Modelo De Predicción De Deserción Estudiantil, Apoyado En Tecnologías De Data Mining, En Un Curso De Primera Matrícula De La Escuela ECBTI De La UNAD

Mario Luis Avila Pérez

Universidad Nacional Abierta Y A Distancia Unad

Escuela De Ciencias Básicas, Tecnología E Ingeniería

Maestría En Gestión De Tecnología De Información

Colombia

2021

Modelo De Predicción De Deserción Estudiantil, Apoyado En Tecnologías De Data Mining, En Un Curso De Primera Matrícula De La Escuela ECBTI De La UNAD

Mario Luis Avila Pérez

Director:

Dr. Javier Medina Cruz

Universidad Nacional Abierta Y A Distancia Unad

Escuela De Ciencias Básicas, Tecnología E Ingeniería

Maestría En Gestión De Tecnología De Información

Colombia

2021

Nota de aceptación

Firma del presidente del Jurado

Firma del Jurado

Firma del Jurado

Barranquilla, 2021

Dedicatoria

A Dios, por ser mi guía inseparable, que nos ayuda a ser mejores en el desarrollo de nuestras vidas, el que nos brinda la sabiduría y la fuerza de voluntad para alcanzar nuestras metas.

A mi querida esposa Leslie, por ser mi amiga y confidente en todo tiempo, por el apoyo que siempre me brinda. Es mi ángel, a quien quiero mucho.

A mis hijos Isaac y Carlo Mario que me brindan la motivación y el apoyo que tanto he necesitado

A mis queridos padres Luis y Carmen, este logro quiero dedicar a ustedes. Quienes me han apoyado día a día. Siempre me han impulsado a seguir adelante, a ser perseverante para alcanzar cada sueño.

Mario.

Agradecimientos

Agradecimiento a mi asesor de tesis Dr. Javier Medina Cruz.

Por haber aceptado guiarme en esta investigación, por compartir sus conocimientos, orientación, persistencia, paciencia y motivación.

A los docentes de la MGTI de la UNAD por haber depositado en nuestras mentes la semilla del conocimiento.

Resumen

A continuación, se presenta la investigación modelo de predicción de deserción estudiantil, apoyado en tecnologías de big data, en un curso de primera matrícula de la escuela ECBTI de la UNAD. Este proyecto se desarrolla como requisito de grado para la maestría en gestión de TI de la Universidad Nacional Abierta a Distancia UNAD. El proyecto incluye el planteamiento del problema en donde se expresa la necesidad de aplicar técnicas de analítica de datos a la información que se almacena como producto de los procesos académicos. En los procesos académicos de la UNAD se producen en cada periodo académico una gran cantidad de datos los cuales son susceptibles de analizar, con el fin de generar un modelo de predicción que coadyuve en la mitigación del problema de la deserción estudiantil en la institución, mediante el pronóstico temprano de los estudiantes en riesgo de deserción. Como parte de esta información se tienen en la UNAD instrumentos como la encuesta de caracterización que se aplica a los estudiantes nuevos, el cual es un instrumento muy valioso que permite conocer información de los estudiantes que inician su proceso en la UNAD. Este estudio aplica técnicas de minería de datos basada en Machine Learning, mediante el uso de algoritmos de aprendizaje supervisado que permitan generar modelos de predicción de la deserción estudiantil que de manera temprana determine si un estudiante probablemente desertará de su proceso de formación. Durante el desarrollo de este proyecto se utilizaron herramientas de software Libre tales como WEKA que permitieron obtener algunos resultados a partir de la aplicación de algoritmos de machine learning. Estos resultados proporcionan un soporte para la toma de decisiones, lo que permite a los directivos de las institución 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.

Abstract

Below is the research model of student dropout prediction, supported by big data technologies, in a first enrollment course of the ECBTI school of the UNAD. This project is developed as a degree requirement for the master's degree in IT management of the Universidad Nacional Abierta y a Distancia UNAD. The project includes the problem approach which expresses the need to apply data analytics techniques to the information that is stored as a product of academic processes. In the academic processes of the UNAD, a large amount of data are produced in each academic period which are capable of analyzing, in order to generate a prediction model that will contribute to the mitigation of the problem of student dropouts in the institution, through the early prognosis of students at risk of dropping out. As part of this information, the UNAD has instruments such as the characterization survey that is applied to new students, which is a very valuable instrument that allows to know information about the students who begin their process in the UNAD. This study applies data mining techniques based on Machine Learning, through the use of supervised learning algorithms that allow to generate models of prediction of student dropout that determine early if a student is likely to drop out of their training process. During the development of this project, Free software tools such as WEKA were used to obtain some results from the application of machine learning algorithms. These results provide a support for decision-making, which allows the managers of the institutions to concentrate efforts or direct them to certain specific areas or areas, which greatly improves the effectiveness in the processes allowing them to approach knowledge more effectively and efficiently.

Keywords: *Data mining, Bigdata, Decision Trees, Prediction, KDD, Aprendizaje supervisado, WEKA*

Tabla de contenido

Pág.

Lista De Tablas

Lista De Figuras

Listado De Anexos

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 reconocer cuáles podrían llegar a ser las causas de dicha deserción y formas efectivas de mitigarla. Las instituciones cuentan información para iniciar procesos exploratorios de 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 a la ejecución de un análisis diagnostico que haga visible los requerimientos para el desarrollo de un proyecto de Data Mining. De este análisis se pretenden generar unos diseños de elementos 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 precisión del modelo mediante la revisión de los indicadores tales como la matriz de confusión, el accuracy, el recall o sensibilidad. Finalmente se presentan las conclusiones del estudio.

Problema De Investigación

Título: Modelo de predicción de deserción estudiantil, apoyado en tecnologías de Data Mining, en un curso de primera matrícula de la escuela ECBTI de la UNAD.

Planteamiento Del Problema

Teniendo en cuenta la problemática de este proyecto denominada “Deserción estudiantil” se afirma que este fenómeno se presenta a una escala global y en los distintos niveles de la educación básica, media y superior por lo que es necesario abordar de manera concreta cómo afecta este fenómeno a otros territorios e instituciones a nivel nacional e internacional. Se tienen referentes internacionales en Latinoamérica como el caso de Bolivia que de acuerdo con (Poveda Velasco et al., 2020) de las 11 universidades públicas de Bolivia se registra una deserción estudiantil promedio del 10,66%. Otro referente es (Salazar et al., 2004) que afirma que la deserción estudiantil universitaria en argentina es del 43%, México 40% y en chile, Chile con un promedio del 54%.

En este mismo sentido, (Rodríguez Núñez & Londoño Londoño, 2011) indica que la tasa de deserción nacional a nivel de pregrado oscila entre el 45% y 50% lo que es de alta preocupación. También , dando una mirada local, se tienen referentes nacionales tales como (Patiño Garzón & Cardona Pérez, 2012) que expone que la tasa bruta acumulada de la deserción estudiantil en la Universidad de Ibagué dentro de los años 2000-2006 fue del 13%. En esta misma línea se tiene el caso específico de la deserción en la Universidad del Rosario, más específicamente en la facultad de Ciencias económicas y Administrativas donde se evidencian los comportamientos de este fenómeno a lo largo del tiempo concluyendo que ha habido una reducción significativa de la tasa de deserción que para el periodo 2001-I era del 25,02% a un 6,25% para el periodo 2006-I (Lopera, 2008)

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 100.000 estudiantes, lo que la convierte en una de las instituciones más grandes del país en cuanto a número de estudiantes. Sin embargo, una gran parte de estos estudiantes desertan por diferentes motivos o circunstancias.

La deserción estudiantil se presenta 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 expresa Facundo, A. (2009) en su trabajo titulado Análisis sobre la deserción en la educación superior a distancia y virtual: El caso de la UNAD-Colombia.

Al analizar la corte 2001-I, a la cual ingresaron 6.011 estudiantes, se encontró que, luego de nueve años, se han graduado 1.432 estudiantes (el 23.82%) y permanecen aún en la institución 321 (el 5.34% de los estudiantes de la cohorte). Es decir que, al realizar el corte para el análisis en el año 2008, la deserción de esta cohorte fue de 70.84%, dato por demás elevado. (Ángel & Facundo, 2009)

Estas cifras han venido mejorando año tras año, pero aún persisten altos índices 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 de recursos y esfuerzos que deja muchos sinsabores y frustraciones a los afectados.

Los costos sociales de la deserción estudiantil son extremadamente altos, lo cual repercute en un incremento de las tasas de criminalidad y decremento de la tasa de crecimiento económico y muchos otros índices como aumento de la inequidad o la brecha entre ricos y pobres (Rodríguez et al., 2016)

Los síntomas de este fenómeno se evidencian en cada periodo académico en las

instituciones de educación observándose que un alto porcentaje de los estudiantes abandonan los estudios. Es un fenómeno que afecta tanto a países desarrollados como subdesarrollados. En la UNAD el fenómeno de la deserción es recurrente, y debido al modelo de educación a distancia es aún más marcado. Un ejemplo de ello es que al iniciar un curso se tiene un determinado número de estudiantes que presentan la actividad de reconocimiento del curso, pero a medida que el curso avanza la taza de participación va disminuyendo. También se aprecia una faceta del fenómeno en los bajos índices de participación en el desarrollo de las actividades y el aumento de los estudiantes incluidos en el listado de alertas tempranas que periódicamente hacen los tutores de los cursos académicos, en un esfuerzo por mejorar la permanencia y retención de los estudiantes.

En este mismo sentido, como posibles causas del fenómeno de la deserción algunos autores como (Torres et al., 2015). Mencionan el nivel socioeconómico, como el principal factor externo asociado a la deserción. Torres afirma que, en todos los estudios revisados, la pobreza o los bajos ingresos familiares son determinantes de la deserción; ligado a la necesidad de buscar trabajo por algunos estudiantes. También afirma que el capital cultural y simbólico-lingüístico, en el contexto familiar del estudiante es significativo, siendo menor la probabilidad de deserción cuando los padres valoran la educación como factor de mejora de la calidad de vida, Torres afirma que es mayor la probabilidad de deserción cuando no se vive con ambos padres, Así también aumentan la probabilidad del fenómeno factores como la vida en pareja del estudiante, la maternidad o paternidad temprana., cuando el estudiante siente que no avanza de acuerdo con lo esperado ya sea por sus condiciones, falta de apoyo, ritmos de aprendizaje, falta de saberes previos, Además incrementa la probabilidad de deserción. La relación con el docente y con los compañeros que también es un factor determinante en la deserción. (Torres et al., 2015).

Continuando lo anterior, también es posible hablar acerca del pronóstico a partir de la problemática donde se tiene que este fenómeno afecta a las instituciones de educación en un mayor o menor grado lo cual origina pérdida o desaprovechamiento de recursos. El fenómeno de la deserción provoca afectación negativa de los procesos sociales, económicos y políticos en las proyecciones de la institución y el país. Este fenómeno tiene un alto costo social que se refleja en el futuro, dando como resultado menor cantidad de personas calificadas y en consecuencia el estado se verá obligado a financiar programas sociales paliativos. Otro de los resultados de la deserción universitaria es que también contribuye al incremento de las desigualdades, con la deserción aumenta el número de desocupados los cuales pueden pasar a conformar grupos delictivos. En este sentido para los estudiantes desertores disminuye la posibilidad de oportunidades de conseguir trabajos mejor remunerados.(Rodríguez et al., 2016)

Como control al pronóstico del problema de la deserción, las instituciones vienen trabajando en investigaciones que permitan disminuir los índices de abandono de los estudios. La UNAD preocupada por el fenómeno de deserción. viene haciendo una serie de esfuerzos, en la búsqueda de estrategias que permitan paliar el fenómeno; ejemplo de ello es el Acuerdo no. 002 del 30 de enero de 2018, “Por el cual se establece la Política Institucional De Retención Y Permanencia Estudiantil y se adopta el plan institucional de acogida y permanencia diferencial en la Universidad Nacional abierta y a distancia – UNAD”. Así mismo, se plantea esta investigación que apoyada en las tendencias disruptivas y de transformación digital, se propone como una medida que coadyuve a aliviar el problema, mediante la utilización de técnicas de Data mining para encontrar patrones que permitan predecir el riesgo de deserción de los estudiantes. El proyecto plantea que mediante análisis de información por medio de técnicas de Big Data, se podrían encontrar formas de predicción de la deserción estudiantil para facilitar la toma de

decisiones en la institución y de esta manera contribuir en la disminución del número de estudiantes desertores.

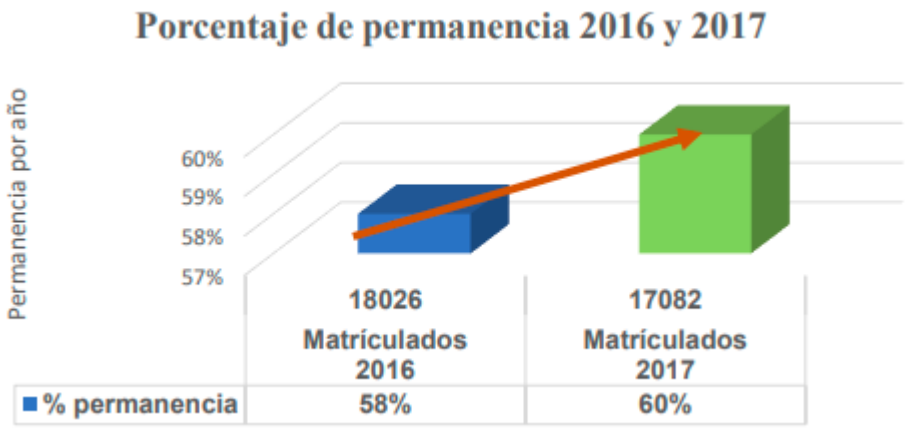
En los cursos de la UNAD se producen en cada periodo académico una enorme cantidad de datos los cuales podrían ser sometidos a técnicas de análisis de datos para extraer información de valor que podría ser útil para predecir la deserción estudiantil en la institución, esta información pudiera ser muy importante para la mejora continua de los procesos.

Los datos recopilados en la UNAD provenientes de los procesos académicos incrementarían su valor en el momento en que se sometan a técnicas de análisis de datos. Estos datos podrían ser recopilados, procesados y analizados de tal manera que beneficien los procesos de formación, los métodos y prácticas del proceso educativo. Se pueden analizar datos que el sistema almacena, producto de las interacciones del estudiante en los diferentes procesos y a partir de estos datos generar información que sirva de soporte para la toma de decisiones en la institución.

La información obtenida a partir del análisis de esta gran cantidad de datos puede facilitar el hallazgo de tendencias o comportamientos que de manera anticipada posibilite la detección temprana de estudiantes en riesgo de deserción, lo que contribuiría a reducir las cifras de deserción.

Figura 1

Porcentaje de permanencia 2016 y 2017

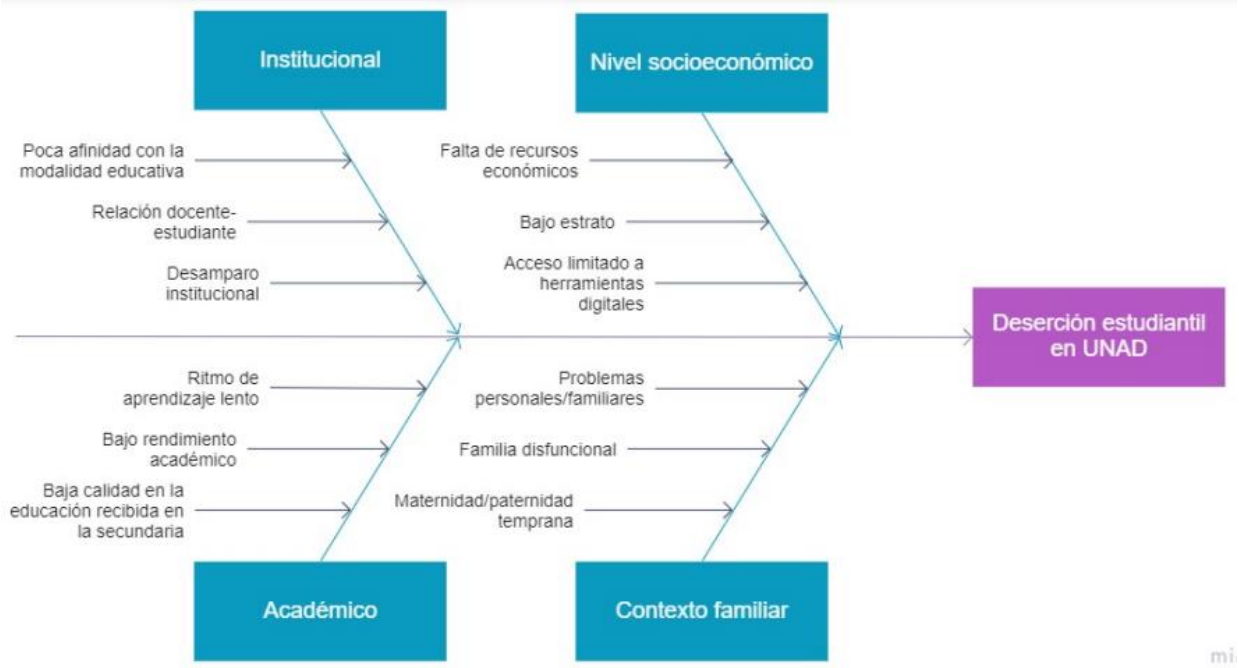


Nota El grafico representa los porcentajes de permanencia en los años 2016 y 2017 en la UNAD. Tomado de: *(Abadía et al., 2018) Pag 42*

Cabe señalar, que la generación de estadísticas universitarias ligadas con esta problemática es limitada por diversas causas, entre ellas, la cantidad enorme de variables que intervienen y son determinantes para este problema. En este sentido para (Torres et al., 2015) no es posible medir de manera absoluta los índices de deserción debido a que a que hay estudiantes que salen del sistema por un tiempo y luego regresan, presentado intermitencias en su proceso de formación, es así como se tiene unos estudiantes que terminan su periodo pero no se matriculan en el siguiente y los que abandonan durante el desarrollo del periodo.

Figura 2.

Problema de Deserción estudiantil



Nota. El grafico muestra los factores que pueden influir en la deserción

Formulación Del Problema

Del planteamiento anterior surge la siguiente pregunta de investigación:

¿En qué medida la implementación de un modelo de predicción de deserción estudiantil, apoyado en técnicas de minería de datos, en un curso de primera matrícula de la escuela ECBTI de la UNAD, contribuirá a mejorar los índices de deserción estudiantil?

Alcance

Con esta investigación se pretende obtener un modelo de predicción de la deserción estudiantil a partir de un instrumento encuesta aplicado a los estudiantes nuevos y sus datos históricos a lo largo del tiempo. A partir de estos datos mediante la aplicación de técnicas de minería de datos se pretende generar un modelo que permita hallar correlaciones entre diferentes variables y a partir de allí poder hacer la predicción de la deserción

Esta investigación abarca la aplicación de técnicas de Data Mining a los estudiantes nuevos ingresados en el periodo 2018 16-01, a través del instrumento encuesta de caracterización

que se aplicó en el mes de febrero del año 2018, a estudiantes del curso de la ECBTI, Herramientas digitales para la gestión del conocimiento.

Esta investigación proporcionará a los stakeholders, una herramienta que ayude a detectar de forma temprana aquellos factores que puedan incidir en la posible deserción y su consecuente permanencia, principalmente en cursos iniciales de la Escuela de Ciencias Básicas, Tecnología e Ingeniería (ECBTI).

Con esta herramienta se pretende contribuir a aliviar este fenómeno de deserción que tanto afecta negativamente las instituciones educativas, teniendo de presente que en la UNAD se vienen haciendo esfuerzos por mejorar estos indicadores, y en este sentido la UNAD ha mejorado su infraestructura tecnológica, además de la tendencia en las mejoras en todos los procesos en los últimos años, todo en procura de mejorar la calidad en los procesos académicos, además del mejoramiento de la cualificación del cuerpo docente, así de esta manera también desde la aplicación de técnicas de minería de datos se pueda intentar detectar elementos que tal vez se han pasado por alto y que permitan tomar nuevas estrategias para evitar la deserción de estudiantes.

Límites Proyecto

En el proyecto del modelo para la predicción para la deserción estudiantil en la UNAD, se plantea que el modelo estará limitado inicialmente al análisis de datos de los cursos de primera matrícula de los programas de la escuela ECBTI, para lo cual se tomarán los datos disponibles, correspondientes al año 2018, para poder contrastarlos con la situación actual de los estudiantes en cuanto a permanencia o abandono del proceso de formación en la institución.

Objetivos

Objetivo General

Desarrollar un modelo prototipo para 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.

Objetivos Específicos

Realizar un análisis diagnóstico para la determinación de los requerimientos del modelo a través de la revisión de las fuentes de información.

Diseñar los elementos del modelo de predicción para proveer un mecanismo de pronóstico de deserción estudiantil mediante el uso de una herramienta de Data Mining.

Evaluar la efectividad del prototipo con el fin garantizar su validez mediante la comparación con datos históricos.

Justificación

Los enormes retos a los que se enfrenta la educación actual motivan la búsqueda de nuevas formas de hacer las cosas, estas nuevas formas de enfrentar las problemáticas propias de la educación deben apoyarse en el uso de las tendencias disruptivas digitales, toda vez que estas tendencias están dando muy buenos resultados en áreas como la salud, la economía, la finanzas, el marketing entre otros, y la educación también es un campo en el que el uso de tecnologías como Big Data, a través del análisis de grandes volúmenes de información procesados a gran velocidad, y datos provenientes de formatos variados, proporcionan un soporte muy importante para la toma de decisiones, lo que permite a los ejecutivos de las organizaciones, en este caso a las vicerrectorías de la universidad concentrar los esfuerzos o dirigirlos a ciertos ámbitos o áreas específicas, lo que mejora enormemente la efectividad en los procesos permitiendo acercarse al conocimiento de manera más efectiva y eficiente. Este proyecto de investigación tiene pertinencias en distintos ámbitos expuestos a continuación.

En primer lugar, este proyecto tiene pertinencia en el ámbito institucional ya que para los usuarios de la Universidad Nacional Abierta y A Distancia, consejeros, tutores, directores entre otros. Resulta muy laborioso consultar el rendimiento académico de un estudiante en cualquier instante del periodo académico. Esto puede dificultar la detección temprana del bajo rendimiento de los estudiantes de un determinado periodo, debido a que normalmente las actividades académicas están dispersas o distribuidas en diferentes cursos, los cuales regularmente se encuentran en diferentes campus, servidores o bases de datos, lo que a menudo dificulta la detección de estudiantes desertores potenciales. Con tecnologías de análisis de datos y técnicas de descubrimiento de conocimiento KDD se propone la detección de patrones, comportamientos o tendencias que permitan predecir de manera temprana los estudiantes que muy probablemente desertarían de la UNAD, lo que permite a la institución la adopción de las medidas pertinentes

para paliar este fenómeno que afecta de manera muy negativa a la institución (Dominguez, 2018).

La Universidad nacional Abierta y a distancia UNAD se esfuerza mucho en mejorar los índices de deserción para lo cual ha adoptado una serie de estrategias como las de retención y permanencia en las cuales se invierten recursos considerables, además viene haciendo grandes esfuerzos en procura de mejorar la calidad en los procesos académicos, estos esfuerzos se pueden evidenciar en la inversión en infraestructura tecnológica, en infraestructura locativa, en mejoramiento de la cualificación en la planta de docentes, entre otras. “Educación con calidad global” era hasta hace unos pocos años el eslogan institucional con el que se identificaba la UNAD. En este sentido el desarrollo de este proyecto le permite a la institución la utilización de tecnología de punta para el abordaje de esta problemática, lo que contribuiría al mejoramiento de los índices de deserción.

El desarrollo de este proyecto es pertinente para la institución ya que aporta a los procesos de investigación de la Escuela de Ciencias Básicas Tecnología e Ingeniería, en particular a las temáticas propias de la Maestría en gestión de TI, involucrando un importante tema como lo es las tecnologías de Big Data aplicadas a la educación, como una tendencia disruptiva digital. La pertinencia se sustenta en que permitirá reunir elementos y aspectos importantes para generar nuevas iniciativas investigativas para el mejoramiento continuo de los procesos académicos. El proyecto también coadyuva a la búsqueda de la calidad de la educación superior, la cual está manifiesta en la ley 30 de 1992, al propender el desarrollo de nuevas estrategias que fortalezcan el impulso de competencias de los estudiantes utilizando diversas herramientas que le ayuden a resolver situaciones problémicas en su ámbito social y profesional.

En este mismo sentido, este proyecto tiene pertinencia en el ámbito social toda vez que a través de los resultados se benefician también los estudiantes en la medida en la que mejoren los procesos académicos de la institución. Las decisiones derivadas de la analítica de datos, las cuales son soportadas por las tecnologías de la información y la comunicación, permitirán ofrecer "Educación 'personalizada'" al estudiante, mejorando de esta forma los índices de deserción o de bajo rendimiento académico en los cursos ofertados por Universidad Nacional Abierta y a Distancia. El uso de técnicas de Big data permiten el análisis y la interpretación de los datos mediante algoritmos de Data Mining. Con lo cual se busca el mejoramiento de las prácticas educativas y la optimización del rendimiento de los alumnos, los docentes y también del modelo educativo.

La pertinencia y relevancia disciplinar del proyecto se fundamenta en la importancia de esta temática debido a que es un tema vanguardista, es una tecnología nueva que ha tenido mucha acogida en diversas áreas, es un tema de actualidad muy importante para el área de TI lo cual beneficiaría mucho a los programas de la cadena de sistemas.

En este sentido esta investigación cobra relevancia toda vez que la sociedad de la información experimenta cambios constantes impulsados por el desarrollo tecnológico, lo que obliga a las organizaciones a adaptarse constantemente, mediante la incorporación de nuevas estrategias. En la actualidad surgen nuevas diciplinas como los son la analítica de datos, Data Mining, Machine Learning entre otras tecnologías sobre las cuales se soportan las técnicas de Big data, las cuales podrían promover disrupciones en el ámbito educativo hacia un nuevo paradigma educativo más efectivo y responsable, dirigido hacia la actividad progresiva y autónoma de los estudiantes.

