implementaciones (algoritmos) propias de estas técnicas, se hizo una búsqueda para establecer los más referenciados en la literatura y se encontró, que al combinar en una cadena de búsqueda, el nombre del algoritmo y EDM (*"algorithm"* and "educational data mining"), se obtienen los siguientes resultados, recopilados en las herramientas bibliográficas Science Direct, Scopus e IEEE Explore, mostrados en la Tabla 11, esto contribuye a ratificar que los árboles de decisión son una de las técnicas más usadas en este campo, así como también las redes neuronales.

Tabla 11. Algoritmos más referenciados en la literatura de EDM

ALGORITMO	SCIENCE DIRECT	SCOPUS	IEEE EXPLORE
CART	494	23	15
Random Forest	367	129	67
Naive Bayes	348	177	147
C4.5	230	52	82
Multilayer Perceptron	141	17	18
ID3	110	25	53
C5.0	35	13	15
JRip	25	8	5
BayesNet	15	2	2
Simple Perceptron	3	1	18

Fuente: Elaboración propia

Después de comparar diversas técnicas de minería de datos educativos, se puede concluir que no existe una técnica única que funcione perfectamente en todas las situaciones. La elección de la técnica adecuada, así como el algoritmo, depende en gran medida de los datos que se tienen y de los objetivos que se quieran lograr. En general, se ha encontrado que los algoritmos de árboles de decisión, y redes neuronales son los más utilizados en la predicción del rendimiento académico y la deserción escolar (Xiao et al., 2022).

Los algoritmos de ensamble, Random Forest, han demostrado ser útiles para mejorar la precisión de las predicciones. Además, las técnicas de clustering y asociación pueden ser útiles para identificar patrones y relaciones ocultas en los datos educativos. Es importante tener en cuenta que el éxito de la minería de datos educativos no solo depende de la elección de la técnica, sino también de la calidad y la cantidad de los datos utilizados, así como de la comprensión y el conocimiento del dominio en el que se está trabajando y de la calibración de los parámetros de dichas técnicas.

2.7 Tendencias y problemática en minería de datos educativos

Al hablar de tendencias en minería de datos educativos, se debe considerar inicialmente que el objetivo principal de este enfoque es analizar los datos de los procesos educativos mediados por computador. En general, los datos que se usan en el proceso de EDM, son producto de los sistemas de e-learning, si bien en algunos trabajos se analizan datos generados en los salones de clase y de herramientas de gestión y administración educativa, son muy pocos.

Desde el punto de vista de los actores a los cuales está enfocada la EDM, se encuentran tres tendencias (Romero & Ventura, 2010):

- EDM orientada a los estudiantes. Para recomendar actividades, recursos, tareas, mejorar su proceso de aprendizaje.
- EDM orientada a los profesores. Para darles retroalimentación de su trabajo, evaluar la estructura de los cursos, contenidos y efectividad en el proceso de aprendizaje.
- EDM orientada a los administrativos y directivos académicos. Para dar parámetros de mejora en la efectividad de los servicios, perfeccionar la organización institucional y de los programas.

La EDM como tal es una tendencia en el campo de las analíticas de datos. Las oportunidades para EDM son significativas, puesto que la educación es una de las

principales prioridades de la sociedad mundial, y como tal requiere de nuevos paradigmas para mejorar el alcance, la calidad, la eficiencia y los logros de los sistemas educativos. Los paradigmas pedagógicos exigen una educación centrada en el estudiante; pero también satisfacer las demandas individuales, grupales y comunitarias (Peña-Ayala, 2014).

En cuanto a las problemáticas que se identifican en la EDM, se destaca la necesidad de lograr un efectivo análisis de datos que provengan de ambientes educativos presenciales. Así mismo, se evidencia que los trabajos de EDM se concentran en los datos producidos desde la educación superior (universidades o instituciones de educación complementaria) con un número bajo de trabajos en educación primaria y secundaria.

Algunas problemáticas más específicas se han podido rescatar desde los trabajos y revisiones presentadas en este capítulo:

- El uso de herramientas de tipo genérico para la minería de datos educativos sigue siendo el patón más encontrado, este fenómeno ocurre igualmente en otros dominios de datos.
- Se requiere de herramientas de minería intuitivas, de fácil uso para profesores o actores del entorno educativo, pero no expertos en minería de datos, informática o programación.
- Falta de estandarización de métodos y datos.
- La integración de las herramientas de EDM con sistemas de e-learning continúa siendo una tarea difícil.
- Se requiere de técnicas de minería de datos específicas. Herramientas de minería más efectivas que integren conocimiento del dominio educativo con técnicas de minería de datos. Los algoritmos de minería tradicionales deben ajustarse para tener en cuenta el contexto educativo.
- La EDM tiene que lidiar con la falta de una teoría particular para fundamentar los elementos esenciales del enfoque.
- Se enfrenta también la falta de reconocimiento y valorización de las contribuciones que la EDM puede proporcionar para extender y mejorar los logros de los sistemas educativos.

Resumen del capítulo

En este capítulo se realizó una revisión, principalmente teórica, de las técnicas de análisis y en particular de la minería de datos educativos. Se inició definiendo la minería de datos y se contextualizó su papel dentro de metodologías como KDD. La mayor parte del capítulo se centró en la EDM y se presentaron sus objetivos,

herramientas, comparación de técnicas, tendencias y problemática. Este capítulo corresponde a la segunda parte de la conceptualización de los elementos esenciales que se toman en cuenta para el desarrollo del modelo de esta tesis, dando cumplimiento al objetivo específico 2.

Como aporte de esta caracterización realizada, encontramos que desde el punto de vista de las técnicas y algoritmos en EDM, no existe un diferencial marcado, es decir, no hay presencia de técnicas que sean solo para tratar este tipo de datos, por el contrario, se trabajan técnicas y algoritmos de DM, el diferencial se da en los datos y en la forma como se preparan estos.

CAPÍTULO 3 – Caracterización de las fuentes y datos educativos

En el contexto educativo se están generando datos desde diferentes acciones que conforman este ecosistema, no solo relacionados con los estudiantes y sus interacciones, sino también con las condiciones necesarias para que exista el desarrollo del aprendizaje. En este capítulo se hace una caracterización de los datos educativos y se muestran algunas fuentes donde se puede acceder a este tipo de datos. Además, se hace una revisión de la organización del sistema educativo colombiano y de algunos de los sistemas de información que lo conforman, ajustándose al alcance definido para esta tesis en términos del nivel educativo (educación básica y media).

3.1 Caracterización datos educativos

Se considera como dato educativo todo aquel que surge como resultado de un proceso educativo, que viene de un sistema o ambiente educativo. A continuación, se presenta, en la Figura 13, los tipos de ambientes y sistemas educativos (Romero & Ventura, 2020).

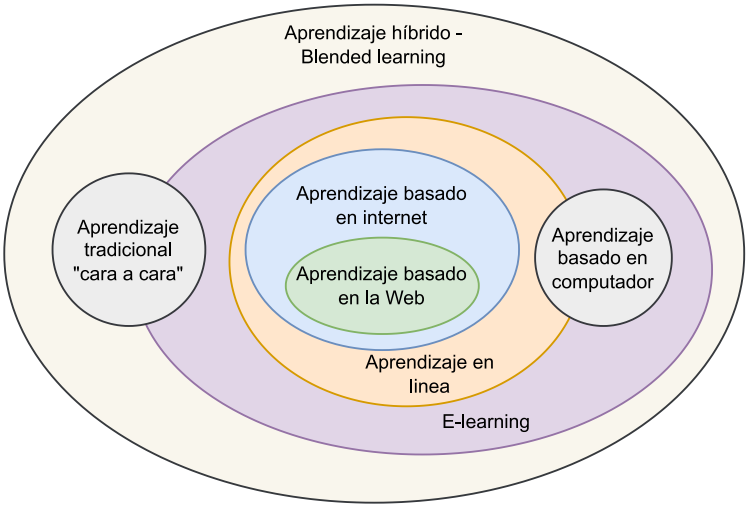


Figura 13. Tipos de ambientes y sistemas educativos
Fuente: Adaptado de (Romero & Ventura, 2020).

En la Figura 13 se representa como el aprendizaje combinado o híbrido, más conocido como Blended Learning, enmarca tanto el aprendizaje tradicional (face-to-face) como el aprendizaje soportado en computador. Sin embargo, el aprendizaje denominado como tradicional también se relaciona y hace uso del e-learning. Ahora bien, en cuanto a la granularidad y cantidad de datos que se pueden generar en los procesos educativos, en la Figura 14 son presentados.

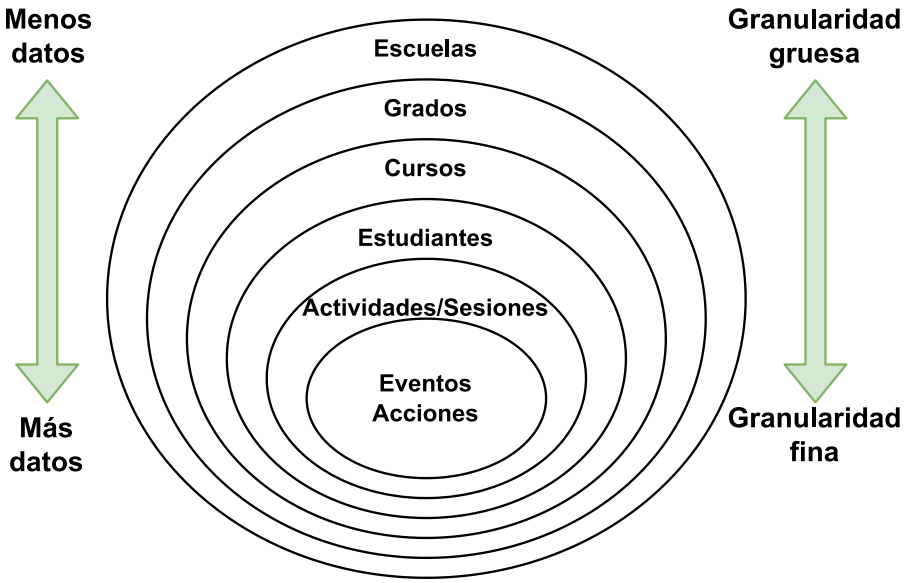


Figura 14. Diferentes niveles de granularidad y su relación con la cantidad de datos
Fuente: Adaptado de (Romero & Ventura, 2013)

Ya sea con datos de una granularidad fina o una granularidad gruesa, la EDM se presenta como una posibilidad para hacer los respectivos análisis en búsqueda de patrones, relaciones o reglas y la construcción de modelos. Sin embargo, la cantidad de datos es uno de los factores que influye como ventaja u oportunidad para la aplicación de muchas de las técnicas y algoritmos de minería de datos. Esto es una de las explicaciones a la presencia de un mayor número de trabajos relacionados con los ambientes educativos a distancia, puesto que, en estos, al tratarse de la generación de datos online y asociados a eventos, interacciones o acciones, el número es más representativo. Tratando de llevar esto a un ejemplo concreto, si se analiza la asistencia de estudiantes a un curso online, en el cual se pueden presentar múltiples interacciones en un mismo día, se ve una clara diferencia frente a la asistencia a una escuela, donde se tendría un único registro diario.

La literatura demuestra que los estudios en EDM y analíticas de aprendizaje se concentran en ambientes de educación virtuales. En Cechinel et al., (2020) se indica que, para América Latina, el 39,9% de los trabajos en Learning Analitycs tienen una aplicación en ambientes de aprendizaje virtual (VLE – Virtual Learning Enviromental). Los datos educativos típicamente obedecen a múltiples niveles jerárquicos y no son independientes de su naturaleza, luego los investigadores

deben ser muy cuidadosos al escoger el algoritmo y técnica de EDM a usar, de manera que los resultados obtenidos sean válidos y de confianza.

De acuerdo a lo anterior, la naturaleza de este tipo de datos permite identificar problemáticas que afectan los procesos de análisis (Baker & Inventado, 2014a), pero que a la vez admiten determinar aspectos propios del dominio, que posteriormente se puede utilizar en la particularización y representación del conocimiento para el modelado. A continuación, se detallan algunos de estos problemas o retos.

Calidad de los datos: Uno de los principales retos en el tratamiento de datos educativos es la calidad de los datos. Para que los datos sean útiles y confiables, es necesario que se recopilen y registren correctamente. Los errores en la recopilación o registro de datos pueden tener un impacto negativo en la toma de decisiones y en la implementación de políticas educativas.

Privacidad y seguridad de los datos: La protección de la privacidad y la seguridad de los datos es un desafío importante en el tratamiento de datos educativos. Es esencial garantizar que los datos sean tratados de manera responsable y que se respeten los derechos de privacidad de los estudiantes y docentes.

Integración de datos: En muchos casos, los datos educativos se recopilan y almacenan en diferentes sistemas y plataformas. Integrar estos datos en una sola fuente puede ser un desafío, pero es importante para tener una visión completa y comprensiva del sistema educativo.

Interpretación de datos: Los datos pueden ser complejos y difíciles de interpretar, especialmente para aquellos que no tienen experiencia en análisis de datos. La interpretación errónea de los datos puede conducir a decisiones incorrectas.

Actualización y mantenimiento de datos: Los datos educativos deben ser actualizados regularmente para ser útiles y relevantes. La falta de actualización y mantenimiento de los datos puede hacer que los datos sean obsoletos y, por lo tanto, menos útiles.

Adicionalmente, los datos educativos se diferencian de los datos de otras áreas de conocimiento en varios aspectos.

Naturaleza de los datos: Los datos educativos están relacionados con el sistema educativo y los procesos de enseñanza y aprendizaje. Por lo tanto, los datos educativos pueden incluir información sobre el número de estudiantes

matriculados, la tasa de graduación, el rendimiento académico, el número de docentes, entre otros.

Características de los datos: Los datos educativos a menudo se caracterizan por ser datos longitudinales, es decir, se recopilan a lo largo del tiempo y se utilizan para hacer seguimiento al progreso de los estudiantes y a la evolución del sistema educativo. Además, los datos educativos pueden ser tanto cualitativos como cuantitativos, lo que significa que pueden incluir información subjetiva, como la percepción de los estudiantes sobre la calidad de la enseñanza.

Uso de los datos: Los datos educativos se utilizan principalmente para mejorar el sistema educativo y para tomar decisiones en cuanto a políticas públicas en educación.

Impacto social: Los datos educativos tienen un impacto directo en la sociedad, ya que la educación es un derecho humano fundamental y un factor clave para el desarrollo social y económico. Por lo tanto, los datos educativos se utilizan para identificar problemas en el sistema educativo y para diseñar políticas y estrategias que permitan mejorar la calidad de la educación y reducir las desigualdades educativas.

3.2 Fuentes de datos educativos

En el apartado anterior se dejó sentado las bases para determinar los ambientes educativos donde se producen los datos. En este apartado se pretende mostrar algunas bases de datos concretas que pueden ser consultadas o accedidas como fuentes para hacer estudios de EDM. A nivel internacional, algunas entidades de reconocido prestigio ofrecen bases de datos en este campo de forma abierta. En la Tabla 12 se presentan algunos enlaces de acceso interesante junto con una corta descripción.

Tabla 12. Enlaces a datasets o fuentes de datos educativos

Nombre	Dirección Web	Descripción
UIS UNESCO	http://www.uis.unesco.org/education	El Instituto de Estadística de la UNESCO (UIS) es la fuente oficial y confiable de datos comparables a nivel internacional sobre educación, ciencia, cultura y comunicación.
BD de educación del Banco Mundial	http://datatopics.worldbank.org/education/	La BD de resultados de aprendizaje del Banco Mundial contiene datos sobre los niveles de aprendizaje de los estudiantes en lectura, matemáticas, ciencias y resolución de problemas en algunas pruebas internacionales

Nombre	Dirección Web	Descripción
BD de educación de la OECD	http://www.oecd.org/education/	Las BD de la OCDE en educación incluyen datos y estadísticos para diferentes países y en temas variados.
BD Gapminder	https://www.gapminder.org/for-teachers/	Gapminder es una fundación sueca independiente que produce recursos de enseñanza gratuitos basado en estadísticas confiables y entre otras cosas comparte algunas BD de educación.
Eurostat base de datos: educación y formación para toda Europa	http://ec.europa.eu/eurostat/web/education-and-training/data/database	Eurostat es una institución que ofrece estadísticas europeas y bases de datos de algunas temáticas, entre ellas de educación
World Education Services	https://www.wes.org/ca/wedb/	Recursos educativos gratuitos para profesionales de la educación superior a nivel internacional. Permite hacer una búsqueda por país y muestra fuentes de datos educativos para dicha nación.
Base de datos del Centro Nacional de Estadísticas de Educación de los Estados Unidos	https://nces.ed.gov/	El Instituto de Ciencias de la Educación (IES) es el órgano de estadísticas, investigación y evaluación del Departamento de Educación de los EE. UU. Su misión es proporcionar evidencia científica para prácticas y políticas educativas y compartir esta información en formatos que sean útiles y accesibles para educadores, padres, formuladores de políticas, investigadores y público en general.
Data shop	https://pslcdatashop.web.cmu.edu/	Uno de los mayores repositorios de datos de aprendizaje, permite acceder rápidamente a informes estándar, como curvas de aprendizaje, así como navegar por los datos mediante la aplicación web interactiva y descargarlos en formatos compatibles con otras herramientas de análisis.
UCI Machine Learning Repository	https://archive.ics.uci.edu/ml/index.php	Repositorio con más de 497 conjuntos de datos como un servicio para la comunidad de machine learning, entre ellos se encuentran de educación.
Programa ICFES de Investigación	https://www.icfes.gov.co/web/guest/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos	Este programa promueve el uso de los resultados de las evaluaciones nacionales e internacionales en las que Colombia participa, en investigaciones que aporten información confiable para orientar la toma de decisiones en políticas públicas para mejorar la calidad de la educación. Propicia el vínculo entre

Nombre	Dirección Web	Descripción
		investigación, práctica educativa y política pública.
Datos abiertos Colombia	https://www.datos.gov.co/	Programa que promueve una política de datos abiertos en Colombia, entre ellos algunos provenientes de instituciones del sector educativo.
ASSISTments Competition Dataset	https://sites.google.com/view/assistmentsdatamining/home	Competencia en la que los mineros de datos pueden intentar predecir un resultado longitudinal importante utilizando datos educativos del mundo real.
Canvas Network dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZZH3UB	Datos identificados de Canvas Network cursos abiertos (de enero de 2014 a septiembre de 2015), junto con la documentación relacionada.
HarvardX-MITx dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147	Datos del primer año de los cursos MITx y HarvardX MOOC en la plataforma edX junto con la documentación relacionada.
Learn Moodle dataset	https://research.moodle.org/	Datos anonimizados del curso “Enseñando con Moodle agosto de 2016” de learn. moodle.net.
MOOC-Ed Dataset	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZZH3UB	Datos de comunicaciones que tienen lugar entre los alumnos del Curso en línea para educadores (MOOC-Eds).
Open University Learning Analytics Dataset	https://analyse.kmi.open.ac.uk/open_dataset	Contiene datos sobre cursos, estudiantes y sus interacciones con Moodle para siete cursos seleccionados.
Student Performance Dataset	https://archive.ics.uci.edu/ml/datasets/Student+Performance	Estos datos se acercan al rendimiento de los estudiantes en la educación secundaria de dos escuelas portuguesas.
xAPI-Educational Mining Dataset	https://www.kaggle.com/aljarah/xAPI-Edu-Data	Conjunto de datos de rendimiento académico de los estudiantes recopilados del sistema de aprendizaje electrónico llamado Kalboard 360

Fuente: Adaptado de (Romero & Ventura, 2020).

Estas fuentes de datos o repositorios de datasets son solo una muestra. Sin embargo, la mayoría de los trabajos de EDM y de analíticas de aprendizaje se concentran en sus propios datos. Estos datasets pueden ser usados más como una referencia para hacer pruebas de algoritmos, para compararse o para extraer indicadores de contexto que orienten los estudios propios. En la mayoría de los procesos investigativos, el acceso a las fuentes de datos y la adquisición de estos suele ser la tarea que más esfuerzo y tiempo toma. Por una parte, no siempre es fácil establecer los convenios de cooperación o vínculos con las instituciones que poseen los datos. Por otra parte, se debe garantizar el cumplimiento de las políticas de uso de los datos, que, en algunos casos, para el ambiente educativo,

implican información de niños y adolescentes, menores de edad. El componente ético juega un papel fundamental en el dominio educativo, tanto en la manipulación de la fuente como en la presentación de los datos y resultados.

3.3 Educación en Colombia

"En Colombia la educación se define como un proceso de formación permanente, personal, cultural y social que se fundamenta en una concepción integral de la persona humana, de su dignidad, de sus derechos y de sus deberes" (Ministerio de Educación Nacional, 2020). La constitución Política es el documento que presenta los fundamentos iniciales del sistema educativo, el cual es regulado, inspeccionado y vigilado por el Estado, para velar por la calidad educativa y la formación (intelectual, moral y física) de los estudiantes. Lo anterior acompañado de la garantía en cobertura, acceso y permanencia en el sistema.

El sistema educativo colombiano lo conforman: la educación inicial, la educación preescolar (al menos un grado obligatorio), la educación básica (primaria cinco grados y secundaria cuatro grados), la educación media (dos grados, culmina con el título de bachiller), y la educación superior o terciaria (duración variable). La educación preescolar tiene como propósito que los niños aprendan a convivir, a integrarse y a jugar con otros. En la educación básica los estudiantes se forman y adquieren las competencias primordiales para para poder desarrollar su vida e integrarse a la sociedad. La educación media es una preparación inicial a la vida productiva y puede inclinarse por un componente académico o técnico. La educación superior se divide en técnica, tecnológica y universitaria. La formación obligatoria normalmente suele requerir de 11 años continuos de escolarización. En la Figura 15 se esquematiza la organización del sistema educativo en Colombia, se presenta un rango de edades promedio para cursar cada grado, pero no tiene que ser así para todos los estudiantes, puede variar.

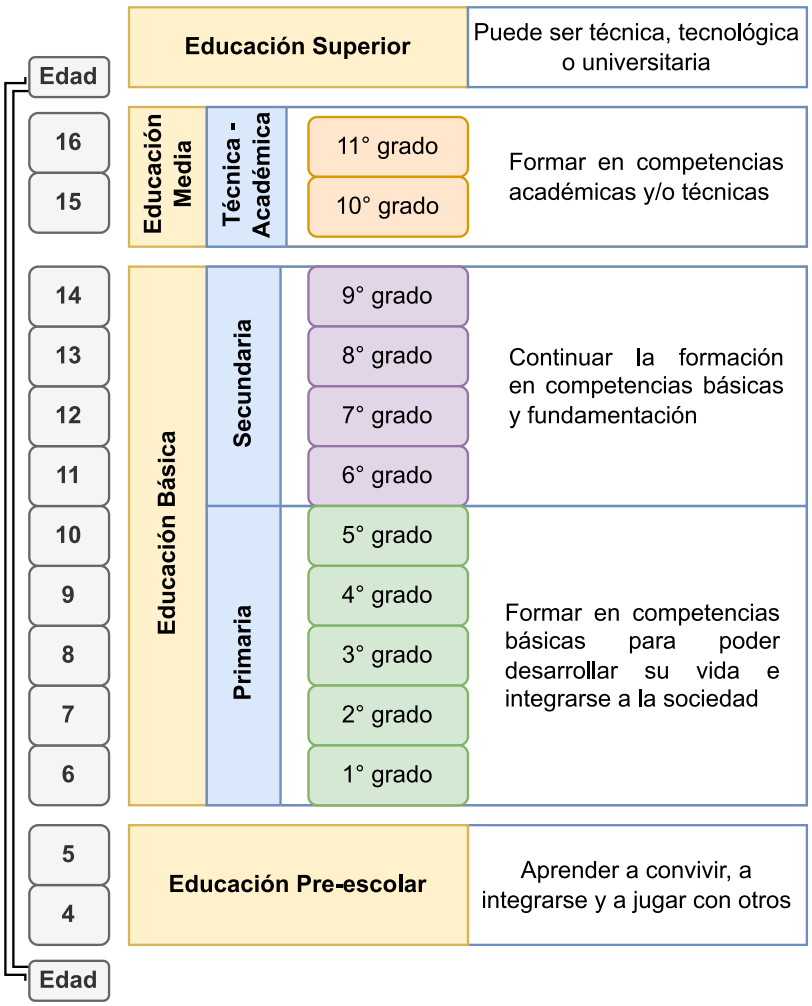


Figura 15. Organización del sistema educativo en Colombia
Fuente: Elaboración propia

Los grados, se refieren a la ejecución ordenada del plan de estudios y están conformados mediante periodos, que son etapas que agrupan objetivos educativos específicos. Un grado consta de cuatro periodos y tiene un año de duración. Respecto a la comunidad educativa, se organiza en cinco estamentos: a) los estudiantes, b) los padres de familia o acudientes, c) los docentes, d) los directivos docentes y personal administrativo, y e) los egresados. Todos los estamentos tienen participación en el gobierno escolar.

Las áreas fundamentales y obligatorias definidas por la ley (Ley 115 de 114) para la educación formal son: a) Ciencias naturales y educación ambiental; b) Ciencias sociales, historia, geografía, constitución política y democracia; c) Educación artística; d) Educación ética y valores humanos; e) Educación física, recreación y deportes; f) Educación religiosa; g) Humanidades, lengua castellana e idiomas extranjeros; h) Matemáticas; e i) Tecnología e informática.

3.4 Sistemas de información educativos en Colombia

El ministerio de Educación Nacional ha venido llevando a cabo iniciativas para tratar de estandarizar los procesos educativos, principalmente en el nivel de educación básica (primaria y secundaria), entre las iniciativas, el uso de tecnologías de la información y las comunicaciones y el apoyo de los sistemas de información para la gestión de los datos. A continuación, se presentan algunos de los sistemas de información que se usan a nivel de educación básica y media y que han sido consultados en el sitio web del Ministerio (Ministerio de Educación, 2020c).

SIMAT: El sistema integrado de matrícula -SIMAT- es una herramienta que permite organizar y controlar el proceso de matrícula en todas sus etapas, así como tener una fuente de información confiable y disponible para la toma de decisiones. Es un sistema de gestión de la matrícula de los estudiantes de instituciones oficiales que facilita la inscripción de alumnos nuevos, el registro y la actualización de los datos existentes del estudiante, la consulta del alumno por Institución y el traslado a otra Institución, entre otros.

Está dirigido a las Secretarías de Educación. Este sistema es una de las estrategias del proyecto de modernización de Secretarías. Mediante la automatización del proceso de matrícula, a través del SIMAT, se logra sistematizar, consolidar y analizar la información. De esta manera, se mejoran los procesos de inscripción, asignación de cupos y matrícula, y por ende el servicio a la comunidad.

url: https://www.sistemamatriculas.gov.co/simat/app

SICIED: El SICIED es el Sistema Interactivo de Consulta de Infraestructura Educativa. Es una metodología que permite cuantificar, evaluar y calificar el estado de los establecimientos educativos en relación con estándares de infraestructura (NTC 4595 ICONTEC). En el año 2006 el Ministerio de Educación entregó de manera gratuita a los encargados de la infraestructura escolar de los entes territoriales este software para inventario de bienes inmuebles escolares. El aplicativo fue desarrollado por la Cámara de Comercio de Barranquilla con el apoyo de Promigas, la Fundación Empresarios por la Educación y ajustado de acuerdo con los requerimientos del Ministerio.

El uso de este software apoya la elaboración y el levantamiento del inventario de infraestructura de los establecimientos educativos en las entidades territoriales certificadas. Esta información permite la organización y el diagnóstico real de la infraestructura educativa, así como la toma de decisiones oportunas para el mejoramiento continuo de los ambientes escolares como apoyo fundamental a las estrategias de cobertura y calidad educativa. El SICIED ayuda a consolidar datos históricos sobre la edificación y facilita el uso de estándares de infraestructura en la edificación e intervención de establecimientos educativos.

url: https://sineb.mineducacion.gov.co/sicied/

Sistema de Seguimiento: Es el sistema con el que se hace seguimiento al reporte de la planta docente y la matrícula de educación básica y media. Se puede consultar los resultados por el total nacional, por secretarías o por establecimientos educativos. Este sistema permite a los gobernadores, alcaldes, directivos de establecimientos educativos y demás interesados hacer un seguimiento al cumplimiento de sus objetivos, a la cobertura y calidad de educación que ofrecen en las instituciones. Con la información de este sistema, se apoya y facilita la toma de decisiones que tienen en cuenta la planta docente y la matrícula.

url: http://menweb.mineducacion.gov.co/seguimiento

EVI: EVI es el Sistema de Información de Evaluación Institucional y Tarifas de Establecimientos Educativos Privados de Preescolar, Básica y Media. En este sistema se gestiona el proceso de evaluación institucional de la calidad del servicio prestado por establecimientos educativos de preescolar, básica y media privados, así como el reporte de información financiera y la fijación de tarifas.

Todos los establecimientos privados registran en línea su autoevaluación institucional (existe un formulario para colegios: el 1A, uno para jardines: 1B, otro para educación de adultos: 1C y para establecimientos nuevos) junto con su información financiera y propuesta de tarifas (formulario 2). Adjuntan las actas de revisión de esta información por el Consejo Directivo y contador o revisor fiscal. Deben reportar 60 días antes de la matrícula. A partir de esta información, las secretarías hacen seguimiento al proceso, generan resoluciones de clasificación y tarifas para cada establecimiento educativo (antes de que este matricule) y organizan las visitas de evaluación externa. Esto se hace una vez al año, para calendario A y B. La información del sistema no sólo se usa en el monitoreo de la calidad educativa del sector privado (complementario a las pruebas SABER), sino que alimenta a la herramienta "Buscando Colegio", diseñada para dar a las

familias información de calidad y tarifas de los establecimientos educativos, además de servir como directorio del sector.

url: https://autoevaluacion.mineducacion.gov.co/autoeval/faces/index.jsp

SINCE: corresponde al Sistema Nacional de Información de Contratación Educativa. Es una estrategia usada en el Proyecto de Ampliación de Cobertura para la Población Vulnerable que lidera el Ministerio de Educación Nacional para facilitar el acceso a la educación formal de niños, niñas y jóvenes de poblaciones afectadas por el conflicto armado, indígenas, con discapacidades y de áreas rurales dispersas.

Este sistema consigna datos sobre las solicitudes realizadas, la ejecución de los convenios interadministrativos, la conformación del banco de oferentes de las entidades territoriales y la ejecución de los contratos de prestación de servicios educativos.

url: https://vumen.mineducacion.gov.co/TMS.Solution.VUMEN/Basica/

Sistema de seguimiento a los recursos de gratitud: Aplicativo para hacer seguimiento a los recursos girados para la gratuidad de la educación preescolar, básica y media. Este sistema de información permite hacer seguimiento a los recursos del Sistema General de Participaciones (SGP) girados por el gobierno nacional a los municipios y distritos, y destinados para la gratuidad de la educación preescolar, básica y media del país. Los alcaldes y directivos de establecimientos educativos deben ingresar y confirmar cuántos recursos llegaron a su presupuesto y cuánto han girado para el pago de matrícula. Los alcaldes tienen la obligación de transferir los recursos de gratuidad a los establecimientos educativos estatales de su jurisdicción y reportar la información a la secretaría de educación certificada correspondiente, sobre los giros realizados.

url: https://www.mineducacion.gov.co/portal/micrositios-preescolar-basica-y-media/Gratuidad-Escolar/

Recursos humanos: Sistema para mejorar la gestión del recurso humano en las secretarías. Es un sistema de Información para apoyar en las secretarías de educación los procesos de administración, organización y control de la información relacionada con la gestión del recurso humano, así como la liquidación de la nómina para el personal docente y administrativo de las Secretarías de Educación. Este sistema de información cubre los alcances de definición de la planta personal, continuando con la selección e inducción del personal, la administración

de la carrera administrativa y el escalafón docente, el desarrollo de procesos de capacitación y bienestar, la administración de las hojas de vida, finalizando con la generación y liquidación de la nómina para los funcionarios docentes y administrativos de la Secretaría de Educación.

url: https://www.mineducacion.gov.co/portal/micrositios-institucionales/Sistemas-de-Informacion/Educacion-Basica-y-Media/168884:Recursos-humanos

SINEB: Sistema de Información Nacional de Educación Básica y Media. Este sistema recoge los datos que deben reportar los establecimientos educativos oficiales y no oficiales a los municipios y/o departamentos. Estos reportes incluyen la información de los establecimientos educativos, la situación académica de los estudiantes al finalizar el año anterior, los resultados de calidad y los datos de los docentes de establecimientos privados, entre otros. El SINEB proporciona los datos necesarios para determinar la cobertura, la calidad y la eficiencia del servicio, brinda a la nación, los departamentos, los distritos y los municipios la información requerida para la planeación del servicio educativo y para la evaluación de sus resultados.

Entre otros objetivos, con el SINEB es posible estimar los costos y determinar las fuentes de financiamiento del servicio público educativo. Además, sirve de base para distribuir entre las entidades territoriales los recursos de la participación para educación del Sistema General de Participaciones, de acuerdo con la población atendida y la población por atender.

url:
https://sineb.mineducacion.gov.co/?josso_back_to=http://sineb.mineducacion.gov.co/sineb/josso_security_check&josso_on_error=http://sineb.mineducacion.gov.co

Buscando colegio: Sistema de consulta de las instituciones educativas del país. En esta herramienta se encuentra un directorio completo de instituciones oficiales y no oficiales de educación preescolar, básica y media. La reseña de cada institución muestra datos como teléfonos, dirección y una ficha técnica. Se pueden hacer consultas de los establecimientos por ubicación geográfica, nombre, sector (oficial o no oficial), entre otros. También se puede ver información sobre la jornada escolar, nombre completo, calendario académico, género (masculino, femenino o mixto) y tarifas, entre otras características. Los datos del sistema son tomados del Directorio Único de Establecimientos Educativos (DUE) que administran las Secretarías de Educación certificadas. Permite que la comunidad educativa tome decisiones y haga veeduría de la información suministrada. Por

ejemplo, los padres y madres interesados pueden buscar una institución educativa que brinde la formación que desean para su hijo o hija.

