Minería de datos para la predicción de la deserción estudiantil en la Universidad Nacional Abierta y a Distancia

Data mining on predicting student dropout at Universidad Nacional Abierta y a Distancia

M. Ávila Pérez[1]
Escuela de Ciencias Básicas Tecnologías e Ingeniarías,
Universidad Nacional Abierta y a Distancia UNAD, Bogotá, Colombia.

J. Medina Cruz[2]
Escuela de Ciencias Básicas Tecnologías e Ingeniarías,
Universidad Nacional Abierta y a Distancia UNAD, *Bogotá, Colombia,* Colombia.

Resumen

Este documento tiene como fin divulgar una propuesta de investigación en el marco de la maestría en gestión de TI de la Universidad Nacional Abierta y a Distancia, el documento presenta un planteamiento donde se expresa la oportunidad de aplicar técnicas de analítica de datos a la información de los estudiantes que se recopila de los procesos académicos de la UNAD. Los cuales, son susceptibles de analizar mediante técnicas de minería de datos para generar un modelo de predicción de la deserción estudiantil, con el propósito de contribuir con la adopción de estrategias que permitan implementar medidas paliativas para disminuir este fenómeno que afecta no solamente a la UNAD sino a todas las instituciones de educación en el país y en el mundo. El análisis de grandes cantidades de información mediante data mining ha permitido afinar las estrategias y campañas en campos como la inteligencia de negocios; en el campo de la educación la aplicación de técnicas de data mining a través del análisis de grandes volúmenes de datos, proporcionan un soporte para la toma de decisiones, lo que permite a los directivos de las instituciones concentrar los esfuerzos o dirigirlos a ciertos ámbitos o área específicas, lo que mejora enormemente la efectividad en los procesos permitiendo acercarse al conocimiento de manera más efectiva y eficiente.

Palabras clave*:data mining, big data, árboles de decisión, predicción, KDD.*

Abstract

This document aims to disseminate a research proposal within the framework of the master's degree in IT management at Universidad Nacional Abierta y a Distancia. The document presents an approach expressing the opportunity to apply data analytics techniques to student information that is collected from UNAD's academic processes. Which, are able to analyze using data mining techniques to generate a model of predicting student dropout in order to contribute to the adoption of strategies to implement mitigation measures to reduce this phenomenon that affects not only UNAD but all educational institutions in the country and the world. Analyzing large amounts of information using data mining has enabled the refinement of strategies and campaigns in fields such as business intelligence; in the field of education the application of data mining techniques through the analysis of large volumes of data, provide support for decision-making, allowing managers of institutions to concentrate efforts or direct them to certain specific

[1] mavilap@gmail.com. https://orcid.org/0000-0002-7834-3578
[2] Javier.medina@unad.edu.co. https://orcid.org/0000-0001-8047-2259

ambits or areas, greatly improving process effectiveness by allowing them to approach knowledge more effectively and efficiently

Keywords: *Data mining, Big data, Decision Trees, Prediction, KDD.*

1. Introducción

La minería de datos o data mining en inglés, es una técnica que ha dado muy buenos resultados en diferentes campos del conocimiento, contribuyendo a que los encargados de la toma de decisiones de las organizaciones tengan un soporte más confiable y asertivo en la administración de estas(Cesarotto & Yuri, n.d.).

Una problemática dentro de las instituciones de educación superior es la deserción estudiantil, la cual afecta de forma negativa los indicadores de gestión, por lo que representa una necesidad el reconocer cuáles podrían llegar a ser las causas de dicha deserción y las formas efectivas de mitigarla. La mayoría de las instituciones cuentan con la información que necesitan para identificar las causas, sin embargo, no han adoptado las técnicas de tratamiento de esta información, pero hoy en día con los avances tecnológicos en el campo de la inteligencia artificial y específicamente a través de técnicas de data mining que permiten el análisis de estos grandes volúmenes de información, a lo cual se le denomina BigData(Dumon, 2014) es posible extraer los datos, y analizarlos aplicando técnicas de analítica de datos.

Es así como se ha propuesto el desarrollo de un modelo para la predicción de la deserción estudiantil en un curso de primera matrícula de la ECBTI de la UNAD, mediante el uso de herramientas de data mining. Este modelo se planea construir con base en la ejecución de un análisis diagnóstico que haga visible los requerimientos para el desarrollo de un proyecto de data mining, para que provea los mecanismos para la predicción de deserción estudiantil. Con el fin de garantizar la validez del modelo se plantea realizar la evaluación de la efectividad del prototipo mediante la comparación con datos históricos obtenidos durante el año 2018, de esta manera, observar los datos arrojados por el modelo y compararlos con los datos reales.

Ahora bien, la Universidad Nacional Abierta y a Distancia es una institución superior del orden nacional, que debido a su modalidad puede llegar a donde otras instituciones no llegan, lo cual le ha permitido crecer y contar hoy con más de 100.000 estudiantes, lo que la convierte en una de las instituciones más grandes de Colombia en cuanto a número de estudiantes. Sin embargo, una gran parte de estos estudiantes desertan por diferentes motivos.

La deserción estudiantil es un fenómeno que afecta la población estudiantil de las instituciones de educación superior, y se da cuando los estudiantes abandonan su proceso académico por diversas razones. La Universidad Nacional Abierta y a Distancia UNAD no es ajena a este fenómeno como lo expresan Ángel & Facundo (2009) en su trabajo titulado “Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD-Colombia”.

Para el estudio de esta problemática hay que ir varios años atrás, Ángel & Facundo (2009) analizaron la corte 2001-I, a la cual hubo ingreso de 6.011 estudiantes, y se halló que, luego de nueve años, 1.432 estudiantes han obtenido grado (el 23.82%) y aún se encuentran en la institución 321 (el 5.34% de los estudiantes de la cohorte). Esto quiere decir que, para el estudio en el año 2008, la deserción fue de 70.84%.