Continuando lo anterior, para el desarrollo de este proyecto se han analizado las distintas clases de viabilidad con el fin de identificar el nivel de factibilidad para llevar a cabo el proyecto. Es así como, en primer lugar, desde el punto de vista tecnológico, es posible desarrollar y cumplir eficientemente con los objetivos del proyecto ya que se cuenta de manera permanente con las herramientas de software libre, las cuales son accesibles, así como con el hardware necesario para llevar a cabo los experimentos.

En este mismo sentido, desde el punto de vista económico, se dice que este proyecto es viable ya que se dispone del recurso financiero necesario ya que este proyecto no supone costos altos ni adicionales lo que permite desarrollar el proyecto sin problemas de presupuesto, toda vez que los costos de tiempo y recursos son asumidos por el investigador del proyecto.

Por último, se afirma que este proyecto es viable desde el punto de vista de recurso humano toda vez que éste se justifica con base a los conocimientos adquiridos en la maestría ya que dotan al autor de este proyecto de las habilidades y conocimientos que requiere la planeación, documentación y ejecución del mismo.

Marco De Referencia

A continuación, se exponen las teorías, antecedentes, experiencias pioneras y demás referentes que son clave para el desarrollo de este proyecto.

Antecedentes

El interés por pronosticar la deserción estudiantil a pesar de que es un tema actual ya ha sido abordado por varios autores desde hace algunos años, en trabajos entre las los cuales se destacan las siguientes investigaciones, las cuales se tomarán como referentes para esta investigación:

En primer lugar, se tiene la investigación de (Mustafa et al., 2012), la cual propone el desarrollo de un modelo dinámico de predicción para institutos de orden superior y básico mediante la utilización de los árboles CART y CHAID obteniendo que El árbol de clasificación y regresión (CART) fue el más exitoso y se concluyó que para los árboles basados solamente en los datos de inscripción no son muy buenos para identificar los estudiantes exitosos de los no exitosos en la deserción.

En el año 2013 se publicó el trabajo de Titulado “Aplicando estrategias y tecnologías de Inteligencia de Negocio en sistemas de gestión académica”. El cual 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 los procesos en las unidades académicas". Este estudio propone un análisis del perfil del estudiante de la Facultad de Informática de la UNLP con ayuda de la tecnología y estrategias de la Inteligencia de negocios. Dicho estudio demostró que se ha afianzado la asignatura “Tecnologías Aplicadas a BI” dado que la matricula va en aumento y la tasa de asistencia es alta evidenciando la mejora en el desenvolvimiento del papel de los estudiantes. (Díaz & Osorio, 2013)

Por otra parte se tiene el estudio 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).En este estudio 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).

En este mismo sentido en Costa Rica, el TEC se convirtió en la primera universidad pública de Costa Rica en implementar una iniciativa que permite pronosticar la deserción estudiantil mediante el uso de técnicas de minería de datos. Esto se debió al esfuerzo conjunto del personal de la OPI y un estudiante (González-Loaiza, 2018)

Continuando lo anterior, (Cuji et al., 2017) en su trabajo titulado "Modelo predictivo de deserción estudiantil basado en árboles de decisión" presenta la construcción de un modelo o predictivo de deserción Estudiantil con el fin de identificar la probabilidad de que un estudiante abandone su programa académico utilizando técnicas de clasificación usando arboles de decisión. La investigación señala que bajo la metodología basada en Knowledge Discovery in Database (KDD) que se compone de 5 etapas y Aplicando el algoritmo, Classification and Regression Tree (CART) de la herramienta R, es posible construir un árbol con cuatro niveles de profundidad y mismo número de reglas, que evalúan a los posibles desertores. Proporcionando la información que permite concluir las variables que tuvieron mayor influencia en la deserción.

Dicha investigación utilizó los datos correspondientes a los estudiantes de la Carrera Docencia en Informática de la Universidad Técnica de Ambato (UTA). a partir del año 2006 sobre la cual se concluyó dentro de las diferentes variables nominales y cuantitativas que la más influyente en la deserción era la variable 'Nivel` de la siguiente manera: mayor tendencia a

desertar: primero, segundo, tercero, cuarto y quinto nivel, mínima tendencia: sexto, séptimo, y nula: octavo, noveno y décimo.(Cuji et al., 2017)

En este mismo sentido en el ámbito nacional se encuentra la investigación de (Romero et al., 2017) quienes realizaron el trabajo de 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. 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. La investigación se enmarcó dentro de un enfoque empírico analítico implementando un clúster computacional basado en Hadoop. La muestra estudiada era la Base de datos de estudiantes desertores en la Universidad de Mariana desde el año 2002 hasta el 2015. El estudio confirmo la factibilidad de las técnicas del Big Data como gestor de un pronóstico de comportamiento dentro de la institución educativa. (Romero et al., 2017).

Al interior de la UNAD se han venido desarrollando algunos proyectos de investigación entre los que se destacan el estudio Factores personales y académicos que inciden en la deserción temprana de estudiantes del programa de psicología adscritos a la UNAD CEAD Ocaña, donde se intenta evidenciar como las variables personales y académicas se convierten en factores que inciden en la deserción estudiantil, además se pone de manifiesto la influencia que tiene el contexto geográfico en la variación de estos factores (Noriega, 2019).

Así mismo se destaca la investigación de (Del Toro Díaz, 2013) titulado Factores que contribuyen en la deserción de los estudiantes de la Escuela de Ciencias Administrativas en la UNAD – CEAD Simón Bolívar – Cartagena. El estudio concluye que los factores que influyen en la deserción estudiantil son el factor económico, la mala administración del tiempo por parte

del estudiante, la falta de apoyo permanente para el desarrollo de su proceso de aprendizaje. También destacan el contacto con los tutores presenciales y virtuales, lo que originaria una menor probabilidad de abandono de sus estudios.

En este mismo sentido también destaca el estudio titulado Influencia de la virtualidad en la deserción de estudiantes en la Universidad Nacional Abierta y a Distancia. UNAD – CEAD, Yopal (Casanare – Colombia) esta investigación presenta los resultados del impacto que ha tenido la mediación virtual en la educación abierta y a distancia, y si su influencia en la deserción de los estudiantes en la UNAD CEAD Yopal,.(Mendoza García & Gómez Orduz, 2012)

En este orden de ideas también se encuentra el estudio titulado Factores asociados a la deserción académica en los programas de las escuelas de la universidad nacional abierta y a distancia – UNAD- CCAV Cartagena, En este estudio aborda la problemática de la deserción analizando varios factores, teniendo como base teórica dos grandes marcos interpretativos, uno enfocado en agentes exteriores de índole económicos, laborales, horas de dedicación al estudio y contexto familiar; y el otro desde el punto de vista de los agentes internos como aspectos motivacionales, personales, desempeño, bajo rendimiento, mala conducta y edad, y se analizan factores individuales, familiares, relacionados con la universidad. Los hallazgos de este estudio manifiestan que los factores más incidentes son en su orden, el factor económico, los métodos inadecuados de estudio, la adaptación al sistema, la clase de trabajo en que se desenvuelve el estudiante y las bases inadecuadas de formación previas al ingreso.(Sánchez-Sánchez, 2018)

En otro estudio titulado Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD – COLOMBIA donde se aborda la información dispuesta en el sistema implementado por el ministerio de educación nacional en el año 2006 denominado Spadies (Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación

Superior), considerado como una herramienta alerta útil para que las instituciones puedan tomar medidas preventivas. En este estudio se halló que la información disponible está incompleta sin embargo se continuó con la metodología propuesta para el estudio donde luego de un tiempo fue posible identificar la diferencia entre los inscritos en una cohorte determinada y la sumatoria de los graduados y quienes aún permanecen en la institución. (Ángel & Facundo, 2009)

Otro estudio al interior de la UNAD es el titulado Deserción académica y agente inteligente para su control y seguimiento: un aporte al sistema de consejería de la UNAD en el que se busca La identificación de la posible relación entre los factores de retención y los factores de deserción académica como base para la construcción del agente inteligente que favorezca la determinación de riesgos de deserción. Los resultados de este estudio permiten plantear un modelo donde, a partir de la relación de los factores, se indique si un estudiante activo está en riesgo de deserción.(Barrera Suárez et al., 2010)

Marco Teórico

A continuación, se presentan los contenidos teóricos abordados a lo largo de esta investigación los cuales se obtuvieron a partir de la revisión de los referentes bibliográficos explorados durante el desarrollo de la misma.

Hadoop

Los orígenes de Hadoop están profundamente relacionados con los antecedentes de Big Data. Para comprenderlo mejor hay que remontarse al año 1958. (Niño & Illarramendi, 2015) afirma que Hans Peter Luhn, investigador de IBM utilizó el término de Inteligencia de negocio en un artículo titulado IBM Journal of Research and Development. Publicado 1958, No obstante, el término estaba un poco alejado de lo que hoy se conoce como inteligencia de negocios.

En los años 90 utilizó el concepto de Machine learning como una herramienta predictiva

para la construcción de modelos en bancos y en compañías de seguros, donde se utilizó para dar soporte en la toma de decisiones, lo que constituyó en lo que conocemos como Data Mining. A partir del 2000 surgió el término Data Science o Ciencia de Datos la cual se ha convertido en una ciencia que integra la estadística, la matemática, la informática y la computación para soportar el análisis de datos y la extracción de conocimiento de estos. (Niño & Illarramendi, 2015)

Con el surgimiento de las empresas ".com" como Google, yahoo entre otras, las cuales se hicieron muy populares con el auge del world wide web, las organizaciones se dieron cuenta que debido al enorme volumen de datos que se estaba generando llegaría un momento que las técnicas tradicionales de procesamiento no serían apropiadas para procesar tal cantidad de información. Es así como Google, para optimizar el procesamiento de su algoritmo PageRank, se vio en la necesidad de explorar nuevas técnicas de procesamiento para reemplazar las técnicas de procesamiento en paralelo High-Performance Computing, HPC. Para resolver sus inconvenientes se orientó hacia el uso de técnicas que no necesitaran máquinas tan potentes como HPC (High Performance Computing), sino más bien técnicas con sistemas de archivos distribuidos en máquinas con potencia media (commodity servers) configuradas como nodos clúster, y un software que trabajara con este sistema distribuido, el cual era capaz de procesar los datos en este entorno. A este software o modelo de programación se le denominó MapReduce. (Niño & Illarramendi, 2015).

Figura 3

Línea de tiempo de Hadoop.

Hadoop Timeline	
2003	Doug Cutting and Mike Cafarella build Nutch
Oct 2003	Google publishes Google File System (GFS) paper
Dec 2004	Google publishes MapReduce (MR) Paper
2006	Yahoo! Builds Hadoop based upon GFS and MR papers with Doug Cutting and team
2007	Hadoop scales out to 1000 nodes at Yahoo!
Jan 2008	Hadoop becomes an Apache Software Foundation (ASF) project
Jul 2008	Hadoop is tested successfully on a 4000 node cluster
2009	Hadoop demonstrates sorting of Petabytes of data
Dec 2011	Hadoop Version 1.0 is available
Aug 2013	Hadoop Version 2.0 is available
Apr 2015	Hadoop Version 2.7.0 is available
Aug 2016	Hadoop Version 2.7.3 is available
Sep 2016	Hadoop Version 3.0.0-alpha1 is available

Nota. En el grafico se muestra una línea de tiempo de la evolución de hadoop Tomado de Shrivastava*, A., & Deshpande, T. (2016)* p11.

Es así como, MapReduce con sus dos funciones Map y Reduce, lo cual propició que se usara para resolver problemas similares. Es así como este trabajo inspiró a Cutting, empleado de Yahoo, para desarrollar un motor de búsqueda a nivel global utilizando MapReduce, para procesamiento distribuido de grandes cantidades de datos, y es así como nació el sistema de código abierto Apache Hadoop. (Niño & Illarramendi, 2015).

En este mismo sentido, Hadoop ha alcanzado una popularidad inusitada, de tal manera que empresas proveedoras de software empresarial tales como IBM, Teradata, Oracle and SAS, ofrecen dentro de su portafolio de productos soluciones Hadoop. Otros proveedores como Cloudera, MapR y Hortonworks también son colaboradores muy activos del código abierto Hadoop y de otras herramientas del framework. Otras empresas como Amazon han optado por el modelo de servicios en la nube, ofreciendo servicios basados en Hadoop.(Shrivastava & Deshpande, 2016) p11.

El modelo se caracteriza por la fiabilidad, escalabilidad y gran potencia. La filosofía en que se sustenta hadoop es distribuir el gran volumen de datos en varios nodos. Hadoop incluye los siguientes módulos que son los esenciales para el framework:

Hadoop Common, HDFS, Hadoop Yarn, Haddop, Map Reduce.

- Hadoop Common

Este módulo está constituido por una colección de utilidades comunes que soportan al resto de módulos, este conjunto de utilidades lo conforman librerías Java (.jar) y scripts que proporcionan abstracción a nivel de sistema operativo y de archivos. El paquete de Hadoop common incluye el código fuente y documentación. (*Apache Hadoop.*, 2018). A continuación, se mencionan las librerías que subyacen en el paquete hadoop common.

- CLI Mini Cluster

Esta utilidad puede arrancar o detener un único nodo de un cluster Hadoop sin la necesidad de configurar variables de entorno o tocar archivos de configuración, puede detener o iniciar Yarn/MapReduce y HDFS cluster. (*Apache Hadoop.*, 2018)

- Natives Libraries

Las librerías nativas de Hadoop contienen implementaciones nativas de ciertos métodos por razones de rendimiento o en los casos en que java no tiene la implementación disponible. (*Apache Hadoop.*, 2018)

- Proxy User

Esta utilidad posibilita que un súper usuario pueda enviar trabajos o acceder a HDFS a nombre de otro usuario, cuando se necesita que el usuario X pueda conectarse al namenode en una conexión autenticada con las credenciales de super. Dicho de otro modo, super se hace pasar por el usuario X. Apache Hadoop. (2018)

- Rack Awareness

La colocación de bloques HDFS utiliza Rack Awereness con la finalidad de otorgar tolerancia a fallos; para esto Haddop coloca una réplica de bloque en un rack diferente. Esto proporciona disponibilidad de datos en caso de una falla del conmutador de red o partición dentro del clúster. (*Apache Hadoop.*, 2018)

- Secure Mode

Hadoop puede ser configurado en modo seguro para lo que se recomienda saber cómo trabaja kerberos y DNS. En modo seguro cada servicio de hadoop y cada usuario deben ser autenticados con kerberos. Apache Hadoop. (2018)

- Service Level Authorization

El Level Authorization es el mecanismo de autorización que garantiza que los clientes que se conectan a un servicio Hadoop tengan los suficientes permisos y autorizaciones para accederlo. El nivel de autorización de servicio se realiza antes que otras comprobaciones de control de acceso y está definida en el archivo $ HADOOP CONF DIR / hadoop-policy.xml. Apache Hadoop. (2018)

- HTTP Authentication

Permite configurar las HTTP web-console de Hadoop (ResourceManager, NameNode, NodeManagers and DataNodes) para solicitar la autenticación del usuario, las cuales por defecto permiten el acceso sin ningún tipo de autenticación. Se pueden configurar para requerir la autenticación Kerberos mediante el protocolo HTTP SPNEGO. Apache Hadoop. (2018)

- Credential Provider API

La API Credential Provider se utilizan para separar el uso de tokens, secretos y contraseñas confidenciales de los detalles de su almacenamiento y administración. La capacidad de elegir

varios mecanismos de almacenamiento para proteger estas credenciales permite mantener esos activos protegidos. Apache Hadoop. (2018)

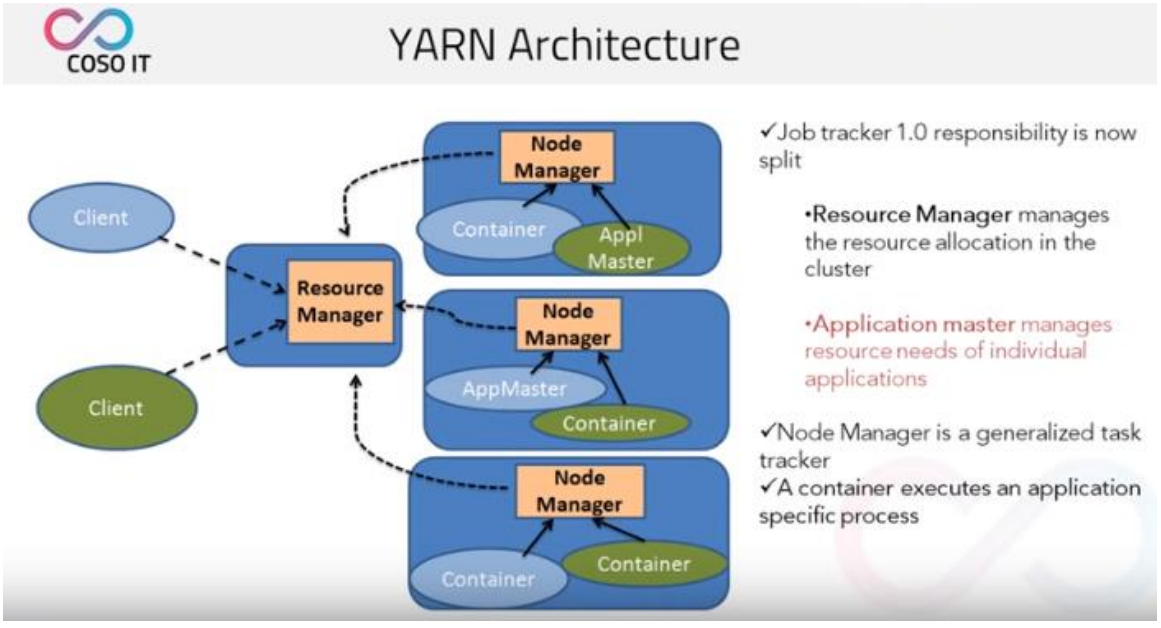
- Hadoop KMS

KMS es una aplicación web de Java y se ejecuta utilizando un servidor Tomcat preconfigurado, consiste en un servidor de administración de claves criptográficas basado en la API de KeyProvider de Hadoop. (*Apache Hadoop.*, 2018)

- Tracing

Es un sistema de rastreo que funciona mediante la recopilación de información en estructuras denominadas 'Spans'. Se puede usar la implementación de la interfaz SpanReceiver combinada con HTrace o implementando uno mismo. Apache Hadoop. (2018)

Figura 4
YARN Architecture



Nota. El grafico muestra la estructura de YARN. Tomado de: *https://www.cosoit.com/*

Big-Data Aplicado Al Sistema De Evaluación Del Comportamiento De Aprendizaje.

Es posible la inclusión del big-data también en los entornos de aprendizaje, lo cual se puede lograr aprovechando las plataformas de medios sociales como un recurso multipropósito, haciendo uso de las redes como agente de discusión y aprendizaje asíncrono aplicado a la enseñanza, alcanzando la función educativa de aprendizaje en todo momento y lugar. Hechos observables en la creación de los MOOCS lo cual ha innovado la educación considerablemente proporcionando interacciones sociales condensadas profesor-estudiante, estudiante-estudiante, propiciando el intercambio de puntos de vista, conocimientos, antecedentes, etc. A través de las técnicas de análisis de big-data es posible explorar el comportamiento del estudiante haciendo uso de las redes sociales y los MOOCS como agentes de recopilación en los foros de discusión que se gestan en ambos dando al docente una recomendación de recursos de aprendizaje mejorando la calidad de enseñanza. (Hai-ling et al., 2018) .

Big-Data Aplicado A La Evaluación De Programas Educativos.

Los autores (Yue & Liu, 2016) propone una nueva estrategia basada en un método de un algoritmo de red neuronal convolucional multi agente para mejorar la funcionalidad de la evaluación en educación superior dentro de un ambiente de BIG-DATA, esto es posible con el estudio del índice de educación innovadora de las universidades, que provee una experiencia eficiente y precisa verificando la efectividad del algoritmo.

Así mismo Hai-Ling afirma que el big-data podría equipar a investigadores en universidades con nuevas herramientas para entender un mundo cambiante y nuevas formas de compartir sus hallazgos con otros, para maximizar el impacto y la eficiencia de su investigación. Big-Data hace posible educar de una forma más personalizada, esta experiencia de aprendizaje

motivará jóvenes para estudiar y equiparse mejor. (Hai-ling et al., 2018)

Big-Data Y El Análisis De Sentimiento Para El Comportamiento De Decisión

Las técnicas del big-data hacen factible el estudio del comportamiento de los estudiantes, más específicamente acerca de cuáles son las propiedades que direccionan las motivaciones de los estudiantes. Con ayuda del big-data es posible proporcionar mayores resultados en la investigación acerca de los antecedentes del estudiante. Esto se logra a través de un muestreo de publicaciones en línea en redes como Twitter, Google, YouTube, etc. Estas grandes cantidades de datos se extraen y se analizan en la nube, buscando patrones que afecten las decisiones de los estudiantes. A partir de las nuevas tecnologías del big-data se amplía la perspectiva introducida en el desarrollo de la investigación educativa haciendo posible que la parte contextual del sistema educativo se presente como un factor clave en el manejo de la percepción del estudiante. (Hai-ling et al., 2018).

Minería De Datos, Motor De Desarrollo De Un Modelo Predictivo De Deserción

Las instituciones educativas superiores cuentan con sistemas académicos de información donde se registran todo tipo de datos de los estudiantes; datos personales, socioeconómicos y derivados del desempeño académico de los estudiantes durante y antes su estancia en la institución. La combinación de dichos datos aplicada a un proceso de ’Data mining’ podría proveer un modelo de predicción de deserción resolviendo la necesidad de las instituciones de disminuir el riesgo de deserción en las distintas carreras.

Las investigaciones realizadas alrededor del mundo muestran distintas técnicas de ‘Data mining’ de la cual el árbol de decisión resulta la alternativa más adecuada ante la predicción ya que abarca numerosas y distintas variables como: edad actual, ciudad de procedencia, estrato,

sexo, ocupación, estado civil, nivel de estudios del padre y de la madre; variables económicas como valor de la matricula e ingresos y de carácter académico como semestre, jornada, materias cursadas, materias perdidas y promedio. Esto abre la posibilidad de que las instituciones educativas superiores tengan una solución analítica que indique los motivos de la deserción estudiantil. (Heredia et al., 2015).

Métodos De Visualización De Datos De Educación En Línea Basados En Idl Y Hadoop.

El contar con diferentes plataformas puede dificultar el seguimiento a estudiantes, al encontrarse la información diseminada, como en el caso de la UNAD hasta hace un año, que solo al finalizar el periodo es cuando se consolida la información de los estudiantes y se obtiene un resultado de cómo va su proceso de aprendizaje. En este artículo los autores (Hai-ling et al., 2018) proponen un método mediante técnicas de big-data que unifica los datos provenientes de diferentes plataformas y que pueden ser desplegadas de forma consolidada.

HDFS facilita El procesamiento normalizado de datos que se obtienen de diferentes plataformas educativas y permite convertirlos en gráficas que faciliten la toma de decisiones. IDL es una especie de lenguaje de programación de 4ta generación utilizado para el despliegue de gráficas. IDL transforma datos en gráficos, para lo cual primeramente se hace el procesamiento estandarizado utilizando Hadoop para obtener la información de cada plataforma, para posteriormente pasar a la etapa de programación de visualización de datos utilizando IDL.(Hai-ling et al., 2018).

Para el análisis de visualización primero que todo se generan los archivos normalizados, tarea que se realiza mediante la importación de datos con Hadoop, de cada una de las plataformas donde el estudiante tenga cursos. Estos datos se unifican utilizando la herramienta MapReduce. Para el Análisis de visualización se realiza la programación visual usando el

lenguaje IDL y se procede al Análisis de visualización de datos mediante gráficas. Hai-ling, L., Jun-huai, L., & Jun, P. (2018)

Clasificación Basada En Árboles De Decisión

Los Árboles de decisión permiten obtener de forma visual patrones de decisión bajo los cuales operan una determinada población, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación. Esta técnica tiene aplicabilidad en diversas disciplinas del conocimiento y es muy útil cuando se tiene una propiedad conocida para un conjunto de elementos, pero, no se conoce esa misma propiedad en un elemento concreto. La clasificación basada en arboles de decisión es tal vez una de las técnicas que más se ha popularizado en Data Mining o minería de datos, los árboles de decisiones se usan 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". p3

TDIDT(Top-Down Induction of Decision Trees)

Es preciso afirmar que el Data mining sobre el machine learning es una herramienta útil para obtener nuevos conocimientos a partir de pequeños conjuntos de datos obtenidos de operaciones educativas que bajo sistemas automatizados son recolectados; dichos datos proveen información suficiente para mejorar los procesos y aportar a la adaptación y personalización de las instituciones. Dentro de las ventajas que se obtienen partir del uso del Data mining basado en machine learning se tiene la facilidad para establecer estrategias o estilos de aprendizaje general

o común, predecir los intereses de un estudiante de acuerdo con sus comportamientos anteriores y formar grupos homogéneos para identificar los conceptos erróneos en los procesos de aprendizaje. De las técnicas más comunes de minería de datos son el uso de algoritmos TDIDT (Top Down Induction of Decision Trees) que se caracterizan por representar a través de árboles de decisión los conocimientos obtenidos. La formación de este conocimiento simple carece de poder expresivo por lo que las metodologías de aprendizaje utilizadas en el TDIDT son menos complejas que aquellas utilizadas en un sistema donde pueden expresarse en un lenguaje más desarrollado, todo esto sin menospreciar la eficacia del uso de esta familia de algoritmos para resolver problemas difíciles. (Kuna et al., 2010).

Herramientas De Minería De Datos.

CRISP-DM, de Cross Industry Standard Process for Data Mining, es un modelo de proceso de minería de datos que describe los enfoques comunes que se recomienda usar al aplicar técnicas de Data mining.

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)

Aprendizaje Supervisado

Es la técnica de aprendizaje aplicada al Machine Learning bajo la cual se capacita al algoritmo dándole características y etiquetas a las clases conociendo la clase objetivo a predecir. De acuerdo con (Witten et al., 2016) se afirma que el aprendizaje supervisado está presente cuando se cuentan con resultados reales, es decir, que cada clase está etiquetada.

Métricas De Evaluación

Las siguientes son las principales métricas que se deben tener en cuenta para la evaluación de un modelo predictivo de minería de datos basado en machine learning.