En cuanto a educación superior, también se encuentran algunos sistemas de información que son suministrados por el Ministerio de Educación. Sin embargo, en ese nivel cada Universidad o Institución suele tener sus propios sistemas. A continuación, se presentan los sistemas de información del Ministerio en cuanto a educación superior.

url: https://sineb.mineducacion.gov.co/bcol/app?service=page/BuscandoColegio

SPADIES: Sistema de Prevención y Análisis a la Deserción en las Instituciones de Educación Superior. Es una herramienta informática que permite hacer seguimiento al problema de la deserción en la educación superior, es decir, a los estudiantes que abandonan sus estudios superiores. El SPADIES fue diseñado por el Centro de Estudios Económicos (CEDE), de la Universidad de los Andes, y está articulado con el Sistema Nacional de Información de la Educación Superior (SNIES), el Instituto Colombiano para el Fomento de la Educación Superior (ICFES) y el Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior (ICETEX). Con este software es posible tener estadísticas sobre la deserción en las instituciones de educación superior, identificar los riesgos que llevarían a un estudiante a abandonar sus estudios y hacer seguimiento y evaluación a las estrategias diseñadas para evitar este problema.

url: https://www.mineducacion.gov.co/sistemasinfo/spadies/

SACES: Sistema de Aseguramiento de la Calidad de la Educación Superior. Este sistema fue creado para que las Instituciones de Educación Superior (IES) realicen de forma automática los trámites asociados al proceso de Registro Calificado y de tipo institucional como: Reconocimiento de Personería Jurídica, aprobación de estudio de factibilidad para Instituciones de Educación Superior públicas, cambio de carácter, reconocimiento como universidad, redefinición para el ofrecimiento de ciclos propedéuticos, autorización de creación de seccionales.

url: https://www.mineducacion.gov.co/portal/micrositios-superior/SACES/

Observatorio Laboral para la Educación: Seguimiento a graduados de la educación superior. Es un sistema que mantiene información sobre sus condiciones laborales y sobre qué tipo de profesionales necesita el mercado (tendencias de la demanda). Sus fuentes de información son las Instituciones de Educación Superior, la Registraduría Nacional (para validar cédulas de

ciudadanía), el Ministerio de la Protección Social y el Ministerio de Hacienda y Crédito Público.

url: https://ole.mineducacion.gov.co/portal/

SNIES: Sistema Nacional de Información de la Educación Superior. En este sistema se recopila y organiza la información relevante sobre la educación superior que permite hacer planeación, monitoreo, evaluación, asesoría, inspección y vigilancia del sector. En el SNIES, los ciudadanos pueden encontrar información sobre todas las instituciones de educación superior registradas y sus carreras, así mismo, reportes de los programas que se ofrecen en el país y tienen el Registro Calificado (que cumplen las condiciones legales para ser ofrecidos).

url: https://snies.mineducacion.gov.co/portal/

Sistemas de información de evaluación externa de la calidad educativa: En temas de evaluación de la calidad educativa, el ICFES (Instituto Colombiano para la Evaluación de la Educación) que es el órgano nacional a cargo de esta tarea, cuenta con un sistema de información robusto y con bases de datos abiertas a la comunidad de investigadores (nacionales e internacionales) garantizado la anonimización de los mismos. Estos datos se encuentran organizados de acuerdos a las pruebas de evaluación que efectúa el estado, las cuales son: SABER 3°, 5°, 9° (en educación básica), SABER 11° (para educación media), SABER TyT (para formación técnica y tecnológica) y SABER PRO (para educación superior). Pueden ser accedidos a través de un sistema FTP, con una solicitud de acceso sencilla. En el sitio web se encuentra el instructivo de acceso a las bases de datos y a los diccionarios de datos correspondientes. De acuerdo con la prueba, se presentan los datos para los periodos en que han sido aplicadas. Los datos incluyen resultados por estudiante, por institución y en algunos casos por municipio. Cabe resaltar que, en el caso de los resultados por estudiante, estos no presentan información personal (como nombres o documentos) que permitan identificar a un individuo.

url: https://www.icfes.gov.co/web/guest/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos

Sistema de Información de la Educación Rural (SIER): Este sistema es utilizado por el Ministerio de Educación Nacional para recopilar información sobre la educación en las zonas rurales de Colombia, incluyendo datos sobre la infraestructura educativa, los programas de estudio y los estudiantes matriculados.

url: no reporta

Sistema de Información de la Educación para el Trabajo y el Desarrollo Humano (SIET): Este sistema es utilizado por el Ministerio de Educación Nacional para recopilar información sobre la educación técnica y tecnológica en Colombia, incluyendo datos sobre las instituciones que ofrecen este tipo de educación, los programas de estudio y los resultados académicos.

url: https://siet.mineducacion.gov.co/siet/

Como se deja ver, existen múltiples sistemas de información que el Ministerio de Educación ha venido poniendo a disposición de la educación básica, media y superior. No se recopilan todos en este documento, además de estos, las secretarias de educación también manejan en cada departamento algunos otros y como se mencionó anteriormente, las instituciones de educación superior también cuentan con sus propios aplicativos y sistemas de información. Incluso, a nivel de las instituciones de Educación Básica oficiales (públicas), los rectores tienen autonomía para contratar servicios con empresas privadas para adquirir software para hacer la gestión de calificaciones y reportes de notas escolares; lo cual en algunos casos hace que los datos queden en manos de operadores privados, ya que los colegios no cuentan con la infraestructura (servidores) para salvaguardar sus datos.

Algunos de los aplicativos del Ministerio presentan política de datos abiertos, pero no todos permiten el acceso a la información, dado que la política de uso y compartición de datos no se encuentra del todo clara y que predomina el manejo de datos personales, que exigen de un cuidado especial y de estándares éticos.

3.5 Datos sistemas de información educativos en Colombia

A continuación, se presenta en la Tabla 13 una recopilación de los principales datos generados por algunos de los sistemas de información (SI) mencionados en el apartado anterior. Es importante tener en cuenta que esta tabla solo muestra algunos ejemplos de los datos que se generan en cada sistema de información educativo, y que la lista no es exhaustiva. Además, cada sistema puede recopilar y generar diferentes tipos de datos según las necesidades específicas de cada institución o región.

Tabla 13. Datos generados principales SI Educativos en Colombia

Sistema de información educativo	Datos generados
Sistema Nacional de Información de Educación Superior (SNIES)	Número de instituciones de educación superior en Colombia, programas de estudio ofrecidos por cada institución, número de estudiantes matriculados en cada programa, número de egresados de cada programa, resultados de las evaluaciones de calidad de las instituciones y programas.
Sistema Integrado de Matrícula (SIMAT)	Número de estudiantes matriculados por grado y nivel educativo, número de estudiantes matriculados en cada institución educativa, número de estudiantes por género y edad, número de estudiantes con necesidades educativas especiales.
Sistema de Información para la Educación Básica y Media (SIEM)	Número de estudiantes matriculados por grado y nivel educativo, número de docentes por institución y grado, número de escuelas y colegios por región, resultados de las pruebas Saber 3°, 5°, 9° y 11°.
Sistema de Información de la Educación Rural (SIER)	Número de escuelas rurales en Colombia, número de estudiantes matriculados en escuelas rurales, número de docentes por escuela rural, programas de estudio ofrecidos en las escuelas rurales.
Sistema de Información de la Educación para el Trabajo y el Desarrollo Humano (SIET)	Número de instituciones de educación técnica y tecnológica en Colombia, programas de estudio ofrecidos por cada institución, número de estudiantes matriculados en cada programa, número de egresados de cada programa, resultados de las evaluaciones de calidad de las instituciones y programas.

Fuente: Elaboración propia

Resumen del capítulo

En este capítulo se ha realizado una caracterización de los datos educativos y se ha contextualizado en particular el caso colombiano, a través de la descripción del funcionamiento del sistema educativo en el país y de los sistemas de información que maneja el Ministerio de Educación Nacional. Con este capítulo se han establecido elementos que permiten determinar posibilidades frente a las diferentes fuentes de datos que alimentarían el modelo de dominio específico para minería de datos educativa. Se han descrito también, algunos retos en temas de educación básica y media en Colombia, en concordancia con el alcance de aplicación y el entorno de validación de esta tesis.

Se identifica que existen problemáticas y retos en el tratamiento de los datos educativos y para abordar estos retos es necesario implementar políticas claras y rigurosas de recopilación y tratamiento de datos, así como utilizar tecnologías y herramientas adecuadas para el manejo y análisis. Además, es importante capacitar a los profesionales en educación en el uso y análisis de estos para garantizar una correcta interpretación y aprovechamiento de la información obtenida. A diferencia de otros tipos de datos, los datos educativos están directamente relacionados con el desarrollo de habilidades y competencias en los estudiantes, y pueden ser utilizados para evaluar la efectividad de los programas y políticas educativas. Además, los datos educativos pueden ser utilizados para identificar patrones y tendencias en el desempeño estudiantil, lo que puede ayudar a informar decisiones sobre cómo mejorar la calidad de la educación. Para ampliar el sistema educativo en Colombia se incluye el Anexo A.

CAPÍTULO 4 – Modelo de dominio específico para EDM propuesto

El objetivo de este capítulo es exponer el modelo de dominio específico para minería de datos educativos, principal aporte de esta tesis doctoral. Para iniciar, se quiere introducir la definición de modelo; según la real academia de la lengua española, modelo es un *"Arquetipo o punto de referencia para imitarlo o reproducirlo*", también se indica que es un *"Esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, … que se elabora para facilitar su comprensión y el estudio de su comportamiento".*

Un modelo busca hacer una representación de un objeto o situación real; por lo tanto, el acto de modelar implica la comprensión de un proceso o fenómeno y de las interacciones entre las partes de este con un posible entorno o sistema. Los principios del modelado permiten la creación de los modelos atendiendo a la rama de la ciencia que intenten explicar. La forma de construir modelos de análisis se estudia en la filosofía de la ciencia, la teoría general de sistemas y en la visualización científica. En casi todas las explicaciones de fenómenos se puede aplicar un modelo u otro, pero es necesario ajustar el modelo a utilizar, para que el resultado sea lo más exacto posible (Maldonado, 2017).

Una construcción inicial de un modelo se puede denotar como *modelo conceptual*, en el cual se incluyen conceptos explícitos y sus relaciones. Los modelos conceptuales son mapas de conceptos abstractos que representan los fenómenos a estudiar incluyendo suposiciones que permitan vislumbrar el resultado del modelo. Tienen un alto nivel de abstracción para explicar el modelo y se consideran como los modelos científicos per se, donde la representación conceptual de los procesos consigue explicar el fenómeno a observar (Acevedo-Díaz et al., 2017).

Luego de un proceso de formalización lógica del modelo conceptual, se puede llegar a un modelo informacional que comprende aspectos matemáticos y/o computacionales. A través de la introducción de datos en el modelo se permite estudiar el resultado final, se puede hacer la operacionalización del modelo, es decir, aplicarlo o traducirlo en un prototipo validable o simularlo mediante una herramienta o software (Acevedo-Díaz et al., 2017).

En minería de datos, la definición de modelo suele estar asociada a las técnicas de aprendizaje supervisado, se tienen modelos de clasificación, de regresión, de

asociación, de frecuencias, entre otros. El concepto está ligado a la forma de extraer el conocimiento de los datos o de la realidad y llevarlo a un esquema repetible. Por su parte, en el enfoque de dominio específico, se asocia el modelo a la representación particular que se puede hacer de un dominio a partir del conocimiento que se tiene de él.

Un modelo de dominio en la resolución de problemas e ingeniería de software es un modelo conceptual de todos los temas relacionados con un problema específico. En él se describen las distintas entidades, sus atributos, papeles y relaciones, además de las restricciones que rigen el dominio del problema. El modelo de dominio se crea con el fin de representar el vocabulario y los conceptos clave del dominio del problema. El modelo de dominio también identifica las relaciones entre todas las entidades comprendidas en el ámbito del dominio del problema, y comúnmente identifica sus atributos. El modelo de dominio proporciona una visión estructural del dominio que puede ser complementado con otros puntos de vista dinámicos, como el modelo de casos de uso.

Después de estas anotaciones, en este capítulo se retoman las caracterizaciones y revisiones realizadas en los tres capítulos anteriores para proponer un modelo de dominio específico que apoye la aplicación de minería de datos en el contexto educativo. Para ello, se presenta en primera instancia la definición del modelo conceptual y luego se lleva a la aplicación para una realidad particular del campo de estudio.

4.1 Justificación del modelo de dominio específico

El modelo se orienta a cubrir las expectativas de la minería de datos desde un enfoque de dominio específico, tomando componentes y características propias del campo de aplicación, en este caso, los ambientes educativos y los datos generados en estos; entiéndase como ambiente educativo todo en el cual se producen datos relacionados con procesos de enseñanza y aprendizaje.

Para ello se tiene en cuenta la caracterización realizada en los capítulos 1, 2 y 3 relacionados con el enfoque de dominio específico, la minería de datos y las fuentes de datos educativos. Así mismo, se involucran los actores del dominio, las problemáticas y objetivos de EDM que pueden ser atacados con el modelo. Cabe anotar en este punto, que el modelo de dominio específico se ve limitado por los datos, la información y el conocimiento disponible en el dominio, pero en todo momento busca adaptarse a los datos educativos y a las posibles entradas que alimenten la búsqueda de nuevo conocimiento.

Existen muchos retos a la hora de incluir un enfoque de dominio específico para una tecnología de análisis de datos como es el caso de la minería de datos educativos, retos que se evidencian en aspectos como la evolución continua del dominio, la construcción de estrategias para la validación del modelo, los requerimientos variables del dominio y el cuestionamiento entre la adaptación de herramientas existentes o el desarrollo de soluciones propias. Así mismo, se deben superar dificultades como: no contar con un marco de referencia del proceso, involucrar al usuario o interesados finales, dar soporte a la evolución del dominio, integración con otros sistemas, esfuerzo y costo de implementación de un prototipo del modelo.

Complementando lo anterior, la implementación de un modelo de dominio específico para minería de datos educativos puede ser un desafío debido a la complejidad de los datos educativos y las técnicas de modelado utilizadas, así como a la necesidad de proteger la privacidad y seguridad de los datos. Es importante abordar estos desafíos de manera efectiva para garantizar la eficacia del modelo y su capacidad para mejorar la calidad de la educación.

No obstante, el enfoque de dominio específico presenta bondades que se desean probar en el campo de la minería de datos educativos, para ello se emprende la propuesta de este modelo y se plantean estrategias que ayuden a cubrir, por lo menos de forma parcial, los retos mencionados. De igual forma, de la mano con la declaración del vacío del conocimiento, esta investigación se concentró en la necesidad identificada de plantear un modelo que pueda cubrir los espacios presentes en los modelos genéricos y que apoye a la minería de datos educativos.

En resumen, el diseño y construcción de un modelo de dominio específico para minería de datos educativos puede justificarse por la necesidad de tomar decisiones más informadas y basadas en datos, identificar patrones y tendencias, personalizar el aprendizaje, ahorrar tiempo y recursos y mejorar la competitividad. Así mismo, brindar la posibilidad de acercar a usuarios no expertos en minería de datos al aprovechamiento de las bondades reportadas por este enfoque.

4.2 Modelo de dominio específico propuesto

El modelo propuesto se orienta a cubrir las etapas de la minería de datos educativos incluyendo un enfoque de dominio específico. Se tienen en cuenta las etapas principales de los procesos de análisis de datos, tales como: recolección y tratamiento de datos, almacenamiento, selección de datasets y algoritmos,

aplicación de técnicas, reporte, interpretación de resultados y visualización. Estas etapas se condensan en cuatro componentes.

Con el ánimo de dar claridad en algunos términos antes de iniciar la descripción detallada de cada uno de los componentes, a continuación, se presentan las definiciones consideradas para algunos elementos usados en el modelo propuesto:

- Usuario estándar: es la persona que presenta interés en el dominio de datos educativos y que desea aplicar el modelo propuesto para hacer la construcción de modelos de minería de datos sobre algún caso particular o problemática asociada al contexto. El usuario no requiere de contar con conocimientos previos fuertes en el campo de la minería de datos.

- Usuario experto: es la persona que tiene un conocimiento significativo del contexto educativo y de las dinámicas del campo de estudio, cuyo conocimiento puede estar concentrado en temáticas como minería de datos educativos, analíticas de aprendizaje, tecnología educativa, ciencia de datos, procesos de dirección y formación en el área educativa. El experto puede ser a su vez usuario del sistema, tanto para compartir su conocimiento por medio del componente de representación del dominio, como para buscar soporte en conocimiento depositado por otros expertos o hacer uso del modelo para sus propios análisis.

- Modelo: cuando se hace uso de la palabra modelo o modelo propuesto, se estará haciendo referencia al modelo de dominio específico para minería de datos, objeto y resultado de esta tesis.

- Modelo de minería: por su parte se puede usar también la palabra modelo para hacer referencia como tal a los resultados de la aplicación de alguna técnica o algoritmo de minería de datos de aprendizaje supervisado; para este caso, se denotará como modelo de minería o modelo DM y evitar confusiones con el modelo propuesto.

En la Figura 16 se muestra el modelo propuesto, detallando las interacciones generales entre los componentes de este. El enfoque de dominio específico es considerado al vincular directamente el conocimiento propio del campo educativo tanto en el componente de representación del dominio como en su conexión con los datos y con los otros componentes ya sea de forma directa o indirecta, porque en todo momento el conocimiento se encuentra transitando por el modelo.

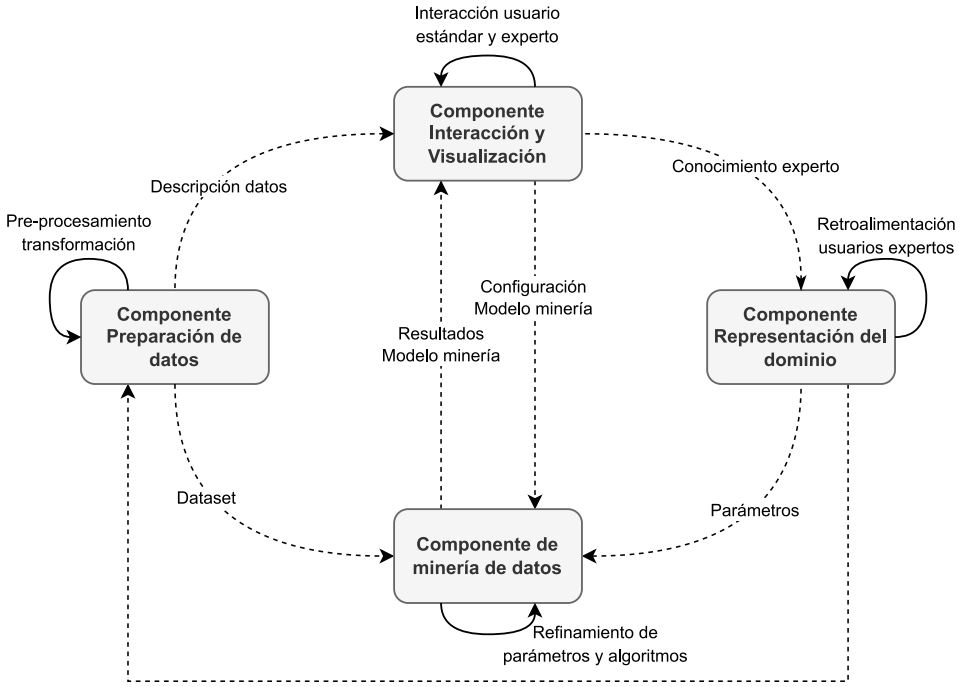


Figura 16. Modelo de dominio específico para Minería de Datos Educativos **Modelo**
Fuente: Elaboración propia

4.2.1 Componente de Preparación de Datos

El preprocesamiento y tratamiento de los datos es una tarea que normalmente exige dedicación y esfuerzo para los procesos de análisis y minería de datos, independiente del dominio que se esté trabajando. Para el modelo de dominio específico planteado, en esta etapa, se incluyen las estrategias necesarias para recolectar y preparar los datos educativos procedentes de diferentes fuentes y estructuras.

Estos datos pueden estar archivados desde acciones previas o ser datos monitoreados por una plataforma. En el componente de preparación de datos se debe permitir abordar los datos de gran variedad de fuentes y formatos. No obstante, garantizar que todas las estructuras sean usables para el procesamiento posterior con las tecnologías disponibles es un gran reto.

Adicionalmente, en este componente se realiza la transformación de los datos al formato requerido por los demás componentes. Para lo anterior se consideran algunos cuestionamientos: ¿qué posibles taxonomías o categorizaciones de los datos educativos se pueden encontrar?, ¿cómo evaluar los datos educativos en estado original (brutos)?, ¿pueden estos datos educativos o sus fuentes incluir

metadatos? Para dar respuesta a esto es necesario realizar un análisis del sistema educativo que genera los datos, los actores y el contexto involucrado.

En la preparación de los datos se toman los datos de la fuente y garantizando la calidad se llevan a un modelo de almacenamiento, sin desechar a priori algunos atributos. Es decir, al almacenamiento van a llegar las variables que acompañen el problema (los datos que han sido generados y captados desde la fuente); pero esto después de haber recibido un procesamiento inicial, que incluye unas fases tradicionales de extracción, filtrado, transformación y finalmente la carga. En la Figura 17 se presenta en detalle el componente de preparación de datos.

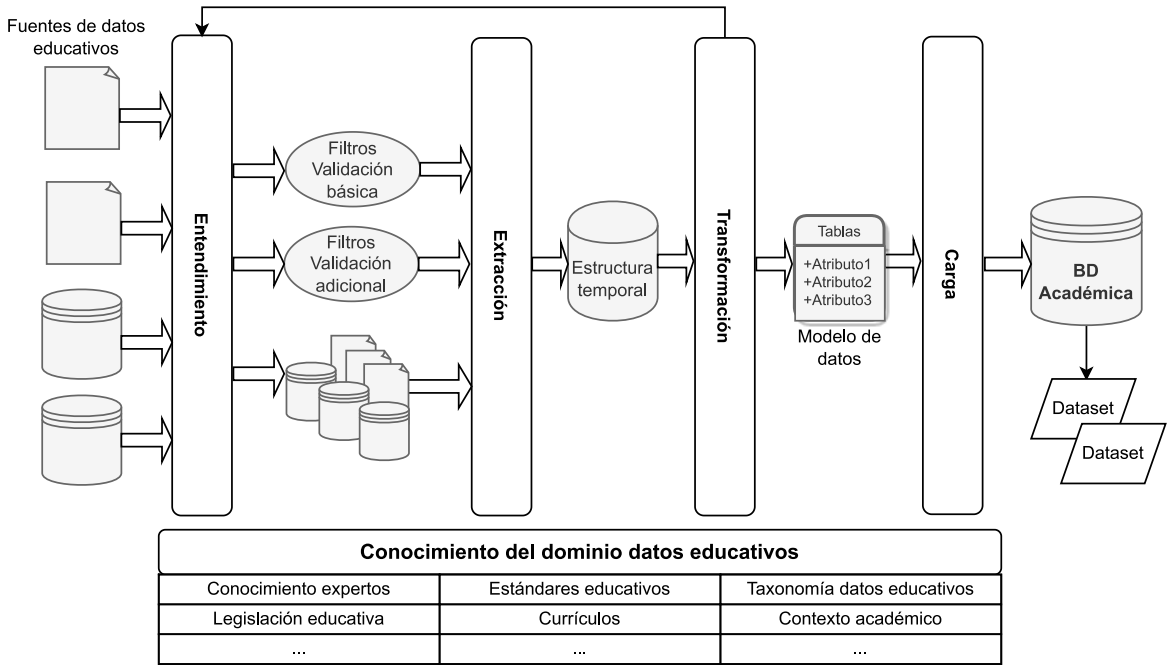


Figura 17. Detalle del componente de preparación de datos
Fuente: elaboración propia

El componente incluye cuatro fases que están representadas en la Figura 17 por las barras verticales. Existe una barra horizontal que interviene en todo el proceso, se trata del conocimiento del dominio de datos educativos, el cual es alimentado por diferentes abstracciones tomadas del ambiente educativo, como es el conocimiento de los expertos, los estándares educativos, las taxonomías de datos educativos, la legislación educativa, los currículos, el contexto académico, los perfiles de estudiante, las políticas institucionales, entre otros.

En el componente de preparación de datos propuesto, además de las típicas fases de extracción, transformación y carga que hacen parte de un proceso de ETL y que pueden ser encontradas en metodologías como KDD, se incluye una fase

inicial denominada Entendimiento, en la cual se contempla la caracterización propia del domino de datos específico. Siendo esto diferencial ante la aplicación de un modelo de tipo genérico, en el que puede que se tenga ese entendimiento, pero que no se deja explícito a la hora de ser abordado por un usuario que podría no ser experto en el tema y que requiera de la guía de los elementos que sería estratégico revisar como base de conocimiento para tomar las decisiones en términos de las intervenciones a realizar en dicha preparación de los datos. Adicionalmente, la denominación de este componente como preparación de datos atiende a que se quiere dar un concepto más amplio que abarca todas las actividades relacionadas con la mejora de la calidad y la utilidad de los datos antes de que sean utilizados para análisis o modelado (Brownlee, J., 2020); a diferencia de ETL que puede llegar a ser un término más técnico y asociarse con la integración de datos a gran escala en entornos empresariales, donde se consolidan datos en modelos de almacenamiento como Datawarehouse y se transforman para cumplir con los estándares y requisitos del negocio. A continuación, se hace una descripción de las fases.

Fase de entendimiento: las fuentes de datos educativos pueden tener diferente estructura y formato, ellas son la entrada a esta fase. Una vez se reciben los datos e identifican las estructuras, formatos y tipología, se empieza a hacer uso del conocimiento del dominio para entender las relaciones existentes y los problemas que se deben enfrentar. En este punto, uno de los aspectos fundamentales es poder contar con la ayuda de los expertos o propietarios de los datos, porque son ellos quienes tienen a priori el conocimiento y dominio de los sistemas que están generando la información. Las salidas de esta fase son los filtros y las fuentes con los datos originales sin sufrir ninguna alteración o preprocesamiento, se hace una copia para conservar el original y luego poder comparar si es requerido una vez son aplicados los filtros, consiguiendo tener una trazabilidad de las alteraciones realizadas.

Fase de extracción: está compuesta por dos actividades: extracción y filtrado. En la primera se toman los datos de interés o que se desean analizar y que requieren ser preparados para una posterior aplicación de técnicas. En la segunda se reciben los datos extraídos con su estructura original, se examinan y detectan posibles problemas por medio de una serie de filtros de validación básica, que se construyen a partir de la documentación de los datos, por ejemplo, la escala de calificaciones, el número de dígitos de un documento de identidad, entre otros. Posteriormente pueden aplicarse unos filtros de validación adicional o compuesta, estos filtros requieren el conocimiento del dominio, información que no necesariamente está a priori en la documentación del sistema o fuente de datos. La salida de la fase son los datos filtrados y llevados a una estructura temporal.

Este filtrado se realiza para que los datos lleguen a la fase de transformación en un estado más operable.

Fase de transformación: la entrada de esta fase es una estructura temporal con los datos filtrados, esta estructura existe para permitir validar y separar los datos que pasaron los filtros y los que no cumplieron las condiciones. La fase de transformación tiene entonces a cargo varias tareas como: normalización, eliminación de duplicados, verificación de faltantes, clasificación de atípicos y ajuste cuando sea posible. En este punto, se verifica con el conocimiento del dominio, para saber si los datos que no pasaron el filtro requieren una transformación para llegar al almacenamiento, o si por el contrario se detecta la necesidad de modificar el filtro o si finalmente se determina que el dato está errado, el problema encontrado no se logra solucionar y debe ser retirado. Esta fase es una fase de refinamiento, porque puede existir tanto transformación de los datos como transformación de los filtros, es por ello, que en el diseño se establece que desde esta fase se puede regresar a la fase de entendimiento de los datos, porque incluso los filtros pueden requerir modificación. Adicionalmente se puede requerir de una transformación de tipos de datos, la cual debe ser detectada en este punto. Es una fase formada por funciones aplicadas sobre los datos extraídos para convertirlos en datos que serán cargados. La salida de esta fase es el modelo de datos y los datos listos para ser cargados.

Fase de carga: esta es la última fase del proceso, consiste en tomar los datos que ya han sido filtrados, transformados si fue el caso y se encuentran limpios en la estructura temporal, para llevarlos, de acuerdo con el modelo de datos, a la carga en una base de datos o bodega de datos académicos. El conocimiento del dominio puede también influir en esta fase de carga, por ejemplo, en la selección del orden en el que tendrán que ser tomados los datos para llevarlos al almacenamiento, dependiendo de las relaciones establecidas en el modelo de datos. Finalmente, la salida de esta fase será la base o bodega de datos poblada y a partir de la cual se podrán extraer los datasets para la aplicación de las técnicas de minería de datos.

4.2.2 Componente de Representación del Dominio

El componente de representación del dominio es la principal fuente de diferenciación del modelo propuesto con los modelos tradicionalmente usados en minería de datos, incluso en minería de datos educativos. Se puede pensar que el conocimiento del dominio finalmente siempre es usado en los modelos, sin embargo, no siempre se cuenta con una estrategia y una guía que permita orientar respecto a los elementos del dominio que deben ser tenidos en cuenta y de qué forma se pueden articular con las demás etapas del proceso de análisis. En

consecuencia, el componente de representación del dominio consiste en metamodelos que definen varios conceptos que pueden ser usados para especificar aspectos de los problemas de análisis de datos educativos.

Este componente puede incluir abstracciones para representar varios escenarios educativos como la deserción o el rendimiento académico. Las abstracciones para calificaciones, estados de aprobación-reprobación o cualquier otra característica que los usuarios interesados quisieren analizar también necesita ser representada. El metamodelo debe también incorporar abstracciones para representar los formatos de datos y operaciones que el modelo debe soportar. Metamodelo está compuesto de dos palabras: "meta", que significa comentario y "modelo". Se pueden seguir, en este punto, algunos de los objetivos de la construcción de metamodelos que se tratan en PNL (Programación Neuro Lingüística), como:

- Concretar el actor que realiza la acción
- Encontrar la información relevante que falta
- Determinar el criterio de la comparación
- Transformar la abstracción en algo concreto
- Determinar la base y el origen de la información
- Encontrar la relación entre la causa y el efecto
- Verificar que la relación es correcta
- Desafiar la presuposición
- Cuestionar la generalización
- Identificar el origen o la causa

En la Figura 18 se presentan los principios de la concepción del componente de representación del dominio, se parte de las principales fuentes de alimentación del conocimiento seguidamente se muestran los productos y resultados; y finalmente se resalta el impacto que entrega el componente.

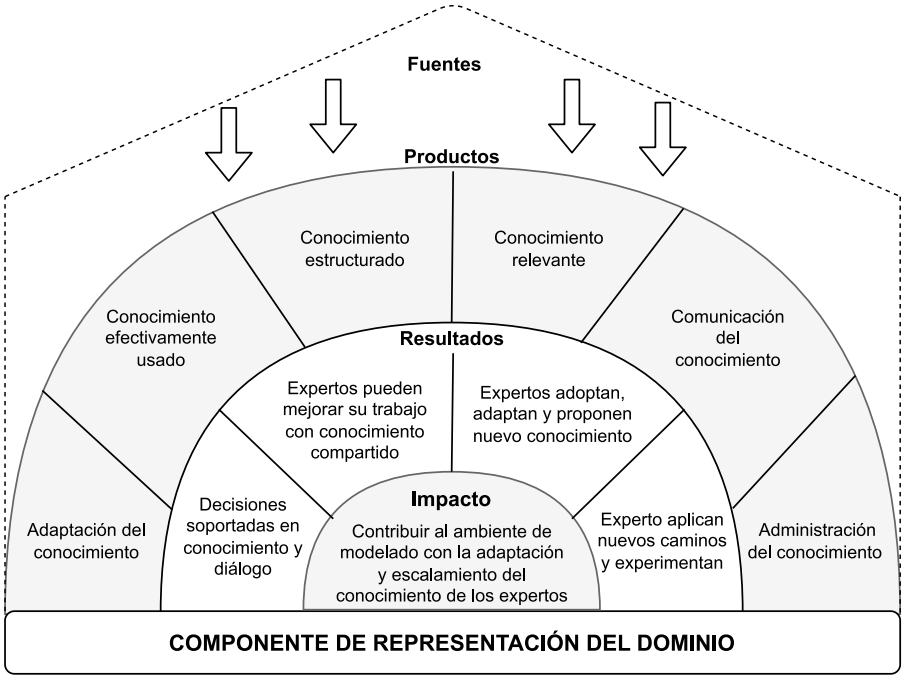


Figura 18. Principios del componente de representación del dominio
Fuente: elaboración propia

Ahora bien, el detalle de las interacciones que se dan dentro de este componente es presentado en la Figura 19, partiendo del dominio del problema, el cual se obtiene del experto y del contexto educativo y que se materializa en elementos como el conocimiento, la legislación y estándares educativos, las taxonomías de datos educativos, los currículos, técnicas de minería de datos probadas previamente en la solución de problemáticas educativas junto con los parámetros definidos para estas, la retroalimentación de los actores del contexto, entre otros. Todo lo anterior ingresa a una fase de extracción o inferencia, que permite obtener desde este cúmulo de conocimientos una serie de soluciones, explicaciones, reglas e instrucciones que se traducen en un conocimiento estructurado y relevante del dominio de datos y de las problemáticas adyacentes a este. Lo anterior llega a hacer parte de una base de conocimiento que representa el dominio.

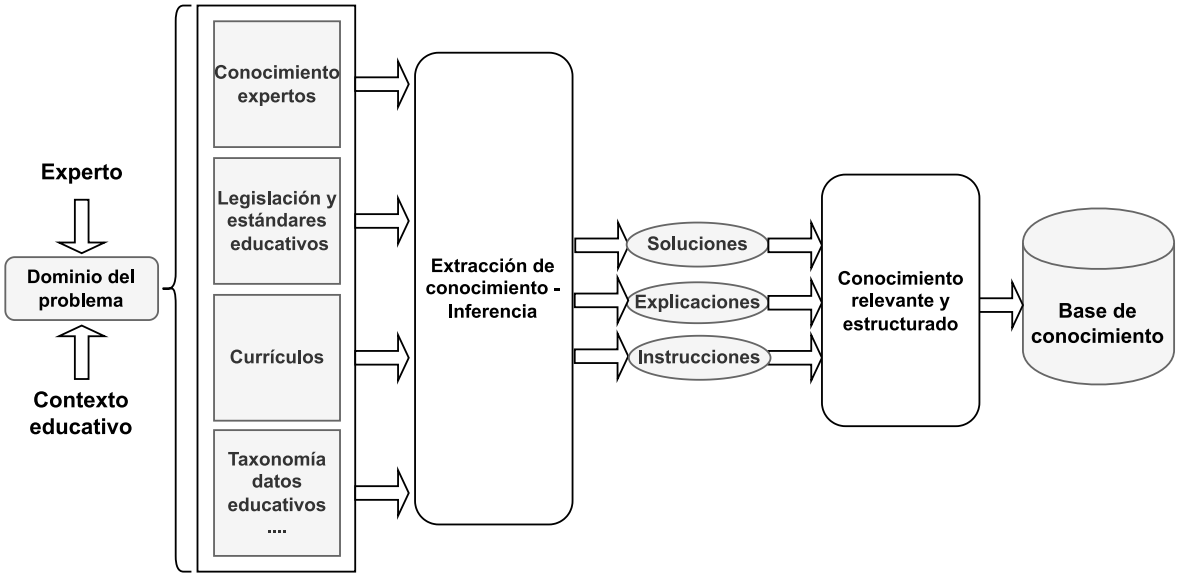


Figura 19. Detalle del componente de representación del dominio
Fuente: elaboración propia

Este componente puede ser visto como una aproximación a un sistema experto, dado que se tienen unas fuentes, que surgen de un contexto particular y con la retroalimentación de expertos pueden ser llevadas a una base de conocimiento estructurado en forma de reglas que podrán contribuir al uso del modelo por parte de los usuarios estándar y a lograr que la aplicación de las técnicas de minería sea orientada y se facilite la construcción e interpretación de los modelos de minería.

4.2.3 Componente de Minería de Datos

En este componente se encuentra la aplicación de las técnicas y algoritmos de minería de datos, es la etapa en la que se procesan los datos para la generación de los modelos de minería. Adicionalmente, en este componente se recomienda realizar un análisis exploratorio inicial, previo a la minería de datos, por medio del cual se puede empezar a intuir relaciones en los datos, además de ayudar a orientar la elección de técnicas y algoritmos.

La selección de las técnicas depende también de los objetivos del análisis y de los datos con los cuales se cuente. El modelo de minería está sujeto a las variables que los usuarios (estándar y expertos) quieran observar y al alcance inicialmente definido. Se pueden definir los objetivos con el uso de abstracciones de los metamodelos o representaciones (indicaciones) que guíen y soporten el proceso de acuerdo con las características del domino.