Estas cifras han venido mejorando año tras año, pero aún persisten altos índices de deserción que, para los últimos años, según datos obtenidos de consejería académica, ronda el 35%. Para Ángel & Facundo (2009) La deserción implica un desperdicio de recursos tanto públicos como privados, así como de esfuerzos, que deja muchos sinsabores y frustraciones en los afectados. En este mismo sentido Rodríguez afirma que los costos sociales de la deserción estudiantil son extremadamente altos, lo cual repercute en un incremento de las tasas de criminalidad y déficit en la tasa de crecimiento económico y afectación de otros índices como el aumento de la inequidad o de la brecha entre ricos y pobres (Rodríguez *et al.*, 2016).

2. Materiales y métodos

Para este trabajo, se realizó una revisión bibliográfica, con el objetivo de sintetizar los conceptos y teorías acerca del uso de Big-Data en

los procesos educativos, mediante la consulta de material bibliográfico en diferentes bases de datos, con la finalidad de proporcionar sustento científico a las temáticas que se abordan en el desarrollo de esta investigación. Para ello, se identifican referentes actualizados pertinentes con el problema: se consulta bibliografía de fuentes primarias de autores que han abordado la problemática de Big-Data aplicada a la educación. Se consultaron varias bases de datos de fuentes bibliográficas, y se aplicaron los criterios de inclusión y exclusión a los resultados obtenidos. Para los criterios de inclusión se tuvieron en cuenta los siguientes:
-Título: ¿Es útil? ¿Es relevante?
-Autores: experiencia en el tema.
-Resumen: ¿Resultados aplicables?
-Resultados
-Año de publicación: se les dio prioridad a los recursos publicados durante los últimos 6 años.

Para los criterios de exclusión se tomaron en cuenta los siguientes:
-Información relacionada pero no exclusivamente necesaria.
-Aunque la temática es de big-data el resumen parece focalizado en temas irrelevantes con relación a los objetivos.
-La temática tratada en el recurso tiene poca aplicabilidad para el problema planteado.
-Contenido poco pertinente con la justificación y las temáticas de la investigación.
Las bases de datos que se consultaron son las siguientes: Google academic, Library & Information Science Source, Business Source Premier, Applied Science & Technology Source, IEEE, EbscoHost.

Para los términos de búsqueda se utilizaron los siguientes:
-El big-data y la educación.
-Técnicas de Big-Data aplicadas a la educación.
-Beneficios del Big-Data aplicados a la educación.
-Mejoramiento de procesos educativos mediante el uso de Big-Data.
-Aplicabilidad del Big-Data en la educación.
-Herramientas de Big-Data, Algoritmos de Big-Data.

Se emplearon las siguientes cadenas de búsqueda para cada una de las bases de datos
-Big-data education
-Big-data in education
-Education big data
-Techniques big data applied to education
-Casos de éxito del big-data aplicado a la educación
-Tesis doctorales big-data en la educación

A partir de esta revisión se exponen los referentes teóricos siguientes:

2.1 Big data

Este concepto se ha vuelto un tema del que muchos hablan y usan para describir grandes cantidades de información, pero también como afirman Boyd y Crawford (2012) lo interesante está en “las capacidades de búsqueda y agregación de grandes cantidades relacionales”. El análisis de big-data tiene un impacto económico fuerte en muchos sectores tanto públicos como privados, permitiendo el aumento de la productividad, la competitividad, la calidad de vida de los ciudadanos y el medio ambiente. El resultado del procesamiento de toda esta información constituye un soporte muy valioso para la toma de decisiones. Mediante técnicas de analítica de datos, se busca hallar correlaciones entre las variables analizadas que permitirán identificar patrones de comportamiento en los datos, que se convierten en ventaja competitiva para las organizaciones(Niño & Illarramendi, 2015)

En el campo de la educación, cuando se fusionan las técnicas del big-data y la educación se habla de un mundo donde los pedagogos pueden analizar el comportamiento de sus estudiantes, dando la posibilidad de dar un enfoque de aprendizaje personalizado y permitiendo que el estudiante alcance su pleno potencial (Malvicino & Yoguel, 2015).

En este mismo sentido, se encuentra que para el análisis de esta gran cantidad de datos, big data se apoya en técnicas de minería de datos mediante el uso de herramientas de data minig. En la actualidad existen muchas herramientas para minería de datos, algunas de ellas son propietarias, como el caso de Oracle Data Mining o IBM SPSS Modeler, estas herramientas de pago son muy robustas y cuentan con mucho soporte por parte del fabricante. Este es un mercado que ha venido en expansión y también existen herramientas de software libre para data mining, las cuales igualmente son muy potentes. Una de las más destacadas es Waikato Environment for Knowledge Analy-sis (WEKA), la cual inició en el año 94 y su primer lanzamiento público fue en

1996. En 1999, fue completamente codificada en JAVA y desde entonces ha tenido varias actualizaciones. Los módulos de WEKA se han integrado en muchas otras herramientas de código abierto como Pentaho, RapidMiner y KNIME (Mikut & Reischl, 2011).

Se pueden clasificar las herramientas de minería de datos, algunas como librerías y otras como suites:

1) ***Librerías***: es necesario tener conocimientos de programación para ser usadas, entre estas podemos mencionar las siguientes: Xelopes y JDMP. El uso de estas librerías agrega funcionalidad y potencia a funcionalidades específicas.
2) ***Suites***: existe un variado número de aplicaciones para data mining dentro de las que se pueden destacar: IBM SPSS Modeler, WEKA, Oracle Data mining, RapidMiner y Statistica Data Miner, las cuales son de las más populares como se aprecia en la Figura 1, que representa una encuesta realizada en el año 2015 (Mikut & Reischl, 2011).