Exactitud (Accuracy)

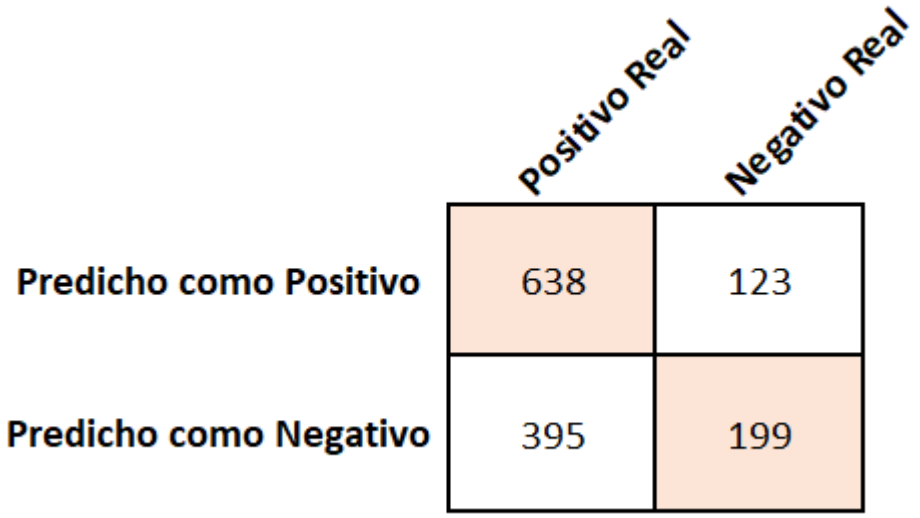
Es una medida que permite evaluar el desempeño del modelo, la exactitud arroja el porcentaje de predicciones correctas dentro del conjunto de datos. Significa que, si se tienen 500 registros y se predicen correctamente 300, se tiene una precisión del 60%. Vista de esta manera es una medida fácil de entender, sin embargo, no siempre es confiable y debe complementarse con otras medidas. Por ejemplo, si se tiene un conjunto de datos muy desbalanceado se pueden obtener indicadores de precisión muy altos que no reflejen la realidad, debido al desequilibrio de los datos, debido a que los algoritmos suelen privilegiar la clase mayoritaria. El accuracy es igual al porcentaje de registros clasificados correctamente. (Witten et al., 2016).

Matriz De Confusión

La matriz de confusión permite visualizar cuantos casos positivos fueron clasificados correctamente y cuantos casos negativos fueron clasificados correctamente. Se trata de una tabla en la cual se muestran los casos positivos que son verdaderamente positivos, casos positivos que en realidad son negativos, casos clasificados como negativos que en realidad son negativos y por último casos clasificados como negativos que en realidad son positivos. (Witten et al., 2016)

Figura 5

Matriz de confusión



Nota. El grafico muestra la estructura de la matriz de confusión

Es posible obtener la precisión del modelo a partir de la matriz de confusión dividiendo los casos clasificados como positivos entre el total de registros. La matriz de confusión describe completamente el desempeño del modelo, pero no es la mejor forma para comparar los modelos. Dentro de los elementos que conforman la Matriz de confusión encontramos los siguientes:

True Positive (TP) son los registros que fueron predichos positivamente y que de verdad eran positivos.

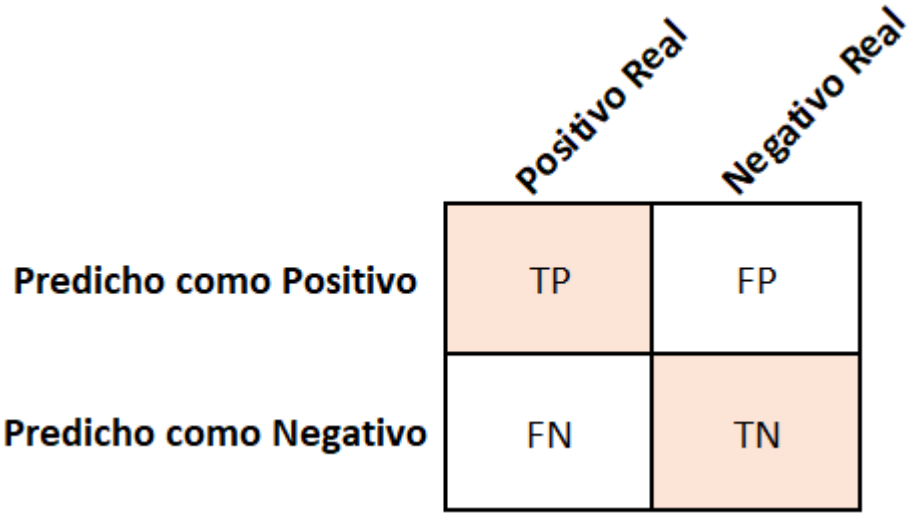
True negative (TN) son las instancias que el modelo clasificó como negativas que en verdad lo eran.

F**alse Positive (FP)** son las instancias que se clasificaron como positivas pero que en verdad eran negativas, es decir, que se clasificaron de manera incorrecta.

F**alse negative (FN)** Son las instancias clasificadas como negativas que e realidad eran positivas

Figura 6

Matriz de confusión interpretación



Nota. El grafico muestra la interpretación de los valores en la matriz de confusión.

Precisión

La precisión es el porcentaje de predicciones positivas del modelo que son correctas.

$$\text{precision} = \frac{\#\text{ positives predicted correctly}}{\#\text{ positive predictions}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Define que tan preciso es el modelo en cuanto a predicciones positivas

Sensibilidad (Recall)

Se define como el porcentaje de casos positivos que el modelo predijo de manera correcta, estos datos también es posible obtenerlos a partir de la matriz de confusión.

$$\text{recall} = \frac{\#\text{ positives predicted correctly}}{\#\text{ positive cases}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Cuestionarios

Según (Sampieri et al., 2014) los cuestionarios son conjuntos de preguntas que se usan para la recolección de datos acerca de una o varias variables que se desean medir. Estos cuestionarios deben ser congruentes con el planteamiento del problema y la hipótesis. Los cuestionarios pueden ser de preguntas abiertas o cerradas. Las preguntas abiertas no restringen las respuestas, sino que permite la posibilidad de un número infinito de posibilidades en la respuesta. Por el contrario, las preguntas cerradas, las respuestas están reducidas a un número finito de opciones de respuestas, estas opciones se caracterizan por haber sido delimitadas por el investigador, y pueden ser de selección múltiple, de tipo dicotómicas entre otras. Se requiere que las preguntas sean claras, comprensibles, breves, que no incomoden a la persona encuestada.

Marco Tecnológico

En los proyectos de minería se requiere de una plataforma de hardware y de software para la ejecución de los algoritmos y de los modelos. Para esta fase de prototipado del modelo no se necesitan recursos de hardware demasiado robustos, no obstante, algunos algoritmos como los de redes neuronales requieren de hardware más robusto, para que la ejecución no consuma demasiado tiempo. El software necesario para la ejecución tanto de algoritmos de machine learning como para los diferentes ejercicios, se pueden ejecutar tanto en sistemas operativos Windows como también en sistemas operativos de software libre como Linux.

En este orden de ideas, los modelos de predicción como técnica de minería de datos son generados a partir de algoritmos de inteligencia artificial, estos algoritmos procesan los datos históricos de algún fenómeno que se pretenda predecir. A partir de estos datos históricos las computadoras aprenden a predecir el fenómeno en estudio. Estos aprendizajes constituyen los

modelos, los cuales pueden ser alimentados con datos actuales y obtener las predicciones más probables sobre el fenómeno en estudio.

Siguiendo lo anterior, es necesario saber que para el desarrollo de este tipo de proyectos se requiere tener acceso a plataformas que permitan el procesamiento de grandes volúmenes de datos a gran velocidad, y en distintos formatos. Este tipo de proyectos requiere de cierta potencia de hardware en función de la cantidad de información a procesar. Así de esta forma si se desea ejecutar la plataforma Cloudera se requiere una máquina con grandes recursos en términos de procesamiento, memoria RAM y capacidad del almacenamiento. Pero también hay plataformas livianas como WEKA que pueden ejecutarse en máquinas con pocos recursos y funcionan de manera adecuada, dependiendo de los algoritmos que se utilicen. Por ejemplo, al ejecutar algoritmos de redes neuronales con 8000 registro, más de 60 variables y validación cruzada con 10 particiones del set de datos, los resultados de la ejecución de algoritmos de redes neuronales en ocasiones suelen pasar de 24 horas de procesamiento en una máquina con un procesador Rizen 3 de 3200 Mhz, y 8 Gb de RAM, donde se hicieron experimentos para el desarrollo de este proyecto.

Las herramientas de minería de datos se pueden clasificar en soluciones independientes o standalone y las que funcionan con el modelo cliente/servidor. Predominan las soluciones cliente/servidor, principalmente orientadas a productos empresariales. están disponibles para diferentes plataformas, incluyendo windows, mac os, linux o servidores más robustos. Ha habido un notable incremento en el número de sistemas basados en java que son independientes de la plataforma para los usuarios orientados a la investigación.(Mikut & Reischl, 2011)

Big-Data

Big data se ha vuelto un tema del que muchos hablan y usan para describir grandes cantidades de información, pero también como dice Boyd y Crawford (Malvicino & Yoguel, 2015) lo interesante está en "las capacidades de búsqueda y agregación de grandes cantidades relacionales" (p.26). Y es por ello que 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.

Es así como, 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.

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. Donde no simplemente se ve al profesor como un instructor si no como un formador e inspiración que estimula a sus estudiantes dando la posibilidad de dar un enfoque de aprendizaje personalizado y permitiendo que el estudiante alcance su pleno potencial, todo esto gracias al big-data; la revolución digital que ha recolectado más información en los últimos dos años que en toda la historia del mundo. A través del aprovechamiento de las grandes cantidades de información, se pueden solucionar los desafíos que el siglo XXI presenta en materia de educación. El internet ya ha tenido un gran impacto en la forma que los estudiantes realizan y adquieren información, ahora la atención está enfocada hacia los grandes volúmenes de datos que proporcionan información valiosa acerca de las habilidades individuales de los estudiantes que hoy en día se traducen en herramientas y

técnicas para desarrollar materiales de aprendizaje interactivos que hacen la educación más efectiva.(Dumon, 2014)

Plataformas Tecnológicas.

Existen diversas plataformas tecnológicas sobre las cuales se ejecutan los algoritmos de minería de datos o machine learning, en estas plataformas conviven los datos y los algoritmos de aprendizaje. las cuales como afirma (Perez et al., 2018) se denominan entornos de desarrollo para ciencia de datos.

Hay muchas plataformas de minería de datos, algunas son de pago y otras son software libre. En las plataformas de pago podemos mencionar IBM Data Modeler, Oracle Data miner. Dentro de las plataformas de software libre se pueden mencionar entre otras WEKA, RapidMiner, Orange. Pandas la cual integra Python y R con sus librerías de machine learning.

Weka

Esta plataforma es de código abierto, fácil de instalar y muy flexible a la hora de ejecutarla. Es liviana y no requiere una gran potencia de procesamiento, dependiendo de la cantidad de registros y de variables a procesar. Sin embargo, para procesar más de 10000 registros con 50 variables y dependiendo del tipo de algoritmos que se ejecuten, podría ser necesaria una maquina dotada con buena capacidad de procesamiento. WEKA Se puede instalar sobre varios sistemas operativos, al estar desarrollado en el lenguaje JAVA es independiente del sistema operativo ya que se ejecuta sobre la JVM de java. WEKA provee implementaciones de algoritmos de aprendizaje que se pueden aplicar fácilmente a los datasets. Permite la realización de proyectos de minería de datos de manera ágil posibilitando al investigador centrarse más en la lógica del negocio que en la programación. De esta forma la plataforma WEKA ofrece una gran gama de herramientas y algoritmos para el tratamiento de los datos permitiendo afinar el set de datos. La

plataforma WEKA se enfoca en las siguientes opciones: Preprocesado, Clasificación, Agrupamiento, asociación, selección de atributos y visualización. El preprocesado de los datos, donde hace la preparación y adecuación del set de datos. El módulo de clasificación de los datos aplicando técnicas de aprendizaje supervisado, con una variedad de algoritmos clasificados en algoritmos de Bayes, Funciones entre las que se destacan algoritmos de regresión logística y redes neuronales, Algoritmos Lazy, algoritmos basados en reglas y árboles de decisión, entre los más populares. (Witten et al., 2016).

Marco Conceptual

En el presente apartado se desarrollan los conceptos con base en su aplicación a este proyecto. Es de suma importancia el reconocimiento de las teorías que son utilizadas en el proyecto mediante la conceptualización. En este sentido, se presentan los conceptos que se aplican a esta investigación.

Data Mining

En el desarrollo de este proyecto, se utilizaron técnicas de data mining aplicadas a los datos obtenidos del sistema de registro y control y la encuesta de caracterización. Las técnicas de minería de datos juegan un papel importante en la exploración y análisis de conocimiento a partir de grandes cantidades de datos. La minería de datos o Data mining es un proceso que consiste en aplicar analítica de datos y algoritmos de descubrimiento para encontrar patrones o modelos. (Mikut & Reischl, 2011)

Es así como, para el desarrollo de este proyecto se aplicaron técnicas de minería de datos a partir del preprocesamiento del set de datos para luego aplicar algoritmos de clasificación de aprendizaje supervisado, esto dio como resultado unos modelos, algunos con mejores métricas que otros dependiendo del tipo de algoritmos utilizados.

Aprendizaje Supervisado

Para el desarrollo de este proyecto se aplican técnicas de aprendizaje supervisado ya que en el set de datos se conocen las variables y el resultado que se espera predecir, que es la deserción estudiantil. De esta manera se procesan datos históricos de matriculados durante el primer periodo del 2018 observando su comportamiento en términos de deserción durante los años 2018,2019 y 2020. A partir de esta ventana de observación empleando técnicas de aprendizaje supervisado se aplican diferentes algoritmos y se observan los indicadores de precisión para cada experimento.(Han et al., 2011)

Algoritmos De Aprendizaje Supervisado

En el desarrollo de esta investigación se utilizaron los siguientes algoritmos de aprendizaje supervisado.

Arboles De Decisión

En el desarrollo del proyecto los árboles de decisión juegan un papel importante toda vez que estos generan unos modelos cuya principal ventaja es la facilidad de explicación. El modelo de bifurcación de las decisiones obtenido de un árbol de decisión es fácilmente explicable, sin embargo, este tipo de algoritmo suele tener el problema de sobredimensionamiento, que ocurre cuando se construye un buen modelo a partir del set de datos, pero no se obtienen buenos resultados en el set de pruebas. Los árboles de decisión son altamente susceptibles a este inconveniente por lo que para solucionarlo se aplican técnicas de poda, quiere decir que, se reduce el árbol para que no haya sobredimensionamiento. Durante el desarrollo de esta investigación se utilizaron algoritmos basados en árboles de decisión como Random forest, Algoritmo J48, Random Tree y LMT.(Fayyad & Piatetsky-Shapiro, Gregory Smyth, 1996)

Algoritmos Bayesianos

En el desarrollo del proyecto se hicieron experimentos con algoritmos bayesianos los cuales parten de la premisa que la clase a predecir se comporta de acuerdo con una distribución probabilística. Para este proyecto se utilizaron algoritmos bayesianos para algunos experimentos con la finalidad de obtener un modelo optimo por medio de estas distribuciones aplicados al con junto de datos. (Witten et al., 2016)

Redes Neuronales

En el desarrollo de este proyecto de minería de datos se hicieron experimentos aplicando redes neuronales buscando obtener los mejores resultados posibles a partir de la aplicación del algoritmo multilayer perceptron disponible en la plataforma WEKA.

Herramientas Y Técnicas De Minería De Datos

Como se mencionó en pasajes anteriores existe una variada gama de estas herramientas lo que abre un abanico de posibilidades de selección de acuerdo con las necesidades y preferencias del investigador. En este sentido puede que alguno opte por utilizar herramientas basadas en Python o algunas otras basadas en Java, de igual manera el factor comercial tiene un rol preponderante en la ecuación y desde esta perspectiva se puede seleccionar una herramienta de software libre o una comercial y ambas tienen sus ventajas y desventajas. (Mikut & Reischl, 2011)

Para el desarrollo de este proyecto se utilizó la plataforma de minería de datos desarrollada por la universidad de Waikato (WEKA) la cual es una plataforma de software libre para aprendizaje automatizado y minería de datos distribuida bajo licencia GNU-GPL.(Cuji et al., 2017)

CRISP-DM

La metodología CRISP-DM aporta mucho al desarrollo de esta investigación. Desde su comprensión y aplicación. Se siguen los pasos a través de los cuales se alcanzan los resultados esperados. Para ello se parte de la comprensión del negocio para determinar los requisitos y objetivos del proyecto, luego se realiza la elección de las fuentes de información, se hacen labores de limpieza y preprocesamiento de los datos, se efectúa la transformación de los datos adecuándolos al formato necesario para aplicarles algoritmos de minería de datos y finalmente se realiza la interpretación y evaluación de los resultados.(Perez et al., 2018)

Preprocesamiento De Datos

El preprocesamiento de los datos en esta investigación se desarrolla mediante la aplicación de técnicas y algoritmos supervisados que posibilitan la revisión de la información, Examinando la validez de los datos, evaluando la importancia de los atributos en el experimento, eliminando atributos irrelevantes para la elección, preprocesamiento, transformación, aplicación de técnicas de minería de datos, interpretación y evaluación intentando reducir el número de atributos, revisión del balance de clases hasta transformar los datos en un conjunto adecuado para el ejercicio de minería de datos.(Witten et al., 2016)

Métricas

Los indicadores de precisión permiten evaluar los modelos de predicción obtenidos. En este sentido se utilizan indicadores como la matriz de confusión la cual se utilizó para evaluar el número de aciertos durante la ejecución de los modelos, así como los falsos positivos y falsos negativos. También se tuvieron en cuenta indicadores como el Accuracy, el recall y el grado de precisión del modelo.(Han et al., 2011).

Diseño Metodológico

Diseño Y Enfoque De Investigación

En este apartado se describe el tipo de investigación adoptado para este estudio, así como el enfoque utilizado y las fases para el diseño de la investigación.

Tipo De Investigación

Para este estudio se adoptó el diseño de investigación descriptivo teniendo en cuenta que se utilizan datos específicos para el diseño de los modelos concebidos a partir de la información recolectada de los estudiantes nuevos que ingresaron en el primer periodo del 2018.

De este modo se definen las fases del diseño planteado para este trabajo de acuerdo con (Rodriguez & Vargas E, 2013) donde en primer lugar se tiene la fase de Delimitación del problema de estudio en el cual se define el objeto de estudio para seleccionar la metodología pertinente. En segundo lugar, se tiene la Revisión teórica en donde se identifica el objeto de estudio dentro del marco de conocimiento que involucra el área. En tercer lugar, se tiene la fase de Elaboración del instrumento, en esta fase, se establecen los criterios de determinación para los datos que se necesitan. En cuarto lugar, se tiene la Aplicación del instrumento en el cual Mediante la recolección de datos se aplica el instrumento. En quinto lugar, se tiene el Análisis de datos en el que se agrupan los resultados a partir de los cuales se puede inferir. Y, por último, se tiene la fase de Redacción de conclusiones donde se seleccionan los resultados obtenidos de la investigación y se genera un informe que contenga de manera coherente los resultados.

Enfoque

Para este proyecto, se toma el enfoque cuantitativo toda vez que mediante la recolección de datos

confiables se busca medir un fenómeno o probar una hipótesis apoyándose en el análisis estadístico y la medición numérica el cual constituirá una base sobre el cual se pueden establecer conclusiones. (Hernández Sampieri & Mendoza Torres, 2018)

La presente investigación se basa en el enfoque cuantitativo y el diseño de esta es descriptivo con la finalidad de detallar los factores que más inciden en la deserción de la población estudiantil en la UNAD. El enfoque cuantitativo se hace visible toda vez que algunos datos de variables son presentados cuantitativamente, sin embargo, se hace la descripción o el significado de esos datos. Este proyecto busca el diseño de una herramienta, que permita la predicción de la deserción a partir de la información recolectada a estudiantes de primera matricula en la UNAD.

Procedimiento

Con base en el diseño de investigación descriptivo se proponen las siguientes fases para alcanzar los objetivos propuestos:

Fase 1: Delimitación del problema de estudio y revisión teórica

En esta fase se llevan a cabo las siguientes actividades:

- Planificación del proyecto
- Se llevará a cabo una revisión bibliográfica en bases de datos especializadas con la finalidad de recabar información sobre la problemática en estudio.
- Revisión de antecedentes.
- Recolectar la información necesaria acerca de las técnicas de minería de datos
- Realizar un análisis de los algoritmos predictivos existentes con el fin de determinar los adecuados.
- identificar las fuentes de donde se tomará a información.

Fase 2: Elaboración del instrumento

En esta fase se llevan a cabo las siguientes actividades:

- Diagnosticar la percepción de la deserción estudiantil en docentes mediante la elaboración de un instrumento tipo encuesta.
- Determinar los algoritmos de machine learning que se consideren adecuados para el tratamiento de los datos.
- Obtención de los datos iniciales.
- Descripción de los datos.
- Verificar la calidad de los datos.

Fase 3: Aplicación del instrumento

En esta fase se llevan a cabo las siguientes actividades:

- Aplicar la encuesta de percepción de la deserción estudiantil en docentes.
- Selección de los datos aplicando criterios de inclusión y exclusión
- Limpieza de los datos.
- Generar atributos derivados a partir de la construcción de datos.
- Integrar los datos a partir de la combinación de las fuentes de información.
- Realizar el cargue de los datos a la plataforma de minería de datos.
- Realizar el preprocesamiento de los datos.
- Ejecutar los algoritmos seleccionados, experimentando con diferentes configuraciones utilizando validación cruzada.
- Identificación de los mejores modelos mediante la comparación de los resultados.

Fase 4: Análisis de datos y redacción de conclusiones

En esta fase se llevan a cabo las siguientes actividades:

- Analizar y redactar las conclusiones de la percepción de la deserción estudiantil en docentes.
- Análisis de los resultados de los algoritmos ejecutados.
- Revisión de los indicadores de precisión de los modelos obtenidos.
- Elaboración de conclusiones.
- Elaboración del informe final del proyecto

Hipótesis

El desarrollo de un modelo de predicción de deserción estudiantil, apoyado en técnicas de minería de datos, permitirá mejorar los índices de deserción estudiantil, en un curso de primera matrícula de la escuela ECBTI de la UNAD.

Variables

En el siguiente apartado se proponen las variables dependientes e independientes planteadas dentro de ese proyecto.

Variables Dependientes

Se identifican las siguientes variables dependientes: El modelo para la predicción de la deserción, el pronóstico de la deserción, el grado de confiabilidad del modelo, la generación de patrones y por último la predicción.

Variable Independientes

Se identifican las siguientes variables independientes: El data set, en análisis diagnóstico de la deserción, diseño de los elementos del modelo, uso de técnicas de data mining, los algoritmos de machine learning a utilizar, y por último el data frame.

Tabla 1.

Operacionalización de variables

Operacionalización de variables:

VARIABLE	DIMENSIÓN	INDICADOR	TIPO	ESCALA
Deserción Estudiantil en la UNAD	Cuantitativa	% de probabilidad	Dependiente	%
Efectividad del pronóstico	Cuantitativa	% de probabilidad	Dependiente	%
El modelo para la predicción de la deserción	Cualitativa	Cantidad de modelos válido	Independiente	Bueno, regular malo
La calidad de los datos	Cualitativa	calidad de los datos	Independiente	Bueno, regular malo
Los algoritmos utilizados	Cualitativa	Validez de resultados	Independiente	Válido, inválido
Los patrones que se lograron identificar a través de la aplicación del modelo	Cuantitativa	Cantidad de patrones encontrados	Dependiente	% de influencia en la predicción

El uso de técnicas de Big Data	Cualitativa	Cantidad de técnicas usadas	Dependiente	% de influencia en la predicción

Nota. Esta tabla muestra la operacionalización de las variables

Población Y Muestra

La identificación de la población y muestra es fundamental dentro del proceso de investigación, de acuerdo con (Sampieri et al., 2014) definir los participantes, delimitar la población y precisar el tamaño de la muestra (en el caso de que sea necesario) constituye una práctica adecuada para el desarrollo de la investigación. Es así como, en primer lugar la población de este estudio está constituida por los estudiantes nuevos que ingresaron en el periodo I del año 2018, los cuales basados en el informe de gestión (Abadía et al., 2018) fueron 15660 Estudiantes nuevos que ingresaron en los periodos 16-01, 16-02 y 8-03. En el semestre I del año 2018. Para determinar la percepción de la deserción, se tiene que la población de docentes de la zona caribe contratados en el periodo 2020 II es de 264 , según (Cuestas, 2020).

Caracterización De La Población

La investigación a nivel práctico y como entorno de obtención de datos primarios, se propone dentro de un sector de la población estudiantil de la UNAD específicamente a los estudiantes de primera matricula del año 2018 con el propósito de poder observar el comportamiento en el tiempo durante los periodos posteriores y de esta manera saber si de verdad el estudiante dejó de matricular o no, pudiendo de esta forma validar el modelo de predicción de deserción.

A partir de los informes de la prueba de caracterización correspondiente al periodo 2018 I, se encuentra que estos estudiantes nuevos que ingresaron en ese periodo se caracterizan por estar distribuidos en todos los centros de la UNAD en el país, de esta población de estudiantes

nuevos el 51% son mujeres y el 49% son hombres. El 85% vive en zona urbana mientras que el 15% vive en zona rural. La mayoría de esta población está ubica en un rango de edades entre los 19 y 30 años. De estos estudiantes el 17% reportó ser desplazado.

Así mismo también para el análisis diagnóstico en esta investigación se necesitaron docentes para aplicar un instrumento a los que pertenecen a la zona caribe. La población de docentes de la zona caribe contratados en el periodo 2020 II es de 264 docentes cifra tomada de Informe De Gestión De Proceso Gestión Del Talento Humano del periodo 2020 II. (Cuestas, 2020). Estos docentes tienen un perfil de formación de diferentes áreas del conocimiento y se encuentran distribuidos en los 10 centros de la zona, distribuidos en los departamentos de Guajira, Magdalena, Cesár, Atlántico, Bolívar, Sucre y Córdoba. De estos docentes 66 son de Hora cátedra ,156 de medio tiempos y 42 de tiempo completo.