Uno de los principales retos al aplicar minería de datos es encontrar las técnicas adecuadas a los datos de entrada y poder interpretar los resultados, llegando realmente a conocimiento con valor. Las técnicas de minería de datos caracterizadas en el capítulo 3 tienen sus fortalezas y debilidades, se aplican a diferentes tipos de datos y pueden funcionar bien de acuerdo con ciertas condiciones y parámetros. Los datos educativos tienen algunas características especiales, son datos jerárquicos y longitudinales, que requieren un tratamiento específico (Romero & Ventura, 2020). En la Tabla 14 se hace un recuento de los métodos más usados en EDM y LA, se destaca el objetivo y las aplicaciones clave.

Tabla 14. Revisión de técnicas y algoritmos de EDM

Método	Objetivo/Descripción	Aplicaciones clave
Minería causal	Encontrar relaciones causales o identificar efectos causales en los datos	Encontrar cuáles características del comportamiento de los estudiantes causan o desenlazan aprendizaje, fracaso académico, deserción, entre otras
Clustering	Identificar grupos de observaciones similares	Agrupar estudiantes o materiales semejantes basados en su aprendizaje y patrones de interacción
Descubrimiento con modelos	Emplear modelos validados previamente de fenómenos como componentes en otro análisis	Identificación de relaciones entre el comportamiento de los estudiantes y las características o variables contextuales, integración de modelos psicométricos en modelos de Machine Learning
Predicción	Inferir una variable objetivo a partir de alguna combinación de otras variables. Clasificación, regresión y estimación de densidad son tipos de modelos de predicción.	Predecir el rendimiento de los estudiantes y detectar los comportamientos de estos
Recomendación	Predecir el rango o preferencias del usuario a través de los ítems dados	Hacer recomendaciones de estudiantes con respecto a sus actividades o tareas, enlaces a visitas o búsquedas, problemas, realización de los cursos, entre otros
Minería de relaciones	Estudiar las relaciones entre variables y reglas codificadas. Reglas de asociación, patrones secuenciales, minería de correlación y minería causal son algunos tipos	Identificar relaciones en el comportamiento del aprendiz y diagnosticar dificultades estudiantiles
Visualización	Mostrar gráficamente la representación de los datos	Producir visualizaciones de los datos que ayuden a la comunicación de los resultados de la minería de datos y de las analíticas de aprendizaje a los educadores

Fuente: Adaptado de (Romero & Ventura, 2020)

Dentro del modelo propuesto, el componente de minería se conecta con los otros tres componentes, existiendo flujos de intercambio de información entre ellos, para

entrar en detalle, en la Figura 20 se define el funcionamiento interno del componente y posteriormente se hace una descripción indicando las entradas, procesos y salidas de este.

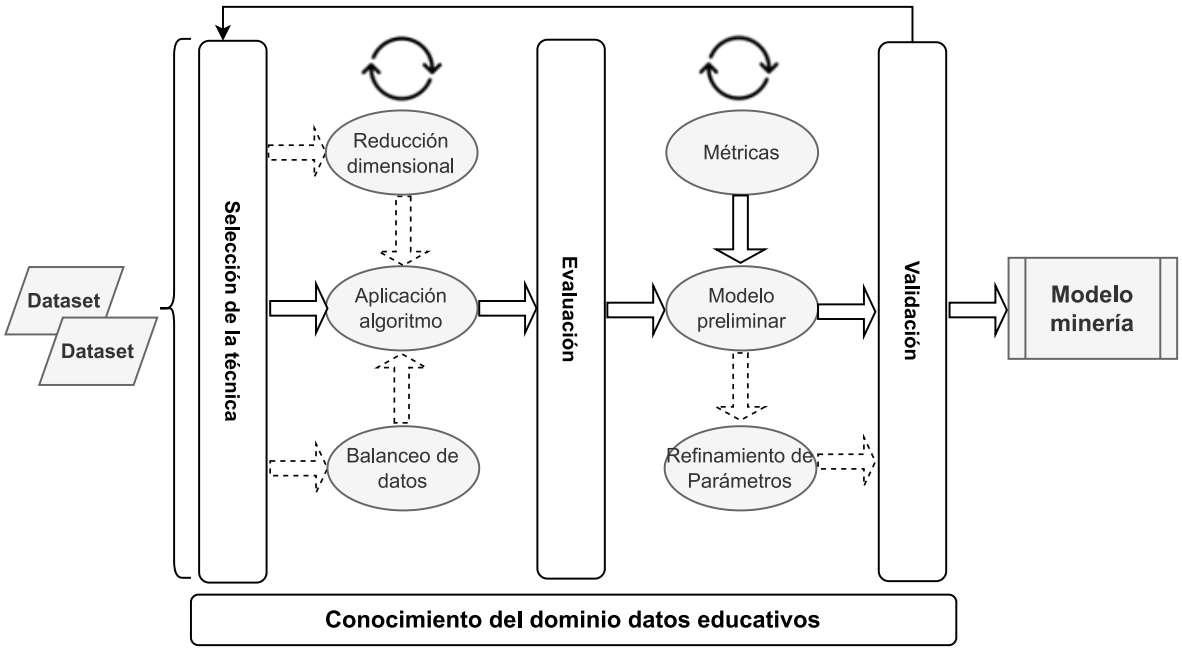


Figura 20. Detalle del Componente de Minería de Datos
Fuente: elaboración propia

Como se aprecia, el componente consta de tres fases, la selección de la técnica, la evaluación y la validación. Seguidamente se describen.

Fase de selección de la técnica: a esta fase ingresa el conjunto de datos que es extraído en el componente de preparación de datos. La selección de la técnica es apoyada con la base de conocimiento del componente de representación del dominio, además se comunica con el componente de interacción y visualización porque se requiere de la retroalimentación del usuario (estándar o experto). La selección de la técnica se apoya en el conocimiento de los métodos o técnicas probados previamente y de los parámetros experimentados con resultados positivos. De esta fase, la salida corresponde al algoritmo aplicado, así como también, de ser necesario, el balanceo de datos o la reducción de la dimensionalidad. Cabe aclarar que esta fase es de retroalimentación interna, puede requerirse de varias iteraciones antes de avanzar a la evaluación.

Fase de evaluación: la entrada de esta fase son los resultados del algoritmo y la salida es el modelo de minería. Hay una retroalimentación que se puede realizar con la aplicación de métricas y la posterior modificación de los parámetros, con lo

que se tiene que volver a la aplicación del algoritmo o incluso a la selección de otra técnica.

Fase de validación: en esta fase se requiere hacer pruebas con un subconjunto que sea extraído con el fin de determinar la validez del modelo de minería y su capacidad para describir datos diferentes a los del entrenamiento. Todas las fases son acompañadas del conocimiento del dominio y de la retroalimentación del usuario.

4.2.4 Componente de Interacción y Visualización

Es el componente que permite llevar a los usuarios (estándar y expertos) los resultados de los modelos de minería con una representación sencilla y fácil de comprender. El componente de interacción y visualización debe permitir la presentación apropiada que corresponde con los elementos del modelo de minería generado o del diagnóstico descriptivo. La función de la visualización de datos es traducir o transformar los resultados de la generación de modelos de minería en información entendible por los usuarios (principalmente estándar). En este componente se pueden encontrar, por ejemplo, reglas de decisión que usen colores para mostrar diferentes estados o árboles de decisión que reflejen modelos predictivos para clasificar instancias o individuos.

Este componente permite a los usuarios la construcción, de forma más sencilla, de modelos de minería y la ejecución de análisis descriptivo. El componente corresponde también al ambiente de modelado (interacción), permite al usuario especificar los problemas del contexto educativo que quiere analizar y de forma asistida generar el modelo de minería desde sus especificaciones.

El componente de interacción y visualización se conecta con el componente de preparación de datos y con el componente de representación del dominio. Es un eje central a partir del cual se asiste la interacción y comunicación entre los otros componentes. En la Figura 21 se detalla el comportamiento de este y los flujos de interacción con los otros componentes.

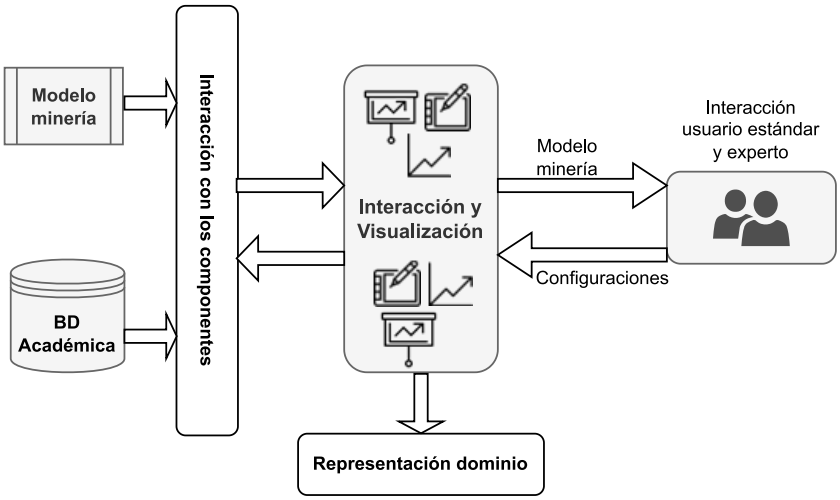


Figura 21. Detalle del componente de interacción y visualización
Fuente: elaboración propia

Anteriormente se hizo una descripción de los componentes asociados al modelo propuesto desde una visión conceptual. El enfoque de dominio específico apoya el modelo de minería de datos educativos en el sentido que puede enriquecer la experiencia de los usuarios al hacer uso del modelo propuesto por medio del conocimiento rescatado de este dominio de datos en particular. Partiendo del modelo conceptual presentado, se hace una formalización para el alcance determinado en esta tesis, es decir, atendiendo al aprendizaje supervisado y para datos generados en la educación básica y media presencial en Colombia, que se presenta en el capítulo siguiente a través de las estrategias de aplicación del modelo.

4.3 Aportes del modelo

Los modelos de minería de datos educativos sólo son útiles si pueden ser interpretados en el contexto de la interacción del estudiante con el sistema. Por ello, para aumentar la posibilidad de hallar modelos de minería útiles por medio del modelo propuesto se buscó extender el proceso o metodología tradicional con representaciones del problema y del dominio específico e incluir la experticia en la detección de configuraciones, parámetros, métodos, técnicas y demás elementos asociados a partir de los usuarios expertos y de conocimiento previo del dominio. El modelo reconoce que el usuario (estándar y experto) debe estar activamente involucrado en el diseño de nuevas representaciones y la búsqueda de los patrones.

Se ratifica que descubrir características útiles en los datos no es una tarea simple puesto que los datos pueden ser una caja negra nada fácil de descifrar, entonces,

para devolver modelos de minería y conocimiento útil a los usuarios, debe seguirse un proceso consistente, en el cual se sigan una serie de pasos asistidos. En el modelo propuesto se adaptaron los pasos de un proceso de minería de datos tradicional con particularidades del dominio y con estrategias de representación del conocimiento y entendimiento del contexto educativo. A continuación, se presentan algunas recomendaciones para abordar los problemas de minería de datos educativos a partir del uso del conocimiento del dominio:

- Definir claramente el problema a resolver: Antes de aplicar técnicas de minería de datos, es importante tener una comprensión clara del problema educativo a resolver. El conocimiento del dominio puede ayudar a identificar las variables relevantes y definir los objetivos de la minería de datos.

- Seleccionar las variables relevantes: Una vez que se ha definido el problema, es necesario identificar las variables relevantes para el análisis. El conocimiento del dominio puede ayudar a seleccionar las variables más importantes para el análisis.

- Realizar una exploración de datos exhaustiva: Antes de aplicar técnicas de minería de datos, es necesario explorar los datos para comprender su estructura y calidad. El conocimiento del dominio puede ayudar a interpretar los datos y a identificar posibles problemas o errores en los mismos.

- Seleccionar técnicas de minería de datos adecuadas: Una vez que se han definido las variables y se ha explorado los datos, es importante seleccionar las técnicas de minería de datos más adecuadas para el problema. El conocimiento del dominio puede ayudar a identificar las técnicas más relevantes y adecuadas para el análisis.

- Interpretar los resultados: Finalmente, es importante interpretar los resultados de la minería de datos para tomar decisiones informadas. El conocimiento del dominio puede ayudar a interpretar los resultados y a tomar decisiones más efectivas y precisas.

Utilizar el conocimiento del dominio en minería de datos educativos puede ayudar a reducir el tiempo de ensayo y error al proporcionar una guía para enfocar el análisis y tomar decisiones informadas. Al tener una comprensión clara del problema educativo y las variables relevantes, se puede reducir la exploración innecesaria de datos y enfocarse en las técnicas de minería de datos más

adecuadas. Además, el conocimiento del dominio puede ayudar a interpretar los resultados de manera más efectiva y tomar decisiones más precisas.

Conjuntamente con lo ya mencionado, el modelo de dominio específico para minería de datos educativos propuesto en este capítulo tiene varias ventajas y contribuciones clave en comparación con un modelo genérico. Por una parte, la posibilidad de mejora en la adaptación al contexto específico puesto que ha sido diseñado y ajustado para abordar desafíos y características particulares del campo particular. Esto permite una mejor adaptación al contexto, ya que pueden capturar patrones y relaciones específicas que un modelo genérico podría pasar por alto. Por otra parte, el modelo contribuye a lograr una mayor precisión y rendimiento de los modelos de minería como tal, al centrarse en un dominio específico, estos modelos pueden optimizarse para las características particulares de los datos provenientes del contexto educativo. Esto conduce a un rendimiento más preciso y eficiente en comparación con un modelo genérico que no está especializado en datos educativos.

También, cabe considerar la reducción de ruido e inconsistencias en los resultados, el modelo de dominio específico puede incorporar conocimientos de expertos y reglas del dominio, lo que ayuda a filtrar el ruido y las variables irrelevantes, viéndose esto reflejado también en la consistencia y confiabilidad de los resultados. En este orden de ideas, en términos de interpretación, al incluir el dominio específico en todos los componentes, se proporcionan elementos que permiten que el usuario pueda llegar a interpretaciones más claras y explicaciones sobre los resultados obtenidos, aspecto esencial en aplicaciones donde se requiere comprensión y confianza en el proceso de toma de decisiones.

Finalmente, el modelo de dominio específico puede soportar la incorporación de conocimiento de expertos externos, reglas del dominio y características específicas que pueden no ser capturadas por modelos genéricos. Esto facilita la integración con el conocimiento existente en el campo de aplicación y le puede dar la capacidad de adaptación a cambios en el mismo dominio, por ejemplo, un cambio en una escala o la definición de una nueva política educativa. Se deja abierta la posibilidad para actualizarse y ajustarse de manera más rápida y efectiva para reflejar nuevas tendencias, regulaciones o situaciones específicas. Utilizar el conocimiento del dominio puede ayudar a reducir el tiempo y los recursos invertidos en el proceso de minería de datos educativos y a aumentar la efectividad del análisis.

Resumen del capítulo

En este capítulo se presentó la justificación del modelo propuesto, el modelo conceptual, la descripción de cada uno de sus componentes y los principales aportes de este. El modelo conceptual busca abarcar las problemáticas generales de la EDM y sus fuentes de datos. Posteriormente, en el capítulo siguiente, se realizará una formalización a través de estrategias de aplicación particulares y acordes con el alcance de esta tesis. El modelo consta de cuatro componentes principales: componente de preparación de datos, componente de representación del dominio, componente de minería de datos y componente de interacción y visualización.

En resumen, incluyendo el conocimiento del dominio a las etapas o componentes de un proceso de minería de datos educativos, se puede reducir el tiempo y esfuerzo dedicado a la experimentación a ensayo y error al enfocarse en los aspectos más relevantes del análisis de datos y aplicar técnicas de análisis de datos adecuadas para identificar patrones y tendencias en los datos de manera efectiva. Los aportes del modelo también son traídos a colación y con esto se da cumplimiento al objetivo específico 4.

CAPÍTULO 5 –Validación y aplicación del modelo

En este capítulo se exponen las validaciones realizadas al modelo revisando para cada componente las posibles intervenciones desde el conocimiento del dominio y de acuerdo con los datos del caso de estudio, el cual se describe también. Así mismo, se presenta la aplicación del modelo para casos particulares y la intervención que se logra realizar con el uso de transfer learning. Se finaliza con una comparación de resultados.

5.1 Validación

Para la validación del modelo se diseñaron una serie de cuestionarios para tomar el conocimiento del dominio presente en diferentes grupos de expertos asociados al medio. Esos grupos de expertos definidos fueron:

- Investigador en minería de datos
- Investigador en analíticas de aprendizaje
- Investigador en tecnologías educativas
- Funcionario del área educativa
- Docente – Directivo docente

El detalle de los cuestionarios para cada grupo se encuentra disponible en el sitio web de documentación del modelo presentado en el Anexo B. Los instrumentos se llevaron a una versión online para facilitar su divulgación y participación de dichos grupos, incluyendo una declaración de consentimiento informado del uso de los resultados y del tratamiento de datos. La búsqueda de estos expertos se realizó a través de redes de grupos de investigación y redes de eventos y publicaciones académicas en el área. Es una muestra de la ventaja que presenta tener este tipo de afiliaciones, no solo en este campo, si no en general cuando se quiere rescatar el conocimiento de un dominio de datos en particular.

Adicionalmente, con el grupo de funcionarios del área educativa y directivos docentes, se hizo un acercamiento particular con los referentes al alcance para el caso de estudio de la tesis. Es así como se sostuvieron algunas reuniones con funcionarios de la secretaria de Educación Departamental de Norte de Santander y con los directivos de algunas Instituciones Educativas de dicho departamento.

En dichas reuniones se dio a conocer el planteamiento conceptual del modelo y se logró establecer una cooperación para caracterizar y utilizar los datos de algunas de las instituciones con las que se llegó a un acuerdo para el tratamiento y anonimización de los datos, dado que en su mayoría se trata de registros de

menores de edad. De allí, se consiguió el acceso a dos fuentes de datos, una administrada por la Secretaría de Educación y que corresponde a los datos de caracterización de matrícula que son suministrados por las familias de los estudiantes al momento de hacer el ingreso a la institución (ver Tabla 15). Manteniendo el anonimato de estas instituciones, a continuación, se caracterizan en la Tabla 16.

Tabla 15. Registros matrícula sistema SIMAT por año

Año	Cantidad atributos	Cantidad registros para las IE del caso de estudio	Total de registros para las IE del departamento
2014	55	6161	146193
2015	58	5994	145196
2016	55	6028	143195
2017	57	6027	145938
2018	60	6104	148527

Fuente: Elaboración propia

Tabla 16. Caracterización Instituciones Educativas Caso de Estudio

Institución Educativa	Ubicación	Años con datos	Cantidad de estudiantes (aprox.)	Cantidad de registros(aprox.)
IE 1	Urbana	2014–2018	2,300	11500
IE 2	Rural	2014–2018	600	3000
IE 3	Rural	2014–2018	500	2500
IE 4	Urbana	2014–2018	3000	15000

Fuente: Elaboración propia

Después de aclarar las fuentes de datos tratadas en la aplicación del modelo, seguidamente se muestran las acciones llevadas a cabo como parte de la validación de cada uno de los componentes.

5.1.1 Validación del Componente de Preparación de Datos

En la Tabla 17 se presentan los datos recibidos desde el sistema de matrícula. Estos datos son recolectados por las instituciones educativas y resguardados por la secretaría de educación departamental de forma anual de acuerdo con el calendario académico. Corresponde a información del estudiante, en la cual, para efectos de una acción posterior, que fue la construcción de una taxonomía, se agrupan en cuatro dimensiones: personales, familiares, socioeconómicos y académicos e información de la institución educativa.

Tabla 17. Datos recibidos del Sistema Integrado de Matrícula -SIMAT

Atributo/Variable	Descripción	Dimensión/grupo
Año reporte	Año al que corresponde la información	Institucional
Código municipio o distrito	Códigos DANE de los Municipios o distritos (3 posiciones) donde se encuentra ubicada la institución	Institucional
Código DANE	Código DANE de la institución educativa (12 posiciones)	Institucional

Atributo/Variable	Descripción	Dimensión/grupo
Institución		
Código Sede	Código DANE que poseía la sede en el año 2001 antes de la fusión establecida por la Ley 715. Para el caso de las Instituciones Educativas del sector no oficial se repite el código DANE asignado, escrito en la variable Código DANE Institución Educativa	Institucional
Consecutivo Sede	Código generado por el DUE (Directorio Único de Establecimientos Educativos) para identificar los establecimientos involucrados en las fusiones. Se asigna el código DANE de la Institución Educativa seguido del número consecutivo de dos dígitos de cada una de las sedes que conforman la institución educativa. La sede principal o donde labora la parte administrativa se codifica como 01, la segunda 02, etc. dando lugar a un número consecutivo de acuerdo con el número de sedes que tenga la institución educativa.	Institucional
Tipo de documento	1 Cédula de Ciudadanía 2 Tarjeta de Identidad 3 Cédula de Extranjería o Identificación de Extranjería 5 Registro Civil de Nacimiento 6 Número de Identificación Personal (NIP) 7 Número Único de Identificación Personal (NUIP) 8 Número de Identificación establecido por la Secretaría de Educación 9 Certificado Cabildo 99 No definido	Estudiante-Personal
Número de documento	Número del Documento de Identidad, es una cadena de caracteres porque algunos tienen documentos extranjeros como pasaporte con letras	Estudiante-Personal
Lugar expedición documento departamento	Código DANE asignado al Departamento donde se expide el documento (2 posiciones)	Estudiante-Personal
Lugar expedición documento municipio	Código DANE asignado al Municipio donde se expide el documento (3 posiciones)	Estudiante-Personal
Apellido 1	Primer apellido del estudiante	Estudiante-Personal
Apellido 2	Segundo apellido del estudiante	Estudiante-Personal
Nombre 1	Primer nombre del estudiante	Estudiante-Personal
Nombre 2	Segundo nombre del estudiante	Estudiante-Personal
Dirección residencia	Ubicación de la residencia donde habita el estudiante	Estudiante-Familiar
Teléfono de ubicación	Número telefónico de contacto del estudiante	Estudiante-Familiar
Lugar de residencia departamento	Código DANE asignado al Departamento donde reside el estudiante (2 posiciones)	Estudiante-Familiar
Lugar de residencia municipio	Código DANE asignado al Municipio donde reside el estudiante (3 posiciones)	Estudiante-Familiar
Estrato socioeconómico del estudiante	0 Estrato 0 1 Estrato 1 2 Estrato 2 3 Estrato 3 4 Estrato 4 5 Estrato 5 6 Estrato 6	Estudiante-Socioeconómico
SISBEN	Valor asignado por el Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales. Valor entre 0 y 100, si el estudiante no tiene SISBEN se asigna el valor -1	Estudiante-Socioeconómico
Fecha de nacimiento	dd/mm/yyyy de nacimiento del estudiante	Estudiante-Personal
Lugar de nacimiento departamento	Código DANE asignado al Departamento donde nació el estudiante (2 posiciones)	Estudiante-Personal
Lugar de nacimiento municipio	Código DANE asignado al Municipio donde nació el estudiante (3 posiciones)	Estudiante-Personal
Género	F Femenino M Masculino	Estudiante-Personal
Población víctima conflicto	1 En situación de desplazamiento 2 Desvinculados de grupos armados 3 Hijos de adultos desmovilizados 4 Victimas de minas 99 No Aplica	Estudiante-Familiar
Último departamento expulsor	Código DANE asignado al Departamento de donde se expulsó (2 posiciones), para víctimas de desplazamiento	Estudiante-Familiar
Último municipio expulsor	Códigos DANE asignado al Municipio de donde se expulsó (3 posiciones), para víctimas de desplazamiento	Estudiante-Familiar

Atributo/Variable	Descripción	Dimensión/grupo
Proviene de sector privado	Estudió anteriormente en Institución educativa privada: S Si N No	Estudiante-Socioeconómico
Proviene de otro municipio	Estudió anteriormente en Institución educativa de otro municipio: S Si N No	Estudiante-Socioeconómico
Tipo discapacidad	Presenta algún tipo de discapacidad: 03 SV-Baja Visión 04 SV-Ceguera 07 Trastorno del Espectro Autista 08 DI-Cognitivo 10 Múltiple 11 Otra 12 SA-Usuario de LSC 13 SA-Usuario de Castellano 14 Sordoceguera 15 Limitación Física (Movilidad) 17 Sistémica 18 Psicosocial 19 Voz y Habla 99 No Aplica	Estudiante-Personal
Capacidades excepcionales	Presenta alguna capacidad excepcional: 1 Capacidades Excepcionales 2 Talento Científico 3 Talento Tecnológico 4 Talento Subjetivo/Artístico 5 Talento Atlético/Deportivo 6 Doble Excepcionalidad 9 No Aplica	Estudiante-Personal
Etnia	Según Anexo C: Listado Etnias	Estudiante-Socioeconómico
Resguardo	Según Anexo C: Listado Resguardos	Estudiante-Socioeconómico
Institución bienestar de origen	Nombre de la institución de bienestar familiar de origen, si proviene de una de estas	Estudiante-familiar
Jornada	La institución educativa tiene jornada académica: 1 Completa 2 Mañana 3 Tarde 4 Nocturna 5 Fin de semana 6 Única	Institucional
Carácter	El carácter de la institución educativa es: 1 Académico 2 Técnico 0 No aplica	Institucional
Especialidad	La institución educativa tiene una especialidad: 05 Académica 06 Industrial 08 Comercial 09 Pedagógica 10 Agropecuaria 16 Promoción Social 07 Otra 00 No aplica	Institucional
Grado	Nivel académico en el que se encuentra el estudiante: -2 Pre-Jardín -1 Jardín I o A o Kinder 0 Jardín II o B, Transición o Grado 0 1 Primero 2 Segundo 3 Tercero 4 Cuarto 5 Quinto 6 Sexto 7 Séptimo 8 Octavo 9 Noveno	Estudiante-académico

Atributo/Variable	Descripción	Dimensión/grupo
	10 Décimo 11 Once 12 Doce - Normal Superior 13 Trece - Normal Superior 21 Ciclo 1 Adultos 22 Ciclo 2 Adultos 23 Ciclo 3 Adultos 24 Ciclo 4 Adultos 25 Ciclo 5 Adultos 26 Ciclo 6 Adultos 99 Aceleración del Aprendizaje	
Grupo/Curso	Grupo del alumno definido en SIMAT	Estudiante-académico
Metodología	Metodología que maneja la institución educativa: 01 Educación Tradicional 02 Escuela Nueva 03 Post Primaria 04 Telesecundaria 05 SER 06 CAFAM 07 SAT 08 Etnoeducación 09 Aceleración del Aprendizaje 10 Programa Para Jóvenes en Extraedad y Adultos 11 Preescolar Escolarizado 12 Preescolar No Escolarizado/Semiescolarizado 13 SAT Presencial 14 Entorno Comunidad 15 Entorno Familiar 16 Entorno Institucional 17 Círculos de Aprendizaje 18 Media Rural 19 Transformemos 20 Grupos Juveniles Creativos 21 Modalidad Virtual Asistida UCN 22 A Crecer 23 Bachillerato Pacicultor 24 A Crecer a través de celulares 25 SENA 26 Ser Humano 27 Vamos a Poder 28 FIMACAF 29 Caminar en Secundaria 30 ESPERE 31 Escuela Indígena Intercultural de Jóvenes y Adultos - ACIN 32 UNAD 33 Formación para la Reintegración	Institucional
Subsidiado	El estudiante pertenece al régimen subsidiado: S Si N No	Estudiante-socioeconómico
Repitente	El estudiante está repitiendo el año escolar: S Si N No	Estudiante-académico
Nuevo en la Institución Educativa	El estudiante ingresa nuevo en la institución educativa: S Si N No	Estudiante-académico
Situación Académica Año Anterior	En el año escolar anterior, el estudiante: 0 No estudió Vigencia Anterior, no haber estudiado año anterior 1 Aprobó 2 Reprobó 8 No culminó Estudios	Estudiante-académico
Fuente de Recursos	Los recursos de la institución educativa provienen de: 1 SGP 2 FNR 3 Recursos adicionales presupuesto nacional MEN 4 Otros Recursos de la Nación 5 Recursos Propios	Institucional
Condición del alumno al finalizar	Al final del año escolar anterior el estudiante: 3 Desertó	Estudiante-académico

Atributo/Variable	Descripción	Dimensión/grupo
el año anterior	5 Trasladado a otra institución educativa 8 Otro motivo de retiro 9 No Aplica	
Zona en que reside el alumno	La residencia donde habita el estudiante pertenece al sector: 1 Urbano 2 Rural	Estudiante-socioeconómico
Alumno madre cabeza de familia	El estudiante está caracterizado como madre cabeza de familia: S Si N No	Estudiante-Familiar
Beneficiario hijos dependientes de Madre Cabeza de Familia	El estudiante está caracterizado como hijo de madre cabeza de familia: S Si N No	Estudiante-Familiar
Beneficiario Veteranos de la Fuerza Pública	El estudiante está caracterizado como beneficiario de veteranos de la fuerza pública: S Si N No	Estudiante-Familiar
Beneficiario Héroes de la Nación	El estudiante está caracterizado como beneficiario de héroes de la nación: S Si N No	Estudiante-Familiar
Internado	La institución educativa maneja la modalidad: 1 = Internado 2 = Semi-internado 3 = Ninguno	Institucional
Valoración desempeño	La institución educativa fue clasificada de acuerdo con su valoración de desempeño como: 1 = Superior 2 = Alto 3 = Basico 4 = Bajo	Institucional

Fuente: Adaptado del diccionario de datos SIMAT 2014-2018

Por su parte también se recibieron datos de las calificaciones/valoraciones de los estudiantes de algunas instituciones educativas públicas del departamento Norte de Santander, para los años escolares de 2014 a 2018. En la Tabla 18 se describen las asignaturas, su correspondencia en los archivos fuente y se hace también una agrupación como asignaturas iniciación, básicas, complementarias y profundización.

Tabla 18. Descripción y agrupación de asignaturas Instituciones Educativas analizadas

NOMBRE ATRIBUTO MODELO DATOS	CORRESPONDENCIAS EN LOS ARCHIVOS FUENTE	DESCRIPCIÓN ASIGNATURA	GRUPO
perym	PERYM	PERCEPCION Y MOTRICIDAD	Iniciación
prvid	PRVID - EXVOC	PROYECTO DE VIDA	Iniciación
dicog	DICOG	DIMENSIÓN COGNITIVA	Iniciación
dicom	DICOM	DIMENSIÓN COMUNICATIVA	Iniciación
dicor	DICOR	DIMENSIÓN CORPORAL	Iniciación
diesp	DIESP	DIMENSIÓN ESPIRITUAL	Iniciación
diest	DIEST	DIMENSIÓN ESTÉTICA	Iniciación
dieti	DIETI	DIMENSIÓN ETICA	Iniciación
disoc	DISOC	DIMENSIÓN SOCIOAFECTIVA	Iniciación
exagr	EXAGR	EXPLORACION AGROPECUARIA	Profundización
natur	NATUR - A.NAT - CINAI	CIENCIAS NATURALES Y EDUCACION AMBIENTAL	Básica

NOMBRE ATRIBUTO MODELO DATOS	CORRESPONDENCIAS EN LOS ARCHIVOS FUENTE	DESCRIPCIÓN ASIGNATURA	GRUPO
fisic	FISIC	FÍSICA	Complementaria
biolo	BIOLO	BIOLOGÍA	Básica
quimi	QUIMI	QUÍMICA	Complementaria
socia	SOCIA - SOCIA2 - ARSOC	CIENCIAS SOCIALES	Básica
histo	HISTO	HISTORIA	Complementaria
cecop	CECOP - CIPOL - CPOLE	CIENCIAS ECONÓMICAS Y POLÍTICAS	Complementaria
geogr	GEOGR	GEOGRAFÍA	Básica
etyva	ETYVA - EDETI	ETICA Y VALORES HUMANOS	Complementaria
filos	FILOS	FILOSOFÍA	Complementaria
human	HUMAN	HUMANIDADES	Complementaria
lengu	LENGU	LENGUA CASTELLANA	Básica
idiom	IDIOM - INGLE	IDIOMA EXTRANJERO (INGLES)	Básica
lecom	LECOM	LECTURA COMPRENSIVA	Complementaria
matem	MATEM	MATEMÁTICAS	Básica
geyes	GEYES	GEOMETRIA Y ESTADISTICA	Complementaria
tecno	TECNO	TECNOLOGÍA E INFORMÁTICA	Básica
edfis	EDFIS	EDUCACION FISICA	Básica
edart	EDART	EDUCACIÓN ARTÍSTICA	Básica
agrop	AGROP	AGROPECUARIA	Profundización
empre	EMPRE	EMPRENDIMIENTO	Profundización
compo	COMPO CONVI - CONVIVENCIA SOCIAL	COMPORTAMIENTO SOCIAL	Básica
culci	CULCI	CULTURA CIUDADANA	Profundización
relig	RELIG - EDREL	EDUCACION RELIGIOSA Y MORAL	Básica
culti	CULTI	CULTIVOS	Profundización
pecua	PECUA	PECUARIAS	Profundización
cosec	COSEC	COSECHA	Profundización
contp	CONTP	CONTROL DE PLAGAS	Profundización
cated	CATED	CATEDRA INSTITUCIONAL	Profundización
efina	EFINA	EDUCACION FINANCIERA	Profundización
agroe	AGROE	EXPLOTACION AGROPECUARIAS ECOLOGICAS	Profundización
copol	COPOL	CONSTITUCIÓN POLÍTICA	Complementaria
edamb	EDAMB	EDUCACIÓN AMBIENTAL	Complementaria
pped	P.PED	PROYECTO PEDAGOGICO DE APRENDIZAJE	Profundización
agric	AGRIC	MODULO AGRICOLA	Profundización
sagec	SAGEC	SISTEMAS AGROPECUARIOS ECOLOGICOS	Profundización

Fuente: Elaboración propia

Siguiendo lo planteado en el componente de preparación de datos, una vez recibidos, entendidos y caracterizados, se procede con el filtrado y limpieza de estos, para ello se plantean los siguientes filtros.

Filtros de validación

Para la selección y aplicación de los filtros en las fuentes de datos, se hizo un trabajo de revisión de la documentación otorgada por la Secretaría de Educación Departamental (SED), en este caso, correspondió a los diccionarios de datos del sistema de matrícula y de valoraciones. Los filtros empleados se presentan en la Tabla 19. Para todos los atributos se realizó un filtro de validación básica, pero también se construyeron algunos filtros de validación adicionales en los cuales se incluía la revisión de más de un atributo y en las que existían relaciones de dependencia.

