INTRODUCCIÓN

La minería de datos educativos (EDM) es una nueva tendencia en el campo de la minería de datos y el descubrimiento de conocimientos en bases de datos (KDD) que se centra en la minería de patrones útiles y el descubrimiento de conocimientos útiles de los sistemas de información educativos, como sistemas de admisión, sistemas de registro, gestión de cursos, sistemas de aprendizaje (Moodle, Blackboard, etc…), y cualquier otro sistema que atienda a estudiantes de diferentes niveles educativos, desde escuelas y colegios hasta universidades. Los investigadores en este campo se centran en descubrir conocimientos útiles, ya sea para ayudar a los institutos educativos a gestionar mejor a sus estudiantes, o para ayudar a los estudiantes a gestionar mejor su educación, sus resultados y mejorar su rendimiento.

Analizar los datos y la información de los estudiantes para clasificarlos, encontrar información que permita tomar mejores decisiones o mejorar el desempeño de los estudiantes es un campo de investigación interesante, que se enfoca principalmente en analizar y comprender los datos educativos de los estudiantes y generar reglas, clasificaciones y predicciones específicas para ayudar a los estudiantes en su desempeño educativo futuro.

La clasificación es la técnica de minería de datos más familiar y eficaz que se utiliza para clasificar y predecir valores. La minería de datos educativos (EDM) no es una excepción a este hecho, por lo tanto, se utilizará en este proyecto para analizar la información recopilada de los estudiantes a través de sus años de estudio y proporcionar clasificaciones basadas en los datos recopilados para predecir y clasificar a aquellos estudiantes que pueden continuar su proceso y/o transición educativa desde sus años de estudio en pregrado hacia los diferentes posgrados de la Pontificia Universidad Javeriana Cali. El objetivo del trabajo que se propone en este documento es identificar las relaciones entre diferentes factores y características de los estudiantes y su continuidad académica (grado educativo). Este conocimiento puede ayudar a las áreas de Promoción Estudiantil a mejorar sus estrategias de atracción, teniendo un mayor enfoque en aquellos estudiantes que presentan una serie de características en común.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Desde el año 2018, la cantidad de estudiantes matriculados en Posgrados en la Pontificia Universidad Javeriana Cali ha tenido una tendencia a la baja; en la Facultad de Ingeniería y Ciencias en los rangos de matriculados desde el segundo periodo del 2018 al segundo periodo del 2020 se ha tenido una caída de casi un 30% en el número de estudiantes. Este es un desafío permanente para las instituciones de educación superior porque la deserción estudiantil trae consigo problemas económicos y de reputación para las universidades. Frente a esta problemática, han sido desarrolladas una variedad de estrategias de retención para revertir esta creciente tendencia regional, nacional y global.

La pandemia de coronavirus que afecta al mundo desde 2020 obligó a muchas instituciones de educación superior a replantear sus metodologías de atracción hacia los nuevos integrantes en los diferentes niveles académicos de pregrado y posgrados. La transformación digital surgió entonces como una alternativa para que las instituciones pudieran reinventarse y seguir ofreciendo las diferentes alternativas de educación a los diferentes públicos. El crecimiento de la virtualidad supuso una alternativa para mantener la productividad, evitar la exposición al contagio, los traslados y gestionar el tiempo de forma eficiente. Sin embargo, también supuso una nueva forma de interacción en el mercado para la oferta de los programas académicos. Esta oferta significaba la explotación de los diferentes medios de comunicación y redes sociales para poder llegar a los clientes de forma oportuna y asertiva.

Según el informe “Big data en educación” desarrollado por la Universidad EAFIT y la Red de Inteligencia Competitiva, el país que más ha investigado el tema es Estados Unidos, mostrando que mucha de esta investigación se ha aplicado a casos reales en sus propias escuelas y universidades[1]. Igualmente, China y Reino Unido son dos países que ya han comenzado a estudiar el tema y a ver su importancia en el mejoramiento de la calidad y pertinencia educativa que le están proporcionando a sus estudiantes. Por otro lado, la aplicación de la minería de datos en la educación viene teniendo un importante lugar dentro del conjunto de investigaciones que buscan desarrollar métodos para explorar la información que se genera dentro de los ambientes educativos con el objetivo de entender la forma en que los estudiantes aprenden, para poder tomar las decisiones adecuadas que garanticen el éxito en el proceso educativo[2].

Por tanto, este trabajo de grado pretende realizar un estudio a partir de la aplicación de técnicas de minería de datos para evaluar y analizar de manera sistemática, la asociación de diferentes aspectos presentes en los estudiantes de pregrado y la continuidad de matricula académica hacia los posgrados de la Pontificia Universidad Javeriana Cali. Además, hacer aprovechamiento de data

mining para revelar las características de los estudiantes y predecir su tránsito de pregrado a posgrado que permita continuar con su formación en conocimientos específicos y especializados de su carrera profesional.