2.2 IBM SPSS ModelerEs

IBM SPSS Modeler, es una plataforma de análisis predictivo diseñada por IBM que permite llevar a cabo procesos de inteligencia predictiva sobre decisiones (Disla & Llaugel, 2015).

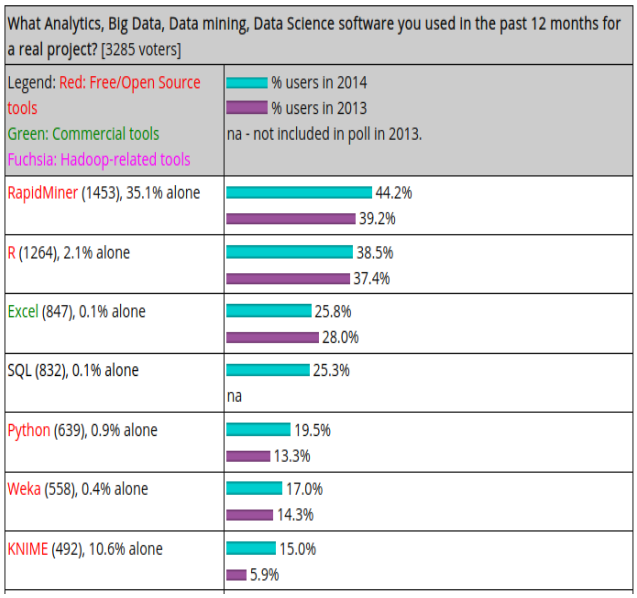

Fig. 1. Herramientas de minería de datos más usadas. Fuente: https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

2.3 Clasificación basada en árboles de decisión

Los árboles de decisión posibilitan la obtención de patrones para tomar decisiones sobre el comportamiento de una población, basados en datos históricos que se han almacenado. Se destaca la facilidad de interpretación como uno de sus principales atributos. Esta técnica tiene aplicabilidad en diversas disciplinas del conocimiento y es de mucha utilidad cuando en un set de datos, se tiene una característica conocida, pero se desconoce esa misma característica en un elemento concreto (Han, Kamber, & Pei, 2011). La clasificación basada en arboles de decisión, es tal vez una de las técnicas que más se ha popularizado en data mining, los árboles de decisiones se han usado en investigación de operaciones para describir modelos jerárquicos de decisión y su consecuencia, en data mining su uso tiene que ver con modelos analíticos encaminados a realizar predicciones (Ramírez & Grandón, 2018). "Esta técnica no paramétrica clasifica una población en un modelo de segmentos de tipo ramas que construyen un árbol invertido, y luego este modelo se utiliza para predecir una variable objetivo" (Ramírez & Grandón, 2018).

El enfoque de esta investigación se considera proyectivo, toda vez que se persigue la utilización o aplicación de tecnología de data mining que permita realizar el procesamiento y análisis de la información de los cursos almacenada en bases de datos y a partir de allí aplicar técnicas de inteligencia artificial que permitan inferir tendencias o comportamientos que puedan ser aprovechados para el mejoramiento de los índices de deserción en la UNAD. Otra razón para considerarse proyectivo, es que se planea desarrollar un proceso metódico de exploración y búsqueda que incluye actividades en las cuales se describe, se compara, se analiza, se explica y predice (Hurtado, 2000).

En otro sentido, en cuanto a la metodología de esta investigación se considera que esta investigación es de tipo mixta debido a que se enfoca en información cualitativa en cuanto a que se describe información, pero también cuantitativa toda vez que algunos datos de variables son presentados cuantitativamente, sin embargo,

prevalece la descripción o el significado de esos datos (Lerma González, 2016).

Para Hurtado (2000), la investigación proyectiva se apoya en la creación de una propuesta ya sea un modelo, plan, etc. como recurso para la solución que responda a una necesidad de tipo práctico, presentada en un espacio geográfico determinado, grupo social o institución. Este tipo de investigación se da en un área específica del conocimiento, en el cual se realiza un diagnóstico de las necesidades, se analizan las causas que las producen y se buscan tendencias futuras. Además, Hurtado (2000) sostiene que "La investigación proyectiva se ocupa de cómo deberían ser las cosas, para alcanzar unos fines y funcionar adecuadamente. La investigación proyectiva involucra creación, diseño, elaboración de planes, o de proyectos".

Para el desarrollo de este modelo de predicción se han planteado las siguientes fases, enmarcadas en la metodología de minería de datos CRISP-DM (Azevedo & Santos, 2008), a través de las cuales se proyecta lograr los objetivos propuestos.

1) ***Fase 1:*** se llevará a cabo una revisión bibliográfica en bases de datos especializadas con la finalidad de recabar información sobre la tecnología big data, sus aplicaciones y tendencias. Asimismo, la recolección de la información necesaria acerca de las técnicas de big data. Además, realizar un análisis de los algoritmos predictivos existentes con el fin de determinar los adecuados. También identificar las fuentes de donde se tomará la información. Y, por último, la recolección, preparación, procesamiento y transformación del set de datos a usar en el análisis.

2) ***Fase 2:*** se diseñará una arquitectura con los componentes de la tecnología big data que se utilizarán, se prepararán los datos para su procesamiento, se determinarán los algoritmos de machine learning e implementarán los que se consideren adecuados o más efectivos para el análisis de la información.

3) ***Fase 3:*** durante esta fase se desplegará el producto mínimo viable PMV del modelo de predicción, con sus componentes, teniendo como guía el documento de diseño, para realizar el análisis de los datos.