Tabla 19. Filtros de validación básica y adicional

ATRIBUTO	FILTRO - VALIDACIÓN BÁSICA	FILTRO - VALIDACIÓN ADICIONAL
Año reporte ***year***	No puede ser nulo	
Código municipio o distrito ***cod_municipio_inst***	Deber tener 3 posiciones	
Código DANE Institución ***dane***	No puede ser nulo - Debe tener 12 posiciones	
Código Sede ***centre***	No puede ser nulo - Debe tener 12 posiciones	
Consecutivo Sede ***educational_centre_cod***	No puede ser nulo	Si la institución tiene una sola sede, debe usarse 01. Si la institución tiene más de una sede, debe continuar con 02, 03, etc
Tipo de documento ***document_type***	No puede ser nulo - Debe corresponder a alguna de las siguientes opciones: 1, 2, 3, 5, 6, 7, 8, 9 o 99	La combinación tipo, número y lugar de expedición de documento debe ser única a nivel nacional. Si el tipo es 1(CC) y 2(TI) debe ser único por número de documento.
Número de documento ***document_number***	No puede ser nulo - es una cadena de caracteres, puede contener números y letras, algunos documentos como pasaporte contienen letras	
Lugar expedición documento departamento ***issuing_state_doc***	No puede ser nulo - Debe tener 2 posiciones	La relación departamento y municipio debe estar acorde con la codificación definida por el DANE
Lugar expedición documento municipio ***issuing municipality_doc***	No puede ser nulo - Debe tener 3 posiciones	
Apellido 1 ***last_name_1***	No puede ser nulo – Solo se deben aceptar letras del alfabeto, ningún otro tipo de carácter	
Apellido 2 ***last_name_2***	Solo se deben aceptar letras del alfabeto, ningún otro tipo de carácter	
Nombre 1 ***name_1***	No puede ser nulo – Solo se deben aceptar letras del alfabeto, ningún otro tipo de carácter	
Nombre 2 ***name_2***	Solo se deben aceptar letras del alfabeto, ningún otro tipo de carácter	
Dirección residencia ***residence_address***		
Teléfono de ubicación ***telephone***	Solo se deben aceptar números, ningún otro tipo de carácter	
Lugar de residencia departamento ***state_residence***	No puede ser nulo - Debe tener 2 posiciones	La relación departamento y municipio debe estar acorde con la codificación definida por el DANE
Lugar de residencia municipio	No puede ser nulo - Debe tener 3 posiciones	

ATRIBUTO	FILTRO - VALIDACIÓN BÁSICA	FILTRO - VALIDACIÓN ADICIONAL
municipality_residence		
Estrato socioeconómico del estudiante ***social_level***	No puede ser nulo - Debe tomar algún valor entre 0 a 6	9 es para identificar los casos especiales que no tienen estrato, puede ser por pertenecer a comunidades especiales (resguardos indígenas), ciudadanos extranjeros o no registraron el campo al momento de la matrícula
SISBEN ***sisben***	Valor entre 0 y 100, si el estudiante no tiene SISBEN se asigna el valor -1	
Fecha de nacimiento ***birth***	Formato: dd/mm/yyyy - No puede ser menor a 3 años, ni mayor a 99	
Lugar de nacimiento departamento ***state_birth***	Debe tener 2 posiciones	La relación departamento y municipio debe estar acorde con la codificación definida por el DANE. Puede ser vacío si el tipo documento es 3 – Identificación de extranjería
Lugar de nacimiento municipio ***municipality_birth***	Debe tener 3 posiciones	
Género ***gender***	Solo acepta las opciones F o M	
Población víctima conflicto ***conflict-affected***	Debe corresponder a alguna de las siguientes opciones: 1, 2, 3, 4 o 99	
Último departamento expulsor ***ejector_state***	Debe tener 2 posiciones	La relación departamento y municipio debe estar acorde con la codificación definida por el DANE. Se debe diligenciar solo si *conflict-affected* (población víctima de conflicto) es 1 (en situación de desplazamiento)
Último municipio expulsor ***ejector_municipality***	Debe tener 3 posiciones	
Proviene de sector privado ***provinence_sector***	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Proviene de otro municipio ***other_municipality***	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Tipo discapacidad ***disability_type***	Debe corresponder a alguna de las siguientes opciones: 03, 04, 07, 08, 10, 11, 12, 13, 14, 15, 17, 18, 19 y 99	
Capacidades excepcionales ***exceptional_capabilities***	Debe corresponder a alguna de las siguientes opciones: 1, 2, 3, 4, o 9 Filtro inicial: 1 - Superdotado 2 - Con talento científico 3 - Con talento tecnológico 4 - Con talento subjetivo 9 - No Aplica	Filtro después de modificación SIMAT en 2016 (filtro final): 1 - Capacidades excepcionales 2 - Talento científico 3 - Talento tecnológico 4 - Talento subjetivo/Artístico 5 - Talento atlético/deportivo 6 - Doble excepcionalidad 9 - No Aplica
Etnia ***ethnic_group***	Debe ser un número, puede tomar un valor de 1 a 100 o las siguientes: 400 - ROM 200 - Negritudes 999 - Otras Etnias 0 - sin etnia	
Resguardo ***resguardo***	Debe ser un número 0 - sin resguardo	
Institución bienestar de origen ***family_welfare_center***		Solo aplica para alumnos que sean nuevos en la institución educativa (new_student = S) y que entren a grado (school_grade): -2 Pre-Jardín -1 Jardín I o A o Kinder 0 Jardín II o B, Transición o Grado 0 1 Primero
Jornada ***course_session***	Debe corresponder a alguna de las siguientes opciones: 1, 2, 3, 4, 5 o 6	
Carácter ***nature***	Debe corresponder a alguna de las siguientes opciones: 1, 2 o 0	Si el alumno no está cursando algún grado de la media

ATRIBUTO	FILTRO - VALIDACIÓN BÁSICA	FILTRO - VALIDACIÓN ADICIONAL
		(school_grade=10 u 11), el carácter debe ser 0 (no aplica), si está cursando media el carácter debe ser académica o técnica (1 o 2).
Especialidad ***specialty***	Debe corresponder a alguna de las siguientes opciones: 05, 06, 07, 08, 09, 10, 16 o 00	Solo aplica a los grados (school_grade): 10,11,12 y 13 Si es diferente de 05,06,07,08,09,10 o 16 debe ser 00
Grado ***school_grade***	Debe corresponder a alguna de las siguientes opciones: -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 21, 22, 23, 24, 25, 26, 99	Para school_grade 99 la metodología debe ser 9
Grupo/Curso ***course***	Solo se deben aceptar números, ningún otro tipo de carácter	
Metodología ***methodology***	Debe corresponder a alguna de las siguientes opciones: 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, 115, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32 o 33,	Para school_grade 99 la metodología debe ser 9
Subsidiado ***supported***	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Repitente ***repeating_student***	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Nuevo en la Institución Educativa ***new_student***	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Situación Académica Año Anterior ***academic_situation_previous_year***	Debe corresponder a alguna de las siguientes opciones: 0, 1, 2 u 8	Si Academic_situation_previous_year es igual a 8, el valor de la condición debe ser 3 o 5. Si Academic_situation_previous_year es diferente 8, el valor de la condición debe ser 9
Condición del alumno al finalizar el año anterior ***student_status_previous_year***	Debe corresponder a alguna de las siguientes opciones: 3, 5, 8 o 9	
Fuente de Recursos ***source_resources***	Debe corresponder a alguna de las siguientes opciones:1, 2, 3, 4 o 5	
Zona en que reside el alumno **zone_residence**	Debe corresponder a alguna de las siguientes opciones: 1 o 2	
Alumno madre cabeza de familia **mother_head_family**	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Beneficiario hijos dependientes de Madre Cabeza de Familia **beneficiary_mother_head**	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Beneficiario Veteranos de la Fuerza Pública **public_force_veterans**	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Beneficiario Héroes de la Nación **beneficiary_heroes_nation**	Solo acepta las opciones S (si) o N (no)	Se lleva a una codificación booleana: 0 para no y 1 para sí
Internado **internship**	Debe corresponder a alguna de las siguientes opciones: 1, 2 o 3	
Valoración desempeño **performance**	Debe corresponder a alguna de las siguientes opciones: 1, 2, 3 o 4	

Fuente: Elaboración propia

Cabe resaltar que, para algunos de los atributos, los filtros se refinaron hasta llegar a la versión presentada, puesto que realizando el filtrado se encontraban datos atípicos, sin explicación evidente y se debía recurrir a los propietarios de la fuente para entender el hecho. Un ejemplo de esto sucedió con los códigos para las capacidades excepcionales (exceptional_capabilities), para este atributo se encontró que en el año 2016 se cambió la codificación y se empezaron a ingresar al sistema de matrícula algunos adicionales, pero no se encontraban

documentados en el diccionario de datos, este hallazgo permitió rescatar datos que podían haber sido borrados por considerarse atípicos.

Para el caso de los datos provenientes del sistema de calificaciones/valoraciones, los filtros realizados corresponden a la escala de calificaciones vigente para los periodos reportados, la cual es cuantitativa, los atributos se validaron como variables continuas en el rango de 0 a 5. Adicionalmente de ese sistema se toma el año, el cual se validaba con el año del reporte de matrícula y el estado del estudiante, que corresponde al atributo tipo clase (etiqueta) para los dataset extraídos y que es de tipo binario, aprobado o reprobado; este también se convirtió a booleano, 0 para reprobado, 1 para aprobado.

Posterior al proceso de limpieza y filtrado, se prosiguió con algunas transformaciones necesarias en los datos.

Transformaciones

La transformación de datos es otro de los procesos que se guía de acuerdo con la necesidad del dominio y a las exigencias de las técnicas. Para los casos abordados de educación básica y media en Colombia, se consideraron las siguientes transformaciones. Cabe anotar que no todas aplican para todos los atributos, esto se realizó de acuerdo con la naturaleza de cada variable y el objetivo de clasificación.

La reducción de dimensionalidad se utiliza para seleccionar un subconjunto relevante de características que capturan la información más importante, lo que facilita el análisis y reduce la complejidad computacional. Esta transformación se utilizó para controlar porcentajes altos de datos faltantes en algunos de los atributos provenientes de la caracterización sociodemográfica, esto dado que no se logró establecer con la fuente la causa de los faltantes y fue preferible retirar todo el atributo que tratar de hacer una imputación en la que sería riesgoso usar estrategias como por ejemplo la moda o un promedio, ya que se podría generar sesgo en el conjunto de datos.

Otra transformación altamente usada es la normalización, este proceso implica ajustar las escalas de los atributos para que tengan una distribución común. La normalización ayuda a evitar sesgos y asegura que diferentes atributos tengan la misma importancia durante el análisis. En particular, para las calificaciones, que corresponden a valores continuos, la normalización puede ser usada, no obstante, es probable que no se evidencien diferencias significativas entre usarla o no, dado que las calificaciones estaban todas en la misma escala, pero puede ser

importante y de ayuda para cuando se trata de calificaciones que provienen de diferentes sistemas y se deben comparar o usar en conjunto.

La codificación de variables categóricas fue la transformación a la que más se acudió. Los datos educativos contienen numerosas variables categóricas, principalmente los de caracterización familiar y social, por ello, es posible que sea necesario codificarlas en forma numérica para que los algoritmos de minería de datos puedan trabajar de mejor manera. Esto puede incluir técnicas como la codificación one-hot, donde se crean variables binarias para cada categoría, fue el caso para las variables que respondían a una distribución de SI o NO.

Para los conjuntos de datos educativos una característica muy común es ser desequilibrados, lo que significa que algunas clases o categorías tienen una representación mucho mayor que otras. Es el caso, del rendimiento académico asociado a una clase de aprobación o reprobación del año escolar. En este caso, para los datos estudiados, prevalece la clase aprobado, ante esto, existen diferentes técnicas que permiten hacer submuestreo, sobre muestreo o combinación de estas dos. Una de ellas es SMOTE, algoritmo que crea nuevas instancias que se encuentren en las regiones de características subrepresentadas para aumentar la representación de la clase minoritaria y equilibrar el conjunto de datos. Es importante tener en cuenta que el balanceo de datos debe aplicarse con cuidado, ya que puede haber riesgos de sobreajuste o pérdida de información. Además, en algunos casos, dependiendo del tamaño del conjunto de datos educativos y la disponibilidad de información, es posible que no sea necesario realizar un balanceo si los algoritmos utilizados pueden manejar el desequilibrio de manera efectiva.

Profundizando en las transformaciones que pueden ser realizadas sobre los datos educativos, la normalización de las calificaciones, es decir, la conversión de las calificaciones de los alumnos a una escala común, como el rango de 0 a 100, para homogeneizar la presentación de los resultados individuales es una posibilidad. También se presenta la normalización a nivel de asignatura, como el ajuste de las calificaciones de cada asignatura para considerar la variabilidad inherente en las dificultades y complejidades de diferentes materias. Esto podría implicar z-score u otros métodos para comparar el rendimiento relativo. Y la normalización a nivel de curso, con el establecimiento de criterios de evaluación uniformes para todos los cursos, garantizando que las calificaciones reflejen el desempeño de los alumnos de manera consistente en todas las clases.

Como se aprecia, algunas de estas intervenciones incluyen intervenciones no solo en los datos si no desde momentos previos a la recolección o extracción de estos,

lo que deja cuenta de la complejidad del dominio y de las particularidades a las que se puede enfrentar quien desea realizar procesos de analítica. Otra transformación tiene que ver con la organización de datos temporales, como notas a lo largo del tiempo, en series temporales para analizar tendencias, trayectorias escolares y patrones a lo largo de los semestres o años académicos.

Desde el punto de vista ético y muy importante, sobre todo teniendo en cuenta que en el contexto se puede abarcar un público de estudiantes menores de edad, es la anonimización de datos sensibles, por medio del enmascaramiento de información para proteger la privacidad de los estudiantes, como sustituir documentos de identidad y/o nombres por identificadores anónimos.

Como último paso del componente de preparación de datos fue concebido el modelo de datos para el almacenamiento.

Estructura modelo de datos para almacenamiento

Para hacer el almacenamiento de los datos se diseñó un modelo de datos relacional con seis tablas: student, socioeconomic, academic, institution, centre y grades, (ver Figura 22). Posteriormente se creó la estructura en el motor de bases de datos PostgreSQL. Los datos fueron cargados siguiendo una secuencia definida para garantizar que las llaves primarias se poblaran en el orden requerido.

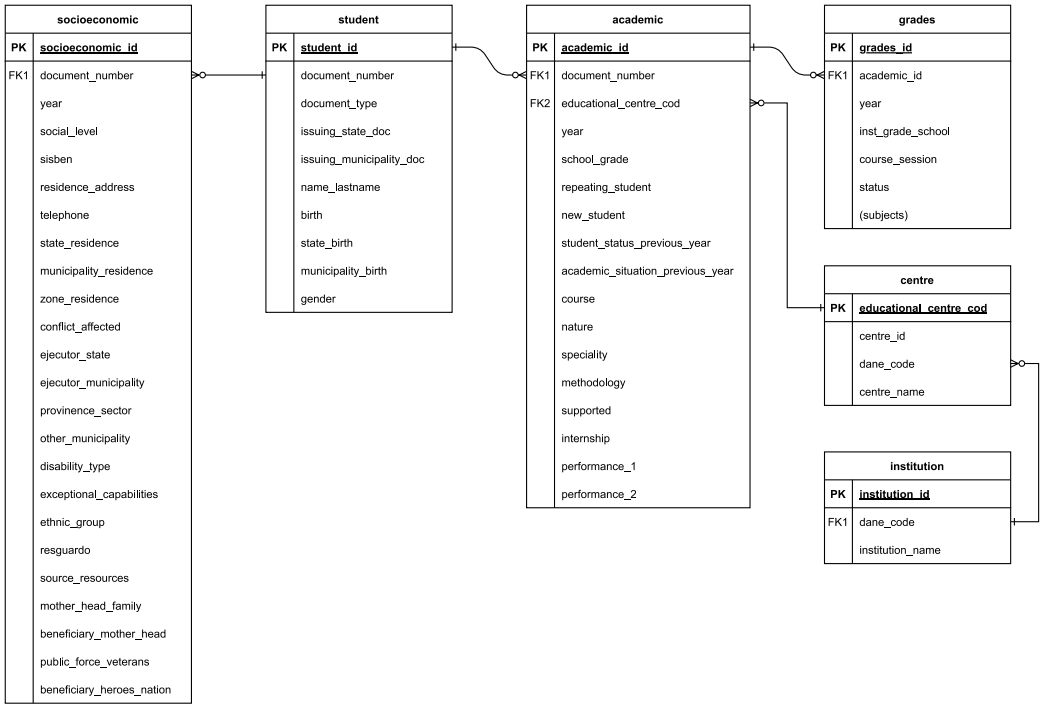


Figura 22. Modelo de datos propuesto
Fuente: elaboración propia

Después de tener todos los datos pre-procesados se realizó la carga de estos en la base de datos, para esto, se identificaron los atributos comunes entre las dos fuentes (sistema de matrícula, sistema de calificaciones/valoraciones), los cuales son referentes al centro de estudio: educational_centre_cod, centre_name, institution_name, y referentes al estudiante: name, document_number, birth y adicionalmente el atributo de año: year. Con estos atributos se realiza una función "join'' y los datos son cargados. En primer lugar, se pobló la base de datos con los provenientes del sistema de matrícula donde se encontraban los datos del estudiante e institución y posteriormente los datos provenientes del sistema de calificaciones/valoraciones.

Como resultado de esto se cargaron un total de 887.874 registros provenientes del sistema de matrícula y 11.744 registros provenientes del sistema de calificaciones, estos registros corresponden a los periodos 2014 a 2018. De estos registros cargados no todos fueron intervenidos, al igual que algunos de los atributos no requirieron de filtros adicionales, solo filtros básicos (de rangos).

Datasets extraídos para pruebas

Finalmente, como salida del componente de preparación de datos del modelo de dominio específico se plantea llegar a extraer los datasets sobre los cuales se puedan aplicar los algoritmos de clasificación, es decir, corresponden a la entrada del componente de minería de datos. Teniendo en cuenta esto, en el proceso de validación se decidió extraer tres datasets de pruebas, los cuales se describen a continuación.

Dataset 1. Datos Socioeconómicos. El primer dataset se denominó socioeconómicos, pero contiene además de los datos estudiante-socioeconómico, algunos de los correspondientes a la clasificación estudiante-personal, estudiante-académico, estudiante-familiar e institucional (clasificación introducida en la Tabla 17 de la sección 5.1.1). No se extrajo en este dataset la totalidad de los datos provenientes del sistema de matrícula dado que para las instituciones educativas de las que se recibieron datos como parte del caso de estudio no aplicaban todos, entonces, con miras a la conformación del dataset 3 (mixto) se buscó tener para el dataset 1 los datos de los estudiantes de los cuales se contaba con las calificaciones/valoraciones. Se incluye en este dataset el atributo clase (estado aprobado o reprobado). Seguidamente se tiene la lista completa de atributos incluidos:

- Social_level

- sisben
- state_residence
- municipality_residence
- zone_residence
- conflict_affected
- ejector_state
- ejector_municipality
- provinence_sector
- other_municipality
- disability_type
- exceptional_capabilities
- ethnic_group
- resguardo
- source_resources
- mother_head_family
- beneficiary_mother_head
- public_force_veterans
- beneficiary_heroes_nation
- gender
- school_grade
- repeating_student
- new_student
- student_status_previous_year
- academic_situation_previous_year
- course
- nature
- specialty
- methodology
- course_session
- status

Dataset 2. Datos Calificaciones. El segundo dataset corresponde al conformado por las calificaciones/valoraciones al finalizar cada año escolar para los estudiantes de cuatro instituciones educativas del departamento Norte de Santander, instituciones que accedieron a participar del caso de estudio. Este dataset incluye adicionalmente el grado y el estado. Para efectos de las pruebas de los algoritmos, se divide en tres grupos, grados 1 a 5 (primaria), grados 6 a 9 (secundaria) y grados 10 y 11 (media). Para efectos de este trabajo no se tuvieron en cuenta los demás grados dado su poca presencia en las instituciones

estudiadas. A continuación, se presenta la lista completa de los atributos extraídos para estos dataset, cabe aclarar, que, para efectos de tratar el problema de datos faltantes, se escogieron las calificaciones de las asignaturas que comúnmente se cursan en los grupos de grados, por ejemplo, para los grados de 6 a 11 se presenta física y química, pero estas no están para los grados 1 a 5, dado que en esos niveles no se abarcan.

Grados 1 a 5: natur, biolo, socia, geogr, lengu, idiom, matem, tecno, edfis, edart, relig, compo, status.

Grados 6 a 11: natur, fisic, biolo, quimi, socia, histo, geogr, etyva, filos, lengu, idiom, matem, tecno, edfis, edart, relig, compo, status.

Dataset 3. Mixto, Socioeconómicos y Calificaciones. Este tercer dataset corresponde a la mezcla del dataset 1 y 2, de acuerdo con el grado se unen los datos del estudiante e institución con los datos de las calificaciones/valoraciones en las asignaturas cursadas.

Con la extracción de los conjuntos de datos que servirán para la aplicación del modelo se finalizó la validación del componente de preparación de datos.

5.1.2 Validación del Componente de Representación de Conocimiento del Dominio

Teniendo en cuenta los componentes planteados en el modelo de dominio específico y en particular el conocimiento que transita por ellos, se tiene la necesidad de representar este. Para ello se parte de la generación de cuestionamientos como: ¿qué posibles taxonomías o categorizaciones de los datos educativos se pueden encontrar?, ¿cómo evaluar los datos educativos brutos?, ¿pueden estos datos educativos o sus fuentes incluir metadatos?

Como parte de la validación del componente de representación del conocimiento del dominio, inicialmente se plantea una taxonomía para los datos educativos provenientes de las fuentes ya caracterizadas. Posteriormente se hace la construcción de una base de conocimiento guía para la aplicación del modelo en los datos del caso de estudio.

Taxonomía datos educativos de educación básica y media en Colombia

Una taxonomía es un sistema de clasificación jerárquica que se utiliza para organizar y categorizar los datos según su naturaleza y características. Una taxonomía de datos educativos puede ser de gran ayuda en la representación del conocimiento de dominio específico (Lu et al., 2019). En el contexto de la minería de datos educativos, una taxonomía puede ayudar a definir las variables e indicadores de interés de manera clara y coherente, lo que a su vez facilita la selección y aplicación de las técnicas de minería de datos adecuadas para el análisis. Además, puede ayudar a identificar las relaciones y patrones entre los datos, lo que a su vez puede conducir a una mejor comprensión y descripción del conocimiento del dominio.

Por ejemplo, una taxonomía de datos educativos puede clasificar los datos según el tipo de estudiante (por ejemplo, género, edad, nivel de estudios), tipo de escuela (pública o privada), el tipo de curso (matemáticas, ciencias, idiomas) y otros factores relevantes del dominio de la educación. Al clasificar los datos de esta manera, se pueden identificar patrones y relaciones entre los diferentes tipos de datos, lo que a su vez puede ayudar a desarrollar modelos de predicción y clasificación más precisos.

Se define hacer uso de una taxonomía dado que este instrumento permite la clasificación de los conocimientos por medio de una forma arbórea donde los términos más generales se ubican en la raíz y los términos más específicos en las ramas y hojas. Adicionalmente, en la validación de este componente, la taxonomía facilita establecer un sistema para clasificar la especificidad de los datos en la medida que se recorre la estructura del árbol, mostrando las categorías de datos de interés y las relaciones entre estos.

Se presenta, a continuación, la propuesta de una taxonomía para datos provenientes de educación básica y media en Colombia, nivel educativo seleccionado en el alcance de la tesis (ver Figura 23).

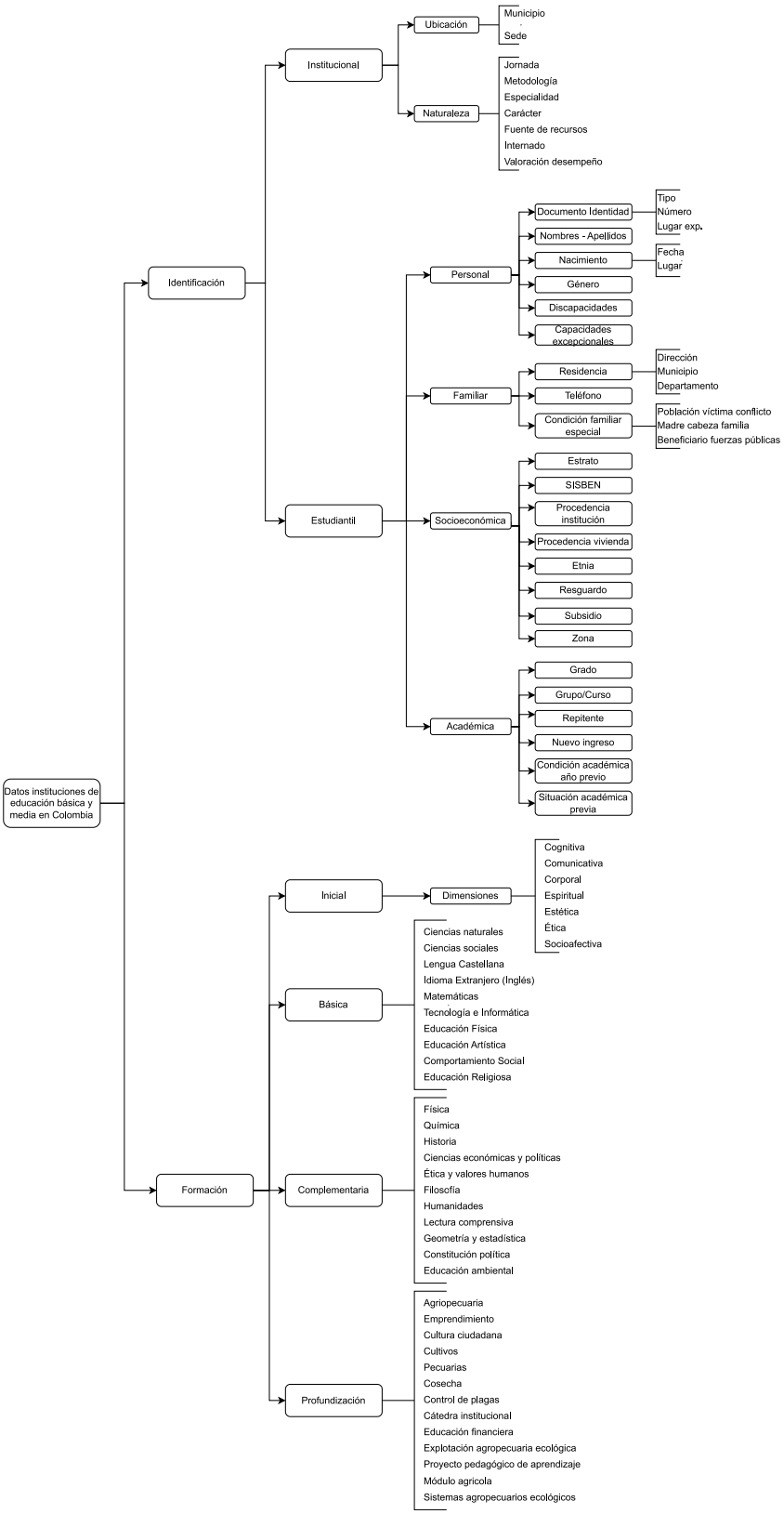


Figura 23. Taxonomía para datos de educación básica y media en Colombia

Fuente: elaboración propia

Una taxonomía de datos educativos puede ser una herramienta útil para la representación del conocimiento de dominio específico ya que puede ayudar a organizar y clasificar los datos de manera clara y coherente, lo que a su vez puede facilitar el análisis y la interpretación de los resultados.

Construcción de la base de conocimiento guía para la aplicación del modelo propuesto

Para la construcción de la base de conocimiento que permita soportar las estrategias de aplicación del modelo se puede recurrir a diferentes medios, como se menciona en el apartado 4.2.2, algunos de estos son estándares y legislación educativa, currículos, taxonomía de datos educativos, y por supuesto, el conocimiento de los expertos. Por ello, para hacer un avance en la construcción de esa base de conocimiento al nivel de los datos tratados según el alcance de este trabajo y poder hacer las pruebas del prototipo y la comparación con la herramienta de minería de datos, se llevaron a cabo varias actividades que permitieran generar una serie de preguntas guía para los procesos de análisis. Estas son:

1. Revisión de trabajos
2. Análisis de algoritmos de clasificación y sus parámetros
3. Preguntas guía
4. Reglas

Una de estas actividades consistió en recopilar desde trabajos previos, una clasificación de las técnicas que mejores resultados reportan para las problemáticas a abordar desde la minería, en este caso la deserción y el rendimiento académico. A continuación, se presentan en la Tabla 20 y Tabla 21 un recuento de trabajos donde se abordaron estas problemáticas y se destaca por medio de algunas métricas las que mejor resultado reportan.

Para el caso de la deserción en la Tabla 20, se encuentra como mejor técnica Random Forest con seis trabajos, luego el algoritmo J48, con 2 trabajos y un trabajo en el que se habla de árboles de decisión sin mencionar exactamente cuál fue el utilizado. En conclusión, los árboles de decisión se destacan por ser la técnica de clasificación con los mejores resultados, en comparación con algoritmos bayesianos, redes neuronales, máquinas de soporte vectorial y regresión.

Tabla 20. Mejores técnicas reportadas para tratar deserción

Autores	Técnicas evaluadas	Mejor técnica	Datos balanceados (SI – NO)	Tamaño dataset	Acurrancy	Área ROC	Instancias correctamente clasificadas (%)
Flores, Heras and Julián (2022)	Random Forest, Random Tree, J48, REP Tree, JRIP, OneR, Bayes Net, Naive Bayes	Random Forest	SI	4365	0.97	0.99	96.78
Behr et al. (2020)	Random Forest	Random Forest	NO	17,910	-	0.86	-
Beaulac and Rosenthal (2019)	Random Forest	Random Forest	NO	38,842	0.79	-	-
Solis et al. (2018)	Random Forest, Neural networks, SVMs, Logistic regression	Random Forest	NO	80,527	-	-	91.00
Hernández-Leal et al. (2018)	Random Tree, J48, REP Tree, JRIP, OneR	J48	NO	655	-	-	95.43
Maya et al. (2017)	Multilayer Perceptron, Random Forest, J48, Random Tree	Random Forest	NO	670	0.88	-	85.50
Miranda y Guzmán (2017)	Neural networks, Decision Trees, Bayesian Nets	Decision Trees	SI	9195	-	0.74	82.00
Torres et al. (2016)	Random Forest, ZeroR, J48, Simple CART, Naïve Bayes, Bayes Net, Multilayer Perceptron	Random Forest	NO	5547	0.89	0.91	-
Eckert and Suénaga (2015)	J48, Bayes Net (TAN), OneR	J48	NO	855	0.79	-	80.23

Fuente: Elaboración propia

Para el caso del rendimiento académico, en la Tabla 21 se muestra que son también los árboles de decisión la técnica que sobresale con mejores resultados y con algoritmos con Random Forest y J48. Es de destacar también como

característica particular de los datos de este dominio, que los datos siempre tienden a estar desbalanceados, es decir, siempre en estas dos problemáticas se tendrán más estudiantes que no reprueban y que no desertan frente a los que aprueban y desertan respectivamente, a menos de que se trate de un caso excepcional.

Tabla 21. Mejores técnicas para tratar rendimiento académico

Autores	Técnicas evaluadas	Mejor técnica	Datos balanceados (SI – NO)	Tamaño dataset	Acurrancy	Área ROC	Instancias correctamente clasificadas (%)
(Chiok & Higinio, 2017)	Regresión logística, árboles de decisión, redes neuronales y redes bayesianas	Naive Bayes	NO	914	-	0.62	71
(Ayala Franco et al., 2021)	J48, RandomForest, LMT (Logistic Model Trees), Logistic y MultilayerPerceptron	LMT	NO	415	-	0.782	71.08
(Orozco Iguasnia et al., 2021)	Árboles de decisión, Redes neuronales, Máquinas de vector de soporte	Árboles de decisión	NO	-	-	-	0.89 (R^2)
(Salgado Reyes et al., 2019)	Redes Neuronales	Redes neuronales	NO	300	-	-	75,28
(Díaz-Landa et al., 2021)	Arboles de Decisión	J48	NO	237	-	-	66
(Candia Oviedo, 2019)	Arboles de decisión J48, Random Forest, KNN, Regresión Logística y Perceptrón Multicapa	Random Forest	NO	12698	-	-	69.35
(Timarán Pereira et al., 2020)	Árboles de Decisión	J48	NO	1.361.495	-	-	66.23
(Contreras et al., 2020)	Árbol de Decisión, KNN, SVC y Perceptrón	SVC	NO	1620	0.61	-	66.24

Fuente: Elaboración propia

Continuando con revisión de los algoritmos y sus parámetros, seguidamente se presenta un mapa conceptual que incluye los datos, técnicas y consideraciones

generales respecto de los parámetros para tener en cuenta para el análisis de deserción escolar (ver Figura 24) y rendimiento académico (ver Figura 25).

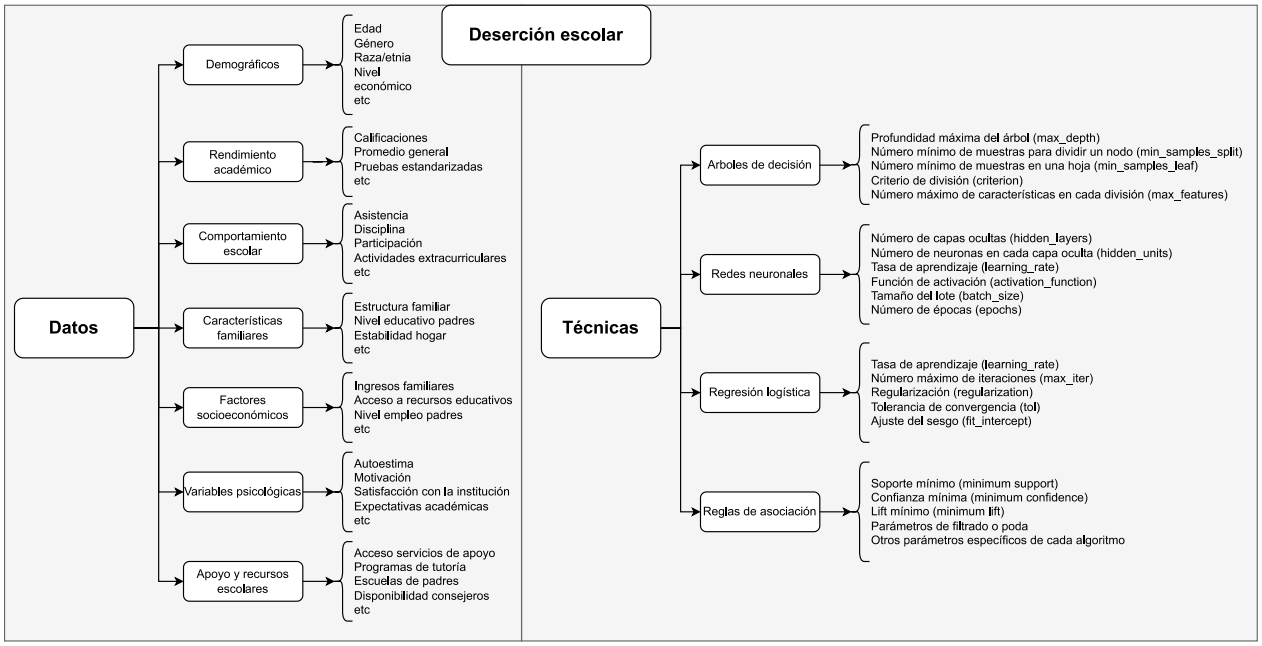


Figura 24. Datos y técnicas para tratamiento de deserción escolar con MD
Fuente: Elaboración propia

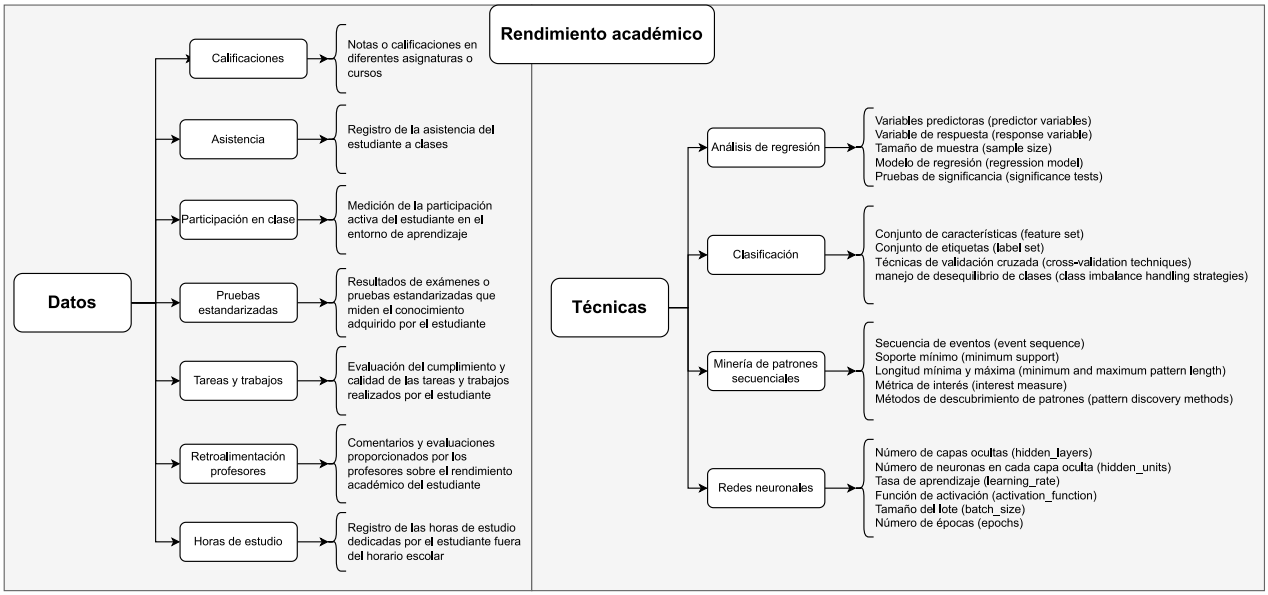


Figura 25. Datos y técnicas para tratamiento de rendimiento académico con MD
Fuente: Elaboración propia

Continuando con el análisis de algoritmos de clasificación y sus parámetros, y antes de presentar las preguntas guía, también se decidió hacer un análisis de las ventajas y desventajas de algunos algoritmos en términos de las datos soportados

y no soportados y también algunos de los usos con buenos resultados (ver Tabla 22).