Muestra

Se considera que es pertinente el uso del cálculo de una muestra representativa ya que según (Sampieri et al., 2014) representa una ventaja por la economía de tiempo y recursos que proporciona. Dentro de esta población de estudiantes se tomó el subgrupo del período 16-01, que se matricularon en el año 2018, Teniendo en cuenta que la población de estudiantes nuevos fue de 15660. Aplicando la fórmula para calcular el tamaño de la muestra con un margen de error del 5%, el tamaño de la muestra debe ser como mínimo 375 estudiantes.

Para el cálculo de la muestra se decidió aplicar la formula del muestreo probabilístico $n = \frac{Z^2 * P * Q * N}{(N-1)E^2 + Z^2 * P * Q}$ que plantea (Sampieri et al., 2014). Donde *n* será el tamaño de la muestra que se desea hallar, Z el grado de confianza,

n= muestra

Z= Nivel de confianza.

Z= 1.96 Nivel de confianza del 95%

N= Tamaño de la población = 15660 estudiantes

E= error de muestreo. El muestreo se conservará en un 5%. E= 0,05

P= Proporción de individuos que poseen en la población la característica de estudio. P=0,5

Q= Proporción de individuos que no poseen esa característica de estudio. Q= 0,5

Cómo el nivel de confianza es del 99%Z^2= 6.6564

(N-1) = 15659

. E^2= 0.0025

n= $(((Z)^2 * (P*N*Q))/(((E)^2 * (N-1)) + ((Z)^2 * (P*Q)))$

n= $((3.8416) * (0.5*15660*0.5))/(((0.0025) * (15659)) + ((3.8416) * (0.5 * 0.5)))$

n= 374.9 ≈ 375 estudiantes a encuestar.

Para calcular el tamaño de la muestra para el instrumento aplicado a Docentes, Se decidió tomar la técnica de muestreo por conveniencia ya que de acuerdo con (Otzen & Manterola, 2017) el muestreo por conveniencia permite seleccionar casos toda vez que estos sean accesibles, que acepten ser incluidos gracias a la conveniencia que posee el investigador con respecto a estos casos, teniendo en cuenta su proximidad y accesibilidad. Es por ello que se seleccionaron 36 docentes a encuestar.

Es de suma importancia mencionar que según (Sampieri et al., 2014) no siempre es necesario emplear el estudio sobre la muestra, si se busca realizar un censo se deben incluir todos los casos. Es así como, a pesar de que se hayan efectuado los cálculos para determinar cuál es el tamaño de la muestra se emplean como un recurso para conocer el mínimo de elementos que se deben incluir dentro de la investigación.

Fuentes De Información

El desarrollo de este modelo requiere sustento de carácter teórico y experimental, de tal manera que se pueda clasificar la información en categorías; y a través de ellas comprender como son las relaciones al interior de los procesos a analizar durante la investigación.

Fuentes Primarias

La fuente primaria de información en la que se pretende basar esta investigación es la encuesta de caracterización que se realizó a los estudiantes nuevos, en el periodo 2018-I. La encuesta es una técnica de investigación cuantitativa que posibilita la recolección de datos de una forma objetiva, sobre una muestra representativa de manera eficiente. En su aplicación "El interés del investigador no es el sujeto concreto que responde el formulario, sino la población a la que pertenece", por lo que mediante las técnicas de muestreo adecuadas, faculta la aplicación masiva de los resultados a una población o universo , del que se pretende "explorar, describir, predecir y/o explicar una serie de características" (Casas et al., 2003)

Las fuentes primarias de información descritas en este apartado hacen referencia a las herramientas para la recolección de datos de primera mano entre estas se encuentran las siguientes fuentes:

- Encuesta de Caracterización de los estudiantes
- Base de datos de estudiantes matriculados durante los años 2018-2020

Fuentes Secundarias

De igual forma este estudio requirió la clasificación de otras fuentes teóricas, que permitieron aumentar teóricamente aspectos como

- Tendencias de matrícula de los estudiantes en el periodo 2019

- Revistas con publicación de artículos acerca de minería de datos
- Tendencias disruptivas en TI

Análisis Diagnóstico Para La Determinación De Los Requerimientos Del Modelo

Requerimientos

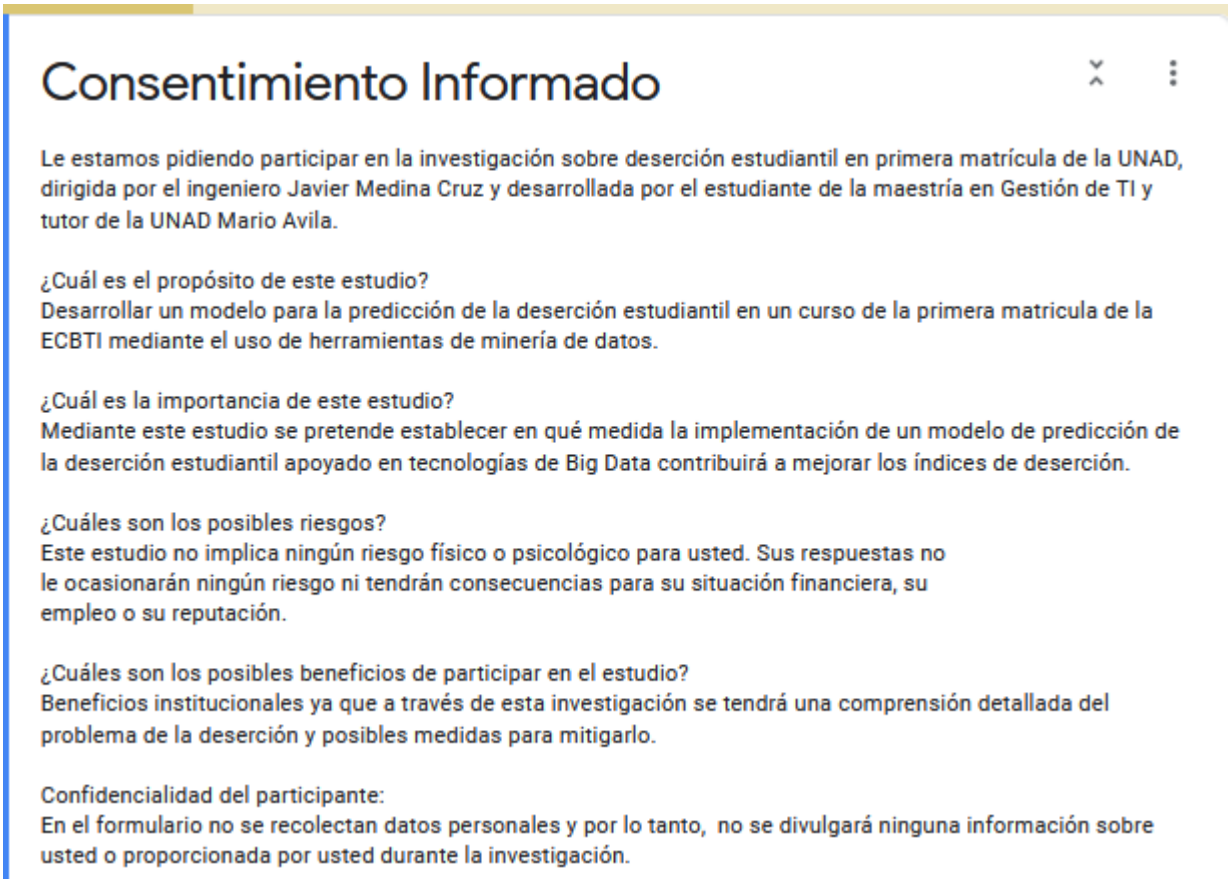
Con la finalidad de hacer una correcta identificación de requerimientos que permitan una adecuada concepción del proyecto que se está abordando, se aplicarán cuestionarios, los cuales serán dirigidos a un público específico para obtener más información.

Requisitos Y Características Del Producto

Los principales requisitos y características del producto están enmarcados en la utilización de herramientas tecnológicas de analítica de datos como, WEKA, que permitan llegar a la construcción de un modelo de predicción para el pronóstico la deserción estudiantil. Esta herramienta debe permitir analizar los datos históricos de los estudiantes Unadistas, como insumo principal para detectar patrones por ejemplo por edades, estratos socioeconómicos, programas académicos y otros relacionados con la deserción. La capacidad por generar corresponde al tratamiento de datos de los estudiantes pertenecientes a los cursos de primera matrícula de los programas de la escuela ECBTI.

Instrumento

Con la finalidad de diagnosticar la percepción que tiene el cuerpo docente frente a la problemática de deserción, se diseñó un instrumento de indagación cerrado para ser aplicado en el cuerpo docente de la UNAD (ver anexo 1). El instrumento encuesta inicia con un consentimiento informado para los participantes que desean participar en la investigación, como se aprecia a continuación



Consentimiento Informado

Le estamos pidiendo participar en la investigación sobre deserción estudiantil en primera matrícula de la UNAD, dirigida por el ingeniero Javier Medina Cruz y desarrollada por el estudiante de la maestría en Gestión de TI y tutor de la UNAD Mario Avila.

¿Cuál es el propósito de este estudio?
Desarrollar un modelo para la predicción de la deserción estudiantil en un curso de la primera matricula de la ECBTI mediante el uso de herramientas de minería de datos.

¿Cuál es la importancia de este estudio?
Mediante este estudio se pretende establecer en qué medida la implementación de un modelo de predicción de la deserción estudiantil apoyado en tecnologías de Big Data contribuirá a mejorar los índices de deserción.

¿Cuáles son los posibles riesgos?
Este estudio no implica ningún riesgo físico o psicológico para usted. Sus respuestas no le ocasionarán ningún riesgo ni tendrán consecuencias para su situación financiera, su empleo o su reputación.

¿Cuáles son los posibles beneficios de participar en el estudio?
Beneficios institucionales ya que a través de esta investigación se tendrá una comprensión detallada del problema de la deserción y posibles medidas para mitigarlo.

Confidencialidad del participante:
En el formulario no se recolectan datos personales y por lo tanto, no se divulgará ninguna información sobre usted o proporcionada por usted durante la investigación.

Y su aceptación con la siguiente leyenda: “He leído satisfactoriamente las explicaciones sobre este estudio. Estoy enterado y deseo participar en este estudio. Autorizo el uso de la información para los propósitos de la investigación”.

El instrumento conta de las siguientes preguntas o categorías con respuestas cerradas:

a) ¿En qué medida considera usted que la implementación de un modelo de predicción de la deserción estudiantil en la UNAD contribuye al mejoramiento de la retención y permanencia de los estudiantes?
b) De las siguientes herramientas de Big Data, ¿Cuáles le son familiares o ha interactuado?
c) ¿Cuál es la percepción de la deserción de los estudiantes en el curso que usted es tutor?
d) ¿Cuáles de los siguientes factores considera usted que son los que más inciden en la deserción estudiantil?

e) ¿Considera usted que los índices de aprobación del curso influyen en la deserción estudiantil?
f) Porcentaje de estudiantes que no hicieron entrega de ninguna actividad en algún curso que usted haya acompañado.
g) ¿Cree usted que aplicando inteligencia artificial es posible mitigar el fenómeno de la deserción estudiantil?
h) ¿A qué nivel ha percibido la deserción estudiantil en el curso que usted es tutor en los últimos 2 periodos académicos?
i) ¿Cree usted que las metodologías educativas aplicadas actualmente fomentan la retención y permanencia estudiantil?
j) ¿Considera usted que tomar en cuenta los ritmos de aprendizaje pueden influir en la deserción estudiantil?
k) De las siguientes consecuencias sociales provocadas en parte por la deserción estudiantil, ¿Cuál cree usted que es en la que más incide?

Es así como, después de la posterior realización de la encuesta, se tomaron los datos arrojados y se llevaron a al software IBM SPSS para realizar el respectivo análisis estadístico de los datos. A continuación, serán presentados el análisis y los gráficos obtenidos del análisis:

Análisis De Resultados Obtenidos De La Aplicación Del Instrumento

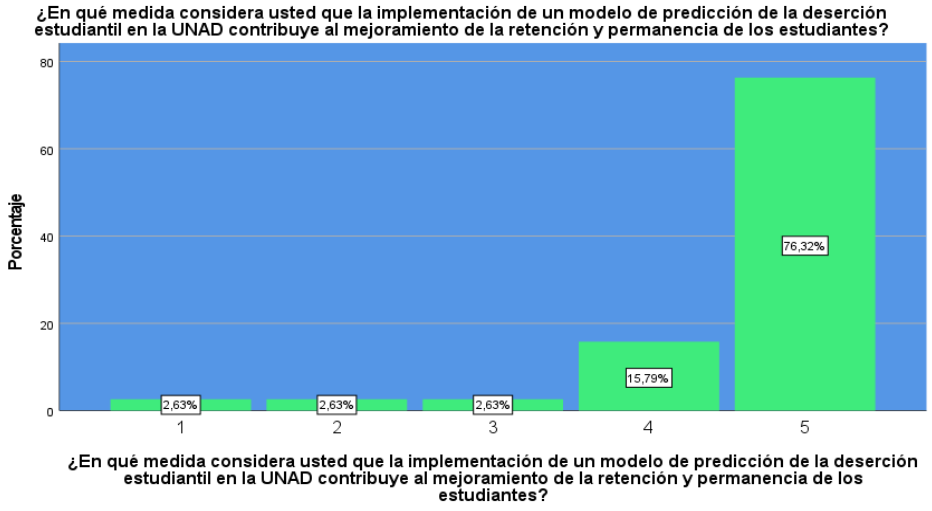
Se observa que para la primera categoría se tuvieron un total de 38 respuestas que responden al interrogante "¿En qué medida considera usted que la implementación de un modelo de predicción de la deserción estudiantil en la UNAD contribuye al mejoramiento de la retención y permanencia de los estudiantes?" de acuerdo con la gráfica se observa que un 76,32% de las respuestas, es decir, 29 tutores consideran que la implementación de un modelo de predicción de

la deserción estudiantil en la UNAD contribuye en gran medida al mejoramiento de la retención y permanencia de los estudiantes. Se observa una respuesta positiva frente a este interrogante, lo cual evidencia una buena percepción acerca de cómo la utilización de tecnologías disruptivas podría contribuir a mejorar problemáticas del contexto como el abandono de los estudios.

De este modo, los resultados obtenidos de esta categoría son una fuente de información que se acerca a la respuesta de la pregunta problema de este estudio, y a su vez coincide con lo planteado en el objetivo general del mismo, relacionándose con lo planteado por el autor (Heredia et al., 2015), en su estudio Modelo predictivo de deserción estudiantil basado en arboles de decisión.

Figura 7

Percepción del mejoramiento a través del modelo de predicción.



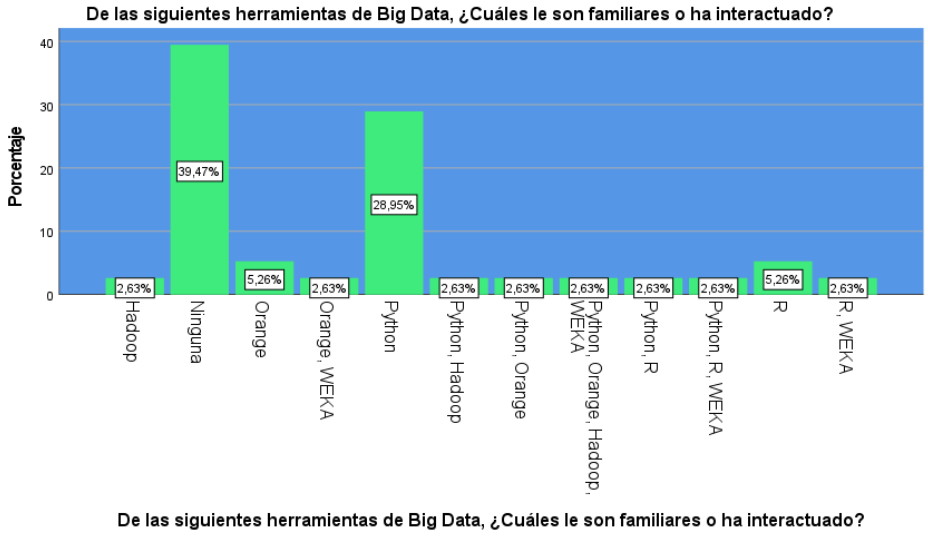
Nota. El grafico muestra la percepción que tienen los docentes acerca del mejoramiento derivado de la aplicación de un modelo de predicción.

Así mismo, se observó en la siguiente categoría, de las cuales hubo 38 respuestas donde se analiza el comportamiento frente al interrogante “De las siguientes herramientas de Big Data,

¿Cuáles le son familiares o ha interactuado?". Observando la figura se infiere que el conocimiento de los tutores acerca de las herramientas de Big Data es bajo, siendo 39,47%, es decir, 15 tutores no tienen conocimiento acerca de ninguna herramienta de Big Data. Esta pregunta permite conocer la percepción que tienen los tutores acerca de las herramientas de Big Data. Estas cifras dan cuenta de lo novedoso del tema para el cuerpo de tutores, y también evidencia un gran potencial de oportunidades que se pueden capitalizar con la aplicación de este tipo de estudios.

Figura 8

Herramientas de Big Data con las que se ha interactuado



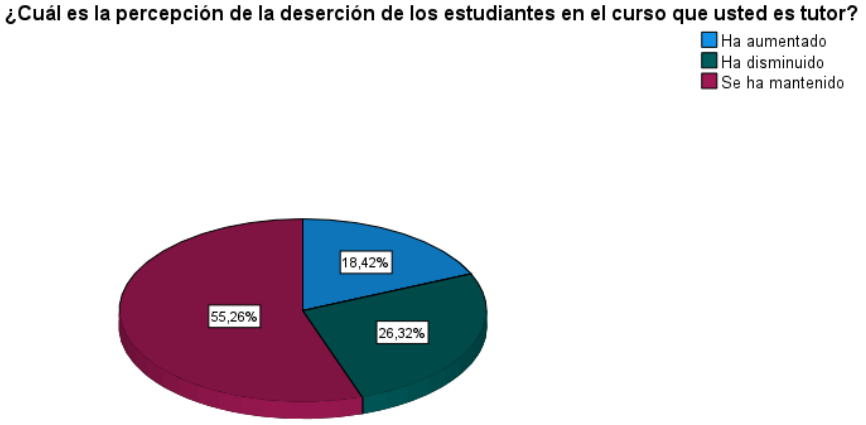
Nota. El grafico muestra las herramientas de big Data cuyo nombre resulta familiar a los encuestados.

Pasando a la siguiente categoría de la cual se obtuvieron 38 respuestas que responden al interrogante: "¿Cuál es la percepción de la deserción de los estudiantes en el curso que usted es tutor?". Pasando a la siguiente categoría de la cual se obtuvieron 38 respuestas que responden al interrogante: "¿Cuál es la percepción de la deserción de los estudiantes en el curso que usted es

tutor?". Dentro de las gráficas se evidencia que la percepción que tienen los tutores con respecto al comportamiento de la deserción en sus cursos revela que la mayoría de los tutores considera que la deserción se ha mantenido y que el menor porcentaje de tutores considera que ha aumentado. Estos resultados coinciden con la problemática planteada ya que el fenómeno de la deserción se hace un hecho perceptible ante los casos evaluados en este estudio. De acuerdo con (Ángel & Facundo, 2009) La Universidad Nacional abierta y a Distancia no se encuentra ajena al fenómeno de la deserción. Esto indica la relevancia que tiene el objetivo general de este proyecto, permitiendo el desarrollo de un modelo para la predicción de la deserción estudiantil toda vez que la deserción es un fenómeno presente en la institución y afecta negativamente los indicadores de gestión institucional.

Figura 9

Percepción de la deserción de los estudiantes.



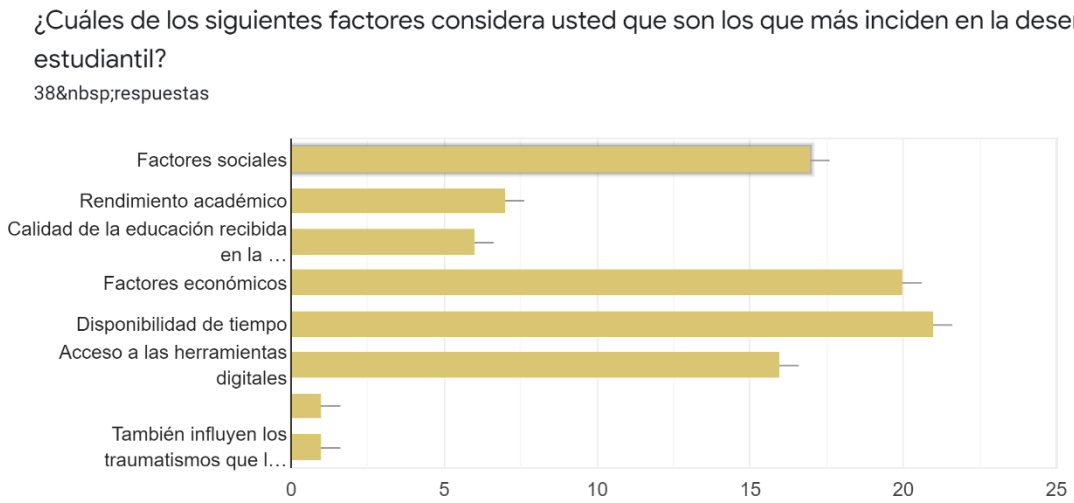
Nota. Esta imagen muestra la percepción de la deserción de los estudiantes.

Dentro de la siguiente categoría se analiza el interrogante del cual se obtuvieron 38 respuestas "¿Cuáles de los siguientes factores considera usted que son los que más inciden en la deserción estudiantil?" observándose gráficamente cual es la percepción de los tutores con

respecto a esta categoría. Los factores más incidentes dentro de la gráfica son: la disponibilidad de tiempo, factores económicos, factores sociales, y acceso a las herramientas digitales. Siendo la disponibilidad de tiempo considerado el factor más incidente. Los resultados de esta categoría coinciden con la problemática planteada y con lo afirmado por (Torres et al., 2015) donde se menciona el nivel socioeconómico, como el principal factor asociado a la deserción. Estos resultados a su vez están relacionados con primer objetivo específico ya que este pretende la obtención de un análisis diagnóstico para la determinación de los requerimientos del modelo a través de la revisión de las fuentes de información, es decir, que estos resultados aportan elementos para la consecución de este objetivo.

Figura 10.

Factores más incidentes en la deserción estudiantil.



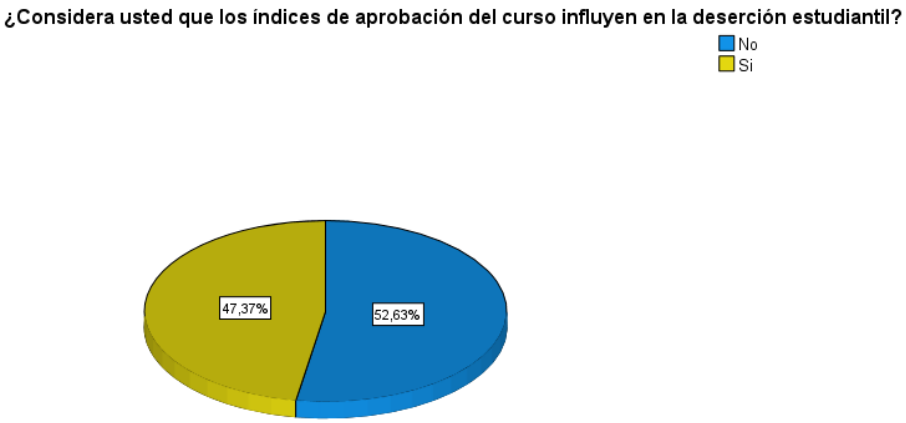
Nota. El grafico muestra lo factores que los encuestados consideran más incidentes en la deserción

En este orden de ideas, se tiene una siguiente categoría donde se registran 38 respuestas

al interrogante “¿Considera usted que los índices de aprobación del curso influyen en la deserción estudiantil?” Donde se observa que hay una percepción dividida ya que el 47,37% de los tutores considera que no influye y un 52,63% considera que los índices de aprobación del curso si influyen en la deserción estudiantil. Sin embargo, a la luz de la problemática de deserción estudiantil este 53% es significativo si se toma en cuenta que el abandono de los estudios comienza con este tipo de síntomas que luego se podrían manifestar en las causas ampliamente descritas en el planteamiento del problema, coincidiendo con lo planteado por (Ángel & Facundo, 2009) que relaciona el rendimiento académico con retención y por ende el abandono o deserción.

Figura 11

Influencia de los índices de aprobación sobre la deserción



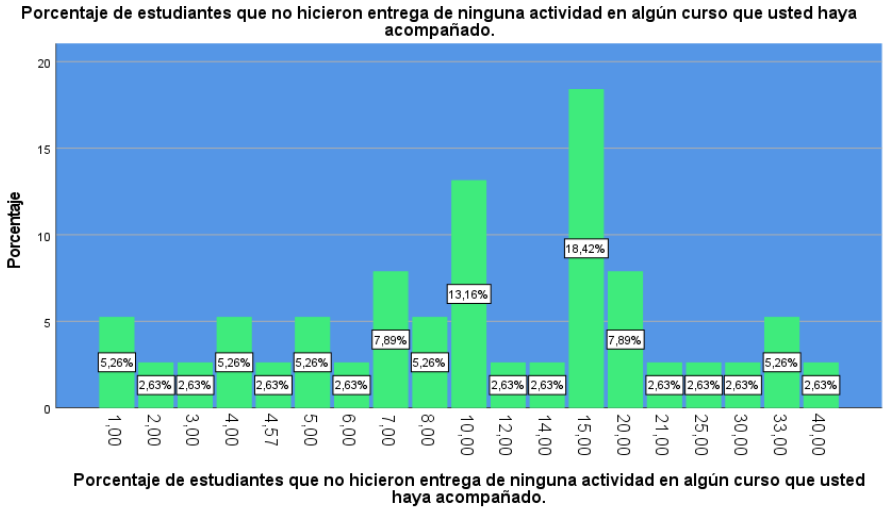
Nota. El gráfico refleja la percepción de los docentes acerca de la influencia de los índices de aprobación en los cursos sobre la deserción.