1.2. FORMULACIÓN DEL PROBLEMA

¿Qué técnicas de minería de datos pueden ser aplicadas para generar un modelo de analítica que responda a la necesidad de fortalecer las estrategias de marketing en el proceso de atracción y mercadeo de la Dirección de Promoción de Programas Académicos de la Pontificia Universidad Javeriana Cali?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Generar un modelo de clasificación o predicción que permita responder a una necesidad detectada en la Dirección de Promoción de Programas Académicos relacionadas con la continuidad en la formación académica de los estudiantes de pregrado de la Pontifica Universidad Javeriana Cali, a través de técnicas de minería de datos.

2.2. OBJETIVOS ESPECÍFICOS

Identificar las necesidades de la Dirección de Promoción de Programas Académicos a partir de los datos obtenidos en el entendimiento del negocio con las áreas involucradas en el proceso de atracción y mercadeo de la Pontificia Universidad Javeriana Cali.

Aplicar técnicas de modelado apropiadas que respondan a una necesidad de la Dirección de Promoción de Programas Académicos identificada para el proyecto de minería de datos específico.

Generar el modelo con las características de los datos resultantes a partir de la preparación de estos, una vez se haya realizado la determinación del método de evaluación del modelo.

Evaluar el modelo generado.

2.3. RESULTADOS ESPERADOS

Un modelo generado aplicando técnicas de minería de datos que permitirán determinar, ampliar el campo de análisis, personalizar o segmentar las estrategias de marketing dirigidas a estudiantes de pregrado de la Facultad de Ingeniería y Ciencias, enmarcado en la estrategia de la unidad organizacional de las Oficinas de Promoción Institucional y Mercadeo de la Pontificia Universidad Javeriana Cali.

3. ALCANCE

El modelo se construirá haciendo uso de la minería de datos con información almacenada en el Sistema de Información Académico, el Sistema de Información del Medio Universitario, el Sistema de Información de la Oficina de Gestión Estudiantil y el Sistema de Finanzas Estudiantiles de los últimos 15 años.

La metodología por emplear será CRISP-DM (Cross Industry Standard Process for Data Mining) para los pronósticos de los algoritmos de clasificación resultantes.

4. JUSTIFICACIÓN

Los métodos y técnicas de minería de datos se han convertido en un área de investigación importante en los últimos años, porque su inclusión en un proceso de análisis de datos puede revelar relaciones ocultas, patrones de comportamiento, perfiles de entidades y regularidades similares en los datos almacenados en grandes bases de datos.

El conocimiento descubierto por la inteligencia artificial difícilmente podría adquirirse por medios tradicionales, como el análisis estadístico, la consulta de datos u otros métodos analíticos, debido a la gran cantidad de datos recopilados y la vaga idea de la existencia del conocimiento. Por lo tanto, el descubrimiento de conocimiento en datos (KDD) y la minería de datos (DM) como parte integral, son indispensables para los analistas de datos. Sin embargo, lo mencionado anteriormente es cierto solo en el caso de datos de entrada de calidad, o datos que podrían transferirse a través del proceso de preprocesamiento[5]. En un artículo publicado por el Servicio de Investigación del Congreso de los Estados Unidos (CRS) declararon que: "La calidad de los datos es un tema multifacético que representa uno de los mayores desafíos para la minería de datos". Se ha reconocido que el éxito de todo un proceso KDD depende de las entradas proporcionadas[6].

En los últimos años, las instituciones de educación han utilizado big data para desarrollar diversas aplicaciones de minería de datos, cuyo objetivo es el análisis generado para resolver problemas de investigación en el campo educativo[7]. A su vez, se ha incrementado cada vez más el uso de análisis de datos para investigar cuestiones científicas dentro de la investigación del dominio educativo para comprender mejor a los estudiantes y sus comportamientos de aprendizaje. Este análisis de datos va muy de la mano del crecimiento de herramientas tecnológicas que se combinan con el aprendizaje tradicional, a fin de generar estrategias para descubrir conexiones ocultas o previamente desconocidas y generar hipótesis en profundidad[8].

Una vez se logra consolidar la información a través del uso de herramientas de análisis de minería de datos, entra a jugar un papel importante el análisis predictivo; analizando datos demográficos y de rendimiento estudiantil, las instituciones de educación pueden predecir diferentes aspectos en el rendimiento de los estudiantes o pueden guiar sus estrategias en consolidar el mercadeo para obtener mejores resultados en la atracción de nuevos estudiantes y aumentar la tasa en la retención de sus estudiantes en las modalidades presenciales y en línea.

Según Gartner[4], “El análisis predictivo es una forma de análisis avanzado que examina los datos o el contenido para responder a la pregunta "¿Qué va a pasar?" o más precisamente, "¿Qué es probable que suceda?", y se caracteriza por técnicas como análisis de regresión, pronóstico, estadísticas multivariantes, coincidencia de patrones, modelado predictivo y pronóstico.”

El análisis predictivo utiliza técnicas de minería de datos, estadística, modelización, aprendizaje automático e inteligencia artificial, para analizar los datos actuales y hacer predicciones sobre el

futuro. La técnica de minería de datos es usada con la finalidad de obtener información específica que se encuentra oculta en grandes volúmenes de información y que aporta características y conocimientos útiles para las instituciones de educación.