Tabla 22. Comparación de algunos algoritmos de acuerdo con datos soportados y usos con buen resultado.

Técnica	Algoritmo de minería	Ventajas	Desventajas	Se ha usado con buenos resultados
Árboles de decisión	ID3 (Iterative Dichotomiser 3)	Discretos, clase binaria	Faltantes, desbalanceados	Predicción rendimiento escolar
	C4.5 (Sucesor de ID3)	Continuos y discretos, faltantes	Atípicos, desbalanceados	Predicción rendimiento escolar, estudio de deserción
	C5.0	Continuos y discretos, faltantes	Atípicos, desbalanceados	Predicción rendimiento escolar, estudio de deserción. En este, sucesor del C4.5, fue mejorado el tamaño de los árboles, intentando ramas más pequeñas con reglas más sencillas
	CART (Árboles de Clasificación y Regresión)	Continuos y discretos, robustez a atípicos	Sobreajuste. Pérdida de información al categorizar variables continuas. Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol	Predicción rendimiento escolar, deserción. Útil en exploración de datos, para identificar importancia de variables. Menos limpieza de datos: outliers y valores faltantes no influyen tanto en el modelo
	Random Forest	Mayor cantidad de variables de entrada, datos faltantes	Se recomienda para conjuntos de datos con alta dimensionalidad	Predicción rendimiento escolar, deserción. Clasificación de datos desbalanceados.
Reglas de asociación	OneR	Se les da un valor especial a los datos faltantes	Clasificador formado por reglas con una única variable en el antecedente	Se utiliza como algoritmo ase para realizar comparaciones de otros clasificadores
	ZeroR		Todas las instancias se clasifican como pertenecientes a la clase mayoritaria	Se usa como caso base para realizar comparaciones (cualquier algoritmo debería al menos igualar su rendimiento)
Redes neuronales	Multilayer perceptron	Atributos numéricos		Predicción en general

Fuente: Elaboración propia

Pasando a las preguntas guía, se establecen la siguientes, teniendo una distinción entre el análisis de rendimiento académico y el análisis de deserción escolar. Si se está interesado en aplicar minería de datos educativos para analizar la problemática de rendimiento académico en instituciones de educación básica y media en Colombia se podrían considera estos cuestionamientos:

Datos disponibles:
¿Qué tipo de datos educativos están disponibles en Colombia para analizar el rendimiento académico?
¿Cuál es la calidad y la cantidad de los datos disponibles sobre el rendimiento académico de los estudiantes?
¿Cómo se recolectan y almacenan estos datos?

Caracterización de rendimiento académico:
¿Con qué escalas se define el rendimiento académico en el contexto colombiano?
¿Existen diferentes indicadores o medidas de rendimiento académico que se deban considerar?
¿Cómo se evalúa y se califica el rendimiento académico en las instituciones de educación básica y media en Colombia?

Atributos relevantes:
¿Cuáles son las variables o factores que se cree que están relacionados con el rendimiento académico?
¿Qué variables demográficas, socioeconómicas o educativas pueden influir en el rendimiento académico de los estudiantes?
¿Existen datos disponibles sobre estas variables y en qué medida son confiables?

Métodos de análisis:
¿Qué técnicas de minería de datos son apropiadas para analizar el rendimiento académico en este contexto?
¿Cuáles son las mejores prácticas en la aplicación de minería de datos a problemas educativos similares?
¿Qué métodos estadísticos o algoritmos pueden ser útiles para identificar patrones o predictores del rendimiento académico?

Aplicabilidad y ética:
¿Cómo se pueden utilizar los resultados del análisis para informar políticas o intervenciones educativas que mejoren el rendimiento académico?
¿Qué consideraciones éticas deben tenerse en cuenta al utilizar datos educativos sensibles?
¿Cómo se pueden abordar las preocupaciones de privacidad y confidencialidad de los estudiantes al trabajar con datos educativos?

Ahora bien, para aplicar minería de datos educativos en el análisis de la problemática de deserción escolar en instituciones de educación básica y media en Colombia, algunas preguntas guía que se podrían considerar son:

Datos disponibles:
¿Qué tipo de datos educativos están disponibles en Colombia?
¿Cuál es la calidad y la cantidad de los datos disponibles sobre deserción escolar?
¿Cómo se recolectan y almacenan estos datos?

Caracterización de deserción escolar:
¿Cómo se define la deserción escolar en el contexto colombiano?
¿Existen diferentes categorías o niveles de deserción escolar que debamos considerar?
¿Qué indicadores se utilizan comúnmente para medir la deserción escolar en Colombia?

Atributos relevantes:
¿Cuáles son las variables o factores que se cree que están relacionados con la deserción escolar?
¿Qué variables demográficas, socioeconómicas o académicas pueden influir en la deserción escolar?
¿Existen datos disponibles sobre estas variables y en qué medida son confiables?

Métodos de análisis:
¿Qué técnicas de minería de datos son apropiadas para analizar la deserción escolar en este contexto?
¿Cuáles son las mejores prácticas en la aplicación de minería de datos a problemas educativos similares?
¿Qué métodos estadísticos o algoritmos pueden ser útiles para identificar patrones o predictores de deserción escolar?

Aplicabilidad y ética:
¿Cómo se pueden utilizar los resultados del análisis para informar políticas o intervenciones en el ámbito educativo?
¿Qué consideraciones éticas deben tenerse en cuenta al utilizar datos educativos sensibles?
¿Cómo se pueden abordar las preocupaciones de privacidad y confidencialidad de los estudiantes al trabajar con datos educativos?

Estas preguntas guía pueden ayudar a desarrollar una estrategia para abordar estas problemáticas utilizando minería de datos educativos y en particular el modelo de dominio específico propuesto, involucrando con esto una reflexión particular del contexto y conocimiento que implica la naturaleza de los datos analizados. Sin embargo, no se puede desconocer que se trata solo de algunos

cuestionamientos guía que deben ser adaptados y ampliados según el contexto y los recursos disponibles en Colombia.

En este orden de ideas, una vez se cuente con la extracción del conocimiento del dominio, se puede pasar a realizar una transformación de este en términos de soluciones, explicaciones e instrucciones, permitiendo tener el conocimiento relevante estructurado en un conjunto de reglas que ayuden a guiar como tal el proceso de minería de datos que conecta en el siguiente componente. De hecho, estas reglas pueden ser consideradas como hipótesis a probar. A continuación, se plantean algunas reglas que pueden servir de guía para el caso de estudio particular que se está analizando, relacionado con educación básica y media en Colombia y para las problemáticas de rendimiento académico y deserción escolar. Estas reglas podrían ser incluso calificadas de acuerdo con partes específicas del aspecto involucrado.

- Regla de la condición familiar: Las condiciones socioeconómicas del entorno familiar y social de un estudiante tienen amplia influencia en sus resultados escolares y por tanto en su rendimiento académico.
- Regla de apoyo familiar: Si un estudiante cuenta con el apoyo activo de su familia, en términos de seguimiento y supervisión, es más probable que tenga un mejor rendimiento académico.
- Regla de retroalimentación del docente: Si un estudiante recibe una retroalimentación regular y constructiva por parte de los docentes, es más probable que experimente mejoras en su rendimiento académico.
- Regla de notas anteriores: Si un estudiante ha obtenido calificaciones sobresalientes en cursos o asignaturas relacionadas previamente, es más probable que tenga un buen desempeño en cursos posteriores.
- Regla de participación en actividades extracurriculares: Si un estudiante participa regularmente en actividades extracurriculares, es más probable que tenga un buen rendimiento académico.
- Regla de tiempo de estudio: Si un estudiante dedica un valor mínimo establecido de horas diarias al estudio fuera del horario escolar, es más probable que tenga un alto rendimiento académico.
- Regla de asistencia: Si un estudiante tiene un porcentaje de asistencia inferior a una cifra establecida, es más probable que tenga un bajo rendimiento académico.

Es importante tener en cuenta también, que se debe adaptar y ajustar estas reglas según las características específicas de las instituciones y los estudiantes que se están analizando, así como las fuentes y datos disponibles. En cuanto a la otra

problemática que está siendo tomada como parte de este caso, la deserción escolar, se podría considerar:

- Regla de asistencia irregular: Si un estudiante tiene una tasa de asistencia inferior a una cifra determinada durante un período prolongado de tiempo, es más probable que esté en riesgo de deserción escolar.
- Regla de bajo rendimiento académico: Si un estudiante tiene un promedio de calificaciones inferior a la nota mínima aprobatoria según la escala usada por la institución, es más probable que esté en riesgo de abandonar la institución educativa.
- Regla de repitencia: Si un estudiante ha repetido más de dos años escolares, es más probable que esté en riesgo de abandonar la escuela.
- Regla de ausencia de apoyo familiar: Si un estudiante proviene de un entorno familiar desfavorecido, con falta de apoyo emocional, económico o educativo, es más probable que esté en riesgo de deserción escolar.
- Regla de desmotivación: Si un estudiante muestra una actitud general de desinterés hacia la educación, falta de participación en actividades escolares y falta de metas académicas claras, es más probable que esté en riesgo de abandonar la escuela.
- Regla de factores socioeconómicos: Si un estudiante pertenece a un grupo socioeconómico bajo, con dificultades económicas y falta de acceso a recursos educativos adicionales, es más probable que esté en riesgo de deserción escolar.

Cabe anotar que estas reglas son ejemplos y es importante adaptarlas a las características específicas de las instituciones y los estudiantes que se están analizando. Además, la deserción escolar, al igual que la predicción del rendimiento académico, es un fenómeno complejo que puede estar influenciado por múltiples factores interrelacionados.

Con reglas de este tipo se justifica la construcción de un sistema experto que permita hacer algunas inferencias en la aplicación de minería de datos educativos. Un sistema de este tipo puede contribuir en la automatización del proceso de análisis de datos educativos, esto permitiría realizar tareas complejas y repetitivas en un tiempo más corto y con menor margen de error que si se hicieran manualmente. Además, con un sistema experto se podría identificar de forma más rápida y precisa las tendencias y problemáticas que se cubren y cómo afrontarlas en términos de las técnicas y parametrizaciones sugeridas.

En este mismo sentido, el uso de un sistema experto en este dominio puede ayudar a los responsables de la toma de decisiones, como directores de instituciones o profesores, a obtener información valiosa y relevante para mejorar aplicar enfoques de análisis sin ser expertos en ciencia de datos ya que el sistema puede proporcionar recomendaciones basadas en reglas establecidas previamente con el conocimiento recogido del dominio. La minería de datos educativos respaldada por un sistema experto ayudaría a realizar un seguimiento continuo y evaluar la efectividad de diferentes intervenciones y análisis.

Como conclusión, una base de conocimiento relevante y estructurado llevada a un conjunto de reglas de aplicación de minería de datos educativos en Colombia puede brindar beneficios significativos, como automatización, eficiencia, toma de decisiones informadas y mejora continua.

5.1.3 Validación del Componente de Minería de Datos

Para la verificación del componente de minería de datos, se consideran dos enfoques, el primero incluye el uso de algoritmos de clasificación destacados en la revisión del componente de representación del conocimiento del dominio. Para este enfoque se hace la descripción de los algoritmos, parametrizaciones y métricas. Para los modelos construidos a partir de aprendizaje automático, la elección de hiperparámetros es un paso crucial para lograr un buen rendimiento. Para un enfoque de dominio específico, obtener buenas configuraciones de hiperparámetros es crucial, particularmente, en el dominio de minería de datos educativos, en el cual se suele estar ante situaciones de capacidades computacionales limitadas, con lo cual poder rescatar los resultados obtenidos de las tareas anteriores para una tarea completamente nueva se convertiría en un factor provechoso para los investigadores e interesados del área de estudio.

En el segundo enfoque se decide usar una técnica innovadora, se trata de *Transfer Learning*, la cual ha sido utilizada en mayor parte para casos de tratamiento y clasificación con elementos multimedia, como imágenes, audios y video, pero que aún no ha sido explorada a detalle para datos de tipo tabular. Se decide probar la transferencia de conocimiento como una forma más eficiente de configuraciones de hiperparámetros entre tareas similares. Para ello se plantea corroborar la efectividad de este enfoque mediante un estudio empírico utilizando una librería que permite implementar la transferencia para algoritmos de árboles de decisión y redes neuronales, usando un conjunto de datos extraído de un contexto educativo con datos similares a los del caso de estudio, esto se ampliará más adelante.

A modo general, para el componente de minería de datos se considera la generación de múltiples modelos predictivos y se busca optimizar el tiempo total de ajuste en lugar de centrarse en una sola tarea, para con esto poder establecer la parametrización y técnica indicada y luego poderla transferir para otros estudios del dominio de datos. Se estable que al seleccionar cuidadosamente una configuración de hiperparámetros se pueden obtener buenos resultados en la optimización en cualquier momento, manteniendo la facilidad de uso e implementación. En la Figura 26, se explica cómo se puede consolidar el aprendizaje en la selección de parámetros para tareas de minería de datos en dominios específicos.

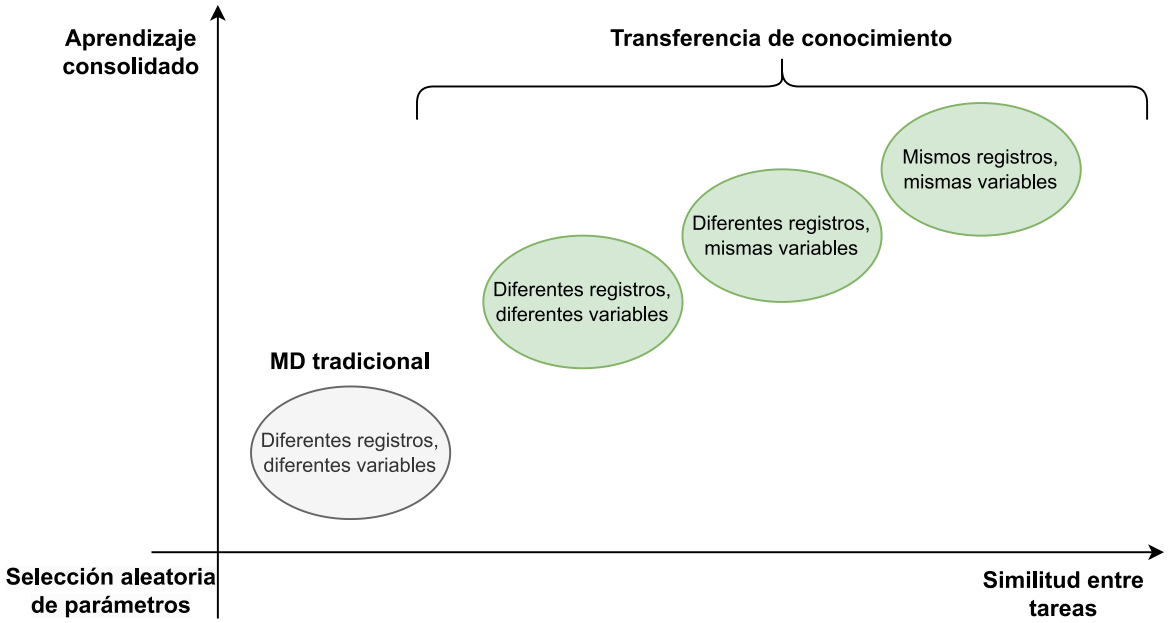


Figura 26. Selección de parámetros y transferencia de conocimiento
Fuente: Elaboración propia

Enfoque 1: Descripción de los algoritmos, parámetros y métricas

En el apartado 5.1.2 dentro de la construcción de la base de conocimiento del dominio, se presentó una revisión de técnicas y algoritmos que han mostrado buenos resultados para las problemáticas seleccionadas como parte de alcance, debido a esto y algunas razones que se presentan a continuación, se consideran para hacer la aplicación en los casos seleccionados los siguientes algoritmos: Decision Tree, Naive Bayes, KNN, Logistic Regression, SVM, Random Forest, XGBoost.

Se toman estos algoritmos para analizar problemas de deserción y rendimiento académico debido a las siguientes razones. La capacidad para manejar diferentes

tipos de datos, estos algoritmos son flexibles y pueden manejar tanto datos numéricos como categóricos, esto es importante en el contexto educativo, donde se pueden tener variables como calificaciones y datos sociodemográficos. La adaptabilidad a problemas de clasificación, los problemas de deserción y rendimiento académico a menudo se abordan como problemas de clasificación (por ejemplo, predecir si un estudiante desertará o no, o predecir el rendimiento académico en términos de aprobación o reprobación de un año escolar). La interpretación de resultados, algunos de estos algoritmos generan modelos que son sencillos de leer, lo cual es crucial para tomar decisiones y desarrollar intervenciones efectivas. La robustez frente a datos desequilibrados o ruido, los problemas de deserción y rendimiento académico a menudo implican conjuntos de datos desbalanceados, donde una clase (por ejemplo, estudiantes que aprueban) puede estar sobrerrepresentada en comparación con la otra. El rendimiento en conjuntos de datos grandes: Algunos algoritmos, como los bosques aleatorios, XGBoost y SVM, son eficientes y tienen un buen rendimiento en conjuntos de datos grandes y complejos. Finalmente, el potencial para mejora del rendimiento, los algoritmos seleccionados son ampliamente utilizados en Minería de Datos Educativos y han demostrado su eficacia en muchos casos, además, se pueden mejorar y optimizar mediante técnicas como la búsqueda de hiperparámetros o la selección de características, lo que permite ajustar su rendimiento en función del problema específico.

Seguidamente se da una breve descripción para tener una visión general de los algoritmos mencionados.

- Decisión Tree: Da como resultado un modelo que permite tomar decisiones a partir de una estructura de árbol. Cada nodo del árbol representa una característica, cada rama representa una decisión y cada hoja representa un resultado o una predicción. Los árboles de decisión son fáciles de interpretar y pueden manejar datos categóricos y numéricos.

- Naive Bayes: Es un algoritmo basado en el teorema de Bayes con la suposición de independencia condicional entre las características. Aunque esta suposición puede no ser realista en muchos casos, Naive Bayes puede ser útil como clasificador de referencia en la comparación de otros algoritmos.

- K-Nearest Neighbors (KNN): Es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión. KNN asigna una etiqueta a un punto de datos basándose en la mayoría de las etiquetas de sus vecinos más

cercanos (definido por el valor de K). Es simple de implementar, pero puede volverse computacionalmente costoso con conjuntos de datos grandes.

- Regresión logística: Es un algoritmo utilizado principalmente para problemas de clasificación binaria. Modela la relación entre las variables independientes y la probabilidad de pertenecer a una clase utilizando una función logística. La regresión logística es rápida, fácil de implementar y proporciona probabilidades interpretables.

- Máquinas de vectores de soporte (Support Vector Machines - SVM): Es un algoritmo que busca encontrar un hiperplano que separe los datos en diferentes clases de manera óptima en un espacio de alta dimensión. Las SVM son efectivas en problemas lineales y no lineales.

- Random Forest: Es un conjunto de árboles de decisión donde cada árbol se entrena con una muestra aleatoria del conjunto de datos. Las predicciones se obtienen a través de votación o promedio. Los bosques aleatorios son eficientes, robustos frente al sobreajuste y adecuados para problemas de clasificación y regresión.

- XGBoost: Es una implementación mejorada del algoritmo de Gradient Boosting. XGBoost utiliza una combinación de árboles de decisión débiles y utiliza regularización para evitar el sobreajuste.

Ahora bien, como el objetivo de este componente es lograr un modelo de clasificación a partir de minería de datos después del refinamiento de hiperparámetros logrado con el conocimiento del dominio, se seleccionaron para cada algoritmo algunos de los hiperparámetros que, soportados en la literatura, se consideran importantes a la hora de lograr mejoras en el rendimiento. Dichos hiperparámetros son los presentados en la Tabla 23.

Tabla 23. Hiperparámetros seleccionados por algoritmo

Algoritmo	Hiperparámetros
Decisión Tree	max_depth min_samples_split min_samples_leaf
Naïve Bayes	var_smoothing
KNN	n_neighbors metric weights
Logistic Regression	Penalty Tol C Solver

Algoritmo	Hiperparámetros
	Max_iter
SVM	Kernel Degree Tol Max_iter C
Random Forest	n_estimators criterion max_depth min_samples_split min_samples_leaf
XGBoost	learning_rate alpha colsample_bytree

Fuente: Elaboración propia

Ahora bien, para el testeo de los diferentes valores que pueden tomar estos hiperparámetros se puede hacer uso de Grid search, técnica utilizada para encontrar los mejores hiperparámetros para un modelo de aprendizaje automático. Dado que los hiperparámetros son configuraciones que no se aprenden directamente del conjunto de datos, sino que se establecen antes de entrenar el modelo y afectan su rendimiento y comportamiento, el Grid search consiste en definir un conjunto de posibles valores para cada hiperparámetro que se desea ajustar. Luego, se entrena y evalúa el modelo para todas las combinaciones posibles de los valores de los hiperparámetros en el conjunto. Cada configuración se evalúa utilizando alguna métrica (para este caso de clasificación) y se selecciona la configuración que obtiene el mejor rendimiento.

Esta búsqueda es una técnica exhaustiva y puede ser computacionalmente costosa, especialmente cuando se tienen muchos hiperparámetros y valores posibles. Sin embargo, es una forma sistemática y objetiva de encontrar la combinación óptima de hiperparámetros para maximizar el rendimiento del modelo.

En cuanto a las métricas, teniendo en cuenta que dentro del alcance de la tesis se definió trabajar con aprendizaje supervisado, el cual tiene como objetivo entrenar un modelo utilizando datos de entrenamiento etiquetados previamente para hacer predicciones precisas sobre los datos de prueba, las métricas utilizadas corresponden a las propias para este tipo de técnicas. A continuación, se describen brevemente las empleadas.

- Precisión: Es la proporción de predicciones correctas realizadas por el modelo en comparación con el número total de predicciones realizadas. La precisión se calcula dividiendo el número de predicciones correctas por el

número total de predicciones. Sin embargo, la precisión no siempre es la mejor métrica para evaluar la precisión de un modelo, especialmente si los datos están desequilibrados.

- Recall: Es la proporción de instancias positivas que se identifican correctamente. En otras palabras, recall mide la capacidad del modelo para encontrar todas las instancias positivas. Recall se calcula dividiendo el número de verdaderos positivos por la suma de verdaderos positivos y falsos negativos.
- F1-score: Es una medida que combina tanto la precisión como el recall en una sola métrica. F1-score se calcula como la media armónica de la precisión y el recall.
- Área Bajo la Curva ROC (AUC-ROC): Es una métrica utilizada para evaluar la capacidad del modelo para distinguir entre clases positivas y negativas. La curva ROC se crea trazando la tasa de verdaderos positivos (recall) frente a la tasa de falsos positivos (1- especificidad) en diferentes umbrales de clasificación. La AUC-ROC es el área bajo la curva ROC y proporciona una medida de la capacidad del modelo para clasificar correctamente las instancias positivas y negativas.

La elección de la métrica depende del tipo de problema y los datos utilizados. Por lo anterior también tiene lugar la inclusión del conocimiento del dominio para la selección y principalmente la interpretación de los resultados, dado que de ello también depende que se pueda dar un refinamiento de parámetros previo a la puesta en producción del modelo de minería.

Enfoque 2: Transferencia de aprendizaje

La validación por medio de este enfoque no se contempla en el alcance inicial del trabajo, sin embargo, se establece a partir de la revisión de la literatura que se consolida como un enfoque novedoso pero que aún son insuficientes las pruebas para datos de tipo tabular (Niu et al., 2020). Por lo cual se vislumbra un espacio de investigación atractivo y que puede conectar muy bien con lo planteado en el modelo de dominio específico para minería de datos educativos.

De acuerdo con esto, se considera que la Transferencia de Aprendizaje puede ser probada con datos educativos (Tsiakmaki et al., 2020), lo que implica transportarla a un tipo de datos particular (tabulares) y a partir de allí generar valor agregado a la propuesta de validación de este componente. Conceptualmente la transferencia de aprendizaje (Yang et al., 2020) incluye una fase de preentrenamiento, se utiliza un modelo de aprendizaje automático para aprender de un conjunto de datos inicial en una tarea determinada. Durante el proceso de preentrenamiento, el

modelo aprende a extraer características relevantes de los datos de entrada, estas características aprendidas son representaciones internas del modelo que capturan información útil sobre los datos. Una vez que el modelo ha sido pre-entrenado en una tarea, se puede utilizar como punto de partida para abordar una tarea relacionada. En lugar de entrenar el modelo desde cero en la nueva tarea, se toma lo aprendido en la tarea inicial y se adaptan para la nueva tarea lo que constituye el proceso de transferencia. Durante dicho proceso de transferencia, el modelo se somete a un proceso de "ajuste fino" en la nueva tarea. Esto implica tomar el modelo pre-entrenado y continuar el entrenamiento utilizando un conjunto de datos más pequeño y específico para la nueva tarea. El objetivo es ajustar los pesos y parámetros del modelo para que se adapte mejor a los nuevos datos y la nueva tarea.

Al transportar este enfoque conceptual al dominio específico, se pueden llegar a asumir varios planteamientos, por ejemplo, la transferencia del aprendizaje en la construcción del modelo para una institución de gran tamaño y transferirlo a una escala o grupo menor de estudiantes. También puede ser considerada la transferencia desde resultados o caracterización socioeconómica realizada por un ente externo hacía lo propio realizado por la Institución educativa. Lo anterior para ejemplificar algunas posibilidades de aplicación y pruebas que se plantean.

Se aclara que aún queda abierta la discusión referente a los resultados que se puedan obtener con este enfoque de transferencia (Niu et al., 2020) (Chan et al., 2023), pero se considera que es un buen mecanismo para llegar a la generación de valor que se espera de este trabajo. Dado que, el uso de transfer learning presenta notables ventajas en términos de esfuerzo y tiempo en el análisis de datos. Este enfoque permite aprovechar el conocimiento previamente adquirido por modelos entrenados en tareas específicas, transfiriendo esa información a nuevos problemas o dominios. Al utilizar modelos preentrenados como punto de partida dentro del modelo de dominio específico, se reduce significativamente la cantidad de datos y el tiempo necesario para entrenar un modelo desde cero. Además, el transfer learning permite aprovechar patrones aprendidos en conjuntos de datos masivos, lo que resulta en modelos más robustos y con un rendimiento superior, especialmente en situaciones de contextos educativos donde los datos de entrenamiento son limitados o costosos de obtener. Esta estrategia además de acelerar el desarrollo de modelos también mejora su capacidad posteriormente ser particularizados, convirtiéndolo en una herramienta valiosa en diversas aplicaciones.

Ahora bien, se seleccionan los siguientes algoritmos para probar el enfoque de transfer learning en este contexto. Dicha selección está de la mano también con el

alcance de aprendizaje supervisado de la tesis y con la selección previa de los clasificadores a utilizar.

- TransferTreeClassifier: Es un algoritmo que utiliza la técnica de transferencia de conocimiento para mejorar la clasificación. Utiliza un árbol de decisión y aprovecha el conocimiento adquirido en un dominio fuente para mejorar el rendimiento en un dominio objetivo relacionado.

- TransferForestClassifier: Similar al TransferTreeClassifier, es un algoritmo que utiliza la transferencia de conocimiento en un conjunto de bosques aleatorios. Aprovecha el conocimiento de un dominio fuente para mejorar la clasificación en un dominio objetivo relacionado.

- RegularTransferNN: el método se basa en la suposición de que se puede obtener un buen estimador objetivo adaptando los parámetros de un estimadorfuente previamente entrenado utilizando unos pocos datos objetivo etiquetados. El enfoque consiste en ajustar una red neuronal en los datos de destino de acuerdo con una función objetivo regularizada por la distancia euclidiana entre los parámetros de origen y destino.

A pesar de que no se ampliará como caso de prueba en este trabajo, se introduce también una posibilidad adicional como enfoque de validación del componente de Minería de Datos, se trata de las herramientas de AutoML (He, X., Zhao, K., & Chu, X., 2021), las cuales pueden llegar a facilitar el desarrollo de modelos predictivos y analíticos en el ámbito educativo. Lo anterior en varios aspectos, uno de ellos, la selección automatizada de modelos basadas en los datos disponibles y en los objetivos de análisis que se persigan, siendo esto un aspecto crucial, ya que, en la minería de datos educativos, los datos pueden tener características específicas que no son siempre evidentes de antemano. Así mismo, la optimización automática de los hiperparámetros del modelo sería una fase que se podría abordar, con alta utilidad en el contexto educativo, donde ajustar los parámetros adecuadamente lleva a obtener modelos precisos y generalizables.

Siguiendo con la idea anterior, en minería de datos educativos a menudo se debe trabajar con conjuntos de datos complejos que pueden contener numerosas variables con diferentes escalas y niveles de granularidad, en este sentido, las herramientas de AutoML pueden ayudar en la selección automática de características relevantes, eliminando aquellas que no contribuyen significativamente al rendimiento del modelo. Estos mismos conjuntos de datos pueden ser desbalanceados con relación a la clase objetivo (ej. rendimiento

académico), para ello las herramientas de AutoML pueden ayudar a aplicar técnicas como submuestreo, sobremuestreo o ponderación de clases (Zeineddine, H., Braendle, U., & Farah, A., 2021).

En general, para los expertos en educación pero que pueden no tener experiencia técnica profunda en minería o ciencia de datos, el AutoML podría ser un camino para llevar al aprovechamiento de las capacidades predictivas de los modelos. Esto es fundamental para la aceptación y confianza en entornos educativos. Otra posibilidad por explorar con este tipo de herramientas es la adaptación continua de los modelos a medida que se recopila nueva información. No obstante, se debe destacar que, a pesar de la automatización, la participación de expertos y su conocimiento del dominio sigue siendo esencial para garantizar que los modelos sean éticos, interpretables y alineados con los objetivos educativos específicos (Karmaker, S. et. al., 2021). Por ello, la combinación de herramientas de AutoML y un enfoque de dominio específico puede conducir a soluciones más efectivas y aplicables en el ámbito de la minería de datos educativos.

5.1.4 Validación del Componente de Interacción y Visualización

Para la validación de este componente se definen dos estrategias de interacción, una para el usuario experto y otra para el usuario estándar. Estas estrategias permiten hacer la ejecución de un proceso de minería de datos educativos de forma trasparente para el usuario (estándar y experto), buscando llegar a resultados relevantes para el contexto analizado. En las estrategias se hace una adaptación de cada uno de los componentes del modelo de manera guiada, donde dependiendo de la información con la que se cuente, primero se caracterizan y preparan los datos disponibles, luego se pasa a la selección y aplicación de las técnicas de minería y por último se presentan los resultados por medio de técnicas de visualización. Todo lo anterior apoyado en el componente de representación del dominio.

El proceso de adaptación realizado con las estrategias de interacción se basa en la disponibilidad de datos y conocimiento del dominio para la formulación de un modelo de minería que permita dar respuesta a un objetivo y problemática asociada al contexto estudiado. En la definición del alcance de esta tesis doctoral se delimitó concentrar los esfuerzos en la aplicación del modelo propuesto para un contexto educativo particular, seleccionando técnicas de aprendizaje supervisado y datos generados en educación básica y media presencial en Colombia, por ello, las estrategias son aplicadas teniendo en cuenta este alcance.

Las estrategias de interacción se basan en adaptar los diferentes componentes del modelo propuesto de forma guiada, esto quiere decir, ir recorriendo el proceso de minería de datos que ha sido intervenido con el enfoque de dominio específico y por medio del conocimiento y la caracterización que se pueda tener de los datos educativos a considerar, llevar a cabo una particularización y mediación en el análisis de estos. En consecuencia, las estrategias inician con la caracterización de la información disponible, se considera la representación del dominio para pasar a la selección de la(s) técnica(s), incluyendo esto también la determinación de algún algoritmo o algoritmos a aplicar junto con las métricas que permitan medir la calidad de los resultados y conectando también con la visualización y reporte. No obstante, teniendo en cuenta el tipo de usuario (estándar o experto) el recorrido varía en cierta medida.

Las estrategias de interacción del modelo son flexibles y están basadas en la habilidad de los dos tipos de usuarios identificados, el sistema debe facilitar a éstos la interacción sobre los componentes del modelo independiente del nivel de conocimiento y experticia de cada uno. Estas podrán ser desarrolladas a través de la representación de una serie de tareas de minería y sobre datos de varias fuentes. Buscan contribuir, por medio de la reducción del nivel de esfuerzo requerido por los usuarios, tanto estándar como experto, en la realización de los procesos de análisis y en la interpretación de los resultados. A continuación, se presenta en la Figura 27, la estrategia de interacción de acuerdo con cada uno de los perfiles o tipo de usuario.