Como sexta categoría se obtuvieron 38 respuestas al interrogante “Porcentaje de estudiantes que no hicieron entrega de ninguna actividad en algún curso que usted haya acompañado.” Con el propósito de identificar el volumen de estudiantes en los cursos que están

presentes pero que sin embargo no hacen las entregas correspondientes. En la gráfica se observan los diferentes porcentajes de los cuales se extrae una media de 13% de estudiantes por curso. Los resultados de esta categoría coinciden con el estudio de la problemática de la deserción toda vez que en esta categoría se evidencian los signos de estudiantes con altas probabilidades de deserción lo que concuerda con lo afirmado por (Ángel & Facundo, 2009) quien en su estudio relaciona el rendimiento académico con el abandono de los estudios, en cierto modo esta categoría está relacionada con la pregunta anterior.

Figura 12.

Estudiantes que no hicieron entrega de ninguna actividad.



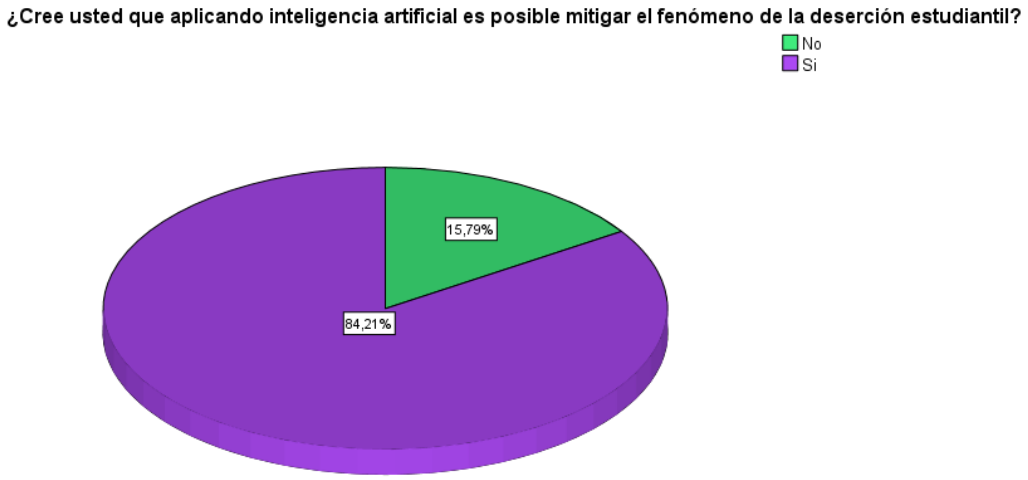
Nota. Al gráfico muestra la percepción que tienen los tutores acerca de la entrega de actividades en sus cursos

En la siguiente categoría de la cual se obtuvieron 38 respuestas al interrogante “¿Cree usted que aplicando inteligencia artificial es posible mitigar el fenómeno de la deserción estudiantil?” mediante el análisis de las gráficas es posible afirmar que la percepción de los tutores es positiva frente al uso de la IA para la mitigación del fenómeno de la deserción siendo

'Si' el 84,2% y no el 15,8%. Los resultados de esta categoría frente a la problemática planteada en este proyecto se consideran pertinentes toda vez que, según (Berlanga, 2016) las técnicas Aprendizaje Automático (campo de investigación de la IA) y la Minería de datos comparten gran cantidad de técnicas, hoy en día apuntan a la construcción de algoritmos que pueden extraer todo tipo de conocimiento a partir de un conjunto masivo de datos. Además, los resultados obtenidos apoyan lo descrito en el objetivo general del proyecto donde se pretende el desarrollo del modelo de deserción mediante las técnicas de Data Mining.

Figura 13

Percepción de la IA para mitigar la deserción.



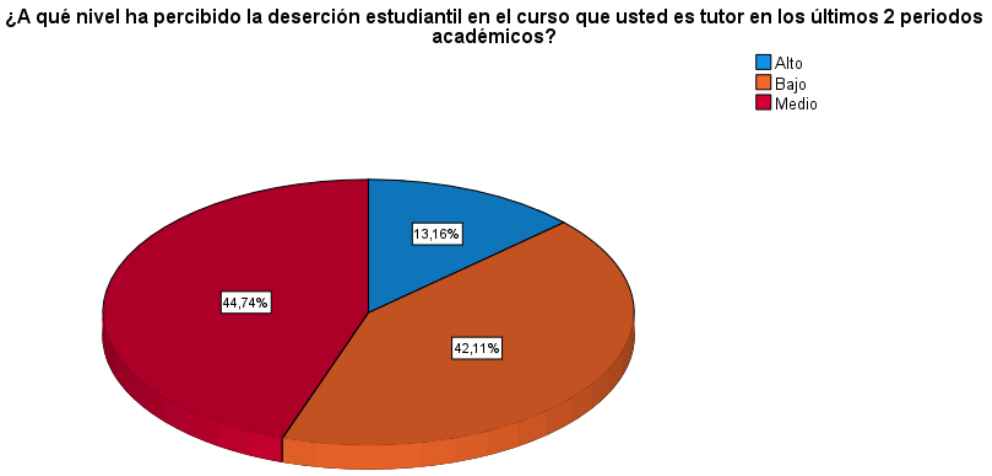
Nota. Esta gráfica muestra la percepción de la IA para mitigar la deserción

Para la siguiente categoría, siendo 38 las respuestas obtenidas, se responde al interrogante "¿A qué nivel ha percibido la deserción estudiantil en el curso que usted es tutor en los últimos 2 periodos académicos?" se obtuvo que se percibe la deserción en su mayoría en un nivel medio arrojando un 44,7%. En este sentido, 42,1% de los tutores lo percibe en un nivel bajo y 13,2% lo percibe en un nivel alto. Es así como, se determina la pertinencia de esta categoría en relación con la problemática que se plantea y la concordancia que mantiene con lo planteado por (Ángel

& Facundo, 2009) cuando afirma que a pesar de que los indicadores de deserción han venido mejorando año tras año, aún son perceptibles altos índices de este fenómeno.

Figura 14

Percepción de la deserción en los dos últimos periodos académicos

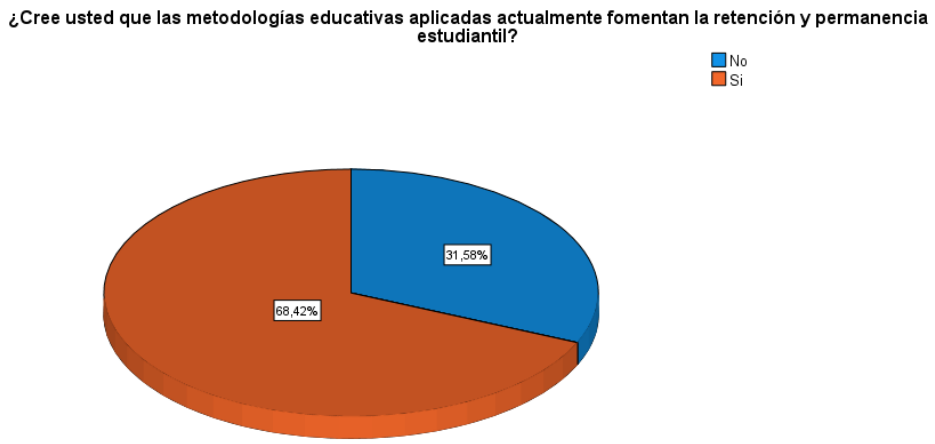


Nota. La grafica muestra la percepción de la deserción en los dos últimos periodos académicos.

En la siguiente categoría, obtenidas 38 respuestas donde se da respuesta al interrogante "¿Cree usted que las metodologías educativas aplicadas actualmente fomentan la retención y permanencia estudiantil?" De la cual se obtuvo que el 68,4% de los encuestados respondió si y un 31,6% no. Este 31,6% es significativo en el sentido que refleja oportunidades de mejoramiento dentro de las cuales el desarrollo de este proyecto tiene cabida.

Figura 15

Metodologías educativas para la retención y permanencia estudiantil.

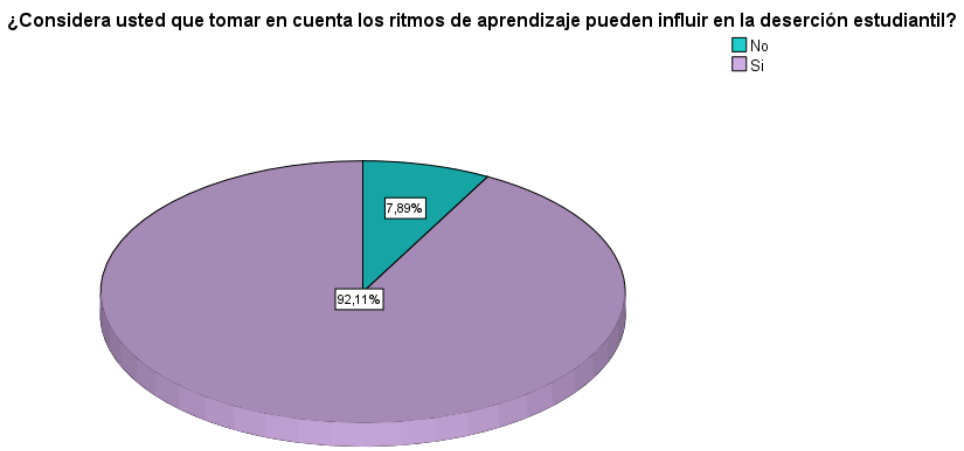


Nota. La grafica muestra la percepción acerca de las metodologías educativas para la retención y permanencia estudiantil

En este orden de ideas, se tiene una próxima categoría correspondiente al interrogante "¿Considera usted que tomar en cuenta los ritmos de aprendizaje pueden influir en la deserción estudiantil?" Donde se obtuvieron 38 respuestas de las cuales el 92,1% de los tutores responde Si y el 7,9% responde No. Se evidencia la pertinencia de esta categoría frente a la problemática planteada ya que se dice que el avance hacía una educación personalizada contribuiría a la disminución de la deserción y para esto las técnicas de minería de datos serían fundamentales en el descubrimiento de los factores particulares de cada estudiante. (Hai-ling et al., 2018) afirma que es posible educar de una forma más personalizada y esta experiencia de aprendizaje motivará jóvenes para estudiar y equiparse mejor.

Figura 16

Influencia de los ritmos de aprendizaje sobre la deserción estudiantil.

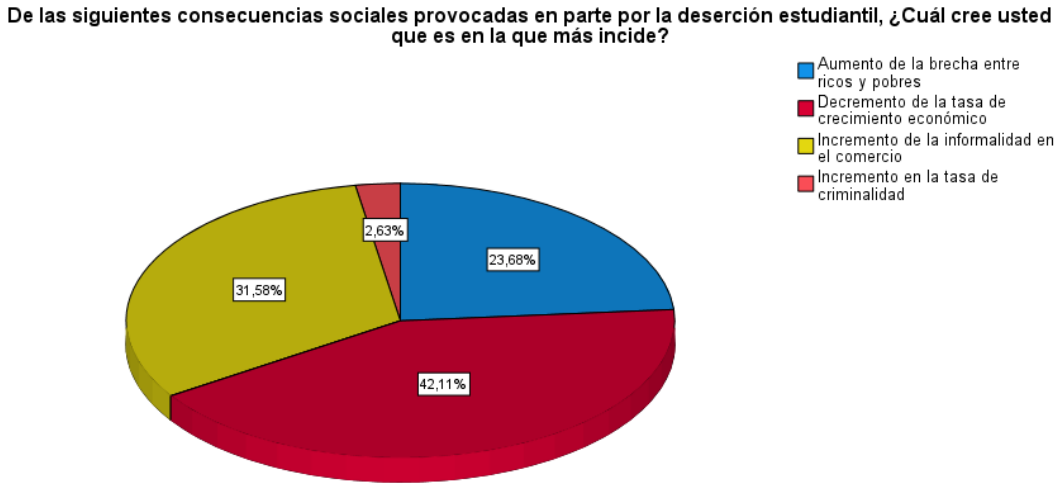


Nota. La figura muestra la percepción de los docentes acerca de la Influencia de los ritmos de aprendizaje sobre la deserción estudiantil.

Como última categoría se registran 38 respuestas al interrogante “De las siguientes consecuencias sociales provocadas en parte por la deserción estudiantil, ¿Cuál cree usted que es en la que más incide?” identificando como la causa más incidente el decremento de la tasa de crecimiento económico con un 42,1%. Además, resaltan dos consecuencias sociales más las cuales son: Incremento de la tasa de informalidad en el comercio con un 31,6% y aumento de la brecha entre ricos y pobres con un 23,7%. Los resultados obtenidos de esta categoría coinciden con la problemática planteada y con lo afirmado por (Rodríguez et al., 2016) acerca de las consecuencias sociales producto de la deserción donde el autor destaca el incremento de las tasas de criminalidad y decremento de la tasa de crecimiento económico, entre otras, como las consecuencias que son más incidentes. En este proyecto, los resultados de esta categoría indican el impacto positivo que tendría el desarrollo del modelo de predicción de deserción estudiantil para aliviar algunas de las consecuencias sociales presentes en el entorno, producto de la deserción.

Figura 17

Consecuencias sociales más incidentes en la deserción. Fuente. El autor



Nota. La grafica muestra la percepción que tienen los encuestados acerca de las consecuencias sociales más incidentes en la deserción.

Predicción Mediante Uso De Técnicas De Minería De Datos

En este capítulo se presentan resultados de la aplicación de Data Mining basada en Machin learning , para lo cual se parte con la identificación y ejecución de algoritmos de Machine Learning, específicamente algoritmos de aprendizaje supervisado debido a que se trata de un conjunto de datos históricos, un set de datos donde se conoce el resultado de deserción de los estudiantes, se pretende que a partir de la encuesta de caracterización a estudiantes nuevos realizada en el año 2018 , en el periodo 16-01, se entrene un modelo capaz de aprender de esos datos y predecir, de manera temprana si un estudiante desertará o no . En este punto del proyecto se ponen en práctica o se llevan a cabo unas actividades que constituyen la metodología utilizada para la obtención del modelo de predicción. Estas etapas o fases están en concordancia con la metodología CRISP-DM.

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, en consecuencia, se han revisado metodologías que suelen usarse para proyectos de minería de datos. La revisión de las metodologías KDD para el descubrimiento de conocimiento, en este sentido (León & Garcia, 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 característica y particularidades (León & Garcia, 2016)

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)

Figura 18.

Etapas del proceso de KDD

Nota. El grafico muestra las fases del proceso de descubrimiento de conocimiento a partir de los datos. Tomado de *(Moine, 2013)*

Continuando 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. 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, 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. Las siguientes fases del modelo pasan por la preparación de los datos, la obtención de los modelos, la evaluación y el despliegue. 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, por lo que el modelo es bastante flexible con lo que se persigue 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 & Garcia, 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 afirma Acevedo, 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 matricula. Es así como para este estudio, se puede contar con una data, que tiene registrada la información acerca de los estudiantes cursando un programa, 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 un porcentaje de confiabilidad que sea un soporte para la toma de decisiones. Con respecto a los recursos de software necesarios, se cuenta con 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 Raizen 3.

Recopilación De Los Datos

En esta etapa se recolectaron los datos de los estudiantes, para ello se tomaron en cuenta los factores que pueden influir en la deserción estudiantil y a partir de allí considerar las fuentes disponibles para la

construcción del set o conjunto de datos, el cual constituye el punto de partida para la construcción del modelo.

Identificación De Las Fuentes De Datos

La identificación de la fuente de datos es una tarea fundamental que se aborda en el desarrollo de este proyecto, esta constituye un hito fundamental toda vez que se parte de entrevista con consejería académica para determinar la disponibilidad de los datos para este estudio. Se logró identificar que se cuenta con información académica de los estudiantes que iniciaron en el año 2018 y la determinación de la calidad de los datos con los que se cuenta para el estudio. También se cuenta con información académica recabada en años anteriores a la cual se tiene acceso. Y a partir de allí hacer una labor de preparación de los datos, para lograr un nivel óptimo del estado de la información.

Para (Márquez, 2015) en la etapa de recopilación de la información se recolectan los datos disponibles de los estudiantes. Para este proyecto se parte por identificar esos factores que estudios previos han determinado que tienen una influencia significativa en la deserción de los estudiantes en la UNAD. Básicamente se trata de hacer una revisión de la información disponible identificar el estado de esta información, determinando si está en óptimas condiciones para ser utilizada para el análisis.

Se han identificado archivos de Excel con información que se considera relevante como base para el desarrollo del modelo de predicción. Se trata de archivos que las unidades académicas del centro CCAV Puerto Colombia de la UNAD han venido elaborando, donde se ha consignado información sobre periodos de matrícula y rendimiento académico de los estudiantes.

Lo ideal en este tipo de estudio es contar con acceso a la información que se encuentra en la base de datos de Registro y control, pero en estos momentos del proyecto por razones de

confidencialidad no es posible acceder a dicha información. Por este motivo para la construcción del modelo se opta por acudir a esos archivos que contienen información muy valiosa para la constitución del set de datos, también se ha solicitado información del sistema inteligente con la finalidad de buscar datos de matrículas durante el año 2020, para cruzarlos con los estudiantes del set de datos y de esta manera identificar los que no tienen matrícula activa.

Actualmente se cuenta con los resultados de un instrumento aplicado al comienzo de cada periodo académico de los estudiantes nuevos, se trata de la encuesta de caracterización, la cual es un instrumento con información muy valiosa de tipo social, académico laboral y económico. Con lo cual se tiene esa valiosa información como punto de partida para la construcción del set de datos y llevar a cabo el estudio.

De acuerdo con estas consideraciones, se han identificado 3 principales fuentes de datos.

- La primera fuente de datos es **la encuesta general de caracterización de los estudiantes**, la cual consiste en un instrumento aplicado a inicios de cada periodo a los estudiantes en la UNAD. Esta información permite la revisión del perfil de ingreso de los estudiantes en factores como: antecedentes educativos, familiares, psicosociales, sociodemográficos, competencias básicas, entre otros. Lo anterior con el propósito que los diferentes actores interesados puedan definir las estrategias disciplinares, curriculares y didácticas apropiadas para las necesidades particulares de los estudiantes.
- La segunda fuente la constituye la información obtenida sobre los estudiantes que no realizaron su proceso de matrícula en el año 2020. Estos datos se obtuvieron del sistema de ryc de la UNAD. Básicamente esta información consiste en determinar si un estudiante tuvo o no matricula en el año 2020, de esta manera se identificaron los

estudiantes que en este momento han desertado de su proceso de autoaprendizaje en la UNAD.

- Una tercera fuente de datos la constituyó una encuesta de caracterización realizada en el 2018-I , para el curso de herramientas digitales, en la cual se obtuvieron 1600 respuestas de estudiantes, principalmente orientadas a identificar problemas para conectarse a la plataforma, disponibilidad de equipos de cómputo para el desarrollo de La actividades, así como las disponibilidad de tiempo y empatía con la modalidad de estudio.

Descripción Del Dataset

Para la construcción de la data set o conjunto de datos se partió de los datos obtenidos provenientes de las fuentes mencionadas en el apartado anterior. En la revisión de estos datos se identificaron los siguientes atributos, los cuales se relacionan en la siguiente tabla:

Tabla 2
Atributos recopilados de los estudiantes en el 2018 - I

Fuente	Atributos
encuesta general de caracterización de los estudiantes	Documento, Nombres, Programa,Cead,Zona,Escuela,Estado Civil,Género,Etnia, Ciudad, Residencia, Area Residencia, Estrato, Desplazado, Disc. Visual, Disc. Auditiva, Disc. Cognitiva, Disc. Fisica, Disc. Emocional, Disc. Mental, Enfermedad, Convenio INPEC, conoció UNAD, Deporte, Instrumento, Actividad artística, Género Danza, Edad, Tipo

	Institución, Modalidad, Enfasís Bachillerato, Rendimiento Académico, Grados Perdidos, Área Dificultad, Área Mejor, Edad Icfes, Último Nivel Alcanzado, Modalidad Estudios, Graduación Estudios Cursados, Razón Abandono Estudios, Cursos Virtuales, Tomado Cursos Virtuales, Aprobado Cursos Virtuales, Tiempo sin Estudiar, Razón Sin Estudiar, Recibió Orientación Vocacional, Primer opción de estudio, No primer Opción, Razón Ingreso al Programa, Razón Ingreso a la UNAD, Persona Convivencia, Tamaño Núcleo, Numero Hijos, Hijos Dependencia, Dependencia Económica, Tipo De Dependencia, Acudiente, Parentesco Acudiente, Escolaridad Padre, Escolaridad Madre, Posición Entre Hermanos, Escolaridad Conyugue, Vivienda Actual, Tipo Vivienda, Familiar en la UNAD, Lectura Critica, NivelLec, Razonamiento Cuantitativo, NivelRaz, Tics, NivelTic, Ingles, NivelIngl

Datos obtenidos de registro y control	Código, Tuvo matricula en el año 2020
Encuesta 2018-I, para el curso de herramientas digitales para la gestión del conocimiento	Fecha, nombreApel, tipoID, ID, discapacitado, freq-Activ-U, prblmaDsllo Act, Genero, FechaNacimiento, Edad , teléfono, residencia, Hijos, centro, programa, empleado, DescDiscapacidad, horasTrbjo PC-propio, freq-uso-PC, internet-residencial, skill-PC, medio-Favorito, usa-Skype, freq-correo, skill-Redes-soc, interes-part-grupo, horario-Preferido, interes-videos-cursol, grupo, tutor HDGC, hizo-induccion, afinidad-met-virtual, exp-Virtual, Raz-ingreso, autonomía, email

Nota. Esta tabla muestra los Atributos recopilados de los estudiantes en el 2018 - I

Con toda esta información se dispone de un conjunto de datos de 77 atributos que caracterizan a cada uno de los 7881 estudiantes con los que cuenta el set de datos.

En la siguiente tabla se muestra el número de atributos por cada fuente de datos, con lo cual se determina que el conjunto de datos presenta una alta dimensión, lo cual no es muy recomendable para este tipo de ejercicios porque puede afectar de manera negativa la efectividad y eficiencia de los algoritmos de clasificación.

Las complicaciones con la alta dimensión son ampliamente conocidas en el campo de la minería de datos y se presenta cuando el número de atributos es demasiado alto, como en este caso que se está por encima de 70 atributos lo que provoca que la efectividad de los algoritmos

no sea muy precisa debido al gran número de atributos para seleccionar. Hay que reconocer que algunos algoritmos son más susceptibles a la alta dimensionalidad que otros y para algunos incluso resulta inviable debido al alto costo computacional que representa el procesamiento de un gran número de atributos.(Han et al., 2011)

Tabla 3
número de atributos por fuentes de información

encuesta general de caracterización de los estudiantes	Datos obtenidos de registro y control	Encuesta 2018-i , para el curso de herramientas digitales para la gestión del conocimiento
72	2	37

Nota. Esta tabla muestra el número de atributos por fuentes de información

Para la construcción del dataset se procede a cargar la información en una base de datos relacional para proceder a realizar el cruce de los datos y de esta manera integrar un set de datos que consolide la información en una única tabla donde se haga el match de los atributos con la variable de salida o target. Durante este proceso se tuvo el reto del ensamble, la integración y la realización de algunas tareas de limpieza de datos eliminando registros incompletos. Para ello se cargó la información de matrículas en una base de datos relacional y mediante el uso de consultas y actualización de con el lenguaje de manipulación de datos DML de SQL se hizo la integración de una especie de datawerehouse o almacén de datos, el cual consiste básicamente de una gran tabla en donde se hace un proceso de desnormalización de la base de datos relacional para integrar toda la información en una única tabla la cual posterior mente se exportó al formato CSV(Witten et al., 2016). Los archivos CSV consisten en un archivo de texto plano, con campos de valores separados por coma, en cuya primera línea se encuentra la descripción o

nombres de los campos. Este archivo CSV es el que se puede cargar a la plataforma WEKA, y una vez solucionados los problemas de cargue los cuales se detallan en el siguiente apartado , el set de datos se guarda o exporta al formato ARFF, que es el formato por defecto soportado por WEKA. En el anexo 2 se puede observar la cabecera del archivo ARFF.

La variable target o variable objetivo en este conjunto de datos, y que algunos autores también suelen llamar como variable de salida es el estado del estudiante, y que tiene dos posibilidades: **desertó**, que se refiere a los estudiantes que no matricularon durante el año 2020, y **no desertó** que corresponden a los estudiantes que actualmente se encuentran matriculados o que durante el año 2020 matricularon al menos un periodo.

El formato de archivos ARFF en WEKA la forma estándar de representar los data set, son los archivos ARFF que es el acrónimo en ingles de attribute-*relation file format*. El archivo inicia con % que es el símbolo usado para comentarios en el archivo, seguido de @relations, que es el nombre de la relación, que en este caso es deserción. Luego viene el bloque de atributos donde se hace la descripción de cada una de las variables que conforman el dataset por ejemplo *@attribute Género {F,M}* En el cual se describe el atributo género, el cual toma 2 valores Fo M. Luego viene el bloque @data, en el cual contiene los registros con los valores de las variables o atributos que se van a procesar.(Witten et al., 2016) A continuación se presenta un ejemplo del data set.