Por otro lado, cuando la información no cuenta con una serie de datos ya etiquetados en diferentes categorías, grupos o clases en base a las cuales se puedan hacer predicciones o aplicar análisis predictivos, es necesario recurrir a técnicas de análisis de clúster o clustering.

El análisis de clúster es una técnica multivariante de minería de datos, encuadrada dentro de la disciplina de inteligencia artificial. Concretamente es una técnica de análisis exploratorio de datos para resolver problemas de clasificación, que intenta identificar de manera automática, agrupaciones de elementos (también llamados conglomerados o clústeres homogéneos) de acuerdo con una medida de distancia o similitud entre ellos.

El análisis de clúster tiene una extraordinaria importancia en la investigación científica, ya que la clasificación es uno de los objetivos fundamentales de la ciencia. En el campo del marketing es aplicable para la segmentación e investigación de mercados, ayudando a descubrir grupos que permitan desarrollar estrategias orientadas a tales grupos.

Finalmente, frente a los cambios en la educación superior y la movilidad estudiantil, haciendo que la oferta de productos y servicios sean cada día más competitivos y se ajusten a las necesidades de los aspirantes de posgrados, es importante aplicar las técnicas mencionadas anteriormente para conocer los aspectos clave que determinan la continuidad del proceso académico del pregrado al posgrado; si bien es cierto que lo que más puede primar en la decisión de un estudiante para continuar realizando posgrados en la universidad es un plan de estudio bien estructurado, que permita satisfacer los objetivos de su vida profesional, pueden existir aspectos que generen valor agregado a la institución y ser claves para elegir donde estudiar. Por esta razón es importante proponer y desarrollar técnicas que permitan tomar decisiones oportunas frente al mercadeo educativo, aunque no se encontraron trabajos que precisen puntualmente sobre investigaciones previas acerca de la continuidad del proceso de formación de estudiantes de pregrado.

5. MARCO TEÓRICO DE REFERENCIA Y ANTECEDENTES

5.1. DATA MINING

La minería de datos (DM) es una técnica dedicada a escanear grandes conjuntos de datos, generar información y descubrir conocimiento a partir de esos datos. El significado del término minería tradicional sesga los motivos de la minería de datos. Pero, en lugar de buscar minerales naturales, el objetivo es el conocimiento. DM busca descubrir patrones de datos, organizar información de relaciones ocultas, estructurar reglas de asociación, estimar valores de elementos desconocidos para clasificar objetos, componer grupos de objetos homogéneos y revelar muchos tipos de hallazgos que no son fácilmente producidos por un sistema de información clásico. Por lo tanto, los resultados de la DM representan un apoyo valioso para la toma de decisiones[2].

En educación, es un nuevo objetivo de aplicación de DM para el descubrimiento de conocimientos, la toma de decisiones y la recomendación. Hoy en día, el uso de DM en el ámbito educativo es incipiente y da origen al campo de investigación de la minería de datos educativos (EDM)[10].

La EDM surge como un paradigma orientado a diseñar modelos, tareas, métodos y algoritmos para explorar datos de contextos educativos. EDM busca descubrir patrones y hacer predicciones que caractericen los comportamientos y logros de los estudiantes, el contenido del conocimiento del dominio, las evaluaciones, las funcionalidades educativas y las aplicaciones.

5.2. MACHINE LEARNING

Machine Learning es un área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos. Hoy en día existen diferentes modelos que utilizan esta técnica y consiguen una precisión incluso superior a la de los humanos en las mismas tareas, por ejemplo en el reconocimiento de objetos en una imagen[11].

La construcción de modelos de Machine Learning requiere adaptaciones propias debido a la naturaleza de los datos o a la problemática a la que se aplica. Así́, surge la necesidad de investigar las diferentes técnicas que permitan obtener resultados precisos y confiables en un tiempo razonable.

Machine Learning permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita[12]. Sin embargo, no es un proceso sencillo. Conforme el algoritmo ingiere datos de entrenamiento, es posible producir modelos más precisos basados en datos. Un modelo de Machine Learning es la salida de información que se genera cuando entrena un algoritmo de Machine Learning con datos. Después del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Por ejemplo, un algoritmo predictivo creará un modelo predictivo. A continuación, cuando proporcione el modelo predictivo con datos, recibirá un pronóstico basado en los datos que entrenaron al modelo.

Machine Learning permite modelos a entrenar con conjuntos de datos antes de ser implementados. Este proceso iterativo de modelos online conduce a una mejora en los tipos de asociaciones hechas entre los elementos de datos. Debido a su complejidad y tamaño, estos patrones y asociaciones podrían haber sido fácilmente pasados por alto por la observación humana. Después de que un modelo ha sido entrenado, se puede utilizar en tiempo real para aprender de los datos. Las mejoras en la precisión son el resultado del proceso de entrenamiento y la automatización que forman parte del Machine Learning[12].

Las técnicas de Machine Learning son necesarias para mejorar la precisión de los modelos predictivos. Dependiendo de la naturaleza del problema empresarial que se está atendiendo, existen diferentes enfoques basados en el tipo y volumen de los datos.