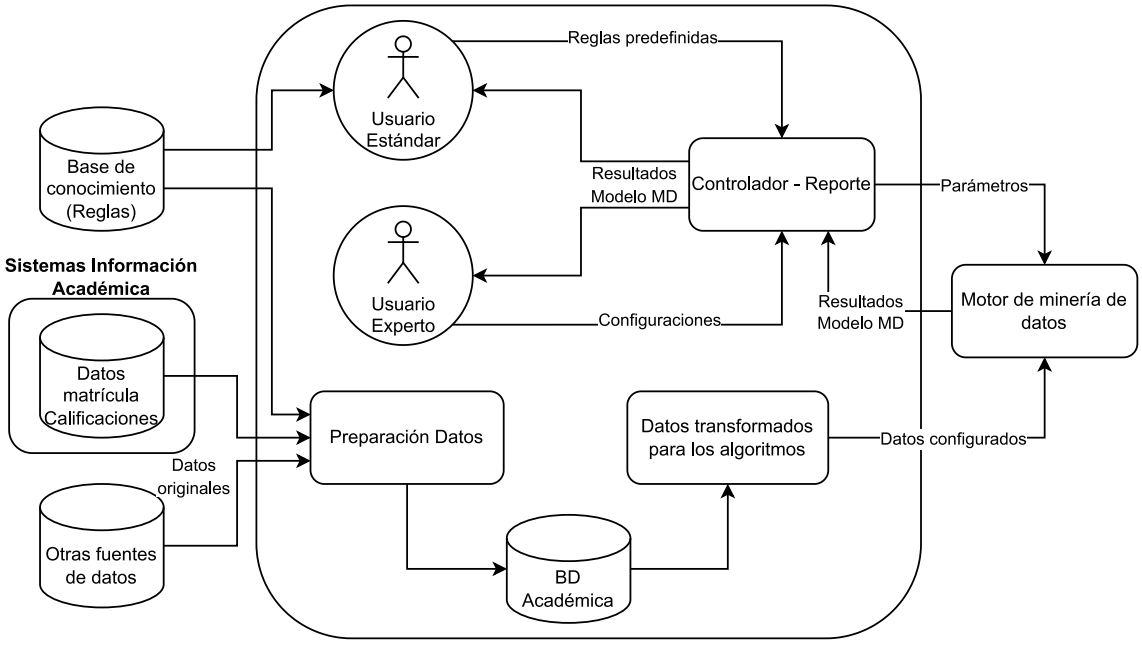


Figura 27. Representación general de las estrategias de interacción
Fuente: elaboración propia

Teniendo en cuenta la disponibilidad de información y conocimiento del dominio y de los casos o contextos educativos particulares, las estrategias de interacción del modelo permiten particularizar los escenarios y dar soporte al usuario, tanto estándar como experto, para descubrir e interpretar el conocimiento oculto en los datos del dominio. Seguidamente se describen las dos estrategias.

Estrategia de interacción para usuario estándar

La estrategia para el usuario estándar busca permitir que, de forma flexible con respecto al nivel de conocimiento y las habilidades del usuario, este pueda aplicar el modelo por medio de una particularización sobre un conjunto de datos específicos, se busca que el usuario estándar no requiera contar con una experticia o con un conocimiento previo profundo sobre minería de datos, ni sobre los parámetros de configuración de los algoritmos. Sin embargo, a pesar de que no se cuente con ese conocimiento, después de aplicar la técnica o algoritmo, se debe proveer la forma de interpretar los resultados y que dichos resultados estén adaptados a los datos y al objetivo que se está persiguiendo.

Para ello la estrategia de interacción para el usuario estándar provee una guía que a través del uso de algunas reglas predefinidas logra rescatar el conocimiento del dominio y seguir el modelo. Estas reglas se construyen posterior a que el usuario responde algunas preguntas que se concentran en la identificación de qué es lo que quiere saber o analizar este, como se mencionó en el apartado 5.1.2.

Estas reglas que pueden guiar al usuario estándar son construidas a partir del conocimiento rescatado del dominio de datos educativo y también del conocimiento que los expertos puedan aportar al modelo o del conocimiento que ha sido rescatado previamente de los procesos de minería y en particular de los procesos de minería de datos educativos. Estas reglas se pueden presentar al usuario estándar en forma de preguntas, quien de forma simple pueda escoger la regla de interés para su objetivo de minería y a partir de esto seleccionar la técnica de minería apropiada y configurar los parámetros usando las características y configuraciones previas que han sido introducidas a partir de los resultados de la representación del dominio.

En la Figura 28 se presenta un diagrama del flujo de interacción que va a tener el usuario estándar para la aplicación del modelo, se busca con esta estrategia dar al usuario estándar la habilidad de explorar y dar valor a los datos sin requerir de tener un conocimiento particular en el proceso o en la técnica de minería aplicada. Para la conformación del banco de reglas que permiten realizar la guía y generar

las reglas se plantea utilizar el conocimiento que se ha recolectado del dominio, pero a su vez se dejan a disposición algunos cuestionarios para que los usuarios expertos e interesados del dominio de datos puedan dejar y retroalimentar a partir de su experiencia, información valiosa para ir ampliando la base de conocimiento (reglas). Dichos cuestionarios se encuentran disponibles en el sitio web de documentación del modelo presentado en el Anexo B.

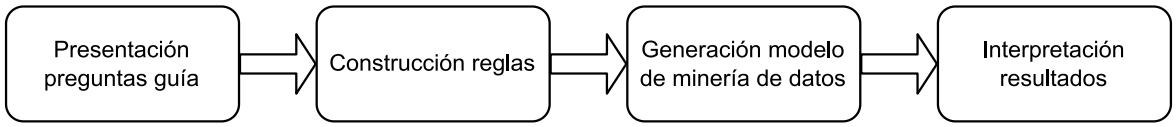


Figura 28. Flujo de interacción para el usuario estándar
Fuente: elaboración propia

Para el usuario estándar se proveen los elementos que le soportan el proceso de construcción de modelos de minería a partir de los datos educativos y también a la interpretación de los hallazgos y patrones presentes en estos. El modelo propuesto a través de la estrategia de interacción retornará para el usuario estándar los resultados del algoritmo, así como también información adicional que indica el rendimiento del algoritmo con los datos suministrados. La información adicional ayudará al usuario estándar a determinar hasta qué punto pueden confiar en la exactitud de los resultados producidos por el análisis.

Estrategia de interacción para el usuario experto

En el caso del usuario que cuenta con conocimiento previo del dominio y que tiene la habilidad de explorar diferentes configuraciones para determinar el valor que existe en los datos, los pasos involucrados en la estrategia de interacción varían un poco respecto con el usuario estándar. El usuario experto puede seguir un flujo de interacción en el cual parte de la selección de la técnica, actividad que él realiza a partir de su conocimiento; posteriormente se tiene la oportunidad de seleccionar las dimensiones de los datos o los atributos sobre los cuáles desea realizar el análisis, para posteriormente permitirle también la selección de filtros y parámetros del algoritmo. Después de estos tres aspectos se presentarán los resultados y el usuario experto tendrá la opción de evaluarlos y decidir si desea reiniciar la aplicación del modelo en busca de nuevos hallazgos. A pesar de tratarse del usuario experto éste también va a ser guiado en el proceso de modelamiento del problema de minería de datos y en la aplicación de los filtros y parámetros con el fin de permitir un buen desempeño en la tarea de minería y en la interpretación de los resultados o en el aprovechamiento del conocimiento obtenido. En la Figura 29 se muestra el recorrido en el modelo propuesto por medio de la estrategia de interacción para el usuario experto.

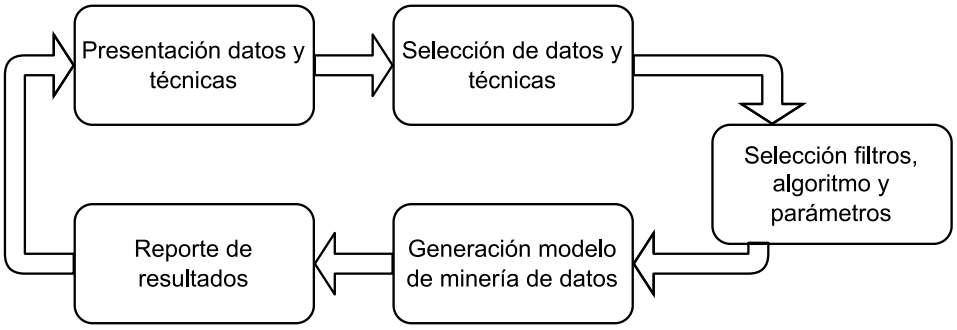


Figura 29. Flujo de interacción para el usuario experto
Fuente: elaboración propia

La estrategia para el usuario experto le provee el control granular sobre los algoritmos de minería de datos, el experto tendrá la posibilidad de elegir las opciones de configuración y los filtros que corresponden de acuerdo con la respectiva tarea. Los resultados del modelo de minería, así como los datos del rendimiento son retornados para permitir que el usuario experto determine si continuará modificando los parámetros de los algoritmos en función de descubrir nuevo conocimiento.

5.2 Aplicación del modelo

Para la aplicación del modelo se tomaron datos de instituciones educativas públicas de básica (primaria y secundaria) y media del departamento Norte de Santander como se ha indicado previamente. Para efectos de este documento y siguiendo los acuerdos de confidencialidad, dichas Instituciones Educativas serán referenciadas como IE1, IE2 e IE3. En el apartado 5.1.1 se reportó la caracterización y proceso de preparación de estos datos, así mismo, se comentó de los tres dataset extraídos para las pruebas, dichos dataset serán los usados para este apartado de aplicación y se han estructurado en tres casos respectivamente. La aplicación se realiza por medio del planteamiento de una serie de experimentos de clasificación, centrados en la clase “aprueba” o “reprueba”, la cual hace referencia al rendimiento académico. Para ello se comparan los clasificadores mencionados en el apartado 5.1.3, siendo estos Decision Tree, Naive Bayes, KNN, Logistic Regression, SVM, Random Forest, XGBoost. Y para el enfoque asociado a transfer Learning, se usarán TransferTreeClassifier y TransferForestClassifier. Para la evaluación de los resultados se utilizan como métricas las descritas en el apartado 5.1.3.

5.2.1 Caso 1: Dataset 1. Datos Socioeconómicos

El primer dataset surge de los datos que fueron suministrados por la Secretaría de Educación de Norte de Santander y contiene información de: institución educativa, identificación del estudiante, ubicación geográfica, nivel socioeconómico y académico. De este conjunto de datos, se seleccionaron solo aquellos registros relacionados con los 6.400 estudiantes presentes de las cuatro IEs de las cuales se cuenta con las calificaciones, lo anterior para conseguir hacer el mapeo para el dataset del Caso 3. En la Tabla 24 se presentan algunos elementos de análisis exploratorio para este conjunto de datos.

Tabla 24. Distribución de la clase para algunos de los atributos del Dataset 1.

Atributo	Valores/categorías	Total	Aprobados	Reprobados
Estrato	0	0.9	92.86	7.14
	1	71.39	90.79	9.21
	2	23.47	90.03	9.97
	3	4.24	90.17	9.83
	4	0.42	0.89	0.11
	5	0.079	100	0
	6	0.003	100	0
Género	Femenino (F)	50.64	92.68	7.32
	Masculino (M)	49.36	88.60	11.40
Discapacidad	Si	1.31	82.84	17.16
	No	98.69	90.79	9.21
Habilidades especiales	Si	0.09	75.00	25.00
	No	99.91	90.66	9.34
Zona residencia	Urbano	83.80	90.64	9.36
	Rural	16.20	90.61	9.39
Total			90.63	9.37

Fuente: elaboración propia

A partir del análisis exploratorio de los datos se puede identificar también algunos patrones educativos en los datos:

Patrón 1: En el grado escolar en el que se da paso del nivel primario al secundario (6° grado) se concentra el mayor porcentaje de estudiantes reprobados, en su mayoría hombres.

Patrón 2: La mayoría de los estudiantes de las IEs analizadas, las cuales dan un panorama general de las IEs de la región, provienen de familias de nivel socioeconómico bajo (estrato 1 y 2), corresponden al 94.86% del total. La proporción de estudiantes de los estratos 3, 4, 5 y 6 es muy baja, 5.14%.

Patrón 3: Hay un alto número de estudiantes reprobados que provienen de familias de nivel socioeconómico 1 (9.21%); sin embargo, en términos porcentuales, el reprobado tiene un comportamiento similar para todos los estratos sociales.

Patrón 4: Los siguientes atributos están altamente correlacionados con el estatus final de los estudiantes: nivel social (estrato), jornada escolar (mañana/tarde), área geográfica (urbana/rural), género y estatus académico del año anterior. Los anteriores atributos obtienen un F-Score superior a 150 en una escala de 0 a 300 a partir de la aplicación de *Feature ranking* con *xgboost* (ver Figura 31).

Patrón 5: La ubicación de la escuela (región de conflicto o no) parece no ser un atributo relevante para determinar el desempeño de los estudiantes para los periodos e IEs revisadas (ver Figura 31).

A la hora de correr los experimentos se utilizó Grid Search y se hizo la conexión entre la implementación en Python y la herramienta MLflow[1], la cual es una plataforma de código abierto, que permite monitorear experimentos, reproducir resultados y hacer la gestión y despliegue de modelos de aprendizaje máquina, con esta herramienta se logra también hacer una gestión del ciclo de vida de los modelos de aprendizaje, lo cual es importante para el propósito de rescatar el conocimiento y poderlo reproducir en eventos posteriores. El código de los experimentos puede ser encontrado en https://colab.research.google.com/drive/1HNDATydw6Oo3UDzJgjGiNZOFqZZAGo_K?usp=sharing

Los hiperparámetros configurados para la búsqueda del mejor modelo para cada clasificador fueron los presentados en la sección 5.1.3. A continuación se presentan, en la Tabla 25, los resultados.

Tabla 25. Resultados de los clasificadores para el Dataset del Caso 1.

Métrica	Naive Bayes	Decision Tree	KNN	Logistic Regression	SVM	Random Forest	XGBoost
Precisión	0.638	0.912	0.822	0.837	0.844	0.924	0.772
Recall	0.51	0.923	0.821	0.915	0.793	0.918	0.711
F1-Score	0.62	**0.899**	0.821	0.874	0.816	0.881	0.707
Acurracy	0.055	**0.96**	0.816	0.955	0.883	0.918	0.673
AUC-ROC	0.66	0.825	**0.906**	0.621	0.444	0.83	0.79

Fuente: elaboración propia

[1] https://mlflow.org/

Aunque varios de los algoritmos de clasificación comparados tienen resultados bastante positivos, revisando a partir de la métrica F1-Score y Acurracy, se selecciona Decision Tree como el mejor clasificador. Para este, los parámetros que lograron los resultados presentados y que corresponde al mejor modelo después de Grid Search son, best_criterion: Gini, best_max_depth: 10, best_min_samples_leaf: 1 y best_min_samples_split: 3. Así mismo, la curva ROC para este modelo se presenta en la Figura 30.

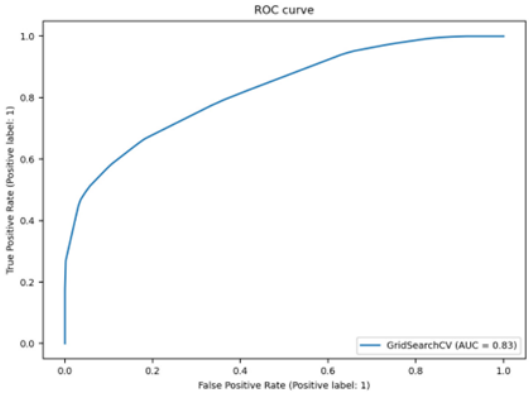


Figura 30. Curva ROC para el Desicion Tree
Fuente: elaboración propia

Adicionalmente, se quiso realizar una selección de las principales características del dataset 1 (datos socioeconómicos), para orientar en los casos que no se pueda contar con todas las variables socioeconómicas, cuáles se debería tratar de priorizar en la formación del dataset (ver Figura 31).

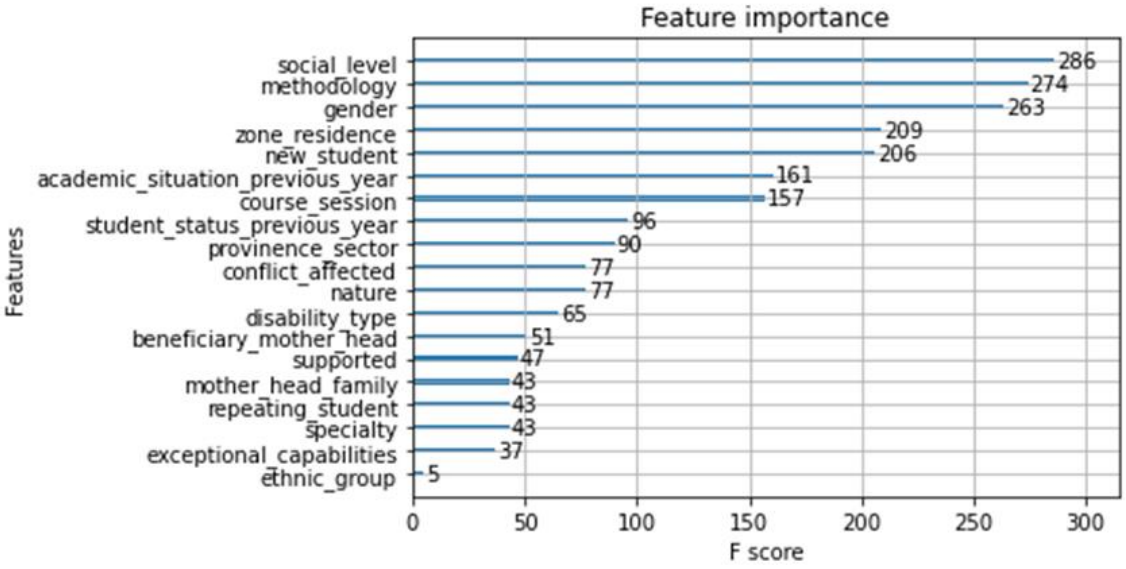


Figura 31. Selección de los atributos más relevantes dataset 1.
Fuente: elaboración propia

5.2.2 Caso 2: Dataset 2. Datos Calificaciones

El segundo dataset corresponde al conformado por las calificaciones/valoraciones al finalizar cada año escolar para los estudiantes de las cuatro IEs analizadas. Este dataset incluye adicionalmente el grado y el estado. Para efectos de las pruebas de los algoritmos, se divide en tres grupos, grados 1 a 5 (primaria), grados 6 a 9 (secundaria) y grados 10 y 11 (media). En la Tabla 26 y Tabla 27 se presentan algunos elementos de análisis exploratorio para este conjunto de datos.

Cabe aclarar que, para este caso, entendiendo que existe una alta correlación entre las calificaciones obtenidas en las asignaturas y el criterio de aprobación o reprobación del año escolar, los experimentos reportan el porcentaje de instancias correcta e incorrectamente clasificadas de cada clase, pero se presentan también los atributos, en este caso asignaturas, con mayor influencia en la generación de las reglas de los árboles de decisión. También los patrones 6, 7 y 8, asociados a este caso mencionan las disciplinas que más contribuyen a determinar el estatus final de los estudiantes para cada nivel educativo.

Tabla 26. Estadísticas para estudiantes de primaria y asignaturas principales

Atributos	Medidas de tendencia central			Medidas de dispersión				
	Media	Mediana	Moda	Varianza	Desviación estándar	Cuartiles		
						25%	50%	75%
Lenguaje	3.72	3.7	3.5	0.31	0.56	3.4	3.7	4.1
Matemáticas	3.71	3.7	3.4	0.32	0.57	3.3	3.7	4.1
Ciencias naturales	3.81	3.8	3.8	0.25	0.50	3.5	3.8	4.1
Ciencias sociales	3.80	3.8	3.6	0.25	0.50	3.5	3.8	4.1
Inglés	3.75	3.8	4	0.24	0.49	3.4	3.8	4.1
Conducta	4.47	4.6	5	0.27	0.52	4.1	4.6	4.9

Fuente: elaboración propia

Tabla 27. Estadísticas para estudiantes de secundaria y media y asignaturas principales

Atributos	Medidas de tendencia central			Medidas de dispersión				
	Media	Mediana	Moda	Varianza	Desviación estándar	Cuartiles		
						25%	50%	75%
Biología	3.51	3.50	3.20	0.30	0.55	3.2	3.5	3.8
Química	3.43	3.40	3.20	0.30	0.55	3.1	3.4	3.8
Conducta	4.03	4.00	4.00	0.37	0.60	3.6	4	4.5

Atributos	Medidas de tendencia central			Medidas de dispersión				
	Media	Mediana	Moda	Varianza	Desviación estándar	Cuartiles		
						25%	50%	75%
Inglés	3.56	3.50	3.50	0.32	0.56	3.2	3.5	3.9
Lenguaje	3.36	3.30	3.20	0.26	0.51	3.1	3.3	3.6
Matemáticas	3.30	3.30	3.00	0.25	0.50	3	3.3	3.6
Ciencias Naturales	3.49	3.50	3.20	0.24	0.49	3.2	3.5	3.8
Física	3.47	3.50	3.30	0.28	0.53	3.2	3.5	3.8
Ciencias Sociales	3.48	3.50	3.30	0.24	0.49	3.2	3.5	3.8

Fuente: elaboración propia

A partir del análisis exploratorio de los datos se puede identificar también algunos patrones educativos en los datos:

Patrón 6: El bajo rendimiento de los estudiantes se concentra principalmente en las materias de matemáticas y lenguaje. Los estudiantes que tienen bajos resultados en estas asignaturas tienden a reprobar el año, en casi todos los grados y para ambos géneros.

Para este segundo caso se realizó la ejecución con todos los clasificadores, se reporta el mejor clasificador que coincide con el del caso anterior, Decisión Tree, y se comparó contra el clasificador con los resultados menos satisfactorios, Naive Bayes. A continuación, se presentan los resultados organizados por grados escolares, dado que como se cursan asignaturas diferentes entre la primaria, secundaria y media, se tuvo que hacer esta segmentación, ver en la Tabla 28.

Tabla 28. Resultados para el Dataset del caso 2

IE	Grados	Raíz/Nodo secundario	% Registros correctamente clasificados Aprobado	% Registros correctamente clasificados Reprobado	F1 Score Decisión Tree	F1 Score Naive Bayes
IE 1	Primaria	Lenguaje/ NA	100%	87.1%	0.995	0.918
	Secundaria	Lenguaje/Biología	99%	81.2%	0.954	0.871
	Media	Matemáticas/ NA	99.8%	60.3%	0.975	0.874
IE 2	Primaria	Lenguaje/ Matemáticas	95.5%	88.4%	0.986	0.908
	Secundaria	Matemáticas/Artes	95.2%	96.1%	0.943	0.861

IE	Grados	Raíz/Nodo secundario	% Registros correctamente clasificados Aprobado	% Registros correctamente clasificados Reprobado	F1 Score Decisión Tree	F1 Score Naive Bayes
	Media	Física/Biología	97.1%	80.5%	0.969	0.857
IE 3	Primaria	Lenguaje/ Matemáticas	97.4%	86.5%	0.961	0.842
	Secundaria	Lenguaje/Biología	94.3%	90.7%	0.954	0.871
	Media	Matemáticas/ Lenguaje	97.3%	80%	0.963	0.915
IE 4	Primaria	Lenguaje/ Matemáticas	99.4%	82.2%	0.979	0.888
	Secundaria	Lenguaje/ Matemáticas	95.6%	90.1%	0.957	0.900
	Media	Ciencias Naturales/ Matemáticas	97.3%	86.7%	0.962	0.931

Fuente: elaboración propia

Patrón 7: La disciplina Lenguaje es el atributo que más contribuye a determinar el estatus final de los alumnos de Nivel Primario. Esto también es cierto para el Nivel Secundario con unas pocas excepciones, donde Matemáticas es el atributo más importante.

Patrón 8: Las asignaturas Matemáticas y Física son las disciplinas que más contribuyen a determinar el estatus final de los estudiantes de Media con una excepción, donde Ciencias Naturales es la más importante.

5.2.3 Aplicación Transfer Learning

Para aplicar el enfoque de transferencia de aprendizaje se tomó un dataset externo a los trabajados en los casos anteriores para hacer el entrenamiento y posteriormente esos modelos se transfieren para hacer la clasificación de los datos del dataset 1 (datos socioeconómicos).

Los atributos presentes en el conjunto de entrenamiento corresponden a la caracterización demográfica realizada por el Instituto Colombiano para la Evaluación de la Educación – ICFES y los resultados de las competencias genéricas medidas para los estudiantes que presentan la prueba SABER 11.

Con esta aplicación se busca detectar la posibilidad de transferir un modelo que puede ser entrenado con un conjunto mucho mayor de datos (dado que los datos del ICFES son recolectados a nivel nacional y bajo un mismo esquema centralizado de almacenamiento y de acceso libre para fines investigativo) a un conjunto menor como lo es los datos que se tienen de las cuatro IEs del caso de estudio.

Con esto surgen varios planteamientos, en un primer momento se hizo un proceso de mapeo entre las varias disponibles en los datos del ICFES y los datos de las IEs, puesto que, para los algoritmos a probar, lo único que se debe garantizar es que tanto el conjunto de entrenamiento como el de validación (transferencia) cuenten con la misma cantidad de atributos. Es así como se decidió finalmente trabajar con los atributos, ver Tabla 29.

Tabla 29. Atributos utilizados transfer learining

Atributos dataset ICFES	Atributos dataset IEs
ESTU_GENERO	gender
ESTU_TIENEETNIA	ethnic_group
FAMI_ESTRATOVIVIENDA	social_level
COLE_NATURALEZA	nature
COLE_AREA_UBICACION	zone_residence
ESTU_MCPIO_RESIDE	municipality_residence
ESTU_GENERACION-E	status

Fuente: elaboración propia

Para este caso se usaron dos algoritmos que permiten aplicar la transferencia de conocimiento y los resultados son los presentados en la Tabla 30.

Tabla 30. Resultados algoritmos Transfer Learning

Algoritmo	Acurracy	AUC-ROC	F1-Score	Precision	Recall
TransferTreeClassifier Modelo entrenado	0.694	0.747	0.687	0.693	0.694
TransferTreeClassifier Modelo transferido	0.915	0.514	0.956	0.917	0.998
TransferForestClassifier Modelo entrenado	0.74	0.806	0.736	0.739	0.74
TransferForestClassifier Modelo tranfserido	0.9147	0.500	0.955	0.915	1

Fuente: elaboración propia

Como se evidencia en los resultados, tanto para el TreeClassifer como para el ForestClassifier hay una mejora en los resultados del modelo entrenado con los datos del ICFES al modelo transferido para los datos socioeconómicos de las IEs del caso de estudio. Sin embargo, aún quedan muchos aspectos por ampliar para poder realizar afirmaciones concluyentes respecto al uso de este enfoque para

datos tabulares, dado que en la literatura también ha sido ampliamente utilizado para otros tipos de datos como los datos multimedia, pero no para tabulares.

5.2.4 Comparación con herramienta genérica

Para hacer la comparación de los resultados del proceso de minería de datos realizado siguiendo el modelo propuesto, frente a una posibilidad de solución con una herramienta disponible en el mercado, se seleccionó Orange Data Mining[2]. Orange es una herramienta que permite construir flujos de análisis de datos de forma visual con una amplia variedad de recursos; la exploración de los datos se puede hacer por medio de una interfaz gráfica que permite a usuarios principiantes ir haciendo una exploración de los datos con algoritmos de minería de datos y machine learning. Orange no se concentra en ningún dominio de datos en particular, es una herramienta genérica. La selección de esta se realiza por los motivos que se enuncian a continuación:

1. Es un software gratuito, actualmente utilizado por más de 300 universidades en todo el mundo.
2. A priori no se requieren conocimientos previos sobre minería de datos por lo cual estaría al alcance de cualquier usuario, incluso el estándar.

Los experimentos planteados incluyen:

EXP1: Dataset 1 (datos socioeconómicos), datos originales (sin preprocesamiento ni transformaciones). Se utiliza la herramienta Orange.

Tabla 31. Resultados EXP1

Algoritmo	Métricas				
	AUC	CA	F1	Precision	Recall
Naive	0,361	0,583	0,437	0,467	0,533
knn	0,399	0,683	0,557	0,591	0,633
tree	0,41	0,691	0,559	0,604	0,641
logistic regression	0,312	0,685	0,534	0,537	0,635
svm	0,137	0,519	0,452	0,548	0,469
Ramdom Forest	0,522	0,698	0,57	0,617	0,648

Fuente: elaboración propia

[2] https://orangedatamining.com/

EXP2: Dataset 1. Se tienen los datos preprocesados y con transformaciones, pero sin balanceo de datos.

Tabla 32. Resultados EXP2

Algoritmo	Métricas				
	AUC	**CA**	**F1**	**Precision**	**Recall**
Naive	0,501	0,603	0,567	0,557	0,603
knn	0,539	0,703	0,687	0,681	0,703
tree	0,55	0,711	0,689	0,694	0,711
logistic regression	0,452	0,705	0,664	0,627	0,705
svm	0,277	0,539	0,582	0,638	0,539
Ramdom Forest	0,662	0,718	0,7	0,707	0,718

Fuente: elaboración propia

EXP3: Dataset 1. Se tienen los datos preprocesados, con transformaciones y balanceo de datos.

Tabla 33. Resultados EXP3

Algoritmo	Métricas				
	AUC	**CA**	**F1**	**Precision**	**Recall**
Naive	0.611	0.713	0.677	0.667	0.713
knn	0.649	0.813	0.797	0.791	0.813
tree	0.66	0.821	0.799	0.804	0.821
logistic regression	0.562	0.815	0.774	0.737	0.815
svm	0.387	0.649	0.692	0.748	0.649
Ramdom Forest	0.772	0.828	0.81	0.817	0.828

Fuente: elaboración propia

EXP4: Dataset 1. Se tienen los datos preprocesados, con transformaciones y balanceo de datos. Se utiliza Grid Search para encontrar los parámetros óptimos.

Tabla 34. Resultados EXP4

Algoritmo	Métricas				
	AUC	**CA**	**F1**	**Precision**	**Recall**
Naive	0.66	0.55	0.62	0.638	0.51
knn	0.906	0.816	0.821	0.822	0.821
tree	0.825	0.955	0.899	0.912	0.923
logistic regression	0.621	0.955	0.874	0.837	0.915
svm	0.83	0.883	0.816	0.844	0.793
Random Forest	0.83	0.918	0.881	0.924	0.918

Fuente: elaboración propia

Se logra conseguir una mejora en los resultados al utilizar los datos preprocesados y transformados a partir de la aplicación del modelo y un leve aumento más con los resultados de los clasificadores después de aplicar Grid Search para la búsqueda de los parámetros óptimos.

Adicionalmente, se quiere dar cierre a este apartado con la presentación, en la Tabla 35, de una ampliación de la Tabla 4, en la cual se comparaban las tres principales metodologías de minería de datos, con una nueva columna en la que se incluye el modelo desarrollado en esta tesis y al que se le asigna el acrónimo de **''2DE-M''**, queriendo hacer referencia con la primera parte del acrónimo, ***2DE***, a Datos Educativos y Domino Específico, y con la ***M***, que cierra el nombre, se hace referencia a Minería.

Tabla 35. Comparativo 2DE-M, KDD, CRISP-DM y SEMMA

Modelo de proceso DM	2DE-M	KDD	CRISP-DM	SEMMA
No. Pasos, fases o componentes	4	9	6	5
Nombre de los pasos/fases/componentes	Representación del dominio	Desarrollo y entendimiento del dominio de aplicación	Entendimiento del negocio	----------
	Preparación de datos	Creación/selección de un conjunto de datos objetivo	Entendimiento de los datos	Muestra
		Limpieza y preprocesamiento	Preparación de datos	Exploración
		Transformación		Modificación
	Minería de datos	Selección de la técnica de minería de datos adecuada	Modelado	Modelo

Modelo de proceso DM	2DE-M	KDD	CRISP-DM	SEMMA
		Selección del algoritmo de minería de datos adecuado		
		Aplicación del algoritmo		
	Interacción y visualización	Interpretación de los patrones/modelo y finalmente	Evaluación	Evaluación
		Uso del conocimiento descubierto	Despliegue	----------

5.3 Aportes

La forma de descubrir la naturaleza y realizar la representación del conocimiento a partir de los datos de un dominio específico es una tarea retadora y significativa. Dar esta respuesta requiere de una gran cantidad de datos relevantes, métodos de análisis y hasta el planteamiento de nuevas teorías. No obstante, se comienza con esfuerzos y propuestas como el modelo planteado en esta tesis y la aplicación de este. En busca de hacer una contribución a este espacio de investigación abierto e ir supliendo vacíos de conocimiento de forma acumulativa, la validación planteada en este capítulo es resultado de una de las etapas más complejas enfrentadas en el desarrollo de este trabajo doctoral y que se podría considerar como una labor de construcción de unas bases de conocimiento para el planteamiento de un modelo de dominio específico de minería de datos educativos y su estrategia de aplicación, en la cual se incluyan las particularidades de los datos y las problemáticas que se deben afrontar en los procesos de análisis de fenómenos del campo educativo.

Se considera que fue una etapa compleja dada la naturaleza dinámica a la que están sujetos los entornos educativos y sus constantes cambios. Los métodos y modelos que funcionan bien en un contexto pueden no ser aplicables en otro. Esto dificulta la generalización y validación de los modelos en diferentes escenarios educativos, así mismo, los datos educativos pueden estar sujetos a variabilidad y sesgo debido a factores como la diversidad de estudiantes, diferentes niveles de granularidad y fuentes de datos heterogéneas. Esta variabilidad puede dificultar la validación de modelos y afectar su rendimiento para diferentes grupos de estudiantes.

Otra limitación corresponde a la ausencia de trabajos previos que reporten resultados de estudios utilizando minería de datos con datos del nivel educativo incluido en el alcance de este trabajo, primaria, secundaria y media, y más aún a nivel de Colombia y con datos provenientes de un entorno educativo tradicional, no asistido por plataformas de apoyo virtuales. Frente a los estudios con datos de educación superior, se identifica a la vez una oportunidad de seguir explorando esta fuente y tipología de datos y revisar opciones en términos de transferencia de aprendizaje entre niveles educativos y comparación de resultados para identificar las causas y la forma de representar esto en la construcción de los modelos de minería.

A partir de los hallazgos es importante ayudar a crear prácticas educativas y políticas sociales adaptadas a este grupo particular de estudiantes. Se puede comprometer que la concentración de reprobados está relacionada con la proximidad con la edad de ingreso a la pubertad y con los cambios que implica la transición a otro tipo de educación. El tránsito de Primaria a Secundaria representa un gran desafío para todos los actores educativos. A los niños les resulta extremadamente difícil adaptarse a las normas, estructura, estilo de enseñanza, desarrollo de tareas y otras actividades requeridas por la Secundaria.
El análisis exploratorio mediante el uso de técnicas de minería ayudó a develar que el reprobado de los estudiantes se concentra mayoritariamente en matemáticas y lenguaje (con aproximadamente un 50% de reprobación) y que el lenguaje es el atributo que más contribuye a determinar el estatus final de los alumnos tanto en primaria como en primaria y niveles secundarios. Este hallazgo corrobora información de SABER que menciona que los estudiantes presentan bajo rendimiento en lectura y escritura y que no están cumpliendo con los estándares mínimos de habilidades lingüísticas (Chica Gómez et al., 2012). Sin embargo, revela el hecho de que los alumnos que presentan tales limitaciones son también los que más fracasan en la IE.

Además, el análisis ayudó a confirmar otras variables asociadas a los problemas de rendimiento académico. Se encontró una clara relación entre el desempeño de los estudiantes y las siguientes condiciones socioeconómicas: nivel social, zona de residencia, jornada, género y nivel académico en el año anterior. Estos hallazgos corroboran estudios anteriores que mencionan que el estatus socioeconómico tiene un fuerte efecto en el rendimiento de los estudiantes en Colombia. La asociación entre el desempeño de los estudiantes y su estatus académico en el año anterior es particularmente interesante ya que permite a las escuelas seguir de cerca a los estudiantes en riesgo con un año de anticipación.