@relation 'Deserción'
@attribute G�nero {F,M}
@attribute Etnia {'No pertenece','Otras comunidades negras','No
@attribute Modalidad {Presencial,Validaci�n,Distancia,Virtual}
@attribute 'Rendimiento Academico' {Sobresaliente,Aceptable,Deficiente}
@attribute 'Grados Perdidos' {Cero,Uno,'Cinco o m�s',Dos,Cuatro,Tres}
@attribute 'Edad Icfes' numeric
@attribute '�ltimo Nivel Alcanzado' {Tecnol�gico,T�cnico,'No Aplica',Profesional}
@attribute Tics numeric
@attribute Campo70 {Sobresaliente,Suficiente,Insuficiente}
@attribute deserto {0,1}

@data
F,'No pertenece','BOGOT\ufffd D.C.','Pagina Web de la UNAD','No Aplica',Presencial,Sobresaliente,Cero,18,Tecnol\ufffdgico,'No Aplica',No,'Insatisfacci\ufffdn con la Universidad',Si,'Poca motivaci\ufffdn por la modalidad','No Aplica',?,'Proyeccion laboral',Cony\ufffdgue,'No Aplica',91,Sobresaliente,0
M,'No pertenece','LA JAGUA DE IBIRICO','Estudiante de la Unad',Bater\ufffda,Presencial,Sobresaliente,Cero,52,T\ufffdcnico,Presencial,Si,'No Aplica',Si,'Poca motivaci\ufffdn por la modalidad','No Aplica',?,'Actualizaci\ufffdn de conocimientos','Cony\ufffdgue e hijos','No Aplica',93,Sobresaliente,0
M,'No pertenece',MEDELLIN,Internet,'No Aplica',Presencial,Aceptable,Uno,26,'No Aplica','No Aplica','No Aplica','No Aplica',No,'No Aplica','Falta de tiempo',?,'Proyeccion laboral',Cony\ufffdgue,'No Aplica',65,Suficiente,1

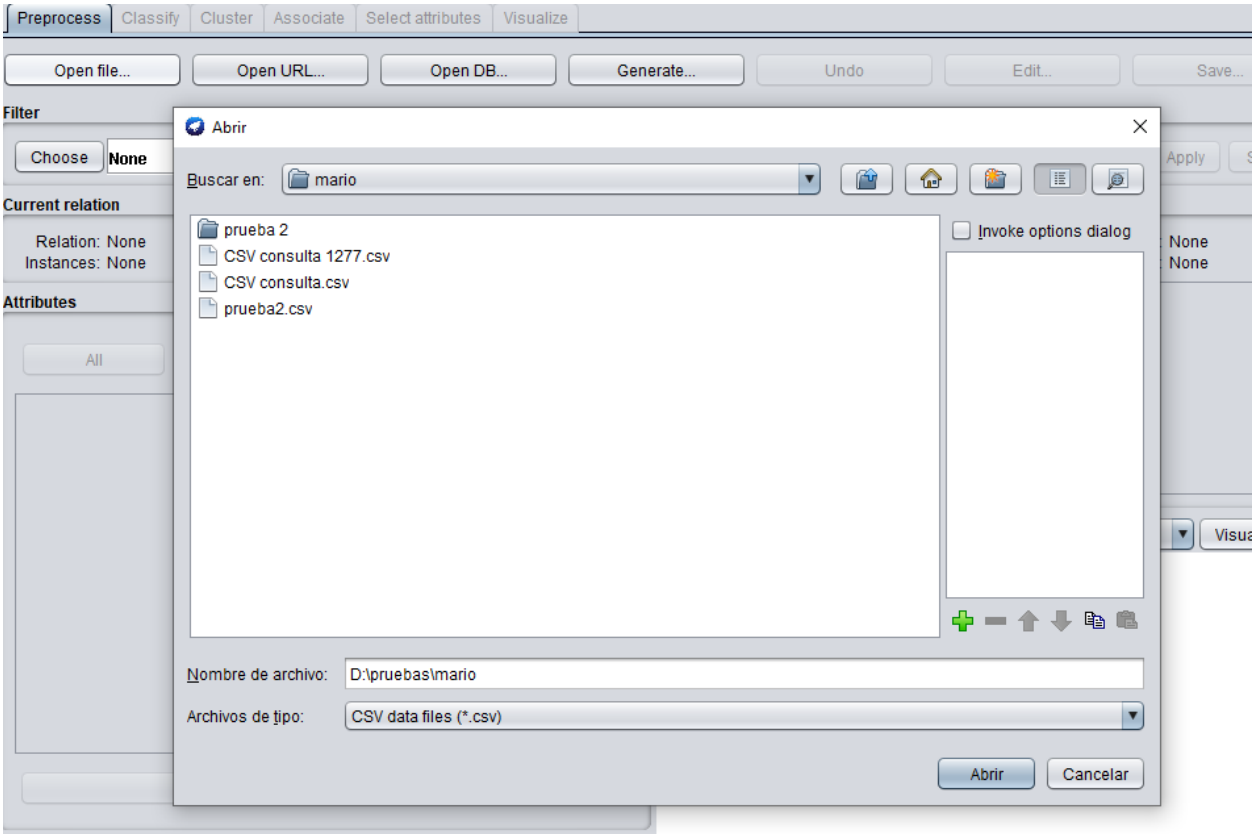
Una vez cargados estos datos los datos a la plataforma WEKA está la estructura de datos de la data set es transformada en el dataframes, la cual contiene la misma información, pero ya de manera más estructurada.

Importación De Los Datos

Para la importación de los datos en WEKA, se realizó desde la opción Explore, luego preprocesamiento, y mediante la opción open file, se seleccionó el tipo de archivo CSS, y se procedió al cargue del conjunto de datos, como se ilustra en la siguiente figura.

Figura 19

Cargue del set de datos.



Nota. La figura muestra el cargue del set de datos desde un archivo CSV.

Preprocesamiento De Datos

Esta etapa consiste en la preparación de los datos para su posterior utilización en la ejecución de los algoritmos de Machine learning. En esta etapa se aplicaron técnicas usuales de pre-procesado de datos. Básicamente estas actividades consistieron en la aplicación de filtros, limpieza de los datos, revisión de los datos en busca de información atípica para que sea subsanada y la selección de los mejores atributos para tratar con el problema de la alta dimensión, además se hace la revisión del balance del conjunto de datos lo cual como se verá más adelante es necesario tomarlo en cuenta en este tipo de ejercicios.

Es muy importante destacar la relevancia que este proceso de preparación de los datos tuvo en el desarrollo del proyecto toda vez que estas actividades impactan directamente en los resultados de la ejecución de los algoritmos de Machine learning, específicamente en este caso

que se están usando algoritmos de clasificación. Dicho de otro modo, estas actividades de preprocesamiento de los datos influyen directamente en la calidad y la confiabilidad de los resultados.

A continuación, se mencionan las actividades llevadas a cabo en este apartado, y corresponden a la integración de los datos, selección de los mejores atributos, la transformación de atributos, y la revisión del balance de los datos.

Limpieza, Transformación e Integración

La limpieza de los datos es un problema que influye de manera significativa en los resultados de minería de datos, esto tiene que ver con la calidad de los datos. Los errores en grandes bases de datos suelen ser más comunes de los que se espera, en ocasiones se encuentran datos incompletos o corruptos. Para este proyecto este problema se abordó tratando de hacer una comprobación minuciosa a través de los datos. Las propias técnicas de minería de datos ayudaron a resolver el problema. (Witten et al., 2016). En este orden de ideas, las labores de limpieza y transformación los datos se iniciaron con la carga de los datos a una base de datos relacional para poder cruzar la información de las diferentes fuentes y de esta manera proceder a su integración en un único conjunto de datos sonde se encuentran incluidos todos los a atributos. A partir de allí se procedió excluir los registros repetidos, con lo cual se redujo el conjunto de datos, y de esta manera quedaron excluidos los registros repetidos. En los datos originales no estaba la edad de los estudiantes, por lo que se procedió a realizar el cálculo de la edad a partir de la fecha de nacimiento. Una vez se integraron los datos a un único set de datos se exportaron los datos en formato CSV. Se procedió a cargarlos en el software WEKA, que es un software de Machine learning, una plataforma de código abierto que incluye una gran variedad de herramientas para estas tareas de limpieza y transformación de los datos.

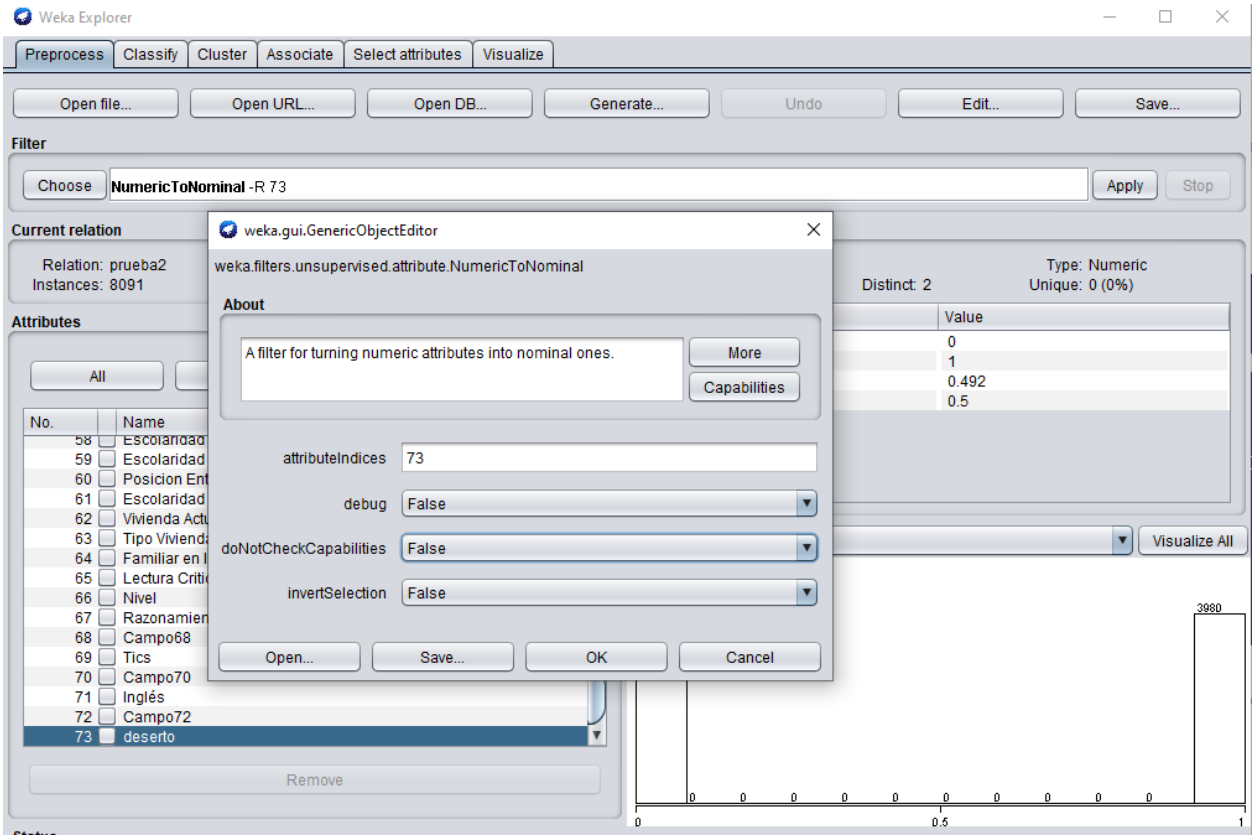
A continuación, se mencionan algunos de los problemas presentados durante el cargue de los datos a la plataforma WEKA:

Al ser exportados lo datos al formato CSV, Excel coloca el separador de campos por defecto con el símbolo “;” lo que suele convertirse en un problema toda vez que WEKA por defecto toma los datos con una estructura de atributos separados por coma (“,”) el inconveniente que se tuvo es que algunos campos que contiene texto en algunas preguntas abiertas del instrumento, en ocasiones suelen tener el signo de puntuación “,” por lo que se debe ser muy cuidadoso y antes de reemplazar la coma por el “;”, se debe tener especial cuidado de primero reemplazarlas comas (“,”) por algún otro signo de puntuación. Una buena forma de hacer estas operaciones que suelen ser tediosas, sobre todo cuando se está trabajando con set de datos muy grandes, es usar un buen editor de texto para ir corrigiendo las inconsistencias que se presentan a la hora de cargar los datos a la herramienta que se va a usar para ejecución de los algoritmos, para este caso se utilizó el editor de texto *Visual Studio Code.*

Una vez resueltos los inconvenientes que se presentaron se cargaron los datos al software WEKA, y se procedió con la revisión de los atributos mediante el uso de las herramientas proporcionadas por la plataforma. Un primer reto para llevar a cabo fue transformar la variable objetivo, (la clase a predecir) a una variable nominal, esto debido a que en los datos originales se utilizó para significar que el estudiante desertó con el número 1, y 0 para indicar que el estudiante no desertó. Durante el desarrollo de esta actividad se procedió a transformar la clase a predecir del set de datos de numérico a nominal, debido a que este valor fue detectado como numérico, para esto se aplicó el filtro NumericToNominal al atributo con Id 73, como se ilustra en la siguiente figura.

Figura 20

Aplicación del filtro NumericTo Nominal.

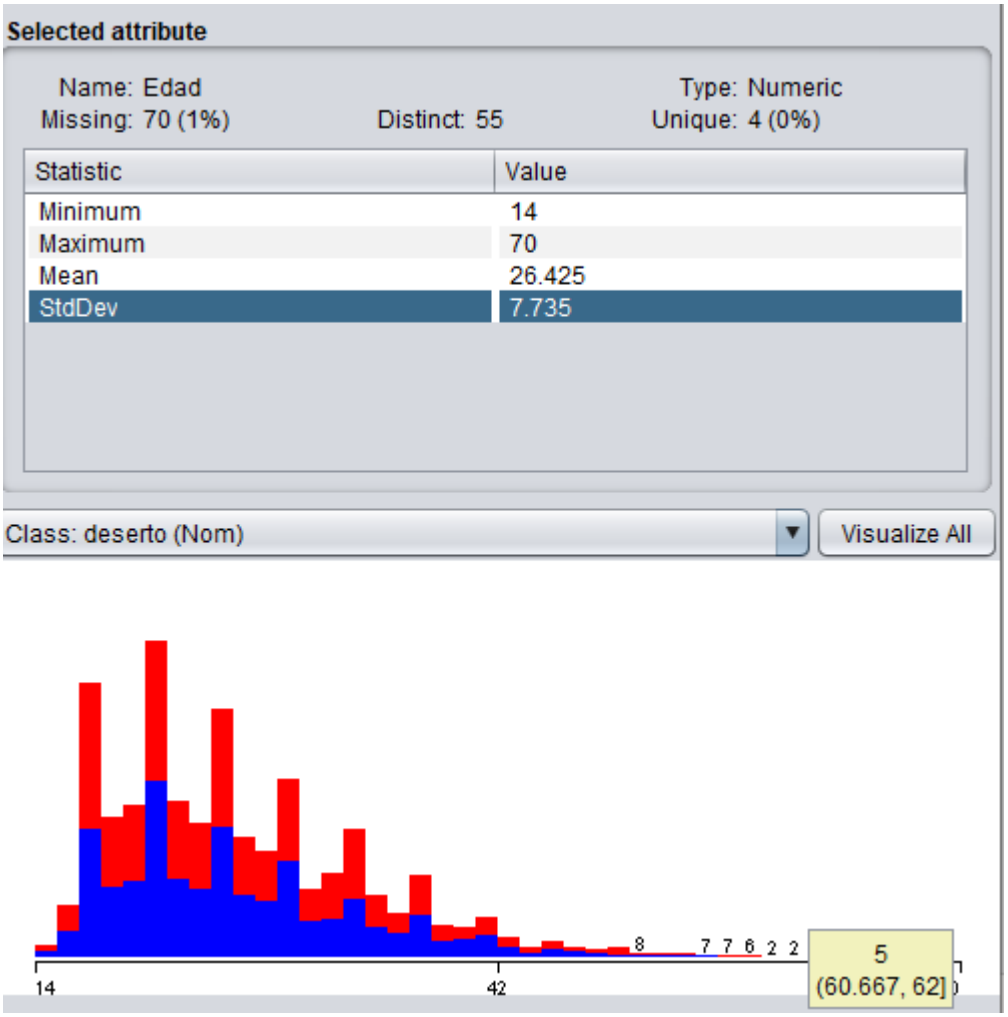


Nota. La imagen muestra la Aplicación del filtro NumericTo Nominal que cambia el tipo de dato de numérico a nominal.

Al hacer un recorrido sobre cada atributo del dataframe, la plataforma brinda información relevante de cada atributo y su comportamiento con respecto a la clase a predecir. En la siguiente figura se muestra el atributo edad, que es de tipo numérico, y la herramienta muestra información muy relevante, como que el 1% de los datos faltan en los registros, que hay 55 valores distintos en el total de datos, los valores de la media, la desviación estándar, el valor máximo y el mínimo, además permite ver gráficamente el comportamiento de este atributo con respecto a la variable objetivo o de salida.

Figura 21

Información del atributo edad



Nota. La gráfica muestra la información del atributo edad.

El siguiente es un resumen en el que se relacionan las principales acciones de limpieza y transformación de los datos realizados en el set de datos:

- Se eliminaron atributos con datos incompletos,
- Se eliminaron atributos irrelevantes como el correo institucional
- Se eliminaron los registros con atributos incompletos o no diligenciados
- 43 registros con fechas inválidas, de 3 años se completaron con el promedio

- 1registro con fecha de -65
- En los datos originales se transformó la fecha de nacimiento en edad
- Reemplazar las “,” por “-“o “.”
- Ajustar los nombres de los atributos toda vez que en los datos originales suministrados los títulos son bastante largos, por lo que fue necesario acortarlos con nemotécnicos.
- Edades de 104 años o con formato de fecha de nacimiento incorrecto
- Se procede a la revisión del atributo edad, donde se evidencian 43 registros con datos inválidos, a estos datos se decidió reemplazarlos por el promedio. Se completaron los datos de edades inválidos colocándoles el promedio de edades de la muestra.
- Ajustar algunas comillas, y datos con algunos caracteres que provocan error en la carga de datos o que el proceso de carga se interrumpa.

Selección De Los Atributos Mejor Rankeados

Se revisó el comportamiento de los atributos con respecto a la clase objetivo, con la finalidad de terminar cuales son los atributos que tienen una mayor incidencia en la variable de salida o variable objetivo. Con esto se intentó paliar el problema de la alta dimensionalidad en el conjunto de datos debido al gran número de atributos presentes en el set de datos. Lo que se intentó fue reducir el número de atributos sin perder la confiabilidad del modelo a la vez que se mejora la eficiencia, teniendo en cuenta que se tiene que interactuar con un menor número de atributos.(Márquez, 2015)

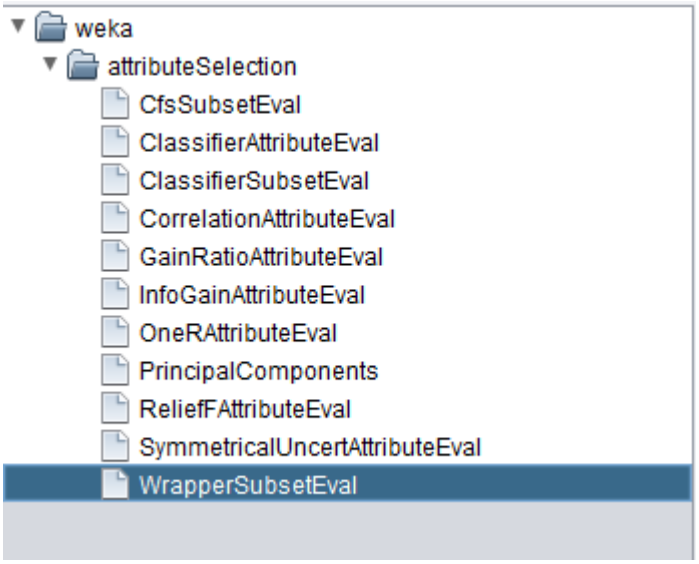
Se inicia por identificar atributos que son irrelevantes, estos atributos que se eliminaron en primera instancia fueron.: el código, el nombre, el correo institucional, el grupo en el curso,

acudiente, el estrato se elimina porque la columna no tiene datos, esta acción también contribuyó a la reducción de la alta dimensionalidad del conjunto de datos. además, se consultó con expertos (decana zonal, consejeros) sobre qué atributos podrían ser incidentes para la deserción estudiantil en la UNAD.

Luego de esto se procedió a la utilización de los algoritmos que posee la plataforma WEKA para determinar los mejores atributos. El software WEKA dispone de distintos algoritmos de selección de atributos, entre los que se eligieron ocho que se listan a continuación.

figura 22

Algoritmos para seleccionar los mejores atributos



Nota. La figura muestra los algoritmos que se pueden usar para seleccionar los mejores atributos.

A continuación, se muestran los resultados obtenidos con los datos preprocesados en donde se observan los atributos mejor rankeados a partir de la ejecución de los algoritmos mostrados en la imagen anterior.

A partir del a ejecución de 7 de estos algoritmos se procedió a sacar la lista top-ten de los 10 mejores atributos de estos 7 algoritmos.

En la siguiente tabla se muestra la clasificación realizada por cada uno de los algoritmos en el set de datos. El número indica la posición del atributo en el set de datos, lo que facilita su ubicación. Esta tarea pudo resultar bastante tediosa debido a la gran cantidad de atributos con los que se lidiaron en el conjunto de datos.

Tabla 4

Los 10 mejores atributos seleccionados por diferentes algoritmos.

Ranker + ClassifierAttributeEval	**Ranker + CorrelationAttributeEval**	**Ranker + InfoGainAttributeEval**	**Ranker+GainRatioAttributeEvalcross**	**Ranker+ ReliefFAttributeEval**	**Ranker+SymmetricalUncertAttributeEval**	**Ranker+OneRAttributeEval**	**Moda**
66 nivelIngl	6 GÃ©nero	8 Ciudad Residencia	13 Disc Cognitiva	6 GÃ©nero	8 Ciudad Residencia	6 GÃ©nero	6 GÃ©nero
17 Enfermedad	39 Aprobado Cursos Virtuales	2 Cead	14 Disc Fisica	3 Zona	2 Cead	29 Grados Perdidos	2 Cead
23 GÃ©nero Danza	29 Grados Perdidos	1 Programa	8 Ciudad Residencia	56 Vivienda Actual	29 Grados Perdidos	39 Aprobado Cursos Virtuales	29 Grados Perdidos
22 Actividad Artistica	38 Tomado Cursos Virtuales	3 Zona	7 Etnia	44 RazÃ³n Ingreso al Programa	6 GÃ©nero	35 GraduaciÃ³n Estudios Cursados	#N/D

21 Instrumento	28 Rendimiento Academico	29 Grados Perdidos	16 Disc Mental	41 RazÃ³n Sin Estudiar	7 Etnia	33 Ãšltimo Nivel Alcanzado	#N/D
20 Deporte	35 GraduaciÃ³n Estudios Cursados	6 GÃ©nero	12 Disc Auditiva	2 Cead	21 Instrumento	1 Programa	#N/D
19 Conocio UNAD	34 Modalidad Estudios	39 Aprobado Cursos Virtuales	29 Grados Perdidos	9 Area Residencia	39 Aprobado Cursos Virtuales	34 Modalidad Estudios	34 Modalidad Estudios
24 Edad	33 Ãšltimo Nivel Alcanzado	21 Instrumento	26 Modalidad	10 Desplazado	38 Tomado Cursos Virtuales	28 Rendimiento Academico	#N/D
25 Tipo InstituciÃ³n	26 Modalidad	33 Ãšltimo Nivel Alcanzado	21 Instrumento	25 Tipo InstituciÃ³n	3 Zona	38 Tomado Cursos Virtuales	25 Tipo InstituciÃ³n
26 Modalidad	63 Tics	35 GraduaciÃ³n Estudios Cursados	6 GÃ©nero	4 Escuela	1 Programa	3 Zona	#N/D

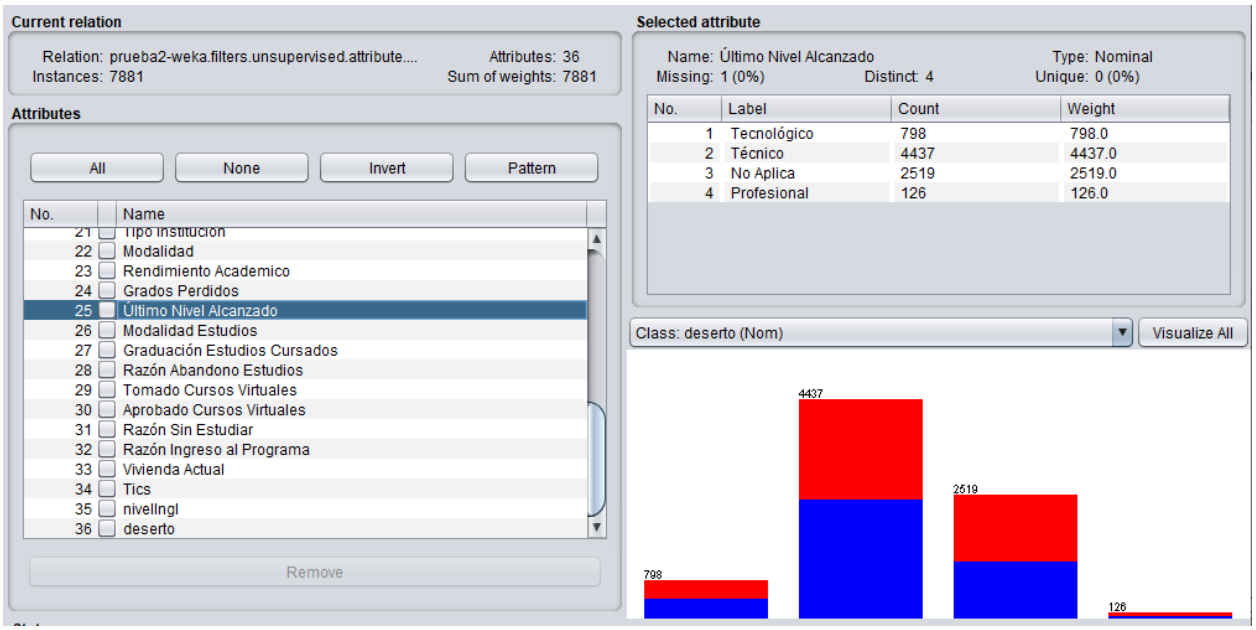
Nota. Esta tabla muestra los 10 mejores atributos seleccionados por diferentes algoritmos.

A partir de la tabla anterior se procedió a tomar los mejores 10 atributos de cada algoritmo de ranqueo y a partir de allí armar una nueva lista de atributos, la cual se conformó con

36 atributos, de esta manera se solucionó el problema de la alta dimensionalidad presentado por el conjunto de datos, en el cual se observaron en primera instancia 77 atributos.