5.2.1. **Aprendizaje supervisado**. El aprendizaje supervisado comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos. Por ejemplo, se puede crear una aplicación de Machine Learning con base en imágenes y descripciones escritas que distinga entre millones de animales.

5.2.2. **Aprendizaje no supervisado**. El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. Por ejemplo, las aplicaciones de redes sociales, tales como Twitter, Instagram y Snapchat, tienen grandes cantidades de datos sin etiquetar. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentra. El aprendizaje no supervisado lleva a cabo un proceso iterativo, analizando los datos sin intervención humana. Se utiliza con la tecnología de detección de spam en e-mails. Existen demasiadas variables en los e-mails legítimos y de spam para que un analista etiquete una cantidad masiva

de e-mail no solicitado. En su lugar, los clasificadores de Machine Learning, basados en Clustering y asociación, se aplican para identificar e-mail no deseado.

5.2.3. **Aprendizaje de refuerzo**. El aprendizaje de refuerzo es un modelo de aprendizaje conductual. El algoritmo recibe retroalimentación del análisis de datos, conduciendo el usuario hacia el mejor resultado. El aprendizaje de refuerzo difiere de otros tipos de aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo. Más bien, el sistema aprende a través de la prueba y el error. Por lo tanto, una secuencia de decisiones exitosas conduce al fortalecimiento del proceso, porque es el que resuelve el problema de manera más efectiva.

5.2.4. **Deep Learning**. El Deep Learning es un método específico de Machine Learning que incorpora las redes neuronales en capas sucesivas para aprender de los datos de manera iterativa. El Deep Learning es especialmente útil cuando se trata de aprender patrones de datos no estructurados. Las redes neuronales complejas de Deep Learning están diseñadas para emular cómo funciona el cerebro humano, así que las computadoras pueden ser entrenadas para lidiar con abstracciones y problemas mal definidos. Las redes neuronales y el Deep Learning se utilizan a menudo en el reconocimiento de imágenes, voz y aplicaciones de visión de computadora.

5.3. TÉCNICAS DE DATA SCIENCE O DATA ANALYTICS

Las técnicas de Data Science o Data Analytics, que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (Knowledge Discovery in Databases) para referirse al (amplio) concepto de hallar conocimiento en los datos. En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los años 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase[13].

En 2015, IBM Corporation, uno de los impulsores tradicionales de CRISP-DM, planteó una nueva metodología llamada Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) que extiende CRISP-DM, y es parte de la metodología general ASUM (Analytics Solutions Unified Method) incorporada en los productos y soluciones analíticas de IBM[13].

5.4. CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM establece un proyecto de minería de datos a través de un modelo de proceso iterativo, desarrollado en 1996, y actualmente es la metodología más favorecida desde entonces[14]. CRISP-DM comprende las siguientes fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue (Fig. 3). Generalmente, estas tareas se establecen como fases secuenciales, pero dentro de esta corriente se presentan muchos ciclos iterativos.

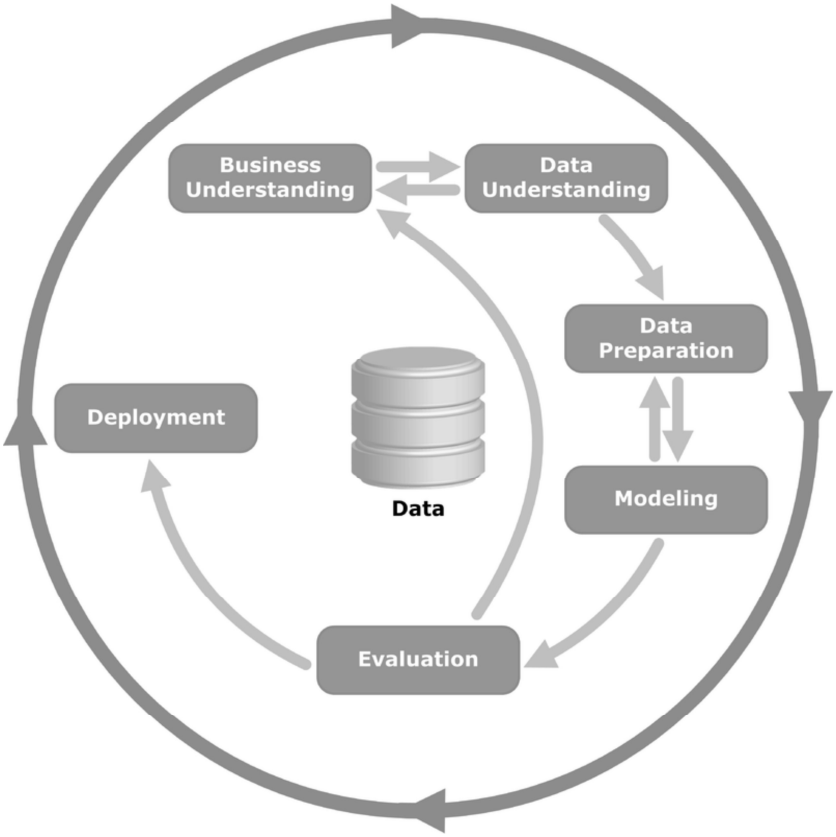

Fig. 1 Ciclo de vida del proceso de CRISP-DM. [15]

5.4.1. **FASE I: ENTENDIMIENTO DEL NEGOCIO**. Esta primera fase es probablemente la más importante y agrupa las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la minería de datos, es necesario entender de la manera más

completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados.