Adicionalmente, un gran aporte de este capítulo fue la implementación y aplicación de Transfer Learning para datos educativos, lograr la construcción de modelos pre-entrenados que pueden ser puestos a disposición de la comunidad académica e investigativa para profundizar y ampliar el panorama de posibilidades en cuanto a enfoques y técnicas complementarias a las genéricas de minería de datos.

Resumen del capítulo

En este capítulo se presentó un esquema de validación conceptual de cada uno de los componentes del modelo propuesto en el capítulo 4 y la aplicación del modelo para datos de instituciones educativas de básica y media de la región seleccionada como caso de estudio. Una vez realizadas las aplicaciones del modelo con los datos mencionados y con el uso de Transfer Learning, se eligió una herramienta de minería de datos presente en el mercado y de uso genérico para hacer la comparación de resultados con el uso del modelo, esto por medio de métricas de clasificación. Se finaliza el capítulo con los aportes del prototipo de validación del modelo y con ello se da cumplimiento al objetivo específico 5.

Conclusiones, recomendaciones y trabajos futuros

Conclusiones

A pesar de que la minería de datos se utiliza en diferentes campos de estudio y que en particular en la educación se ha consolidado una línea particular denominada minería de datos educativos, las metodologías y herramientas que el medio presenta están dirigidas a dominios de datos genéricos y hacer transferencia de estrategias usadas para el análisis de tipos de datos particulares requiere de un trabajo arduo. La EDM tiene como objetivo ayudar a entender por qué ocurren algunas de las problemáticas comunes a las instituciones y niveles educativos, como son el rendimiento académico, el avance en la trayectoria escolar, el alcance de las competencias o resultados de aprendizaje, la deserción, entre muchas otras. El principal objetivo de esta tesis fue hacer frente a las limitaciones encontradas en los procesos de análisis y minería de datos educativos, en particular la falta de un modelo de dominio específico para EDM que pudiese apoyar en el entendimiento de los datos y la predicción de comportamientos presentes en datos provenientes de educación básica y media en Colombia.

El modelo de dominio específico propuesto incluye cuatro componentes a saber, el componente de preparación de datos, donde se reciben las fuentes de datos y soportado en el conocimiento del dominio se construyen filtros de validación, se extraen, transforman (si es requerido) y llevan a un almacenamiento; el componente de representación del dominio, que pretende la consolidación de una base de conocimiento, reglas para soportar a los usuarios en la construcción e interpretación de los modelos de minería; el componente de minería de datos como tal, que en comunicación con el componente anterior permite llegar a un modelo de minería de datos que se pueda validar con datos propios del dominio e interpretar también soportado en la base de conocimiento; finalmente el componente de Interacción y visualización que constituye un mediador de todo el modelo propuesto. La arquitectura por componentes usada en el diseño de dicho modelo permite que se pueda seguir refinando y extendiendo con facilidad al incluir otros componentes o subcomponentes.

Para realizar la validación del modelo se utilizó un caso de estudio que permitió reflejar para cada componente la forma de utilizar el conocimiento del dominio en las tareas particulares del proceso de minería de datos. Por ejemplo, para el componente de interacción, se diseñaron dos estrategias, una pensada en usuarios estándar, entendiéndose por usuario estándar aquel que no cuenta con

un conocimiento especializado y/o profundo de minería de datos y programación, estrategia en la cual se realiza un recorrido guiado por los componentes del modelo propuesto y llegando hasta la interpretación de los resultados del modelo de minería. La segunda estrategia se enfoca en un usuario llamado experto, entendiéndose como experto aquel que tiene un dominio y conocimiento mayor de las técnicas de análisis, algoritmos y en particular de los datos del campo de estudio, por lo cual puede llevar a cabo un proceso un poco más independiente, pero persiguiendo un mismo final, la construcción e interpretación de un modelo de minería de datos. Así mismo, se logró hacer también una aplicación del enfoque Transfer Learning, como una alternativa que vislumbra posibilidades interesantes a la hora de conseguir modelos pre-entrenados y que pueden dejarse disponibles para el uso de otras instituciones o interesados del área educativa.

El modelo de dominio específico propuesto resultado de este proceso de formación doctoral realiza un aporte a la comunidad que trabaja con la EDM y en particular se concentró en llevar un caso de aplicación a datos de educación básica y media de la región Norte de Santander en Colombia, dado que esta tesis fue apoyada por la gobernación de este departamento a través de un crédito educativo condonable para la formación de capital humano de alto nivel.

Finalmente se concluye que el uso de un modelo de dominio específico soportado por unas estrategias de interacción y aplicación, ayuda a los usuarios no expertos y puede soportar también a los expertos, en el proceso de análisis de datos educativos, incluso cuando no se mejoren los resultados de los modelos de minería en términos de métricas de clasificación tradicionales, pero si respondiendo a la necesidad de llevar a cabo procesos de análisis más conscientes y en los cuales los resultados puedan ser aprovechados, ya sea a nivel institucional o regional, en la construcción de estrategias, políticas, planes de acción o de mejora, entre otros. Dado que se necesitan muchos esfuerzos en el domino específico para extraer información significativa de las interacciones y resultados de los estudiantes.

Recomendaciones

Teniendo en cuenta que la aplicación de técnicas de análisis a datos provenientes de los sistemas educativos es un ciclo iterativo de formación de hipótesis, pruebas y refinamiento, se recomienda el uso de estrategias y modelos que permitan que los miembros expertos e interesados del sistema educativo puedan participar para guiar, facilitar y mejorar el aprovechamiento de los resultados del proceso de

análisis, dado que, no se trata solo de transformar los datos en conocimiento, sino de realmente llevar el conocimiento extraído a la toma de decisiones.

Desde la aplicación del modelo sobre los datos objeto de estudio que corresponde a datos de procesos de matrícula y valoraciones de instituciones de educación básica y media del departamento Norte de Santander, surgen algunas recomendaciones para las autoridades regionales y nacionales en términos de facilitar el preprocesamiento y por ende el análisis de los datos provenientes de este nivel educativo. Dichas recomendaciones se resaltan a continuación.

- Uno de los principales factores que puede limitar los procesos de análisis y minería de datos frente a la problemática de desempeño académico es el hecho de que no existía una regla estandarizada para llegar a la condición de aprobado/reprobado, puesto que cada IE puede contar con un criterio particular.
- La ausencia de políticas claras referentes al uso y compartición de los datos educativos, aún más en el nivel educativo del caso de estudio en el cual se trata con información de menores de edad. Seguir avanzando en políticas de uso y compartición de datos abiertos es un aspecto fundamental.
- La integración de datos es otro de los grandes retos y alcances del proceso, en este sentido uno de los aportes dados a las instituciones educativas y entidades gubernamentales es definir un identificador único de estudiante que pueda ser utilizado para seguir su trayectoria dentro y entre instituciones. Este es un paso importante dado que el documento de identidad en Colombia cambia con los años y hay problemas adicionales con las diferencias en la digitalización del nombre y apellido.
- Lo anterior se une también al hecho de tener sistemas de información descentralizados en este nivel educativo, dificultando el acceso y también la integración de los datos. Dado que, diferentes partes interesadas (SED, Instituciones Educativas) conocen solo una parte de los datos. Por ello, se recomienda la creación de mecanismos que permitan una política de acceso a los datos desde escalas como la institucional.
- Otro aspecto está asociado a la propiedad de los datos y las dinámicas de datos abiertos. Esto tiene que ver con el hecho de que muchas IEs tienen sus datos administrados por empresas privadas, lo que compromete el acceso y uso de estos para propósitos de análisis. Por ello se recomienda contar con una política de protección de datos cuando estos son manejados por entidades privadas y no directamente por instituciones educativas, ya que se pierde mucha trazabilidad y se desperdician oportunidades de análisis.

- Poner a disposición modelos pre-entrenados para transferir aprendizaje permitiendo que investigadores e incluso personas no expertas puedan aplicar EDM, democratizando el uso y llevándolo a instituciones que no cuentan con expertos. Ante esto, se da un primer paso dejando a disposición los modelos logrados en la fase de validación y aplicación (http://2de-m.emilcyjuliana.com/mineria/).

Trabajos futuros

El trabajo futuro para seguir mejorando el modelo propuesto y su aplicación se estructura desde dos frentes. El primero en relación con cada uno de los componentes del modelo y el segundo referente a las estrategias de aplicación. A continuación, se detallan.

- **Componentes del modelo propuesto:**

  - Componente de preparación de datos: revisión en detalle de taxonomías para datos educativos, refinamiento de la construida para los datos del caso de estudio de forma que permita resguardar el conocimiento de este dominio. Implementación de una plataforma web que brinde estos estándares, representados en las taxonomías, ofreciendo un medio de consulta fácil y práctico, lo cual puede ser usado para lograr interoperabilidad en los sistemas de información académicos de cada nivel de estudio y en particular para el caso colombiano.
  - Componente de representación del dominio: refinar la concepción del componente en cuanto a la capacidad para asumir las múltiples granularidades de datos que se pueden dar para este dominio particular. Recoger las reglas planteadas como cierre de la construcción de la base de conocimiento en un sistema experto que permita que el usuario ingrese los datos con los que cuenta y la problemática a analizar y se le devuelva una recomendación de la técnica, algoritmos, hiperparámetros, posibles hipótesis y ruta a tomar.
  - Componente de minería de datos: Ampliar la exploración de técnicas asociadas a transferencia de aprendizaje (Transfer Learning) para revisar sus bondades para este tipo de datos y dominio. Ampliar la experimentación asociada a este enfoque y también profundizar en otros tipos de medidas de desempeño del modelo, adicionales al rendimiento de los algoritmos.

- Componente de Interacción y visualización: En la parte de visualización ampliar la revisión sobre las diferentes formas de presentar la descripción de los datos y los resultados de la predicción. Construir también una base de conocimiento sobre la forma más adecuada de presentar los resultados para el dominio de datos educativos.

- **Estrategias de interacción:**

  - Estrategia usuario estándar: Mejorar la base de preguntas guía por medio de la recolección de una mayor cantidad de conocimiento del dominio, así mismo, el refinamiento de la forma de generación de las reglas por medio del diseño e implementación de un sistema experto formal para tal fin. Además, se plantea realizar cuestionarios para validar la usabilidad del modelo propuesto por parte de usuarios no expertos.

  - Estrategia usuario experto: fortalecer la interacción con el usuario y lograr que sea aprovechado al máximo el conocimiento que pueda aportar con sus propias interacciones. Establecer un mecanismo para almacenar y compartir los modelos pre-entrenados y transferir el conocimiento que logra como usuario experto.

  - En el sitio web de documentación del modelo se presentan los enlaces a los cuestionarios diseñados para recolectar conocimiento del domino de datos educativos a través de diferentes grupos de expertos. Como parte del trabajo futuro, se plantea la validación de estos cuestionarios y la realización de un estudio particular para la compilación y publicación de respuestas por medio de un artículo científico.

Producción académica

A continuación, se presentan los productos académicos elaborados y presentados para la divulgación y discusión con la comunidad académica durante el desarrollo de este trabajo doctoral. Esta producción académica está compuesta por artículos presentados como ponencias en eventos internacionales y artículos de revista.

Artículos en revista

E.J. Hernández-Leal, N.D. Duque-Méndez & C. Cechinel, “Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions”, Heliyon, vol. 7, no. 9, 2021 https://doi.org/10.1016/j.heliyon.2021.e08017

E.J. Hernández-Leal, J. Costa & N.D. Duque-Méndez, “Educational data pre-processing from a domain-specific approach”, Respuestas, 27(1), 22–37. https://doi.org/10.22463/0122820X.3113

E.J. Hernández-Leal & N.D. Duque-Méndez, “Transferencia de aprendizaje en datos educativos: aplicación y generación de modelos preentrenados para predecir el rendimiento académico”, (por enviar)

Artículos en eventos

E. J. Hernández-Leal & N. D. Duque-Méndez, “Modelo de dominio específico para análisis y minería de datos educativos: Planteamiento modelo conceptual” Second Ibero-American Symposium of Master and Doctorate in Artificial Intelligence SIMDIA’2023 (virtual)

J.D. Atehortúa Zapata, S. Cano Duque, S. Forero Hincapié & E.J. Hernández-Leal, (2024). “Towards the Construction of an Emotion Analysis Model in University Students Using Images Taken in Classrooms”. In: Tabares, M., Vallejo, P., Suarez, B., Suarez, M., Ruiz, O., Aguilar, J. (eds) Advances in Computing. CCC 2023. Communications in Computer and Information Science, vol 1924. Springer, Cham. https://doi.org/10.1007/978-3-031-47372-2_25

E. J. Hernández-Leal & N. D. Duque-Méndez, “Towards the Proposal of a Specific Domain Model for Educational Data Mining: Problem Identification” XVI Latin American Conference on Learning Technologies LACLO, Arequipa, Perú, 2021 (virtual)

E. J. Hernández-Leal & N. D. Duque-Méndez, “Application of data mining algorithms in a set of academic data and interaction of students in virtual education platforms” III Conferencia Latinoamericana de Analíticas de Aprendizaje LALA, Cuenca, Ecuador, 2020 (virtual)

E. J. Hernández-Leal, N. D. Duque-Méndez & J. Moreno-Cadavid, “Analysis of data in an educational platform with MMOG features” II Conferencia Latinoamericana de Analíticas de Aprendizaje LALA, Valdivia, Chile, 2019

E. J. Hernández-Leal & N. Darío Duque-Méndez “Web application for the learning of Colombian Sign Language” 2018 Edutech – (LACLO), Sao Paulo, Brasil, 2018

Hernández, Emilcy; Duque, Néstor; Quintero, Diana; Escobar Naranjo, Juan & Ramirez, Juan. “Educational data mining for the analysis of student desertion”. I Conferencia Latinoamericana de Analíticas de Aprendizaje LALA, Guayaquil, Ecuador, 2018

Hernández, Emilcy & Duque, Néstor. “Modeling of the phases of the educational data mining through workflow networks”. I Conferencia Latinoamericana de Analíticas de Aprendizaje LALA, Guayaquil, Ecuador, 2018

E. J. Hernández-Leal, N. D. Duque-Méndez, M. Giraldo-Ocampo and P. A. Rodríguez-Marín, “Construction of learning objects with Augmented Reality: An experience in secondary education,” 2017 Twelfth Latin American Conference on Learning Technologies (LACLO), La Plata, 2017, pp. 1-7, doi: 10.1109/LACLO.2017.8120948

J. S. Espinosa Trejos, N. D. Duque Méndez, and E. J. Hernández-Leal, “EduTools — Authoring tool for creating HTML learning objects,” 2017 Twelfth Latin American Conference on Learning Technologies (LACLO), La Plata, 2017, pp. 1-4, doi: 10.1109/LACLO.2017.8120940

Participación en Pasantía

Institución: Universidade Federal de Santa Catarina. Campus Araranguá, Brasil
Período: agosto – noviembre de 2019

Productos académicos en desarrollo

Artículo de investigación sobre minería de datos educativos y analíticas de aprendizaje en Colombia: Análisis bibliométrico.

Artículo de investigación sobre aplicación de un modelo de minería de datos de dominio especifico con datos educativos.

Sitio web de documentación del trabajo doctoral (ver Anexo B).

Referencias

Acevedo-Díaz, J. A., García-Carmona, A., Aragón-Méndez, M. del M., & Oliva-Martínez, J. M. (2017). Modelos científicos: significado y papel en la práctica científica- Scientific models: meaning and role in scientific practice. *Revista Científica*, *3*(30), 155. https://doi.org/10.14483/23448350.12288

Agrawal, A. (2003). Metamodel based model transformation language to facilitate Domain Specific Model Driven Architecture. *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications, OOPSLA*, 118–119. https://doi.org/10.1145/949344.949379

Agt-Rickauer, H., Kutsche, R. D., & Sack, H. (2019). Automated recommendation of related model elements for domain models. Communications in Computer and Information Science, 991, 134–158. https://doi.org/10.1007/978-3-030-11030-7_7/COVER

Arslan, S., & Kardas, G. (2020). DSML4DT: A domain-specific modeling language for device tree software. Computers in Industry, 115, 103179. https://doi.org/10.1016/J.COMPIND.2019.103179

Ayala Franco, E., López Martínez, R. E., & Menéndez Domínguez, V. H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. Revista de Educación a Distancia (RED), 21(66), 1–36. https://doi.org/10.6018/RED.463561

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In *Learning Analytics* (pp. 61–75). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security. http://arxiv.org/abs/1201.3418

Begum, S. H. (2013). Data Mining Tools and Trends – An Overview. *International Journal of Emerging Research in Management &Technology*, 6–12.

Bellifemine, F., Fortino, G., Giannantonio, R., Gravina, R., Guerrieri, A., & Sgroi, M. (2011). SPINE: a domain-specific framework for rapid prototyping of WBSN applications. *Software: Practice and Experience*, *41*(3), 237–265. https://doi.org/10.1002/spe.998

Bermanis, A., Averbuh, A., & Coifman, R. (2013). Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, *34*(1), 15–29. https://doi.org/10.1016/J.ACHA.2012.03.002

Bettini, L. (2016). *Implementing domain-specific languages with Xtext and Xtend : learn how to implement a DSL with Xtext and Xtend using easy-to-understand examples and best practices* (2nd ed.). Packt Publishing.

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security*. http://arxiv.org/abs/1201.3418

Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Candia Oviedo, D. I. (2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando

algoritmos de aprendizaje automático.
Cao, L. (2010). Domain-Driven Data Mining: Challenges and Prospects. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 755–769. https://doi.org/10.1109/TKDE.2010.32
Cao, L., & Zhang, C. (2007). The Evolution of KDD: Towards Domain-Driven Data Mining. *International Journal of Pattern Recognition and Artificial Intelligence*, *21*(04), 677–692. https://doi.org/10.1142/S0218001407005612
Castellanos, C., Varela, C. A., & Correal, D. (2021). ACCORDANT: A domain specific-model and DevOps approach for big data analytics architectures. Journal of Systems and Software, 172, 110869. https://doi.org/10.1016/J.JSS.2020.110869
Castro, M., & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos lineales. Revista Española de Pedagogía, 70(251), 131–148.
Cechinel, C., Ochoa, X., Lemos dos Santos, H., Carvalho Nunes, J. B., Rodés, V., & Marques Queiroga, E. (2020). Mapping Learning Analytics initiatives in Latin America. *British Journal of Educational Technology*, 1–23. https://doi.org/10.1111/bjet.12941
Chan, J. Y. Le, Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: a review of sentiment analysis based on sequential transfer learning. Artificial Intelligence Review, 56(1), 749–780. https://doi.org/10.1007/S10462-022-10183-8/TABLES/7
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *The CRISP-DM User Guide*.
Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. In B. Hong, X. Meng, L. Chen, W. Winiwarter, & W. Song (Eds.), *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, April 22-25, 2013. Proceedings* (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40270-8_1
Chiok, M., & Higinio, C. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos - Dialnet. Anales Científicos, 78(1), 26–33.
Chen, G., Iwen, M., Chin, S., & Maggioni, M. (2012). A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements. *Visual Communications and Image Processing*, 1–6. https://doi.org/10.1109/VCIP.2012.6410789
Cios, K. J., Pedrycz, W., Swiniarski, R., & Kurgan, L. (2007). *Data mining : a knowledge discovery approach*. Springer.
Clark, T., Van Den Brand, M., Combemale, B., & Rumpe, B. (2015). Conceptual model of the globalization for domain-specific languages. In *Globalizing Domain-Specific Languages* (Vol. 9400, pp. 7–20). Springer Verlag. https://doi.org/10.1007/978-3-319-26172-0_2
Csikszentmihalyi, M., & Wolfe, R. (2014). New Conceptions and Research Approaches to Creativity: Implications of a Systems Perspective for Creativity in Education. In *The Systems Model of Creativity* (pp. 161–184). Springer Netherlands. https://doi.org/10.1007/978-94-017-9085-7_10

Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. Formación Universitaria, 13(5), 233–246. https://doi.org/10.4067/S0718-50062020000500233

ANE. (2018). *Boletín Técnico Educación Formal (EDUC) 2018*.

De Almeida Neto, F. A., & Castro, A. (2017). A reference architecture for educational data mining. *Proceedings – Frontiers in Education Conference, FIE*, *2017-October*, 1–8. https://doi.org/10.1109/FIE.2017.8190728

De Lara, J., & Guerra, E. (2012). Domain-specific textual meta-modelling languages for model driven engineering. *ECMFA'12: Proceedings of the 8th European Conference on Modelling Foundations and Applications*, *7349 LNCS*, 259–274. https://doi.org/10.1007/978-3-642-31491-9_20

Desolda, G., Ardito, C., & Matera, M. (2017). Empowering end users to customize their smart environments: Model, composition paradigms, and domain-specific tools. *ACM Transactions on Computer-Human Interaction*, *24*(2), 1–52. https://doi.org/10.1145/3057859

Díaz-Landa, B., Meleán-Romero, R., & Marín-Rodriguez, W. (2021). Rendimiento académico de estudiantes en Educación Superior: predicciones de factores influyentes a partir de árboles de decisión. Revista de Estudios Interdisciplinarios En Ciencias Sociales, 23(3), 616–639.

Dsilva, C. J., Talmon, R., Gear, C. W., Coifman, R. R., & Kevrekidis, I. G. (2015). Data-Driven Reduction for Multiscale Stochastic Dynamical Systems. *Applied Dynamical Systems*. http://arxiv.org/abs/1501.05195

Edwards, G., & Medvidovic, N. (2008). A methodology and framework for creating domain-specific development infrastructures. *ASE 2008 – 23rd IEEE/ACM International Conference on Automated Software Engineering, Proceedings*, 168–177. https://doi.org/10.1109/ASE.2008.27

El Ghosh, M., Naja, H., Abdulrab, H., & Khalil, M. (2017). Ontology learning process as a bottom-up strategy for building domain-specific ontology from legal texts. *ICAART 2017 – Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, *2*, 473–480. https://doi.org/10.5220/0006188004730480

Evans, E. (2004). *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley Professional. http://books.google.com/books?hl=en&lr=&id=7dlaMs0SECsC&pgis=1

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in knowledge discovery and data mining. In *Advances in knowledge discovery and data mining*. AAAI Press. https://dl.acm.org/citation.cfm?id=257942

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*, 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Fernandez, D. B., & Lujan-Mora, S. (2017). Comparison of applications for educational data mining in Engineering Education. *2017 IEEE World Engineering Education Conference (EDUNINE)*, 81–85. https://doi.org/10.1109/EDUNINE.2017.7918187

Gerbig, R. (2017). *Deep, seamless, multi-format, multi-notation definition and use of domain-specific languages*. Verlag Dr. Hut.

Hübscher, R., Puntambekar, S., & Nye, A. H. (2007). Domain Specific Interactive Data Mining. *Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling*, 81–90. http://www.educationaldatamining.org/UM2007/Hubscher.pdf

Hand, D. J. (David J.), Mannila, Heikki., & Smyth, Padhraic. (2001). Principles of data mining. MIT Press.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622.

Inmon, W. H., & Linstedt, D. (2014). Data architecture: a primer for the data scientist: big data, data warehouse and data vault. Morgan Kaufmann.

Iung, A., Carbonell, J., Marchezan, L., Rodrigues, E., Bernardino, M., Basso, F. P., & Medeiros, B. (2020). Systematic mapping study on domain-specific language development tools. Empirical Software Engineering, 25(5), 4205–4249. https://doi.org/10.1007/S10664-020-09872-1/TABLES/9

Jaramillo Valvuena, S., & Londoño, J. M. (2014). Sistemas para almacenar grandes volúmenes de datos. *Revista Gerencia Tecnológica Informática*, *13*(37), 17–28. http://revistas.uis.edu.co/index.php/revistagti/article/view/4689

Jiménez Celorrio, S., & de la Rosa Turbides, T. (2009). Learning-Based Planning. In *Encyclopedia of Artificial Intelligence* (pp. 1024–1028). IGI Global. https://doi.org/10.4018/978-1-59904-849-9.ch151

Jindal, R., & Borah, M. D. (2013). A Survey on Educational Data Mining and Research Trends. *International Journal of Database Management Systems*, *5*(3), 53–73. https://doi.org/10.5121/ijdms.2013.5304

Kantardzic, M. (2011). *Data mining : concepts, models, methods, and algorithms* (Segunda). John Wiley & Sons, Inc.

Karagiannis, D., Mayr, H. C., & Mylopoulos, J. (2016). Domain-specific conceptual modeling: Concepts, methods and tools. In *Domain-Specific Conceptual Modeling: Concepts, Methods and Tools*. Springer International Publishing. https://doi.org/10.1007/978-3-319-39417-6

Kautz, H., & Selman, B. (1998). The Role of Domain-Specific Knowledge in the Planning as Satisfiability Framework. In *American Association for Artificial Intelligence* (pp. 181–189). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.1259

Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2021). Automl to date and beyond: Challenges and opportunities. ACM Computing Surveys (CSUR), 54(8), 1-36.

Krahn, H., Rumpe, B., & Völkel, S. (2014, September 22). Roles in Software Development using Domain Specific Modeling Languages. *6th OOPSLA Workshop on Domain-Specific Modeling (DSM' 06)*. http://arxiv.org/abs/1409.6618

Larose, D. T., & Larose, C. D. (2014). Discovering knowledge in data : an introduction to data mining (Second). John Wiley & Sons, Inc.

Leédeczi, Á., Bakay, Á., MarÓti, M., Völgyesi, P., Nordstrom, G., Sprinkle, J., & Karsai, G. (2001). Composing domain-specific design environments. *Computer*, *34*(11), 44–51. https://doi.org/10.1109/2.963443

Lin, Y., Gray, J., & Jouault, F. (2017). DSMDiff: a differentiation tool for domain-specific models. European Journal of Information Systems, 16(4), 349–361.

https://doi.org/10.1057/PALGRAVE.EJIS.3000685
Liu, Y., Li, J., Ming, Z., Song, H., Weng, X., & Wang, J. (2019). Domain-specific data mining for residents' transit pattern retrieval from incomplete information. Journal of Network and Computer Applications, 134, 62–71. https://doi.org/10.1016/J.JNCA.2019.02.016
López, C. P. (2007). Minería de datos: técnicas y herramientas. In *Paraninfo*. Paraninfo. https://books.google.com.pe/books?id=wz-D_8uPFCEC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
Lu, W., Zhou, Y., Yu, J., & Jia, C. (2019). Concept Extraction and Prerequisite Relation Learning from Educational Data. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 9678–9685. https://doi.org/10.1609/AAAI.V33I01.33019678
Maldonado, C. E. (2017). Tipología de modelos científicos de explicación. Ciencia y complejidad. *Sociología y Tecnociencia*, *7*(2), 58–72. https://doi.org/https://doi.org/10.24197/st.2.2017.58-72
Matignon, R. (2007). *Data mining using SAS enterprise miner*. https://books.google.com/books?hl=es&lr=&id=h7FAN3lijRMC&oi=fnd&pg=PP1&dq=semma+data+mining&ots=C9RYcqzLYC&sig=vk2668sC7hkUa2Y97SpkPybfLfc
Ministerio de Educación. (2019). *Marco Estratégico 2019 – 2022*. https://www.mineducacion.gov.co/1759/articles-382974_recurso_3.pdf
Ministerio de Educación. (2020ª). *Funciones y deberes – Ministerio de Educación Nacional de Colombia*. https://www.mineducacion.gov.co/portal/Ministerio/163oodle163ción-Institucional/85252:Funciones-y-deberes
Ministerio de Educación. (2020b). *Normatividad y documentos – Ministerio de Educación Nacional de Colombia*. https://www.mineducacion.gov.co/1759/w3-article-357542.html?_noredirect=1
Ministerio de Educación. (2020c). *Sistemas de Información*. https://www.mineducacion.gov.co/portal/micrositios-institucionales/Sistemas-de-Informacion/
Ministerio de Educación Nacional. (2020). *Sistema educativo colombiano* . https://www.mineducacion.gov.co/portal/Preescolar-basica-y-media/Sistema-de-educacion-basica-y-media/233839:Sistema-educativo-colombiano
Mitrofanova, Y. S., Sherstobitova, A. A., & Filippova, O. A. (2019). Modeling smart learning processes based on educational data mining tools. In V. Uskov, R. Howlett, & L. Jain (Eds.), *Smart Education and e-Learning 2019. Smart Innovation, Systems and Technologies* (Vol. 144, pp. 561–571). Springer. https://doi.org/10.1007/978-981-13-8260-4_49
Nguyen, B.-A., & Yang, D.-L. (2012). A Semi-Automatic Approach to Construct Vietnamese Ontology from Online Text. *INTERNATIONAL REVIEW OF RESEARCH IN OPEN AND DISTANCE LEARNING*, *13*(5, SI), 148–172.
Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A Decade Survey of Transfer Learning (2010–2020). IEEE Transactions on Artificial Intelligence, 1(2), 151–
Nordstrom, G., Sztipanovits, J., Karsai, G., & Ledeczi, A. (1999). Metamodeling-rapid design and evolution of domain-specific modeling environments.

Proceedings – ECBS 1999, IEEE Conference and Workshop on Engineering of Computer-Based Systems, 68–74. https://doi.org/10.1109/ECBS.1999.755863

Oliveira, C., & Da Silva, J. C. (2009). *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. www.inf.ufg.br

Orozco Iguasnia, W. A., Villao Balón, A. J., Orozco Iguasnia, J., & Villarroel Sánchez, M. V. (2021). Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.' Revista Científica y Tecnológica UPSE, 8(2), 68–75. https://doi.org/10.26423/RCTU.V8I2.637

Parekh, V., Gwo, J., & Finin, T. (2004). Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. *International Conference of Information and Knowledge Engineering*. http://aisl.umbc.edu/resources/94.pdf

Patwari, P., Chaudhuri, S. R., Banerjee, A., Natarajan, S., & Pandey, S. (2016, November 22). A complementary domain specific design environment aiding SysML. *ISSE 2016 – 2016 International Symposium on Systems Engineering – Proceedings Papers*. https://doi.org/10.1109/SysEng.2016.7753164

Peinado, H. S. (2013). *Legislación educativa colombiana*. Editorial Magisterio. https://www.magisterio.com.co/libro/legislacion-educativa-colombiana

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4), 1432–1462.

Pérez Marqués, M. (2014). Minería de Datos a través de ejemplos (1aed.). RC Libros.

Probert, D. R., & Ridgman, T. W. (2013). Structuring technology and innovation management executive education: the research contribution. *ISPIM Conference Proceedings*. www.ispim.org.

Recker, M., & Lee, J. E. (2016). Analyzing Learner and Instructor Interactions within Learning Management Systems: Approaches and Examples. In *Learning, Design, and Technology* (pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_7-1

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Revista Iberoamericana de Inteligencia Artificial*, *29*, 11–18. http://www.aepia.org

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. https://doi.org/10.1016/J.ESWA.2006.04.005

Romero, C. & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Romero, C. & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Romero, C. & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Salgado Reyes, N., Beltrán Morales, J., Guaña Moya, J., Escobar Teran, C., Nicolalde Rodriguez, D., & Chafla Altamirano, G. (2019). Modelo para predecir el rendimiento académico basado en redes neuronales y analítica de aprendizaje. RISTI, 17, 258–266.

Scheuer, O., & McLaren, B. M. (2012). Educational Data Mining. In *Encyclopedia of the Sciences of Learning* (pp. 1075–1079). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_618

Selvaraj, P., Burugari, V. K., Sumathi, D., Nayak, R. K., & Tripathy, R. (2019). Ontology based Recommendation System for Domain Specific Seekers. *Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, 341–345. https://doi.org/10.1109/i-smac47947.2019.9032634

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. https://www.researchgate.net/publication/268770881_A_Comparative_Study_of_Data_Mining_Process_Models_KDD_CRISP-DM_and_SEMMA

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2016). Tools for Educational Data Mining: A Review. *Educational and Behavioral Statistics*, *42*(1), 85–106.

Timarán Pereira, R., Hidalgo Troya, A., & Caicedo Zambrano, J. (2020). Factores asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11° con árboles de decisión - Dialnet. Investigación e Innovación En Ingenierías, 8(3), 29–37.

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. Applied Sciences, 10(6), 2145. https://doi.org/10.3390/APP10062145

Vaisman, A., & Zimányi, E. (2014a). Data Warehouse Concepts. In *Data Warehouse Systems* (pp. 53–87). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54655-6_3

Vaisman, A., & Zimányi, E. (2014b). Extraction, Transformation, and Loading. In *Data Warehouse Systems* (pp. 285–327). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54655-6_8

Varde, A. S., & Tatti, N. (2014). A Panorama of Imminent Doctoral Research in Data Mining. *ACM SIGMOD Record*, *43*(3), 71–74. https://doi.org/10.1145/2694428.2694442

Winch, C., Oancea, A., & Orchard, J. (2015). The contribution of educational research to teachers' professional learning: philosophical understandings. *Oxford Review of Education*, *41*(2), 202–216.

Winner, E., & Veloso, M. (2003). DISTILL: Towards Learning Domain-Specific Planners by Example. *Itwentieth Nternational Conference on Machine Learning*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.3010

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. Engineering Reports, 4(5), e12482. https://doi.org/10.1002/ENG2.12482

Xu, Y., Rajpathak, D., Gibbs, I., & Klabjan, D. (2019). Automatic Ontology Learning from Domain-Specific Short Unstructured Text Data. *Computer Science*. http://arxiv.org/abs/1903.04360

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). Transfer Learining (1st ed.). Cambridge University Press.

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E-Learning*, *17*(1), 118–133. https://doi.org/10.2478/EURODL-2014-0008

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering, 89, 106903.

Zhu, Y., Zhu, H., Liu, Q., Chen, E., Li, H., & Zhao, H. (2016). Exploring the procrastination of college students: A data-driven behavioral perspective. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9642). https://doi.org/10.1007/978-3-319-32025-0_17

Anexo A. Sistema Educativo en Colombia

La Ley General de Educación en Colombia es la Ley 115 de 1994, columna vertebral de las normas educativas que ordenan y orientan el sistema educativo colombiano. Después de esta ley, se han emitido otras leyes, resoluciones y decretos que ayudan a precisar, orientar y desarrollar aspectos particulares como educación inicial, educación para adultos, cobertura para zonas con conflicto armado, entre otras. El conocimiento de estos lineamientos es de gran importancia para los educadores y directivos docentes (Peinado, 2013). A continuación, se presenta un resumen de las principales normas que rigen el sistema educativo colombiano (Ministerio de Educación, 2020b).