Figura 23

Los 36 mejores atributos



Nota. La gráfica muestra los 36 mejores atributos seleccionados del total del set de datos.

Se escogieron los primeros 30 atributos que fueron seleccionados como mejores, como resultado de la ejecución de 7 algoritmos de ranqueo. Con esto se logró reducir la alta dimensión del conjunto de datos, pasando de 69 atributos a 30 mejores atributos.

Balance Del Conjunto De Datos

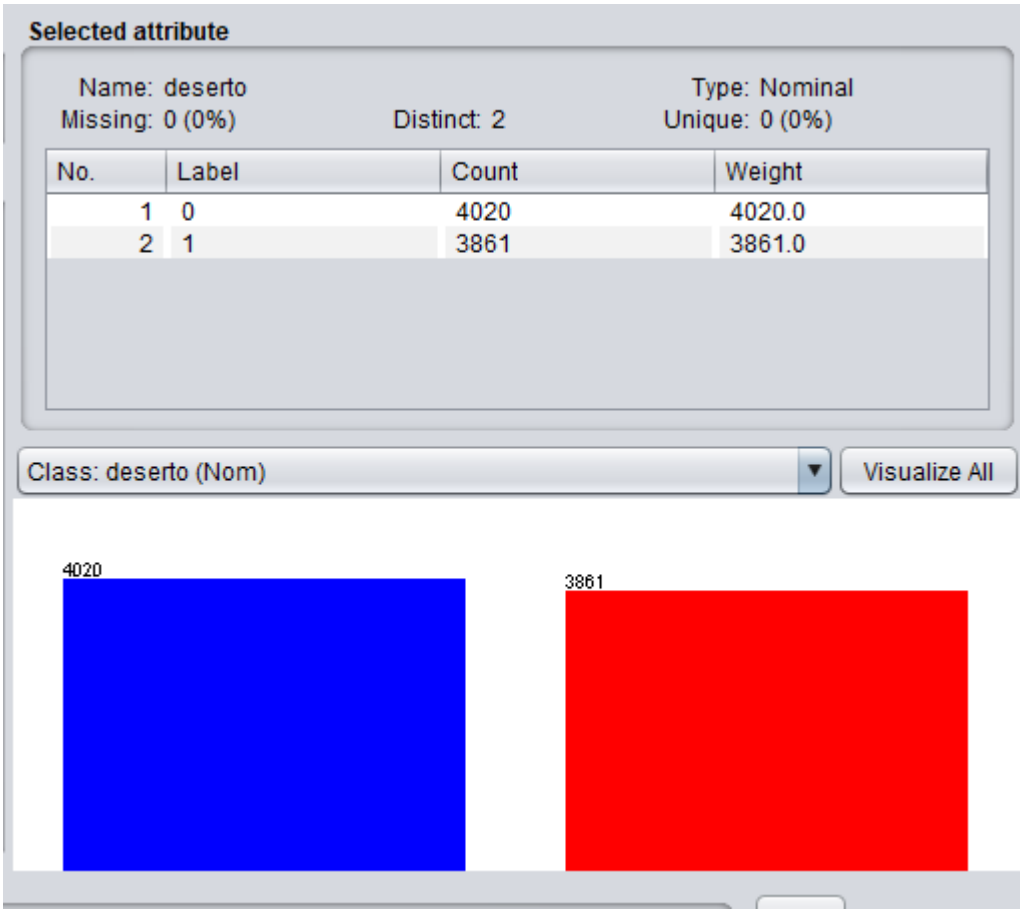
El balance de los datos tiene que ver con que en algunos ejercicios de predicción con Machine learning existe mucha diferencia entre el número de registros que corresponden a un valor de la variable y otro en el conjunto de datos que se viene trabajando esos valores son desertó o no desertó, por ejemplo, en el caso que se tuviera un set de datos de pagos con tarjetas de crédito fraudulentos y pagos legal o verdadero, esto dejaría muy pocos pagos fraudulentos en

comparación con los legales, esta situación hace que los algoritmos privilegien la clase mayoritaria y afecten negativamente la exactitud o ACCURACY del modelo.(Witten et al., 2016)

En la siguiente imagen se puede apreciar que el conjunto de datos con el que se trabaja en este estudio no presenta un desbalance considerable o un gran desequilibrio con respecto a la clase a predecir, por lo tanto, se consideró trabajar con los datos sin balacear y revisar los resultados. Esto teniendo en cuenta que la diferencia es de 1 punto porcentuales dado que son 7881 registros, lo que deja a la clase desertores con un 49%, mientras que los no desertores quedan con 51%. Esta diferencia es relativamente baja por lo que no se considera la aplicación de técnicas de balanceo y se procede a la realización de los experimentos sin aplicar técnicas de balanceo de las clases.

Figura 24

Distribución de la clase a predecir en el set de datos



Nota. La figura muestra la distribución equitativa de la clase a predecir en el set de datos

Modelo De Predicción

En esta etapa, ya con el conjunto de datos preparados en la fase anterior se procede a cargar el conjunto de datos a la plataforma WEKA desde donde se ejecutan los algoritmos de Machine Learning.

La revisión del conjunto de datos permitió identificar que se está frente a un problema de aprendizaje supervisado, toda vez que los atributos del conjunto de datos están identificados y se tiene una clase o atributo conocido que va a ser objeto de aprendizaje en un modelo, y a partir de allí hacer predicción en los experimentos que se ejecutarán a continuación. (Witten et al., 2016)

Basados en estas consideraciones iniciales se procede a la realización de experimentos de predicción de la deserción estudiantil en la UNAD, con base en el conjunto de datos proveniente de la encuesta y prueba de caracterización de los estudiantes realizada en febrero de 2018 y que corresponde al periodo 16-01 de ese año., como ya se había mencionado en apartados anteriores.

Cabe anotar que existe un variado número de paradigmas que se utilizan para abordar problemas de clasificación en aprendizaje supervisado, como el que se afronta en este estudio. Algunos de estos paradigmas son los árboles de decisión y las reglas de inducción, los algoritmos de Bayes, las funciones. Se adoptan estas técnicas de clasificación debido a que son técnicas de caja blanca en el sentido en que aportan una explicación de los resultados de clasificación y podrían utilizarse como soporte en la toma de decisiones por usuarios que no sean expertos o con poco conocimiento en analítica de datos.

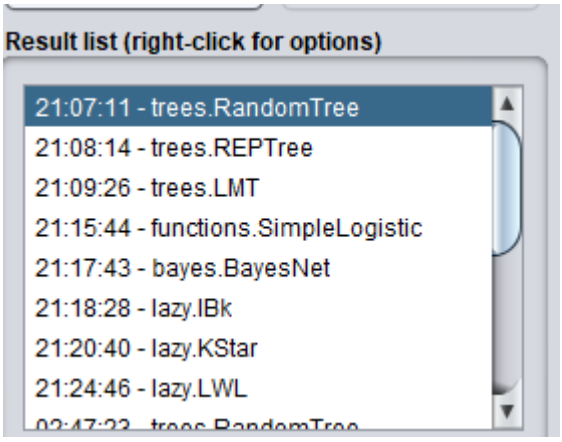
Para esto se proponen algoritmos de clasificación basados principalmente en arboles de decisión y en reglas de clasificación comparando sus resultados para descubrir los mejores modelos.

Ejecución Algoritmos de Machine Learning

Se ejecutaron diversos algoritmos disponibles en la plataforma de Machine learning WEKA. De estos algoritmos disponibles se experimentó con los agrupados en los siguientes grupos: Bayes, Functions, Lazy, Arboles de decisión (Trees). En la siguiente figura se aprecian algunos de los algoritmos que se utilizaron en los experimentos.

Figura 25

Algunos de los algoritmos experimentados



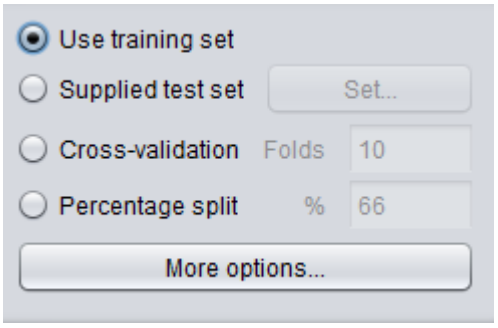
Nota. La imagen muestra algunos de los algoritmos experimentados.

No fue posible para este ejercicio, experimentar en esta fase con algoritmos de redes neuronales, debido a que el conjunto de datos es bastante grande (7881 registros) y los costos de procesamiento y tiempo de máquina lo hicieron inviable, sin embargo, se aclara que para pruebas se ejecutó con un conjunto de datos reducido, sin que los resultados superaran lo obtenidos por otros algoritmos que se van a relacionar más adelante.

En la plataforma WEKA es posible utilizar las opciones de test que se aprecian en la siguiente figura.

Figura 26

Opciones de test en WEKA.

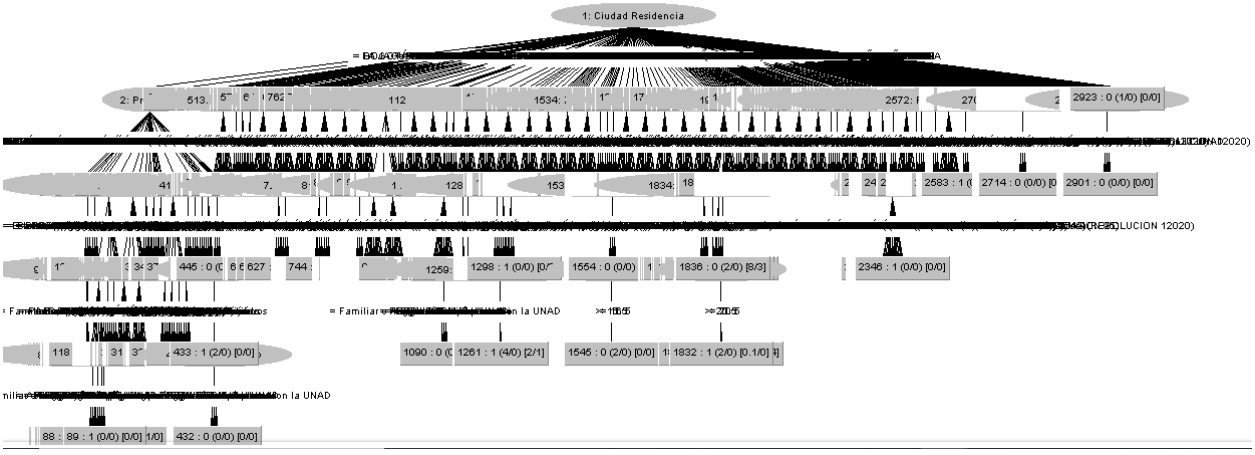


Nota. La figura muestra las diferentes opciones de test en WEKA.

Los experimentos se hicieron con las opciones de training test y Cross Validation o validación cruzada. La opción de Training test, consiste en usar todo el conjunto de datos para la construcción del modelo, la ejecución de este modelo implica que se debe proporcionar un con junto de pruebas (supplied test set) para probar la precisión del modelo. Los resultados que se presentan son producto de aplicar validación cruzada (cross-validation) con particiones en el set de datos de 10 (folds) y percentage split del 70%. Con esto los algoritmos toman el 70% para construir el modelo de aprendizaje y el 30% de datos para probar el modelo.

figura 27

Vista del árbol de decisión.



Nota. La grafica muestra l representación de un árbol de decisión.

Resultados De La Ejecución De Algoritmos De Clasificación Supervisados

A continuación, se relacionan los resultados de la ejecución de los algoritmos de clasificación supervisada con los que se obtuvieron los mejores indicadores de precisión.

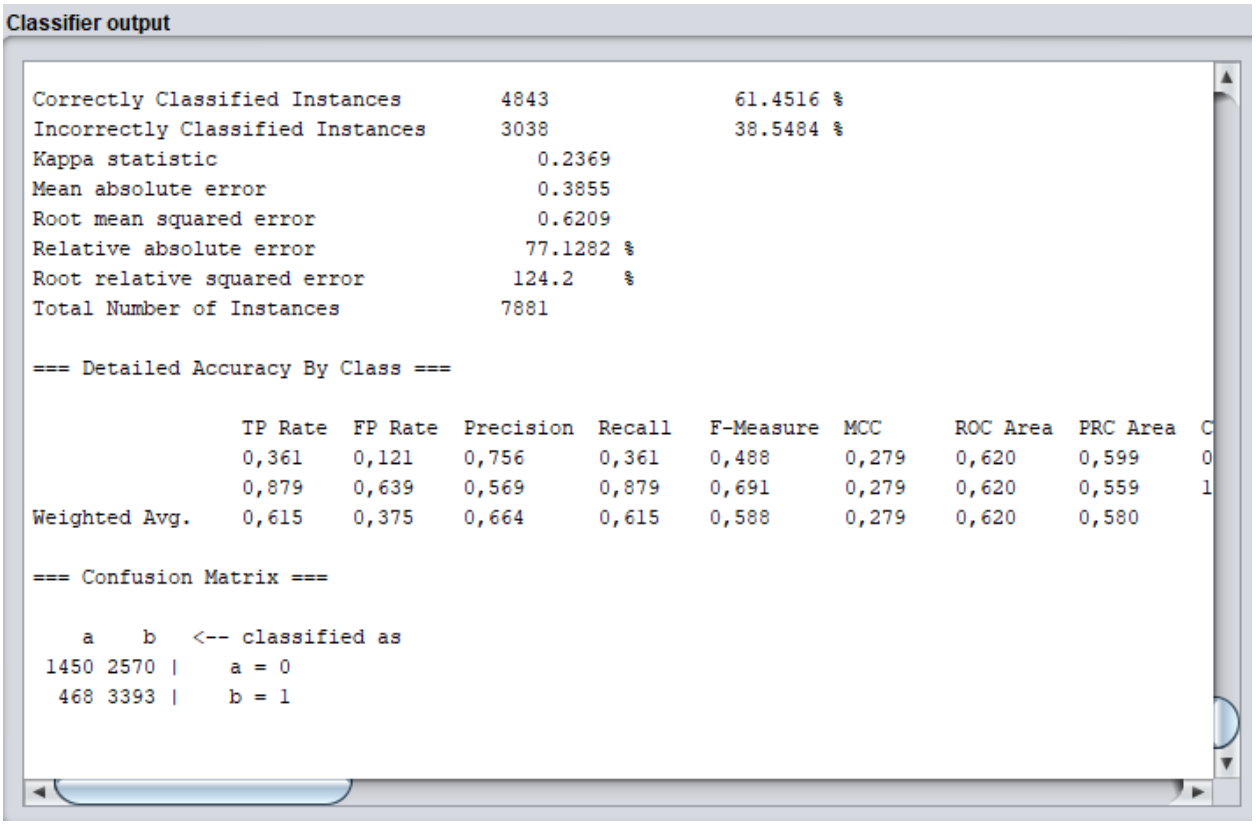
Algoritmo Función SGD

Este algoritmo que implementa Stochastic Gradient Descent en varios modelos lineales. El cual es útil para transformar atributos nominales en binarios y el balanceo del conjunto de datos, con la finalidad que los coeficientes de salida se basen en datos normalizados.

Con este modelo se logró una precisión de 0.66, clasificando correctamente el 61.45% de los datos de manera correcta, como se puede ver a continuación:

Figura 28

Ejecución del algoritmo SGD



Nota. La grafica muestra la salida de la ejecución del algoritmo SGD

Algoritmo NiveBayes

Este algoritmo que se basa en la regla del teorema de Bayes. funciona muy eficazmente cuando se prueba en conjuntos de datos reales, particularmente cuando es combinado con algunos de los procedimientos de selección de atributos, que elimina atributos redundantes (Witten et al., 2016).

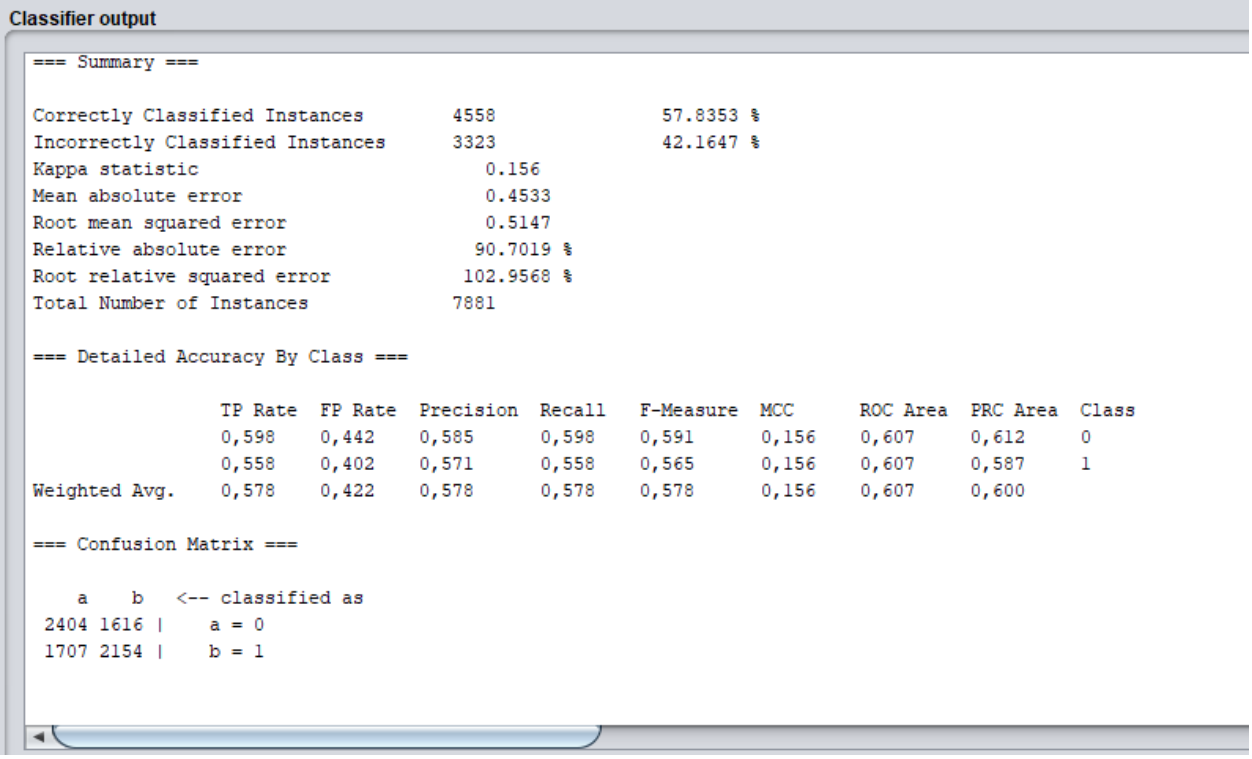
Con este algoritmo se obtuvieron indicadores de 57.8%, lo que quiere decir que de cada 100 instancias o registros suministrado al modelo de clasificación 58 se clasificaron de manera correcta.

La precisión del modelo fue de 0.578 y un Recall igualmente de 0,578.

A continuación, se aprecia la salida de este modelo de aprendizaje supervisado:

Figura 29

Salida del algoritmo Naybe Valles



Nota. La grafica muestra salida de la ejecución del algoritmo Naybe Valles.

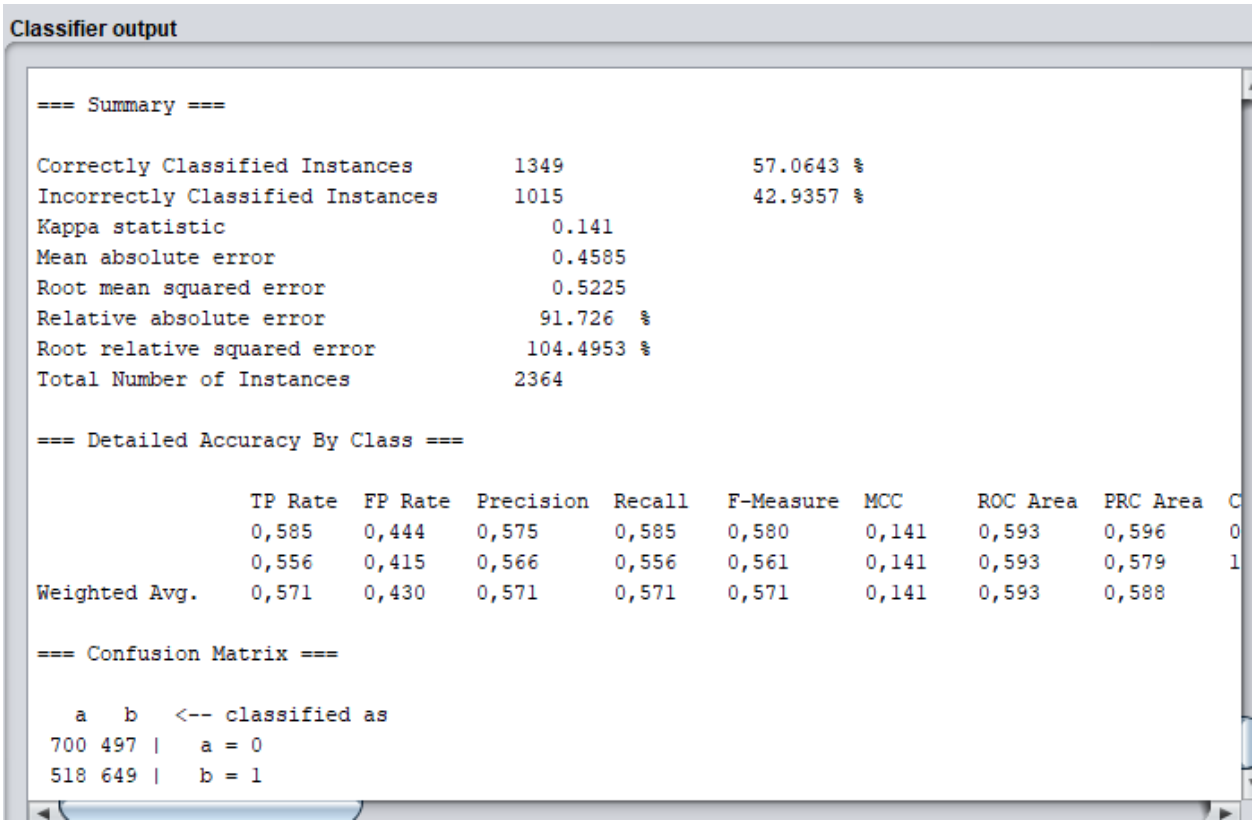
NaiveBayesUpdatable

Esta es la versión actualizable del algoritmo NaiveBayes. Este clasificador usa una precisión predeterminada de 0.1 para atributos numéricos cuando se llame a buildClassifier con cero instancias de entrenamiento.

Con este algoritmo se obtuvo un modelo de clasificación de aprendizaje supervisado, en el que se observa una precisión de 0,571, con igual indicativo en la medida de recall y una tasa de predicciones correctas del 57%.

Figura 30

Ejecución del algoritmo NaviBayesUpdatable



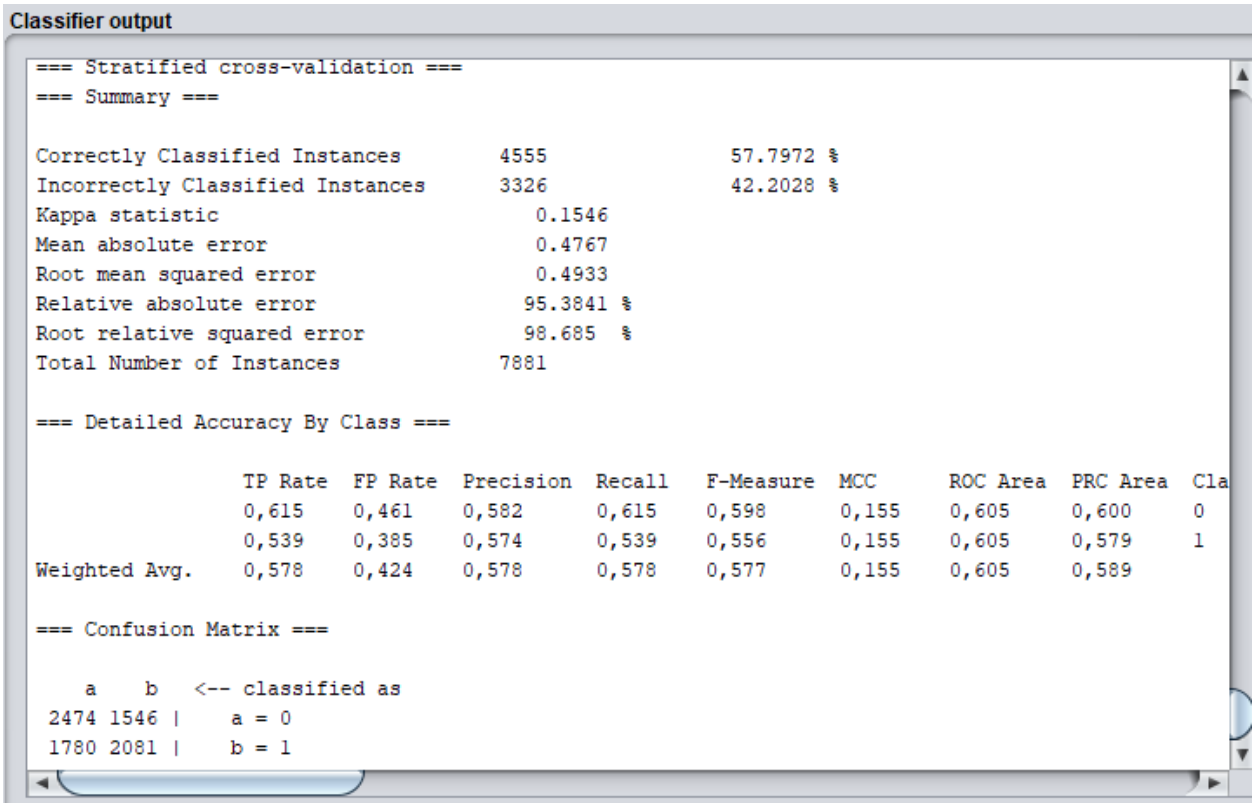
Nota. La grafica muestra la salida de la ejecución del algoritmo NaviBayesUpdatable

SimpleLogistic

Este es un algoritmo de clasificación supervisado en el que obtuvo una precisión del 57.8 %. SimpleLogistic crea modelos de regresión logística, ajustándolos con funciones de regresión simples como aprendices básicos y determinar cuántas iteraciones realizar mediante validación cruzada, que admite selección automática de atributos (Witten et al., 2016)

Figura 31

Salida del algoritmo SimpleLogistic



Nota. El grafico muestra la salida del algoritmo SimpleLogistic.