5.4.2. **FASE II: ENTENDIMIENTO DE LOS DATOS**. Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (Data Mining), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas.

5.4.3. **FASE III: PREPARACIÓN DE LOS DATOS**. En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, estas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, ya que, en función de la técnica de modelado elegida, los datos requieren ser procesados de una manera o de otra, por esta razón las fases de preparación y de modelado interactúan de forma permanente.

5.4.4. **FASE IV: MODELADO**. En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas para utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de los datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos se debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno de ellos. Después de concluir estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

5.4.5. **FASE V: EVALUACIÓN**. En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Hay que considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

5.4.6. **FASE VI: DESPLIEGUE**. En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o por ejemplo aplicando el modelo a diferentes conjuntos de datos o como parte del proceso (en análisis de riesgo de créditos, detección de fraudes, etc.). Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que se deben documentar y presentar los resultados de manera comprensible para el usuario con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

5.5. BIG DATA

Big data es el tipo de datos que tiene una escala muy grande y requiere un tipo especial de almacenamiento y técnicas para almacenarlos. Requiere algoritmos y técnicas especiales para procesarlo y obtener información útil a partir de datos sin procesar. Como este dominio es muy dominante en la era actual, tiene muchos desafíos, oportunidades y hay muchas tecnologías emergentes para el análisis de Big data. Dado que el Big data es de gran escala y los datos provienen de diversas fuentes y con diferentes formatos (Fig. 1), los sistemas de gestión de datos convencionales son incapaces de procesar Big data[9].

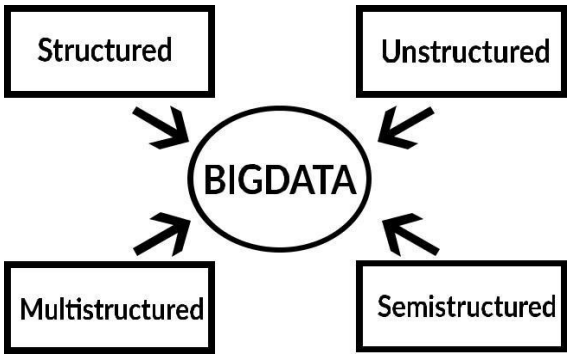

Fig. 2 Tipos de datos en Big data[9]

5.5.1. **Datos estructurados.** El tipo de datos en forma de tablas y filas con los nombres de atributos adecuados y estos datos se pueden mostrar fácilmente en hojas de cálculo.

5.5.2. **Datos sin estructura**. El tipo de datos que no son tablas y filas con nombres de atributos adecuados y es realmente difícil mostrar datos en una hoja de cálculo.

5.5.3. **Datos multiestructurados**. El tipo de datos que tienen diferente tipo de estructura significa que tienen algunos datos numéricos, algunos datos en forma de imágenes, etc.

5.5.4. **Datos semiestructurados**. El tipo de datos que tienen algunos datos estructurados y algunos datos sin la estructura adecuada puede deberse a una mezcla de datos de base de datos y archivos de registro. Los datos de todas estas categorías se almacenan juntos en Big data para análisis, por lo que es realmente difícil almacenar estos datos juntos.

5.5.5. **Dimensiones**. Big data tiene 3 dimensiones principales por sus características que se pueden describir de la siguiente manera: volumen, variedad y velocidad (Fig. 2). Volumen hace referencia a la cantidad de datos que llegan para análisis y es un desafío realmente difícil mantenerlos, almacenarlos en algún lugar y analizarlos. La variedad en Big data significa que se tienen diferentes tipos de datos, puede tener datos de video, puede tener datos textuales, datos binarios, audio, imágenes, numéricos, etc. En Big data, los datos tienen diferentes variedades, depende de las fuentes de donde vienen los datos. Y finalmente velocidad de los datos significa que los datos llegan a gran velocidad, lo que significa que hay solicitudes ilimitadas en los registros de los sitios web.