Normatividad y documentos que orientan la Educación en Colombia

Normativa	Descripción
Constitución Nacional de 1991	Especialmente el capítulo 1 que trata sobre los derechos fundamentales; y, los artículos 41 y 67 a 72, sobre conceptos, derechos y deberes sobre la educación.
Ley 115 de 1994 (Ley General de Educación)	Es quizás la norma de mayor contenido para orientar los procesos educativos y de prestación del servicio en desarrollo de la Constitución Nacional. Define los fines de la educación y el tipo de ser humano que es objeto de la educación colombiana; los objetivos de aprendizaje en cada uno de los niveles y ciclos de la educación formal, la educación de adultos, y en general las pautas sobre los establecimientos educativos en relación con el currículo, el plan de estudios, el calendario escolar y el proyecto educativo institucional, entre otros.
Decreto 1860 de 1994	Por el cual se reglamenta la Ley 115 de 1994, en los aspectos pedagógicos y organizativos. Describe las etapas a cumplir en el proceso de modificación del PEI, aspecto necesario para la articulación del Modelo dentro del proceso de institucionalización.
Decreto 1075 de 2015	Decreto Reglamentario Único del Sector Educación. Por el cual se establecen normas para el ofrecimiento de la educación de adultos y se dictan otras disposiciones. Esta es de obligatoria consulta por parte de los operadores de los Modelos Educativos Flexibles de educación básica secundaria y media que admiten estudiantes adultos.
Decreto 2247 de 1997	Por el cual se establecen normas relativas a la prestación del servicio educativo del nivel preescolar. Esta norma junto con los Lineamientos pedagógicos de la educación preescolar ofrece orientaciones para la implementación del modelo de Preescolar Escolarizado y No Escolarizado en los establecimientos educativos.
Decreto 2562 de 2001	Por el cual se reglamenta la Ley 387 de 1997, en cuanto a la

Normativa	Descripción
	prestación del servicio público educativo a la población desplazada por la violencia y se dictan otras disposiciones. Esta norma permite a los operadores de los MEF tener claridad en la caracterización de sus estudiantes a beneficiar, así como las responsabilidades que en materia del servicio educativo tienen las Secretarías de Educación, los establecimientos educativos, entre otros.
Decreto 1290 de 2009	Evaluación y promoción. Determina los componentes del sistema institucional de evaluación de los estudiantes, dentro de los cuales se cuentan las estrategias flexibles que determinarán las pautas para la evaluación, promoción, informes, y certificación de los estudiantes de los modelos. Este proceso se verifica con la articulación al PEI
Sentencia T25 de 2004 y Autos de la Corte Constitucional	Ordenan a las autoridades pertinentes, la restitución inmediata de los derechos a la educación de las personas desplazadas y en condición de alta vulnerabilidad.
Ley 387 de 1997	Por la cual se adoptan medidas para la prevención del desplazamiento forzado; la atención, protección, consolidación y estabilización socioeconómica de los desplazados internos por la violencia en la República de Colombia
Ley 1448 de 2011 - Ley de Víctimas	Por la cual se dictan medidas de atención, asistencia y reparación integral a las víctimas del conflicto armado interno. Esta Ley complementa a la anterior (387).

Fuente: Elaboración propia

Conformación administrativa del sistema educativo en Colombia para los niveles de Educación preescolar, básica y media

El sistema educativo colombiano está bajo la dirección del Ministerio de Educación Nacional, quien tiene como objetivo fundamental garantizar a través de la educación el desarrollo integral de los individuos y de la sociedad, esto con el impulso de la atención integral durante la trayectoria educativa con la función de generar políticas, lineamientos, directrices y estándares que orienten el servicio educativo con principios de equidad, calidad y cobertura. En detalle, las funciones del Ministerio son las siguientes (según el artículo 2 del decreto 5012 de 2009) (Ministerio de Educación, 2020a):

- Formular la política nacional de educación, regular y establecer los criterios y parámetros técnicos cualitativos que contribuyan al mejoramiento del acceso, calidad y equidad de la educación, en la atención integral a la primera infancia y en todos sus niveles y modalidades.
- Preparar y proponer los planes de desarrollo del Sector, en especial el Plan Nacional de Desarrollo Educativo, convocando los entes territoriales, las

instituciones educativas y la sociedad en general, de manera que se atiendan las necesidades del desarrollo económico y social del país.

- Dictar las normas para la organización y los criterios pedagógicos y técnicos para la atención integral a la primera infancia y las diferentes modalidades de prestación del servicio educativo, que orienten la educación en los niveles de preescolar, básica, media, superior y en la atención integral a la primera infancia.
- Asesorar a los Departamentos, Municipios y Distritos en los aspectos relacionados con la educación, de conformidad con los principios de subsidiaridad, en los términos que defina la ley.
- Impulsar, coordinar y financiar programas nacionales de mejoramiento educativo que se determinen en el Plan Nacional de Desarrollo.
- Velar por el cumplimiento de la ley y los reglamentos que rigen al Sector y sus actividades.
- Evaluar, en forma permanente, la prestación del servicio educativo y divulgar sus resultados para mantener informada a la comunidad sobre la calidad de la educación.
- Definir lineamientos para el fomento de la educación para el trabajo y el desarrollo humano, establecer mecanismos de promoción y aseguramiento de la calidad, así como reglamentar el Sistema Nacional de Información y promover su uso para apoyar la toma de decisiones de política.
- Dirigir la actividad administrativa del Sector y coordinar los programas intersectoriales.
- Dirigir el Sistema Nacional de Información Educativa y los Sistemas Nacionales de Acreditación y de Evaluación de la Educación.
- Coordinar todas las acciones educativas del Estado y de quienes presten el servicio público de la educación en todo el territorio nacional, con la colaboración de sus entidades adscritas, de las Entidades Territoriales y de la comunidad educativa.
- Apoyar los procesos de autonomía local e institucional, mediante la formulación de lineamientos generales e indicadores para la supervisión y control de la gestión administrativa y pedagógica.
- Propiciar la participación de los medios de comunicación en los procesos de educación integral permanente.
- Promover y gestionar la cooperación internacional en todos los aspectos que interesen al Sector, de conformidad con los lineamientos del Ministerio de Relaciones Exteriores.
- Suspender la capacidad legal de las autoridades territoriales para la administración del servicio público educativo y designar de forma temporal

un administrador especial de acuerdo con lo establecido en el artículo 30 de la Ley 715 de 2001.
- Dirigir el proceso de evaluación de la calidad de la educación superior para su funcionamiento.
- Formular la política y adelantar los procesos de convalidación de títulos otorgados por Instituciones de Educación Superior extranjeras.
- Formular políticas para el fomento de la Educación Superior.

El Ministerio de Educación también es quien da apoyo a los entes territoriales para una adecuada gestión de los recursos del sector. Como entes territoriales se tienen las Secretarías de Educación Departamentales y Municipales. Las entidades territoriales son certificadas mediante lo establecido en la ley 715 de 2001 y tienen como función administrar el servicio educativo en cada jurisdicción para garantizar la cobertura, calidad y eficiencia educativa. Cada departamento cuenta con una Secretaría de Educación Departamental (SED) y puede tener municipios certificados, que son aquellos que cumplen con unas condiciones para tener sus propias Secretarías de Educación Municipales. Algunas otras funciones de las Secretarías de Educación son:

- Planificar, organizar, coordinar, distribuir recursos y ejercer control para garantizar la eficiencia, efectividad y transparencia en el servicio ofrecido.
- Organizar y distribuir la planta docente, directiva docente y administrativa de acuerdo con las necesidades del servicio.
- Fortalecimiento de los establecimientos educativos que implica la asistencia técnica, asesoría permanente, capacitación y asignación de recursos para operación.
- Coordinar los municipios no certificados para establecer las acciones que permitan la organización de los recursos y el logro de las metas definidas en los planes sectoriales.
- Mejoramiento continuo en el servicio prestado a los estudiantes.

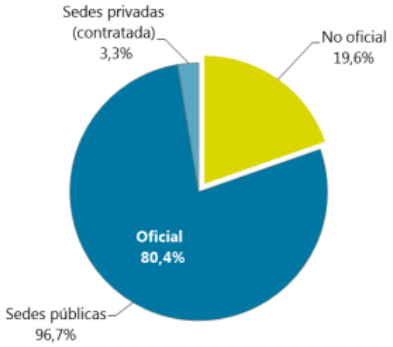
Como instancia final, en la organización administrativa, se encuentran los establecimientos o instituciones educativas, que pueden ser de carácter público o privado y cuya finalidad es prestar el servicio de educación preescolar, educación básica y educación media. La misión de las instituciones educativas se concentra en la tarea de formar y enseñar para que los estudiantes aprendan.

El Ministerio de Educación cuenta con un aplicativo en el cual se pueden hacer consultas de los establecimientos por ubicación geográfica, nombre, sector (oficial o no oficial), entre otros. También se puede ver información sobre la jornada escolar, nombre completo, calendario académico, género (masculino, femenino o

mixto) y tarifas, entre otras características. Este aplicativo se llama *Buscando Colegio* y su enlace es: https://sineb.mineducacion.gov.co/bcol/app.

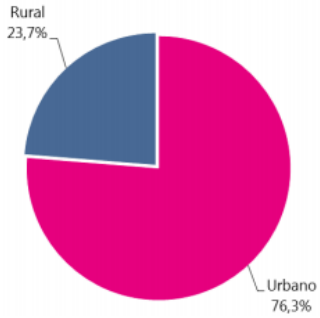
Estadísticas del sistema educativo colombiano en Educación preescolar, básica y media

Dado que el alcance de esta tesis se enmarca en los niveles educativos de básica (primaria y secundaria) y media, las estadísticas que se presentarán a continuación son de dichos niveles. Estas estadísticas han sido tomadas del Boletín Técnico del DANE para Educación Formal 2018 (DANE, 2018). Inicialmente se presenta la distribución de las instituciones educativas, la mayoría son de carácter público (80,4%) y solo un 19,6% de carácter privado.



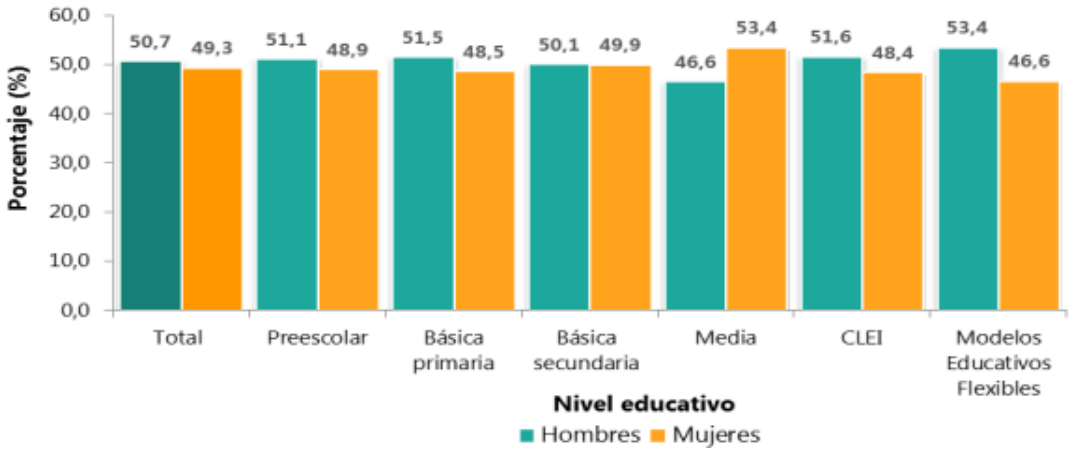
Distribución porcentual de matriculados por sector
Fuente: tomado de **(**DANE, 2018)

Otro factor de interés es la distribución de estas instituciones entre el sector urbano y el sector rural, cabe anotar que una de las premisas que persigue el gobierno nacional es lograr llevar la educación a todos los rincones del país, encontrando que el sector rural suele tener deficiencias para acceder a la educación. Sin embargo, es lógico que se mantiene una preponderancia de instituciones en el sector urbano (76,3%).



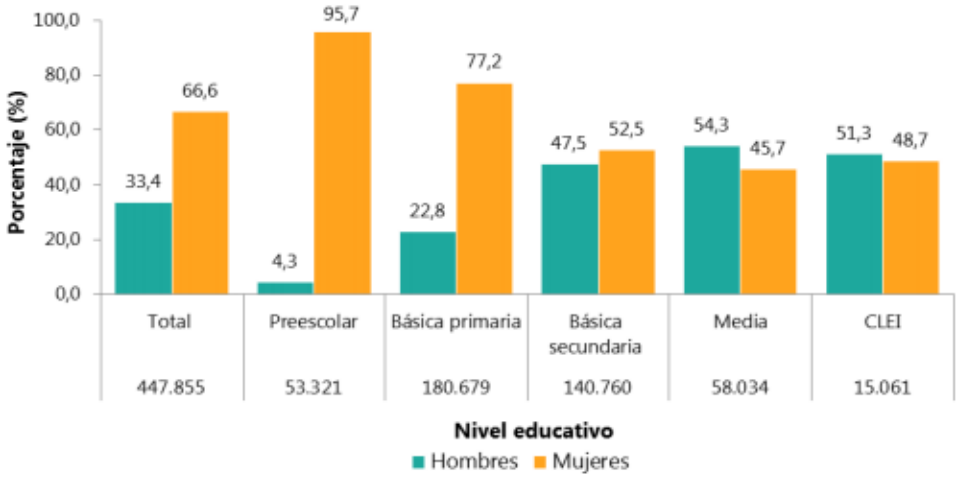
Distribución porcentual de alumnos matriculados por zona
Fuente: tomado de **(**DANE, 2018)

Ahora bien, en temas de matrícula, para todos los niveles: preescolar, básica primaria, básica secundaria, media, CLEI (Ciclos Lectivos Especiales Integrados) y modelos educativos flexibles; se mantiene una proporción equilibrada entre géneros (hombre y mujeres), solo en la media se aprecia una mayoría de mujeres, para los otros niveles los hombres superan, pero de manera muy sutil.



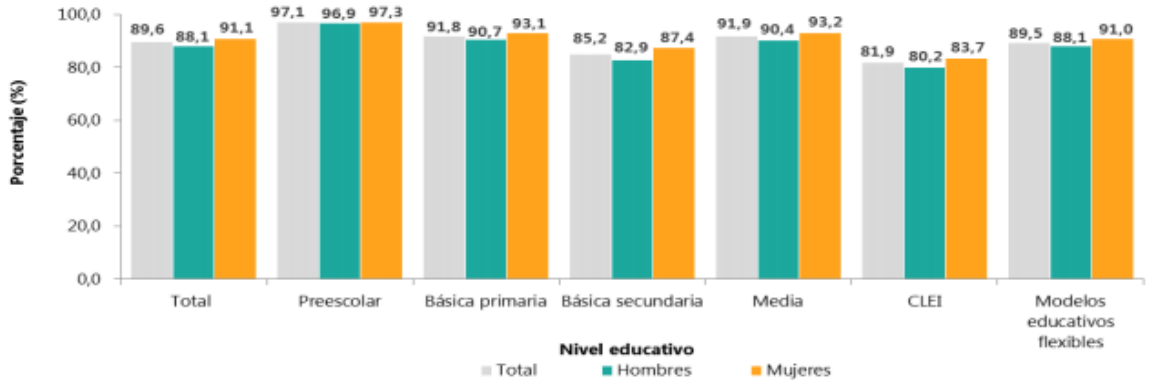
Distribución porcentual de la matrícula, por niveles educativos y sexo
Fuente: tomado de **(**DANE, 2018)

En temas de docentes, existe una preponderancia muy marcada en preescolar hacia las docentes mujeres (95,7%), lo mismo que en básica primaria (77,2%). En básica secundaria, media y CLEI se maneja una distribución semejante para los dos géneros. Pero haciendo un promedio general, las mujeres terminan a la cabeza con una presencia del 66,6% frente al 33,4% de hombres.



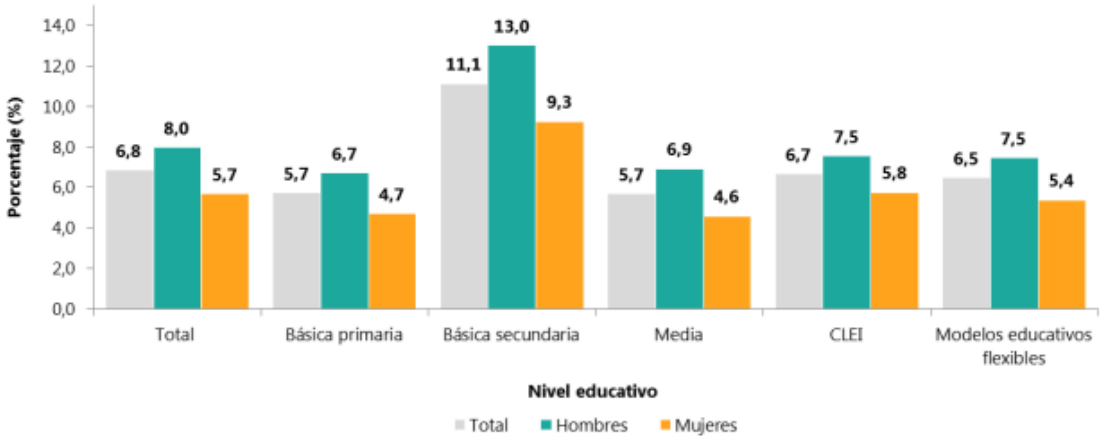
Número y distribución porcentual de docentes con asignación académica, por niveles educativos y sexo
Fuente: tomado de **(**DANE, 2018)

Otro aspecto de interés corresponde a la tasa de aprobación, esta tasa muestra que para el nivel de básica primaria está ligeramente por encima del 90%, lo mismo que para media. Por su parte CLEI y básica secundaria presentan una tasa menor, que ronda el 80%, pero que es un poco mayor para mujeres. Para todos los niveles se cumple que la tasa de aprobación es mayor para las mujeres.



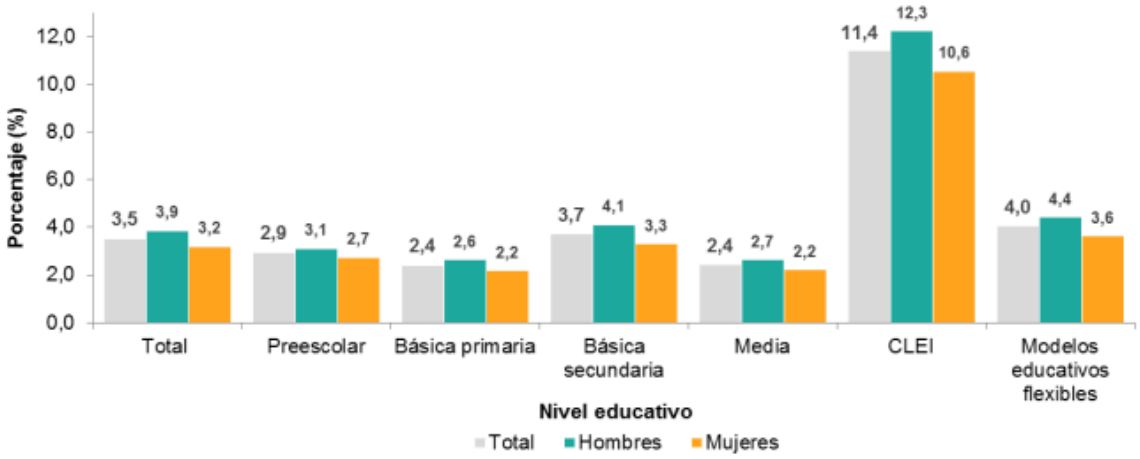
Tasa de aprobación por nivel educativo y sexo
Fuente: tomado de **(**DANE, 2018)

De forma complementaria la tasa de reprobación muestra que, claramente, en la básica secundaria es donde mayor tasa existe y los hombres superan en casi 4 puntos porcentuales a las mujeres. Para todos los niveles la tasa es mayor en hombres, la básica primaria y la media son los niveles con menor reprobación (5,7% en total para los dos casos).



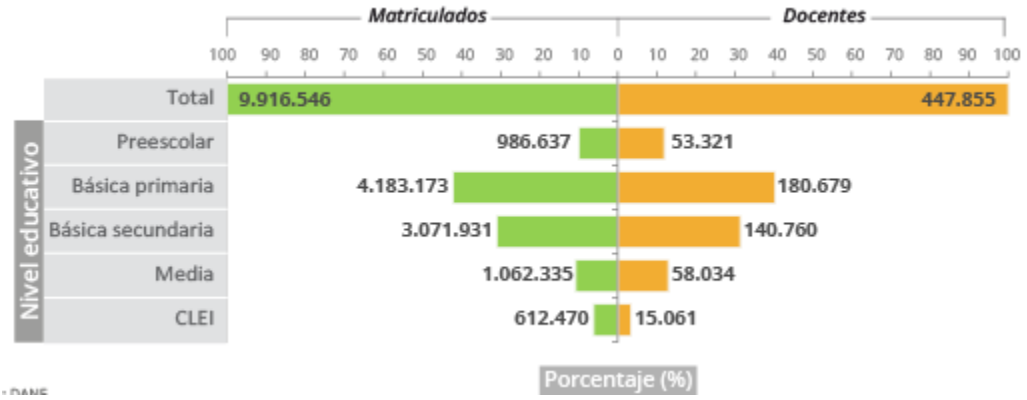
Tasa de reprobación por nivel educativo y sexo
Fuente: tomado de **(**DANE, 2018)

En términos de deserción, los CLEI superan significativamente a los demás niveles, seguidos de los modelos de educación flexibles; para los otros niveles, la básica secundaria sobresale con un 4,1% para hombres y un 3,3% para mujeres. EN todos los niveles la deserción es mayor en hombres.



Tasa de deserción por nivel educativo y sexo
Fuente: tomado de **(**DANE, 2018)

En términos de escolarización, para el 2018, el mayor porcentaje de matriculados se registró en básica primaria (42,2%); mientras, la menor participación en educación tradicional fue para preescolar (9,9%). Paralelamente, los docentes representaron, para básica primaria el 40,3% y para preescolar 11,9.



Porcentaje de escolarización por nivel educativo
Fuente: tomado de **(**DANE, 2018)

Retos educativos en Colombia

El Plan Nacional Decenal de Educación 2016-2026 "*El camino hacia la calidad y la equidad*", traza la ruta de Colombia en Educación hasta el 2026 y establece diez retos para el sector educativo colombiano en el periodo 2017-2026 (Ministerio de Educación, 2019):

- **Regular el alcance del derecho a la educación:** garantizar, por medio de la ley, el acceso y permanencia a una educación de calidad.

- **Mayor articulación de los niveles educativos:** se carece de un sistema educativo que permita la articulación de los diversos niveles y subsistemas. El desafío consiste en trabajar para que los distintos subsistemas interactúen de manera participativa y descentralizada tanto a nivel horizontal como vertical.
- **Lineamientos curriculares pertinentes:** para construir la identidad nacional y aprender de otras experiencias y contextos es necesario que el país construya lineamientos curriculares generales, pertinentes y flexibles.
- **Una política pública para la formación de docentes**: la formación docente es un punto álgido, los egresados de las facultades de educación no tienen buenos puntajes entre los egresados del sistema universitario, en pruebas como la Saber pro. La calidad educativa depende en alto grado de los niveles alcanzados en formación por sus docentes.
- **Hay que dejar de enseñar lo mismo**: hacer un replanteamiento pedagógico y de los currículos. Los modelos que enfatizan la transmisión de informaciones, vigente en la mayoría de las instituciones educativas, han demostrado que no logran promover el desarrollo humano e integral de los estudiantes.
- **El problema de la educación no es tecnológico, sino pedagógico: l**os cambios tecnológicos, por sí solos, no logran transformaciones pedagógicas. La tecnología apoya los procesos de cambio pedagógico, pero no los genera como tal.
- **La sociedad colombiana ha enfermado emocionalmente:** el escenario que dejaron los tiempos de guerra en Colombia obliga a los docentes a trabajar por consolidar las competencias ciudadanas de manera que fortalezcamos la convivencia sana, el trabajo en equipo y la interacción respetuosa con los demás.
- **Superar el atraso en los niveles educativos del sector rural:** actualmente se requiere hacer un énfasis especial en la educación dirigida a la ruralidad.
- **Se requiere de mayor inversión:** para elevar la calidad hay que destinar recursos suficientes a la formación, educación inicial, rural, salarios y a la jornada completa.
- **Más apoyo a la ciencia y la investigación:** se requiere de la ciencia para aumentar la capacidad de respuesta a las demandas sociales, basándose en la investigación de los propios problemas de la nación.

Desafíos Plan Nacional Decenal de Educación 2016-2026
Fuente: tomado de (Ministerio de Educación, 2019)

Anexo B. Sitio web de documentación y disposición del modelo y sus estrategias

Como parte de la documentación del modelo y buscando dejar el conocimiento disponible para la comunidad académica e investigativa, se realizó la construcción de un sitio web que permitiera dejar disponibles las estrategias de interacción para el usuario estándar y para el usuario experto, y también abordar datos del alcance definido en la tesis, es decir, de educación básica y media presencial de una región de Colombia. Cabe anotar que el desarrollo de este sitio se hizo como una forma de mostrar la aplicación del modelo y de sus estrategias, cubriendo datos del nivel educativo y de las técnicas de minería consideradas en el alcance de este trabajo doctoral, pero como tal el modelo, puede ser aplicado en otros casos con el desarrollo de sistemas que tomen los elementos conceptuales en los que se basa el aporte de esta tesis.

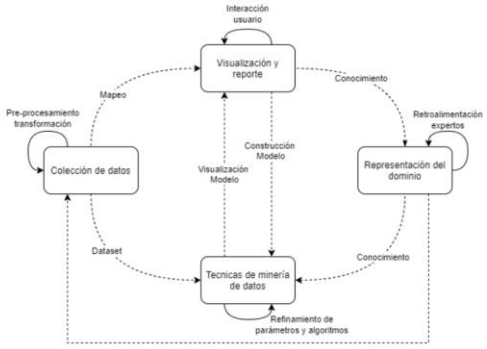
El sitio web desarrollado se llamó 2DE-M, acrónimo definido para el modelo y que hace referencia a Datos Educativos y Domino Específico, de allí la primera parte de la sigla y Minería representada en la M que cierra el nombre. A continuación, se muestran algunos pantallazos de las interfaces e información dejada a disposición en el sitio web. El enlace de acceso es: http://2de-m.emilcyjuliana.com/

En el sitio web también se presentan cuestionarios diseñados para recolectar conocimiento del domino de datos educativos a través de los expertos en diferentes áreas asociadas al modelo (http://2de-m.emilcyjuliana.com/representacion-dominio/).

2DE-M

- Home
- Datos
- Dominio
- Minería
- Reporte

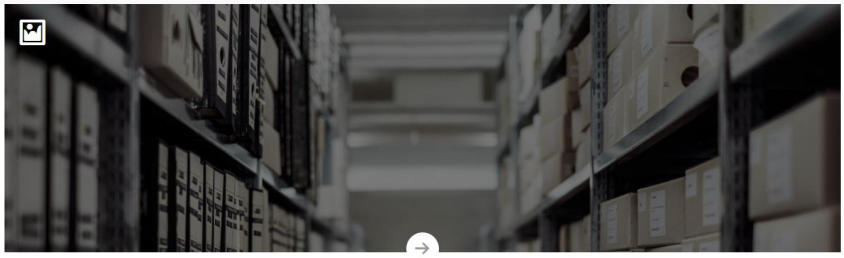
MODELO DE DOMINIO ESPECÍFICO PARA ANÁLISIS Y MINERÍA DE DATOS EDUCATIVOS



El modelo se orienta a cubrir las etapas de los procesos de minería de datos educativos incluyendo un enfoque de dominio específico. Se tienen en cuenta las etapas principales de los procesos de análisis de datos, tales como: recolección y tratamiento de datos, almacenamiento, selección, aplicación de técnicas, interpretación, reporte y visualización. Estas etapas se condensan en cuatro momentos, que son presentados a grandes rasgos en la figura superior. En esta perspectiva, se considera el enfoque de dominio específico incluyendo una etapa para la representación del conocimiento propio del campo educativo. Esta etapa de representación del dominio no solo se conecta con los datos, sino que también con las técnicas de minería y la visualización de los resultados.

2DE-M

- Home
- Datos
- Dominio
- Minería
- Reporte

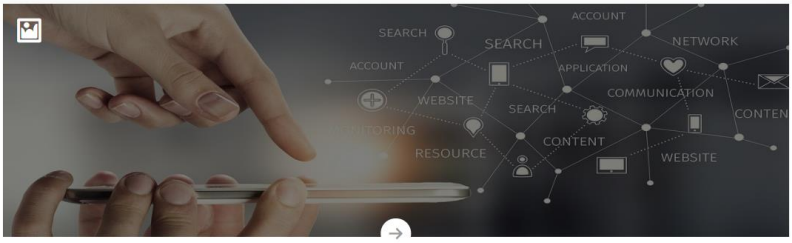


Tratamiento de Datos Educativos

- Cómo se pueden tratar datos educativos
- Taxonomía de datos educativos
- Dataset disponibles
- Carga de datos

2DE-M

- Home
- Datos
- Dominio
- Minería
- Reporte



Representación del Dominio

El módulo de representación del dominio es la principal fuente de diferenciación del modelo. Se puede pensar que el conocimiento del dominio finalmente siempre es usado en los modelos, sin embargo, no siempre se cuenta con una estrategia y una guía que permita orientar respecto a los componentes del dominio que deben ser tenidos en cuenta y de qué forma se pueden articular con las demás etapas del proceso de análisis. Para ayudarnos en la construcción de la base de conocimiento del dominio educativo, si usted es experto en alguno de los siguientes temas, por favor conteste un cuestionario corto relacionado a su área de experticia.

Investigador en analíticas de aprendizaje

Investigador en minería de datos

Investigador en tecnología educativa

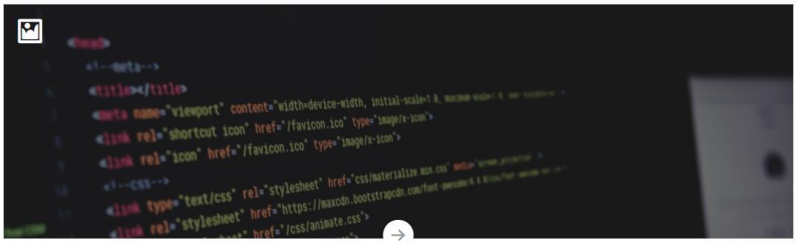
Científico de datos

Docente – Directivo docente

Funcionario del área educativa

2DE-M

- Home
- Datos
- Dominio
- Minería
- Reporte



Minería de Datos Educativos

En este módulo se ubica como tal la aplicación de las técnicas y algoritmos de análisis, es la etapa en la que se procesan los datos para la generación de los modelos de minería, se pueden encontrar técnicas descriptivas o predictivas. Hacer un análisis diagnóstico inicial previo a la minería de datos, es recomendable; por medio de un análisis diagnóstico se puede empezar a intuir relaciones en los datos, además puede ayudar a orientar la elección de técnicas y algoritmos.

Apoyo para la selección de técnica

La elección de las técnicas depende también de los objetivos del análisis y de los datos con los cuales se cuente. El modelo de minería está sujeto a las variables que los usuarios quieran observar y al alcance inicialmente definido. Los expertos del contexto educativo pueden definir sus objetivos con el uso de abstracciones de los meta-modelos o representaciones (indicaciones) que guíen y soporten el proceso de acuerdo a las características del dominio. A continuación se muestra algunas recomendaciones para la selección de técnicas:

(Contenido en construcción)

Prueba de algoritmos

Se pone a disposición los siguiente algoritmos para su prueba y se mencionan algunos casos de éxito y recomendaciones para usarlos.

(Contenido en construcción)

2de-m es el sitio web que se utilizó como uno de los medios de documentación y difusión del modelo de dominio específico para minería de datos educativos. Se busca que 2de-m se convierta en una herramienta para el almacenamiento y análisis de datos históricos, de granularidad fina y longitudinales generados por el proceso de ingreso y la medición del desempeño de los estudiantes en educación básica (primaria y secundaria) y media en Colombia.

2de-m es una tentativa inicial para el desarrollo de un estándar común que se espera sea útil para el campo; así mismo, se espera que 2de-m se pueda convertir en una fuente de datos y un lugar donde los investigadores, expertos y en general los usuarios interesados en el dominio de datos educativos pueden depositar sus conocimientos a través de los cuestionarios de recolección y obtener ayuda a partir de los conocimientos compartidos por otros investigadores sobre temáticas de interés común y relacionadas con los datos objeto de estudio. Otro objetivo de la plataforma es proveer una guía para los usuarios interesados en la construcción de modelos de minería de datos educativos, por medio de construcciones basadas en el conocimiento recolectado de la literatura, el contexto y compartido por los expertos en el dominio.

Anexo C. Listado de Etnias y Resguardos

Listado Etnias

NO.	CÓDIGO EN RESOLUCIÓN 166	ETNIAS REGISTRADAS EN RESOLUCIÓN 166
	00	No Aplica
1	01	Achagua
2	02	Amorúa
3	03	Andoque o Andoke
4	04	Arhuaco (IJKA)
5	05	Awa (CUAIKER)
6	06	Barea
7	07	Barazana
8	08	Barí (Motilón)
9	09	Betoye
10	10	Bora
11	11	Cabiyari o Kawiyarí
12	12	Carapana
13	13	Carijona o Karijona
14	14	Chimila (ETTE E' NEKA)
15	15	Chiricoa
16	16	Cocama
17	17	Coconuco
18	18	Cofán o Kofán
19	19	Pijaos
20	20	Cubeo o Kubeo
21	21	Cuiba o Kuiba
22	22	Curripaco o Kurripako
23	23	Desano
24	24	Dujos
25	26	Embera Catio o Embera Katío
26	27	Embera Chamí

NO.	CÓDIGO EN RESOLUCIÓN 166	ETNIAS REGISTRADAS EN RESOLUCIÓN 166
27	28	Eperara Siapidara
28	29	Guambiano
29	30	Guanaca
30	31	Guayabero
31	33	Hitnu
32	34	Inga
33	35	Kamsa o Kamëntsá
34	36	Kogui
35	37	Koreguaje o Coreguaje
36	38	Letuama
37	39	Macaguaje o Makaguaje
38	40	Nukak (Makú)
39	41	Macuna o Makuna (Sara)
40	42	Masiguare
41	43	Matapí
42	44	Miraña
43	45	Muinane
44	46	Muisca
45	47	Nonuya
46	48	Ocaina
47	49	Nasa (Páez)
48	50	Pastos
49	51	Piapoco (Dzase)
50	52	Piaroa
51	53	Piratapuyo
52	54	Pisamira
53	55	Puinave
54	56	Sáliba
55	57	Sikuani
56	58	Siona

NO.	CÓDIGO EN RESOLUCIÓN 166	ETNIAS REGISTRADAS EN RESOLUCIÓN 166
57	59	Siriano
58	60	Siripu o Tsiripu (Mariposo)
59	61	Taiwano (Tajuano)
60	62	Tanimuka
61	63	Tariano
62	64	Tatuyo
63	65	Tikuna
64	66	Totoró
65	67	Tucano (Desea) o Tukano
66	68	Tule (Kuna)
67	69	Tuyuka (Dojkapuara)
68	70	U' wa (Tunebo)
69	71	Wanano
70	72	Wayuu
71	73	Uitoto
72	74	Wiwa (Arzario)
73	75	Waunan (Wuanana)
74	76	Yagua
75	77	Yanacona
76	78	Yauna
77	79	Yukuna
78	80	Yuko (Yukpa)
79	81	Yuri (Carabayo)
80	82	Yuruti
81	83	Zenú
82	200	Negritudes
83	400	Rom
84	999	Otras etnias

Listado Resguardos

NOMBRE DEL MUNICIPIO	CÓDIGO DEL RESGUARDO DANE	NOMBRE DEL RESGUARDO	TOTAL POBLACIÓN CORTE JUNIO 2018
CHITAGÁ	1,066	UNIDO U'WA	502
TOLEDO	1,066	UNIDO U'WA	715
EL TARRA	1,384	CATALAURA	175
TEORAMA	1,384	CATALAURA	128
CONVENCIÓN	1,385	MOTILON-BARI	609
EL CARMEN	1,385	MOTILON-BARI	714
TEORAMA	1,385	MOTILON-BARI	389
TOLEDO	1,82	KUITUA	64