4) *Fase 4:* plan de pruebas del prototipo; se ajusta el proceso, se mejora la funcionalidad a medida que se prueba el modelo. Así como la prueba del modelo de deserción.

La población para esta investigación se toma dentro de un sector de la población estudiantil de la UNAD, específicamente a los estudiantes de primera matrícula del año 2017 con el propósito de poder observar el comportamiento en el tiempo durante los periodos 2018 y 2019 y de esta manera validar el modelo de predicción de deserción.

Para la muestra, dentro de esta población de estudiantes se pretende tomar el subgrupo de estudiantes del curso Herramientas digitales, que se matricularon en el mismo en el primer periodo de 2017.

3. Resultados

Poniendo en contexto los resultados de esta investigación, teniendo en cuenta que se encuentra en una fase de formulación, se han conseguido avances, los cuales se encuentran enmarcados en el primer objetivo específico, y están contemplados en la Fase 1. Para esto se ha revisado bibliografía existente en lo que tiene que ver con el abordaje de las metodologías para la aplicación en proyectos de minería de datos como se evidencia en los apartados anteriores.

Otro de los logros tiene que ver con la revisión de las metodologías KDD para el descubrimiento de conocimiento, en este sentido León & García (2016) afirman que las tres metodologías que marcan la pauta para el proceso de la minería de datos son: KDD, CRISP-DM y SEMMA. Cada una con sus características.

En este orden de ideas, KDD es una metodología compuesta por 5 fases: selección, preprocesamiento, transformación, minería de datos y evaluación e implantación, esta metodología plantea un proceso iterativo e interactivo en cada fase (Moine, 2013).

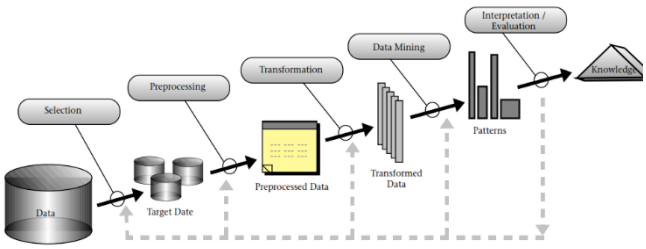

Fig 2. Etapas del proceso de KDD. Fuente: https://wvvw.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131

Siguiendo lo anterior, la metodología CRISP-DM define un modelo de proceso de minería de datos que describe enfoques comunes recomendados para aplicar técnicas de data mining. La metodología plantea seis fases de las cuales en este momento se avanza con la consecución de las dos primeras. La primera, la comprensión del negocio, la cual pretende determinar los requisitos y objetivos desde la perspectiva de negocio a fin de materializarlos en un punto de vista técnico y en un plan de proyecto; y la segunda es la comprensión de los datos donde a partir de una recolección inicial de datos se establecen premisas acerca del problema y de los datos a procesar. Estas fases están compuestas por distintas tareas que permiten su realización (Azevedo & Santos, 2008).

Es así como, CRISP-DM (Cross-Industry Standard Process for Data Mining) tiene como objetivos fomentar la interoperabilidad en cada una de las etapas de todo el proceso de minería de datos, además busca la reducción de malas experiencias derivadas de altos costos en la minería de datos.

Por otro lado, SEMMA fue propuesta por SAS Institute Inc, y se define como un "proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones desconocidos". SEMMA parte datos estadísticos y a partir de estos "pretende facilitar la exploración estadística, las técnicas de visualización, seleccionar y transformar las variables más significativas en la predicción, modelar las variables para predecir salidas y finalmente confirmar la precisión del modelo" (León & García, 2016).

En este orden de ideas, se considera que, SEMMA y CRISP-DM se pueden percibir como una implementación de KDD a primera vista, sin embargo, en la medida en que se profundiza en estas metodologías de minería de datos se encuentra que la metodología CRISP-DM es más completa, toda vez que como lo afirman Azevedo & Santos, las metodologías CRISP-DM guían hacia cómo se puede aplicar la minería de datos en la práctica, en sistemas reales(Azevedo & Santos, 2008)

En cuanto a la comprensión del negocio, se tiene que la aplicación de la minería de datos en este proyecto tiene como objetivo predecir aquellos estudiantes que sean más susceptibles a la deserción en cursos posteriores a la primera matrícula. Con respecto a este proyecto, se puede afirmar que se cuenta con una base de datos que tiene registrada la información necesaria acerca de los estudiantes cursando una titulación, los que han desertado y de los que ya la han terminado. Ahora bien, como criterio de éxito, se establece la realización de las predicciones con alto porcentaje de confiabilidad que sea un soporte para la toma de decisiones. Con respecto a los recursos de software con que se cuentan tenemos herramientas de Open Source, Linux CENTOS, Orange, WEKA, máquina virtual de Virtual Box, en cuanto a recursos de hardware tenemos PC HP all in one con 8Gb de ram 1 terabit de disco, procesador Intel core 7.

Como resultado del trabajo de la revisión bibliográfica se obtuvo el siguiente estado del arte a través del cual se han podido revisar trabajos previos que la comunidad científica viene desarrollando.

La investigación de Romero, Toledo & Paredes (2017), quienes realizaron el estudio de Implementación de un sistema de análisis de datos en la deserción estudiantil mediante la utilización de técnicas de big data, con el fin de facilitar la estructuración de planes de mejoramiento de la Universidad Mariana. En este trabajo se manejan las técnicas del big data como herramienta para soportar un sistema de análisis y predicción de datos para obtener un pronóstico del comportamiento en la deserción estudiantil (Romero, Toledo & Paredes, 2017).