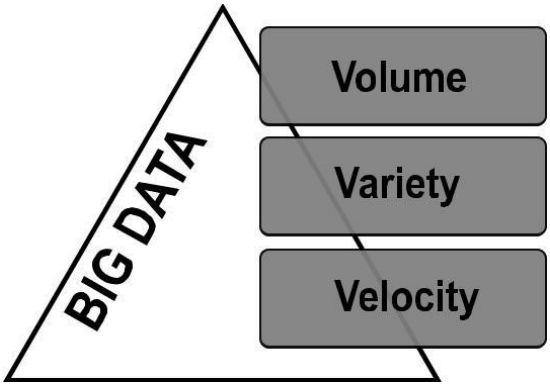

Fig. 3 Dimensiones en Big data[9]

5.6. TRABAJOS RELACIONADOS

La Universidad Estatal de Arizona (la más innovadora de Estados Unidos, según los últimos tres rankings publicados por U.S. News & World Report), implementa un sistema muy particular para monitorear a sus estudiantes[1]. La lógica es bastante simple, pero tiene implicaciones enormes: cada semestre, los alumnos "primíparos" reciben un carnet institucional que les sirve para todo. Con él ingresan a los edificios de clase, a la biblioteca, a las residencias y al centro recreativo; compran, como con una tarjeta de crédito, víveres en las tiendas del campus y entradas para el teatro más cercano, y hasta lo usan en las máquinas de gaseosas y dulces de la universidad. A cambio, la institución recibe información constante de dónde están, a qué hora y con quién.

"Es como un sensor que llevan puesto y que podemos usar para seguirles la pista. Aunque los carnets no están hechos para monitorear sus interacciones sociales, cruzando el tiempo exacto que están dos estudiantes en una ubicación puedes hacerlo", dijo en un comunicado oficial Sudha Ram, directora del Centro de Inteligencia y Analítica de Negocios (Insite, por sus siglas en inglés) de la universidad.

Ram ideó este programa de análisis de big data y empezó a aplicarlo hace tres años. Su objetivo es disminuir el factor de riesgo de deserción, pues, como encontró la directora, la integración social y la rutina de estudios de los estudiantes predicen mejor que las notas cuando un alumno de primer año va a desertar. Y, hasta ahora, lo ha logrado. "Tras solo 12 semanas de clase, somos capaces de identificar previamente el 85% de los estudiantes que desertarán a final del primer

semestre", señala Ram. En parte gracias a este dato, y a una intervención focalizada en los alumnos bajo riesgo, en el año 2017 la institución rompió su récord de retención estudiantil con un 85,6%, superando el promedio nacional (76%). Y, aunque este no es solo un logro del proyecto del Insite, sí demuestra lo que puede llegar a lograr el uso de learning analytics. Desde el punto de vista técnico es interesante, pero desde la privacidad es cuestionable.

Middle East College (MEC) es una de las instituciones de educación superior privadas más grandes de Omán. Student Success Center (SSC) es un departamento de apoyo en MEC para monitorear el progreso de los estudiantes[1]. El objetivo principal de este departamento es identificar los factores que afectan el desempeño de los estudiantes académicamente débiles y brindar intervenciones adecuadas en asociación con otros departamentos. MEC está afiliado a la Universidad de Coventry, Reino Unido y la Universidad de Wolverhampton, Reino Unido. Los módulos en varios programas de pregrado se clasifican en diferentes niveles como Nivel 1, Nivel 2 y Nivel 3 para incorporar habilidades de pensamiento de orden inferior y superior de acuerdo con la taxonomía de Bloom.

El estudio se realizó para determinar la asociación de la edad y el género con el desempeño en matemáticas entre estudiantes de secundaria en EE. UU. El estudio reveló que los puntajes de GPA en matemáticas disminuyeron con la edad. El trabajo de investigación utilizó técnicas de minería de datos para predecir los resultados de los estudiantes nuevos mediante el análisis del género, las notas obtenidas en el examen de calificación anterior y los exámenes de ingreso. El análisis utilizó atributos como género, raza, ciudad natal, ingreso familiar y modo de ingreso a la universidad para determinar el desempeño del estudiante utilizando varias técnicas de minería de datos. Esta investigación notó que la raza era el factor más importante que determinaba el desempeño del estudiante, seguido por el ingreso familiar.

AltSchool, una start up educativa de Estados Unidos ha puesto en marcha un ambicioso proyecto de big data con el fin de mejorar la educación de los estudiantes de 0 a 12 años[16].

El grupo escolar cuenta para ello con aplicaciones que controlan la asistencia de los alumnos y ordenadores y otras herramientas tecnológicas que registran permanentemente su actividad académica. Además, disponen de cámaras de vídeo para grabar constantemente lo que sucede en las aulas desde múltiples ángulos, con el fin de capturar las expresiones faciales de los niños, registrar su forma de hablar, el vocabulario, qué cosas les llaman más la atención, etc. El análisis de toda esa información proporciona una comprensión integral de cada alumno basada en sus patrones de conducta, estados de ánimo, rendimiento, etc., que permite darle a cada uno la educación que necesita atendiendo a sus necesidades y diferencias.

En Colombia, las pruebas de Estado Saber-Pro han sido diseñadas para apoyar la evaluación y el

mejoramiento de la educación superior en el país. Aplicando metodologías de minería de datos, se realizó un estudio de los resultados obtenidos en las pruebas Saber-Pro[17] de estudiantes de la Facultad de Ingeniería en Antioquia (Colombia). Este estudio obtuvo como resultado que se encuentra que algunas de las variables más influyentes sobre el resultado de las pruebas son: el número de personas a cargo, método de enseñanza, si el hogar es permanente, el carácter académico de la institución y facilidades económicas como tener horno micro gas y motocicleta.