El artículo publicado en el 2016 titulado "Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico" (Hernandez Gonzalez *et al.*, 2016), presenta un estudio en el cual se comparan los algoritmos de regresión logística, algoritmos de clústeres, árboles de decisión y la red neuronal de Microsoft. El

estudio muestra que “sí es posible predecir los alumnos que tienen altas posibilidades de desertar de sus estudios a nivel superior. En este caso, fue un análisis sobre los alumnos del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones” (Hernandez Gonzalez *et al.*, 2016).

El trabajo de Díaz & Osorio (2013) Titulado "Aplicando estrategias y tecnologías de inteligencia de negocio en sistemas de gestión académica". Presenta una línea de investigación resultado de actividades relacionadas con la "aplicación de herramientas y técnicas de inteligencia de negocio a datos almacenados en los sistemas académicos de gestión universitaria, que se utilizan para la operativa diaria en las unidades académicas" (Díaz & Osorio, 2013).

Con esta información presentada también se ha propuesto un plan de trabajo para el desarrollo del proyecto y se ha realizado una estimación de los costos del mismo, todo esto buscando llevar este proyecto de data mining a feliz término.

4. Conclusiones

La aplicación de la metodología CRISP-DM aporta al desarrollo de este proyecto las siguientes ventajas:

Es una metodología no–propietaria, ofrece mucha libertad o independencia con respecto a las herramientas que se utilicen, la aplicación o la industria. CRISP-DM es imparcial con relación a las herramientas, y está encaminado al análisis técnico, así como en problemas de negocio. Asimismo, brinda una plataforma guía, una experiencia piloto y plantillas dispuestas para análisis.

El uso de herramientas de minería de datos como RapidMine o WEKA representan un soporte importante para este tipo de investigaciones, toda vez que al ser herramientas con licencia open source abarata los costos en el desarrollo de proyectos de esta naturaleza, máxime cuando se trata de proyectos que no tienen patrocinio económico, y que son desarrollados en entidades de carácter público en las que los presupuestos económicos son muy limitados en materia de recursos monetarios, de tal manera que estas herramientas se vienen a convertir en el medio disponible a través de las cuales es posible alcanzar los objetivos de un proyecto de data minig.

La información es uno de los activos más importantes para las organizaciones, y el hecho de poseer o almacenar datos de los clientes (estudiantes) no necesariamente indica que se tenga un soporte robusto para la toma de decisiones. No obstante para que esos datos verdaderamente sean explotados en su máxima capacidad es necesario realizar procesos que los organice con la finalidad de convertirlos en información, una vez organizados estos datos, se convierten en información y esta información aporta a la consecución de los objetivos estratégicos de la universidad, pero a esa información es necesario agregarle conocimiento y es a través de la agregación de este conocimiento que se crea la cadena de valor como lo explica Michel Porter.

Agregar conocimiento a la información se logra mediante la capacidad de análisis de ésta, que posee la organización, este análisis es posible hacerlo o se logra aplicando técnicas de big data, en este caso minería de datos, a esta información que se tiene almacenada en diferentes medios y formatos.

Mediante el análisis de grandes cantidades de información almacenada en múltiples formatos se pueden descubrir patrones, que, a través de la aplicación de algoritmos como los árboles de decisión, se logra predecir el comportamiento de algunas variables; estas predicciones constituyen una base robusta para orientar las decisiones estratégicas de la organización.

En la UNAD, en cada periodo, se realiza una encuesta de caracterización muy robusta, la cual recolecta una enorme cantidad de información, que constituye una enorme fuente de datos, que mediante este proyecto se propone analizar, y a partir de allí, hacer predicciones acerca de cómo probablemente será el comportamiento en términos de retención y permanencia de los estudiantes en la institución , con lo cual se está avanzando en la dirección de la innovación y la aplicación de inteligencia artificial para la solución de problemas cotidianos, aplicando los avances tecnológicos, logrando así, la optimización de la toma de decisiones, y, maximizando de esta manera los

recursos. En este orden de ideas esta investigación permitirá identificar patrones que podrían marcar el inicio hacia una educación personalizada que permita, por ejemplo, brindar un acompañamiento más preciso y acorde con las necesidades de los estudiantes, presentándose este como un modelo que pretende proporcionar un mecanismo de predicción para la deserción estudiantil mediante la aplicación de técnicas de minería de datos.

Agradecimientos

Agradezco a Dios por permitirme escribir este artículo, a mi esposa Leslie y a mis hijos por darme todo el apoyo necesario y en especial a mi hijo Isaac quien ha sido una ayuda invaluable en todo momento.

Referencias

Para incluir citas

Acevedo, Y. V. N., Quintero, J. F. L. & Clavijo, C. C. G. (2016). Recorrido virtual en tercera dimensión de la sede principal en una universidad de Bogotá. *Publicaciones e Investigación*, *10*, 83-93.

Abello Mendoza, E. N., & Bernal Suárez, W. F. (2017). Prototipo para la orientación automática de paneles solares. https://repository.unad.edu.co/handle/10596/29750

Agreda, F. U. P. & Castrillón, J. H. (2017). Aplicación de la técnica smed en el procedimiento de cambio de tintas de la referencia bolsa kraff colanta entera 3c a bolsa kraff amtex tannus 2c. *Publicaciones e Investigación*, *11*(1), 113-124.

Alegría, Y. M., Collazos, C. A., Granollers, T. & Gil, R. (2014). Propuesta de valoración del comportamiento como complemento a la evaluación emocional de los usuarios mientras interactúan con sitios web. *Publicaciones e Investigación*, *8*, 185-201.

Ángel, H., & Facundo, D. (2009). Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD - Colombia. *Revista de Investigaciones UNAD*, *8*(2), 117. https://doi.org/10.22490/25391887.639

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM. http://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Barragán, F. M. M. (2017). Formulación y elaboración de productos de panificación con yacón (*Smallanthus sonchifolius*) como endulzante, para la población con deficiencias en el metabolismo de los disacáridos. *Publicaciones e Investigación*, *11*(1), 127-139.

Bastidas, S. E. C., Cabrera, A. A., Mez, H. E. C. & Cervelion, A. J. (2019). Sistema en tiempo real para el monitoreo de variables médicas en pacientes hospitalizadas con redes WSN. *Publicaciones e Investigación*, *13*(1), 27-44.

Bastidas, S. E. C., & Peláez, J. M. L. (2015). Algoritmos de planificación para la transmisión de datos en tiempo real con IEEE 802.15. 4. https://hemeroteca.unad.edu.co/index.php/publicaciones-e-investigacion/article/view/1443/1883

Bautista, E. A. S., Roa, J. R. V., & Ortega, J. A. T. (2015). Estimación de la huella hídrica para un cultivo de pitahaya amarilla (Selenicereus megalanthus). *Publicaciones e Investigación*, *9*, 135-146.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://www.tandfonline.com/doi/full/10.1080/1369118x.2012.678878?casa_token=pc46CbAX1dYAAAAA:F9DsdKTfi4WakGoFJOOrUOmWVjL7b-fKt_7kmZitJJadJMFVeWR-88PF48Xw6IUjQ3wzOuEPHPWiOATZ5w

Bríñez, J. A. B., Cuevas, M. M. & Torres, M. (2014). Análisis de parámetros objetivos y subjetivos en pre-amplificadores de audio. *Publicaciones e Investigación*, *8*, 13-24.

Castañeda, C. C. C. (2016). Ros-gazebo. una valiosa Herramienta de Vanguardia para el desarrollo de la robótica. *Publicaciones e Investigación*, *10*, 145-160.

Cerra Escobar, I. L., & Villarreal Padilla, J. E. (2017). State of art: utilizing social network analysis in diverse fields. *Publicaciones e Investigación*, *11*(1), https://doi.org/10.22490/25394088.2257

Cesarotto, Y., & Yuri. (n.d.). El debate académico en curso sobre 'big data' y su incidencia en la comprensión de la comunicación mediática contemporánea. Recercat (Dipòsit de La Recerca

de Catalunya). http://recercat.cat/handle/2072/335833

Cifuentes, A. F. M. & Clavijo, C. C. G. (2015). Marco de referencia para la gestión de TI centrada en la creación de valor compartido, aplicado a una propuesta de formación en maestría. *Publicaciones e Investigación, 9*, 163-176.

Cruz, A. V., Cordero, L. A. & González, A. P. (2014). Evaluación energética de los generadores de vapor F1-2 y BH-109 de una refinería cubana de petróleo. *Publicaciones e Investigación*, *8*, 89-96.

Delgado, Á. D. G., Ruiz, Y. Y. P., Córdoba, L. S., López, L. M., & Kafarov, V. (2014). Experimentación y optimización conjunta de la disrupción celular de microalgas y extracción soxhletde aceite para alimentación y biocombustibles. *Publicaciones e Investigación*, *8*, 127-136.

Díaz, F., & Osorio, M. (2013). Aplicando estrategias y tecnologías de inteligencia de negocio en sistemas de gestión académica. Sedici, UNLP. http://sedici.unlp.edu.ar/handle/10915/27157

Díaz, J. M. G., Díaz, N. G., & Cuellar, A. M. Q. (2010). Comparación entre los índices de agua potable IAP y los índices de riesgo de la calidad de agua para consumo humano IRCA utilizados para la determinación de la calidad del agua para consumo humano. *Publicaciones e Investigación*, *4*, 53-59.

Disla, R., & Llaugel, F. (2015). Un modelo predictivo de deserción escolar para la República Dominicana. https://www.researchgate.net/profile/Renato_Gonzalez-Disla/publication/311951602_Paper-Un_Modelo_Predictivo_de_Desercion_Escolar_en_la_Republica_Dominicana_v2/links/5864906208ae8fce490b7666/Paper-Un-Modelo-Predictivo-de-Desercion-Escolar-en-la-Republica-D

Dumon, O. (2014). Big data and Education: The Power of Transformation. *Research Information, 75,* 10. https://search-ebscohost-com.bibliotecavirtual.unad.edu.co/login.aspx?direct=true&db=lxh&AN=99617894&lang=es&site=ehost-live

Fernández, M. F. C., Casallas, D. M. D., & Marín, C. E. M. (2015). Análisis de la calidad del agua del río Bogotá durante el periodo 2008–2015 a partir de herramientas de minería de datos. *Publicaciones e Investigación*, *9*, 37-50.

Fisco, J. A., & Sabogal, D. P. (2014). Reconstrucción de atmósferas sonoras tridimensionales. *Publicaciones e Investigación*, *8*, 27-33.

Fuentes, L. F. Q., & Castelblanco, S. G. (2011). Perfil del sabor del clon CCN51 del cacao (Theobroma cacao L.) producido en tres fincas del municipio de San Vicente de Chucurí. *Publicaciones e Investigación*, *5*, 45-58.

Fuentes, L. F. Q., Pinilla, M. G., & Mendoza, L. J. (2014). Estandarización de la fase de fermentación "fase i" en la obtención de un licor de mandarina utilizando levadura "*Saccharomyces cerevisiae*". *Publicaciones e Investigación*, *8*, 139-149.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Retrieved from https://www.academia.edu/download/43034828/Data_Mining_Concepts_And_Techniques_3rd_Edition.pdf

Hernandez Gonzalez, A. G., Melendez Armenta, R. A., Morales Rosales, L. A., Garcia Barrientos, A., Tecpanecatl Xihuitl, J. L., & Algredo, I. (2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico. *IEEE Latin America Transactions*, *14*(11), 4573–4578. https://doi.org/10.1109/TLA.2016.7795831

Hurtado, J. (2000). *Metodología de la investigación holística.* Caracas: Fundación Sypal.