6. METODOLOGÍA

Con el fin de dar cumplimiento al objetivo de poder predecir la continuidad de un estudiante de pregrado hacia posgrado bajo un conjunto de características relevantes de los estudiantes en su proceso de formación en la Pontificia Universidad Javeriana Cali, se aplicará la metodología CRISP-DM para el desarrollo de este proyecto siguiendo las fases que define la metodología y aplicando algunas actividades particulares de cada una de las fases

6.1. FASE I: ENTENDIMIENTO DEL NEGOCIO

En esta primera fase se indagará y se validarán las necesidades del negocio y de las áreas involucradas en el proceso de atracción de los estudiantes de la Pontificia Universidad Javeriana Cali con el objetivo de poder establecer un plan de proyecto, recolectando la información adecuada y alineada hacia el cumplimiento de los objetivos. Se definirán un conjunto de actividades que permitirán el entendimiento del negocio:

- Determinar los objetivos del negocio.
- Evaluación de la situación.
- Determinar los objetivos de la minería de datos.
- Realizar el plan del proyecto.

6.2. FASE II: ENTENDIMIENTO DE LOS DATOS

En esta segunda fase se realizará la recolección inicial de los datos que servirán como base para el desarrollo de la minería de datos, se determinará la cantidad de datos involucrados, el tipo de datos, su descripción y se establecerán las primeras relaciones más relevantes. En esta etapa se trabajará con una base de datos copia de la original con la información especifica para el proyecto que se desarrollará. Para esta etapa se contará con estructuras de datos de diferentes orígenes de datos almacenados desde los diferentes sistemas de información. Se definirán un conjunto de actividades que permitirán el entendimiento de los datos:

- Recolectar los datos iniciales.
- Descripción de los datos.
- Exploración de los datos.
- Verificar la calidad de los datos.

6.3. FASE III: PREPARACIÓN DE LOS DATOS

En esta fase se procederá a la preparación de los datos con el objetivo de seleccionar los datos a los que se va a aplicar la técnica de modelado seleccionada, se hará una limpieza de estos datos y se pueden generar de variables adicionales en caso de ser necesario. Se definirán un conjunto de actividades que permitirán el entendimiento de los datos:

- Seleccionar los datos.
- Limpiar los datos.
- Construir los datos.
- Integrar los datos.
- Formateo de los datos.

6.4. FASE IV: MODELADO

En esta fase se seleccionarán las técnicas de modelado más apropiadas para el proyecto dependiendo de las características de los datos y de la precisión que se requiera establecer con el modelo. Se definirán un conjunto de actividades que permitirán el entendimiento de los datos:

- Escoger la técnica de modelado.
- Generar el plan de prueba.
- Construir el modelo.

6.5. FASE V: EVALUACIÓN

En esta fase se realizará la evaluación del modelo, teniendo en cuenta el cumplimiento de los criterios de éxito que se definieron del problema y sobre los datos sobre los cuales se realizará el análisis. Se hará una revisión del proceso a fin de determinar los próximos pasos que pueden desencadenar en la explotación del modelo o en una nueva interacción desde el entendimiento del negocio. Se definirán un conjunto de actividades que permitirán el entendimiento de los datos:

- Evaluar los resultados.
- Determinar los próximos pasos.

6.6. FASE VI: DESPLIEGUE

En esta fase se producirá un informe final que resuma los resultados obtenidos en la aplicación del modelo y en la utilización de la metodología. Así como también se hará una revisión general del proyecto con aquellos puntos que pudieran presentar dificultad y sobre aquellos puntos en los que se acertará según el objetivo propuesto. Se definirán un conjunto de actividades que permitirán el entendimiento de los datos:

- Producir el informe final.
- Revisar el proyecto.

7. RECURSOS A EMPLEAR

7.1. HUMANOS

7.1.1. **Director**. María Constanza Pabón Burbano. Se desempeña como directora del Departamento Electrónica y Ciencias de la Computación de la Facultad de Ingeniería y Ciencias de la Pontificia Universidad Javeriana Cali. Cuenta con un Doctorado en Ingeniería de la Universidad del Valle y con una Maestría en Administración de Empresas en la misma institución. Graduada del Pregrado de Ingeniería de Sistemas y Computación de la Pontificia Universidad Javeriana Cali. Las investigaciones de la profesora se centran en proponer y desarrollar lenguajes y mecanismos de consulta de bases de datos con el objetivo de facilitar a los usuarios finales la formulación de consultas, en particular con modelos de datos de grafos.

7.1.2. **Estudiante**. Juan Felipe Mosquera García. Estudiante de la Maestría en Ingeniería de Software de la Pontificia Universidad Javeriana Cali y autor del presente trabajo de grado. Ingeniero Informático de la Universidad Autónoma de Occidente de Cali, desempeñándose en la actualidad como Ingeniero de Software del Centro de Servicios Informáticos de la Pontificia Universidad Javeriana Cali.

8. CRONOGRAMA DE ACTIVIDADES

Para el desarrollo del proyecto se plantea en el diagrama Gantt las diferentes actividades a desarrollar para el cumplimiento de los objetivos propuestos. El proyecto tiene un tiempo de duración de 6 meses y estará dividido en fases y actividades acorde a la metodología CRISP-MD. También contará con reuniones de seguimiento periódicas con el director de proyecto para la presentación de los avances; las reuniones se darán cada 3 semanas para verificar los resultados de las actividades realizadas. El cronograma del proyecto se muestra en la Tabla I.

TABLA I
Cronograma del proyecto

Fase	Actividad	Mes					
		1	2	3	4	5	6
Entendimiento del negocio	Determinar los objetivos del negocio						
	Evaluación de la situación						
	Determinar los objetivos de la minería de datos						
	Realizar el plan del proyecto						
Entendimiento de los datos	Recolectar los datos iniciales						
	Descripción de los datos						
	Exploración de los datos						
	Verificar la calidad de los datos						
Preparación de los datos	Seleccionar los datos						
	Limpiar los datos						
	Construir los datos						
	Integrar los datos						
	Formateo de los datos						
Modelado	Escoger la técnica de modelado						
	Generar el plan de prueba						
	Construir el modelo						
Evaluación	Evaluar los resultados						
	Determinar los próximos pasos						
Despliegue	Producir el informe final						
	Revisar el proyecto						

9. REFERENCIAS BIBLIOGRÁFICAS

[1] Gobernación de Antioquia *et al.*, "OBSERVATORIO CT + i," *Cons. Nac. Política Económica Y Soc.*, vol. 53, no. 1, pp. 34–55, 2015.

[2] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[3] S. Basu, "Data Mining," *Georgia State University Fall*, 1997. http://www6.uniovi.es/hypvis/applicat/data_mining/data_mining.html (accessed Apr. 05, 2021).

[4] M. Jose, P. S. Kurian, and V. Biju, "Progression analysis of students in a higher education institution using big data open source predictive modeling tool," *2016 3rd MEC Int. Conf. Big Data Smart City, ICBDSC 2016*, pp. 113–117, 2016, doi: 10.1109/ICBDSC.2016.7460352.

[5] Z. Bošnjak, O. Grljević, and S. Bošnjak, "CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data," *Proc. - 2009 5th Int. Symp. Appl. Comput. Intell. Informatics, SACI 2009*, no. 114, pp. 509–514, 2009, doi: 10.1109/SACI.2009.5136302.

[6] J. W. Seifert, "CRS Report for Congress Data Mining," *Reading*, pp. 1–16, 2004.

[7] X. Yu and S. Wu, "Typical Applications of Big Data in Education," *Proc. - 2015 Int. Conf. Educ. Innov. Through Technol. EITT 2015*, pp. 103–106, 2016, doi: 10.1109/EITT.2015.29.

[8] S. Yu, D. Yang, and X. Feng, "A Big Data Analysis Method for Online Education," *Proc. - 10th Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2017*, vol. 2017-Octob, pp. 291–294, 2017, doi: 10.1109/ICICTA.2017.71.

[9] A. Jamil, M. Abdullah, M. A. Javed, and M. S. Hassan, "Comprehensive Review of Challenges Technologies for Big Data Analytics," *2018 IEEE Int. Conf. Comput. Commun. Eng. Technol. CCET 2018*, pp. 229–233, 2018, doi: 10.1109/CCET.2018.8542219.

[10] A. Anjewierden, H. Gijlers, N. Saab, and R. De Hoog, "Brick: Mining pedagogically interesting sequential patterns," *EDM 2011 - Proc. 4th Int. Conf. Educ. Data Min.*, pp. 341–342, 2011.

[11] C. Russo, H. Ramón, N. Alonso, B. Cicerchia, L. Esnaola, and J. P. Tessore, "Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning," *XVIII Work. Investig. en Ciencias la Comput. (WICC 2016, Entre Ríos, Argentina)*, p. 131, 2016.

[12] A. González, *¿Qué es Machine Learning?* 2014, p. 14.

[13] J. Villena Román, "CRISP-DM: La metodología para poner orden en los proyectos," *Sngular*, 2016. https://www.sngular.com/es/data-science-crisp-dm-metodologia/ (accessed Apr. 18, 2021).

[14] P. C. Ncr *et al.*, "Step-by-step data mining guide," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[15] J. F. Vallalta, "CRISP-DM: una metodología para minería de datos en salud - healthdataminer.com." https://healthdataminer.com/data-mining/crisp-dm-una-

metodologia-para-mineria-de-datos-en-salud/ (accessed Apr. 17, 2021).

[16] B. Herold, “The future of big data and analytics in K-12 education,” *Ed Week*, 2016. https://www.edweek.org/policy-politics/the-future-of-big-data-and-analytics-in-k-12-education/2016/01 (accessed Mar. 20, 2021).

[17] A. I. Oviedo Carrascal and J. Jiménez Giraldo, “Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO,” *Rev. Politécnica*, vol. 15, no. 29, pp. 128–140, 2019, doi: 10.33571/rpolitec.v15n29a10.