Garzón, L. J. R., & Jiménez, V. L. L. (2017). Vulnerabilidad hídrica de la cuenca del río Blanco, en el municipio de La Calera, considerando los escenarios de cambio climático propuestos por la corporación autónoma regional de Cundinamarca-Car. *Publicaciones e Investigación*, *11*(1), 77-88.

Giraldo, R., Vargas, T., & Gil, H. (2009). Mejoramiento del proceso de deshidratación de uchuva. *Publicaciones e Investigación*, *3*, 37-49.

Jiménez-García, W. G., & Rentería-Ramos, R. R. (2020). Contributions of complexity for the understanding of the dynamics of violence in cities. Case study: the cities of Bello and Palmira, Colombia (Years 2010-2016). *Revista Criminalidad*, *62*(1), 9-43.

Jiménez, V. L. L., Ramos, J. J. M., & Guio, D. P. A. (2016). Análisis del índice de riesgo de la calidad del agua para consumo humano -Irca- y su relación con variables meteorológicas y ubicación Geográfica para el departamento del

Tolima en los años 2012–2013. *Publicaciones e Investigación*, *10*, 69-81.

Laverde, W. E. M., & Bernal, O. A. V. (2015). Herramientas de gestión ambiental para las carreteras de cuarta generación (4g) en Colombia. *Publicaciones e Investigación*, *9*, 87-98.

León, C. R., & García, M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. *Universidad y Sociedad*, *8*(4) 42–52. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005

Lerma González, H. D. (2016). *Metodología de la investigación: propuesta, anteproyecto y proyecto*. México: Ecoe Ediciones.

Malvicino, F., & Yoguel, G. (2015). Big data: Avances Recientes a Nivel Internacional y Perspectivas para el Desarrollo Local. *Centro Interdisciplinario de Estudios en Ciencia tecnología e Innovación (CIECTI)*. http://www.ciecti.org.ar/wp-content/uploads/2017/07/DT3-BigData-avances-y-perspectivas-de-desarrollo-local.pdf

Martínez, J., & Pino, F. J. (2016). Definición de un modelo de calidad de servicios soportado por tecnologías de la información (TI). *Publicaciones e Investigación*, *10*, 49-67.

Masso, J., & Pardo, C. (2015). Hacia una ontología para el gobierno de desarrollo de software en pymes. *Publicaciones e Investigación*, *9*, 99-112.

Mesa Angulo, O. P., Gabriel, F. J., Ostos Ortiz, O. L., & Rentería, R. R. (2020). Modelo de vigilancia tecnológica e inteligencia estratégica: evaluación de nuevos programas académicos de la Universidad Santo Tomás. https://repository.usta.edu.co/handle/11634/28934

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: data mining and Knowledge Discovery*, 431–443. http://tarjomefa.com/wp-content/uploads/2017/10/7879-English-TarjomeFa.pdf

Milquez-Sanabria, H. A. A. (2017). Digestión anaerobia en dos fases, hidrólisis y metanogénesis, de la semilla de mango (*Mangifera indica*). *Publicaciones e Investigación*, *11*(1), 91-100.

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*. Diss. Universidad Nacional de La Plata. http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1

Molina, L. D., & Lozano, L. P. (2016). La desertificación del suelo, aspectos y estrategias de lucha. *Publicaciones e Investigación*, *10*, 117-127.

Montañez Carrillo, L., & Lis Gutiérrez, J. P. (2016). Medición de la madurez de la gestión del conocimiento en la Escuela de Ciencias Básicas Tecnología e Ingeniería de la UNAD. *Publicaciones e Investigación*, *10*, https://doi.org/10.22490/25394088.1595

Niño, M., & Illarramendi, A. (2015). Entendiendo el big data: antecedentes, origen y desarrollo posterior. *Dyna New Technologies*, *2*(3). https://doi.org/10.6036/nt7835

Ochoa, N. E., Cruz, I. M., Gil, C. E., Chaves, C. C. S., Grajales, S. K., Vargas, L. L. V., & Páez, A. (2015). Estrategias en la construcción de un prototipo como modelo integral en la gestión investigativa orientado hacia el esquema de negocio. *Publicaciones e Investigación*, *9*, 113-134.

Orozco, L. G., & Urrego, A. I. C. (2016). Modelos de ensuciamiento en intercambiadores de calor tubulares en sistemas indirectos en procesos uht en la industria láctea. *Publicaciones e Investigación*, *10*, 95-114.

Ortega, J. A. T., Rubio, O. F. C., & Orozco, I. H. (2017). Análisis de ciclo de vida para una biorefinería derivada de residuos agrícolas de palma aceitera (*Elaeis guineeeensis*). *Publicaciones e Investigación*, *11*(1), 13-36.

Ortiz, I. A. L., & Angulo, H. M. (2016). Percepción de los estudiantes sobre la utilización de videojuegos en cursos de la Universidad Nacional Abierta ya Distancia-UNAD. *Publicaciones e Investigación*, *10*, 163-175.

Parra, C. A. C., & Espinal, J. M. M. (2014). Parámetros técnicos de captura en instrumentos musicales percutidos del folclor colombiano para su uso en bancos virtuales de sonidos. *Publicaciones e Investigación*, *8*, 35-53.

Pérez, L. A., & Vera, C. A. (2015). Método para medir indirectamente la velocidad de fase en sensores *surface acoustic wave*. *Publicaciones e Investigación*, 9, 65-72.

Ramírez-del Río, D., Soto-Mejía, J. A., & Rentería-Ramos, R. R. (2018). Diseño de un modelo bajo el enfoque de dinámica de sistemas para estudiar

comportamiento de la dinámica socioeconómica basada en la atención de primera infancia, infancia y adolescencia. *Investigación Operacional*, *39*(2), 220-233.

Ramírez, P. E., & Grandón, E. E. (2018). Prediction of student dropout in a Chilean public university through classification based on decision trees with optimized parameters. Formacion Universitaria, 11(3), 3–10. https://doi.org/10.4067/S0718-50062018000300003

Reina, C. B., Jiménez, L. N. R., & Pedraza, N. M. (2014). Obtención de biodiesel (etil-éster) mediante catálisis básica a nivel planta piloto derivado de aceites usados de la industria alimenticia. *Publicaciones e Investigación*, *8*, 99-116.

Rentería-Ramos, R. R. & Alfonso, A. V. (2015). Construcción de una red compleja para el estudio de la selectividad de Santiago de Cali por parte de las víctimas desplazadas del conflicto armado en Colombia. *Investigación Operacional*, *36*(1), 60-69.

Rentería-Ramos, R.R., Hurtado-Heredia, R., & Urdinola, B. P. (2019). Morbi-mortality of the victims of internal conflict and poor population in the Risaralda Province, Colombia. *International Journal of Environmental Research and Public Health*, *16*(9), 1644.

Rentería-Ramos, R. R. & Mejía, J. A. S. (2018). Diseño de una sociedad artificial para estudiar la migración forzada por cónflicto armado interno en el suroccidente colombiano. *Investigación Operacional*, *39*(2), 206-219.

Rentería-Ramos, R. R. & Soto Mejía, J. A. (2016). Design agent based model to study the impact of social cohesion and victimization in the criminal behavior. *Ingeniería y Ciencia*, https://repository.eafit.edu.co/handle/10784/11294

Rentería-Ramos, R., Velasco Bonilla, A., María Burbano, J., & M Vitale, A. (2017). Construcción de clústeres empresariales en el sector de la salud en Santiago de Cali a través del algoritmo Multivariate Fuzzy C-Means. *Economía y Desarrollo*, *158*(2), 129-140.

Rodríguez, J. F. G., Ramírez, A. A., Pérez, L. M., Meza, J. R., & Rentería-Ramos, R. R. (2019). Relación entre la innovación y la productividad laboral en la industria manufacturera de México. *Investigación operacional*, *40*(2), 249-254. http://www.invoperacional.uh.cu/index.php/InvOp/article/view/667

Rodríguez, P., Truffello, R., Suchan, K., Varela, F., Matas, M., Mondaca, J., … Allende, C. (2016). Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile. Ministerio de Educación. http://disde.minedu.gob.pe/handle/123456789/4463

Rojas, M. O. A., & Arboleda, L. C. T. (2015). Simulación de redes de sensores inalámbricos: un modelo energético a nivel de nodo-sensor bajo las especificaciones Ieee 802.15. 4tm y Zigbee. *Publicaciones e Investigación*, *9*, 13-24.

Rojas, Y. S. V., Ramírez, L. M. V., & Ortega, J. A. T. (2014). Evaluación de la huella hídrica del lirio japonés (*Hemerocallis*). *Publicaciones e Investigación*, *8*, 79-87.

Romero, J. J. V, Toledo, R. A. J., & Paredes, L. E. (2017). *Implementación de un sistema de análisis de datos en la deserción estudiantil utilizando técnicas de big data para facilitar la estructuración de planes de mejoramiento de la Universidad Mariana*. (Vol. 4). Informativo, C E I.

Sáenz, L. M. B. (2014). Una Visión del sistema de certificación en inocuidad de alimentos. *Publicaciones e Investigación*, *8*, 151-159.

Samper, J. J. C., & Bolaño, M. R. (2015). Seguridad informática en el siglo XX: una perspectiva jurídica tecnológica enfocada hacia las organizaciones nacionales y mundiales. *Publicaciones e Investigación*, *9*, 153-162.

Sanabria, A. E. R., & Pérez, J. R. R. (2015). Catalizadores organometálicos en la industria química. *Publicaciones e Investigación*, *9*, 51-64.

Sánchez, I. C. N., & Alfonso, J. N. M. (2019). Revisión: estimación de deficiencias en la calidad del huevo. *Publicaciones e Investigación*, *13*(1), 103-110.

Sánchez, N. J. Z. (2014). Simulación de un sistema de desodorización de aceite vegetal por medio de un control industrial automatizado. *Publicaciones e Investigación*, *8*, 119-125.

Sendoya, D. F. (2013). ¿Qué es el control predictivo y hacia dónde se proyecta? *Publicaciones e Investigación*, *7*, 53-59.

Sierra, G. I. L., & Gonzalez, N. V. Y. (2014). Estudio descriptivo mediante análisis multicriterio de la cadena agroalimentaria de la panela. *Publicaciones e Investigación*, *8*, 161-183.

Tangarife, J. H., & Acevedo, Y. V. N. (2015). Video juego interactivo mediante Sdk Kinect 1.6 para apoyar la educación básica primaria de niños entre 5 a 10 años de edad. *Publicaciones e Investigación*, *9*, 25-36.

Toro, R. O. (2017). Biocompuestos a base de almidón termoplástico, ácido poliláctico y cascarilla de arroz: efecto del aceite epoxidado de soya. *Publicaciones e Investigación*, *11*(1), 49-55.

Waltero, H. E. P. (2015). Arquitectura de un laboratorio remoto desde el enfoque de la formación de ingenieros en ead. *Publicaciones e Investigación*, *9*, 147-152.