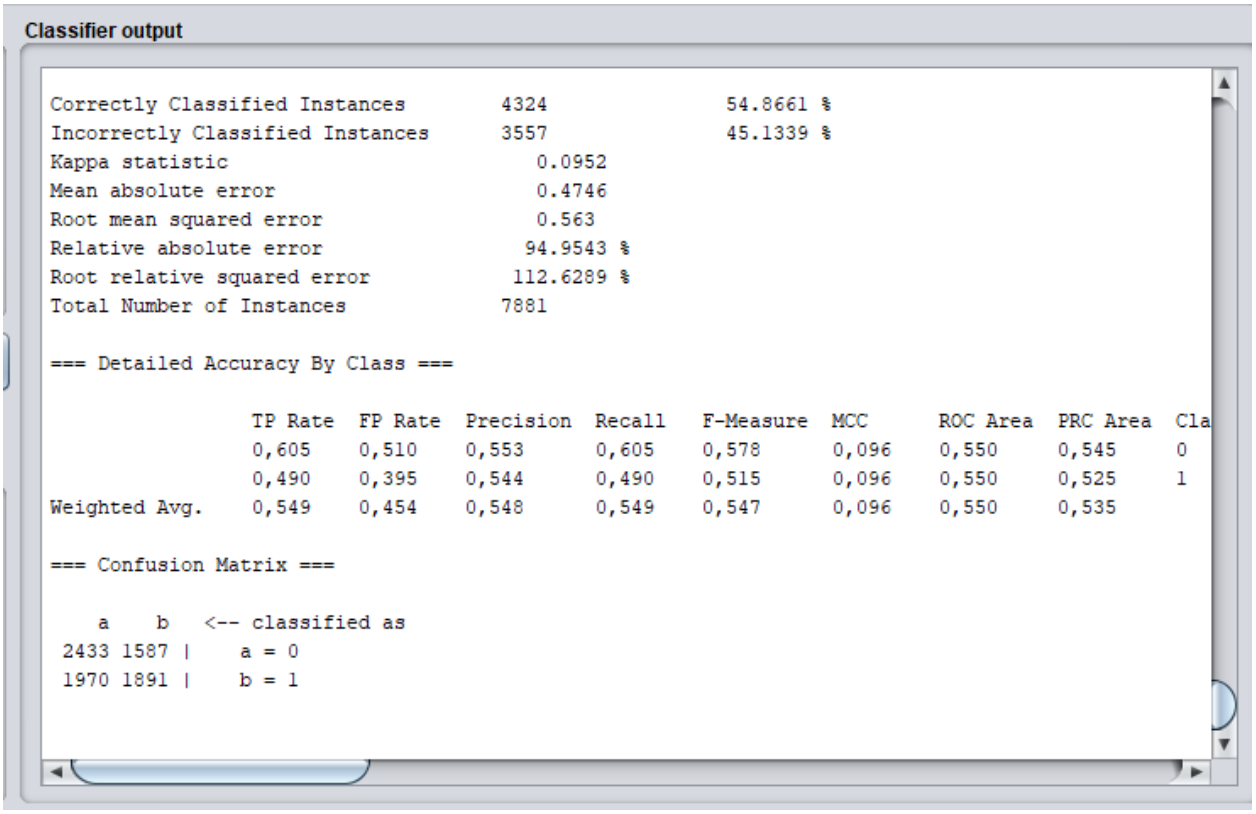
Algoritmo J48

Este es un algoritmo de Árbol de decisión, es la versión de código abierto del algoritmo C4.5 J48 realmente implementa una versión posterior y ligeramente mejorada llamado C4.5 revisión 8, que fue la última versión pública de esta familia de algoritmos antes de la implementación comercial. (Witten et al., 2016)

Este algoritmo goza de cierta reputación en predicciones de clasificación supervisada, y para este ejercicio arrojó unos guarismos de 54.8% de predicciones correctas, una precisión de 0,548 y un Recall de 0,549.

Figura 32

Salida del Algoritmo J48



Nota. El gráfico muestra la salida del Algoritmo J48.

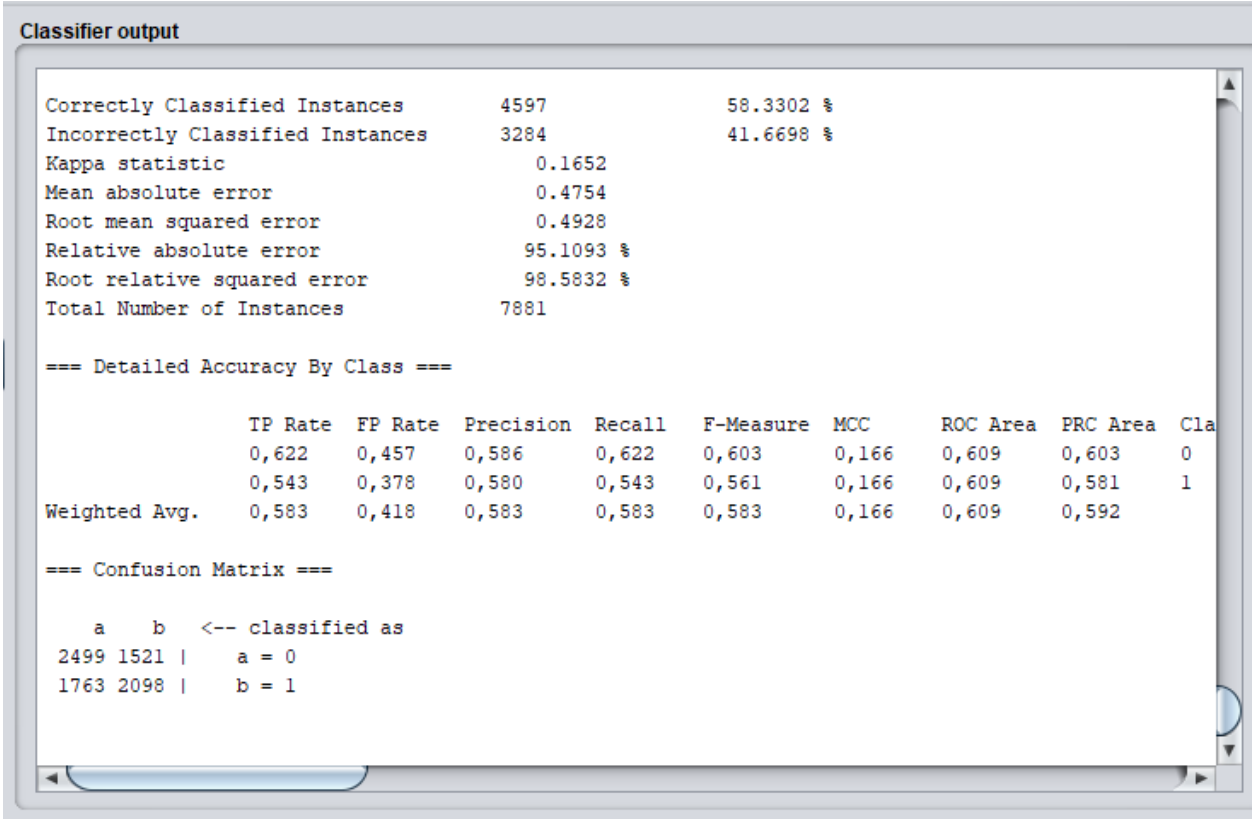
Este algoritmo arrojó altos indicadores en las medidas de errores relativos y absolutos

Tree LMT

Este es un algoritmo del grupo de arboles de decisión con el que se obtuvieron esultados bastante cercanos a los mas altos obtenidos durante los experimentos. De esta manera se obtuvieron tasas de aciertos en la predicción del 58.3%, con medidas de precisión y recall del 0,583 respectivamente como se puede apreciar en la siguiente figura.

Figura 33

Salida del algoritmo Tree LMT.



Nota. En la imagen se muestra la Salida del algoritmo Tree LMT.

Comparación de los Algoritmos de predicción

Tabla 5
Comparación de modelos Supervisados

Algoritmo / Medida	Acuraccy	Recall	Precisión
Función SGDC	61,45%	0,61	0,66
NiveBayes	57,83%	0,57	0,57
NiveBayesUpdatable	57,06%	0,57	0,57
SimpleLogistic	57,79%	0,57	0,57
J48	54,86%	0,54	0,54
Tree LMT	58,33%	0,58	0,58

Nota. Esta tabla muestra la Comparación de modelos Supervisados

Evaluación De La Efectividad De Los Modelos

Pruebas Del Modelo Con Datos De Periodo 2018 16-02

Se configuró un nuevo set de datos de las mismas características usado en la fase anterior, a pesar de que la ejecución de los algoritmos en la fase anterior del estudio se elaboró con la opción de validación cruzada, con 10 particiones (folds) del conjunto de datos y un porcentaje para training del 30%. Se quería experimentar con datos de un periodo posterior, en este caso 2018- 16-02. Haciendo remembranza, en la fase anterior se utilizó el conjunto de datos correspondiente al periodo 2018- 16-01, el cual es un conjunto de datos que consta de 7881 registros de estudiantes nuevos de ese periodo. Lo que se presenta a continuación son unas pruebas de esos mismos modelos en los que la maquina aprendió a predecir la deserción, aplicados a otro periodo, del que también se conocen los resultados actuales en términos de abandono de los estudios en la universidad.

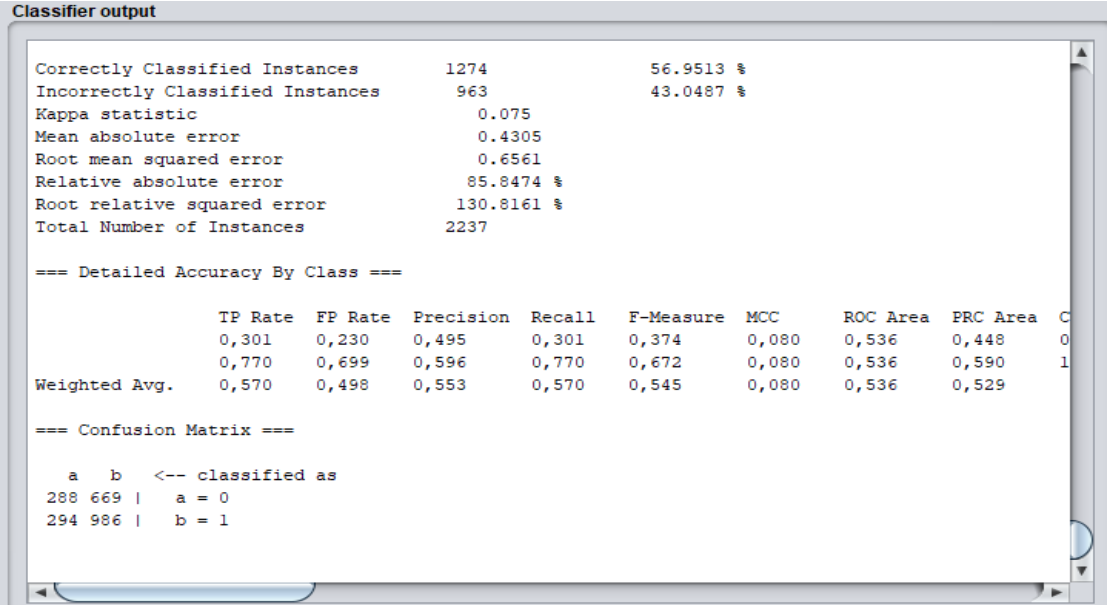
Los siguientes son los resultados de la ejecución de esos modelos concebidos en la fase anterior, ejecutados con un set de pruebas de otro periodo (2018 16-02)

Prueba en SDG

Con este modelo se logró una precisión de 0.55, clasificando correctamente el 56.95% de los datos de manera correcta, como se puede ver a continuación:

Figura 34

prueba SGD



Nota. La figura muestra la salida de la prueba con el algoritmo SGD.

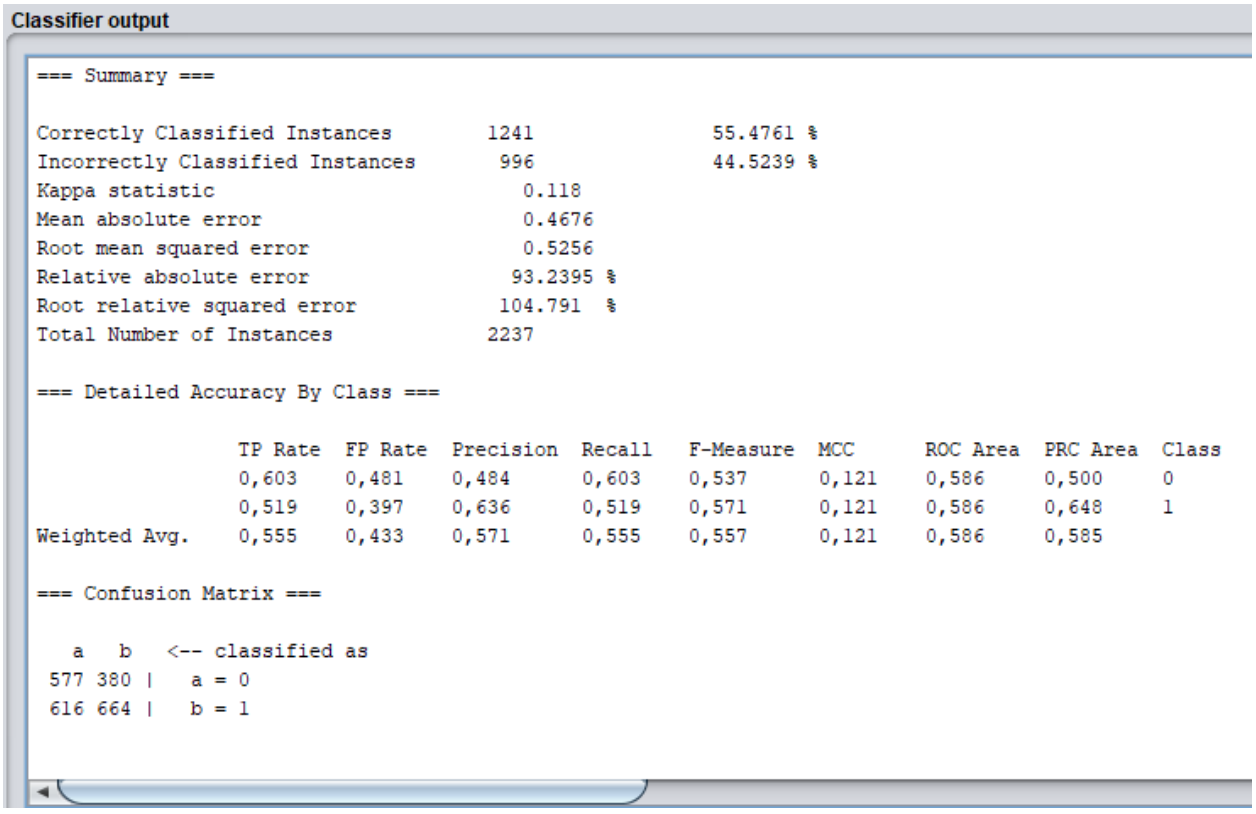
NiveBayes

La prueba del modelo NiveBayes obtenido en la fase anterior con el set de datos 2, correspondiente al periodo 16-02 del 2018, arrojó los siguientes resultados que se observan en la siguiente figura:

Accuracy de 55,47, una precisión de 0,57 y Recall de 0,55. También se puede apreciar la matriz de confusión desplegada por el modelo.

Figura 35

Salida del modelo NaiveBalles.



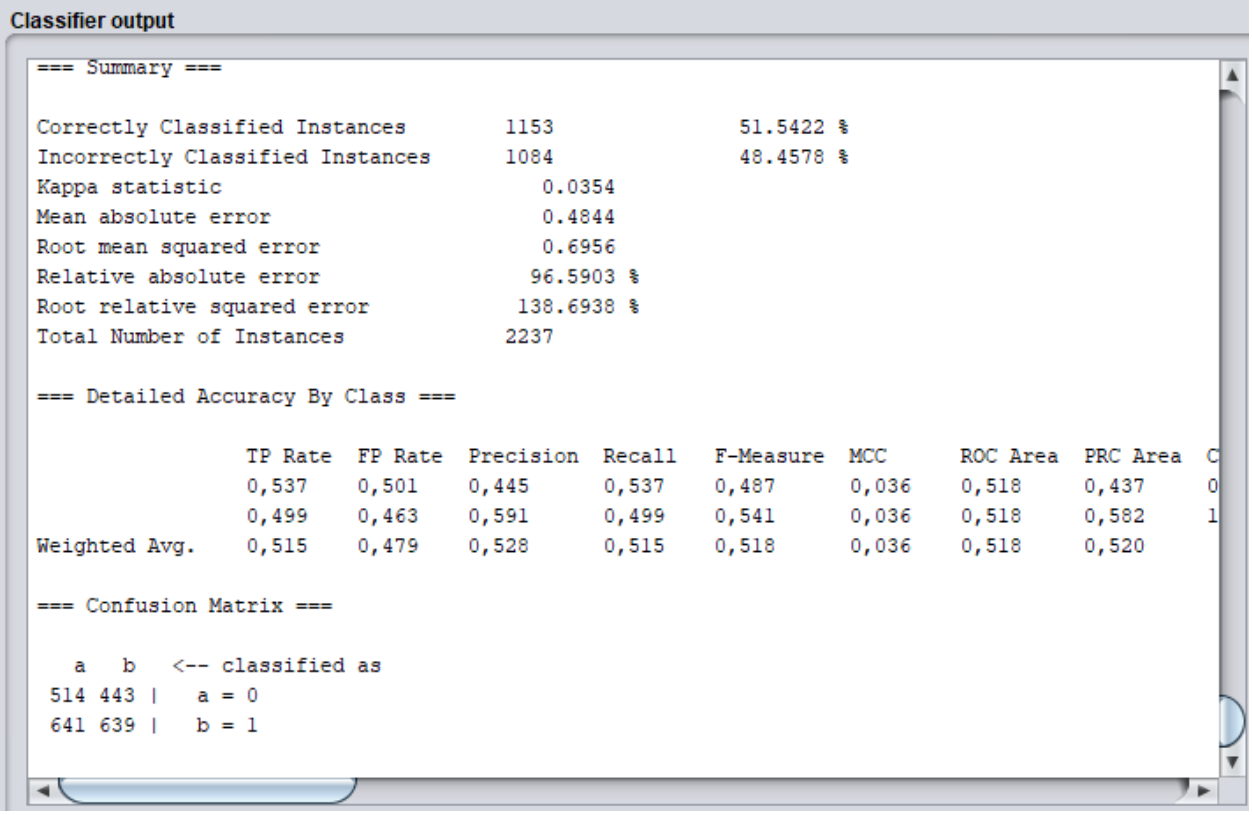
Nota. La figura muestra la Salida del modelo NaiveBalles, con el set de datos 2

Prueba con IBk con datos de 2018 16-02

Este modelo de IBK no se incluyó en los modelos anteriores. El modelo arroja una precisión del 52% y un recall de 51%.

Figura 36

Salida del modelo IBK sobre el set de datos 2.



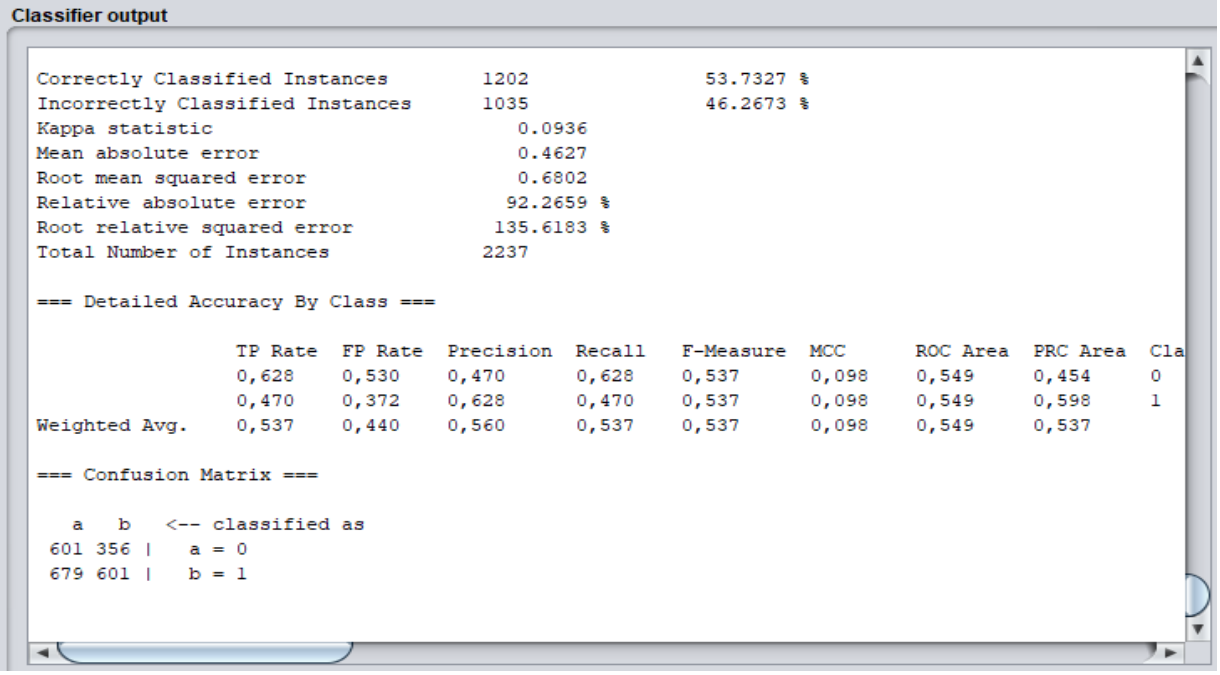
Nota. La Figura muestra la Salida del modelo IBK sobre el set de datos 2

SimpleLogistic

Con la ejecución de este modelo se obtuvieron las siguientes medidas de confiabilidad, los cuales se pueden apreciar en la siguiente figura. Se obtuvo un Accuracy de 53,7 %, así mismo una precisión de 0,56 y Recall de 0,53. Entre las medidas más relevantes, otras medidas se pueden apreciar en la siguiente figura.

Figura 37

Prueba del modelo Simple Logistic, set de datos 2



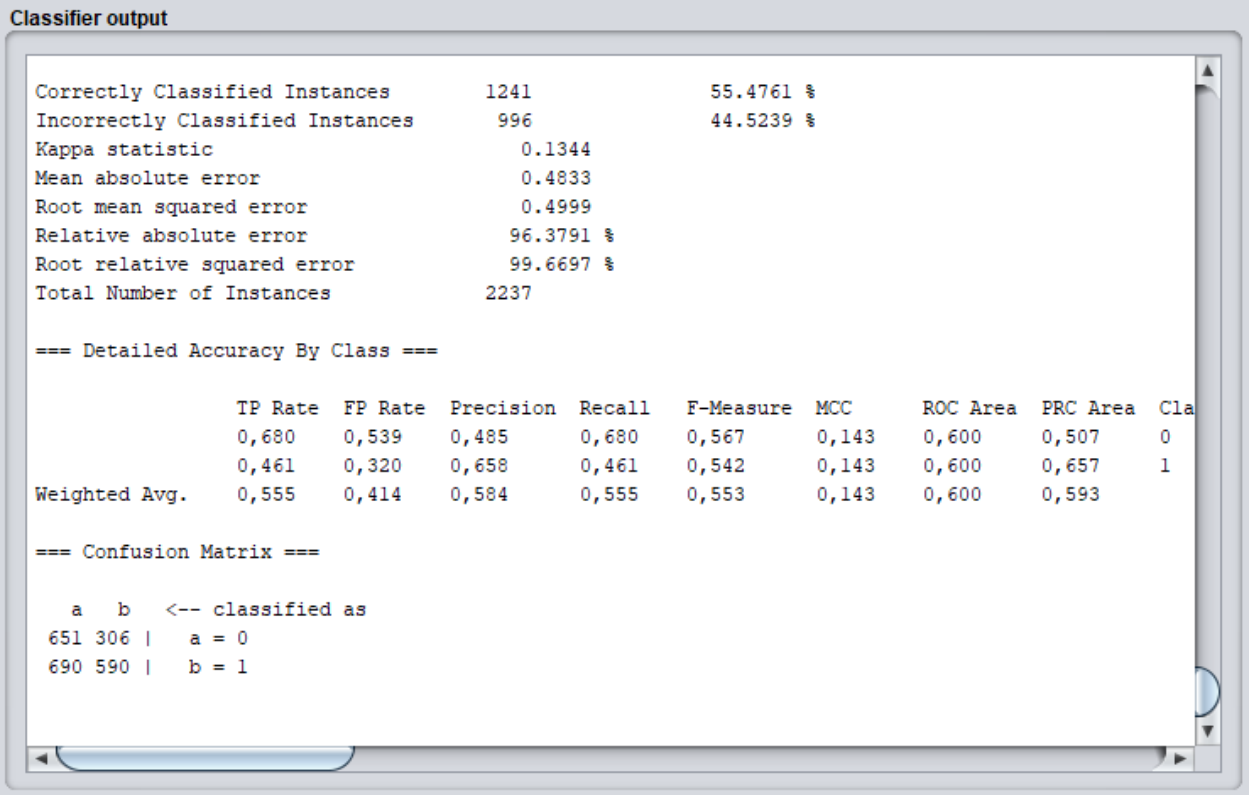
Nota. En la gráfica se muestra la salida del algoritmo del modelo SimpleLogistic ejecutado sobre el set de datos 2.

Tree LMT

La ejecución del modelo obtenido del algoritmo Tree LMT, ejecutado desde la opción *Suplied test set* de WEKA, arrojando resultados de Accuracy de 55,47%, una precisión de 0,584 y Recall de 0,555.

Figura 38

Salida del modelo Tree LMT set de datos 16-02



Nota. El gráfico muestra la salida de la ejecución del algoritmo Tree LMT.

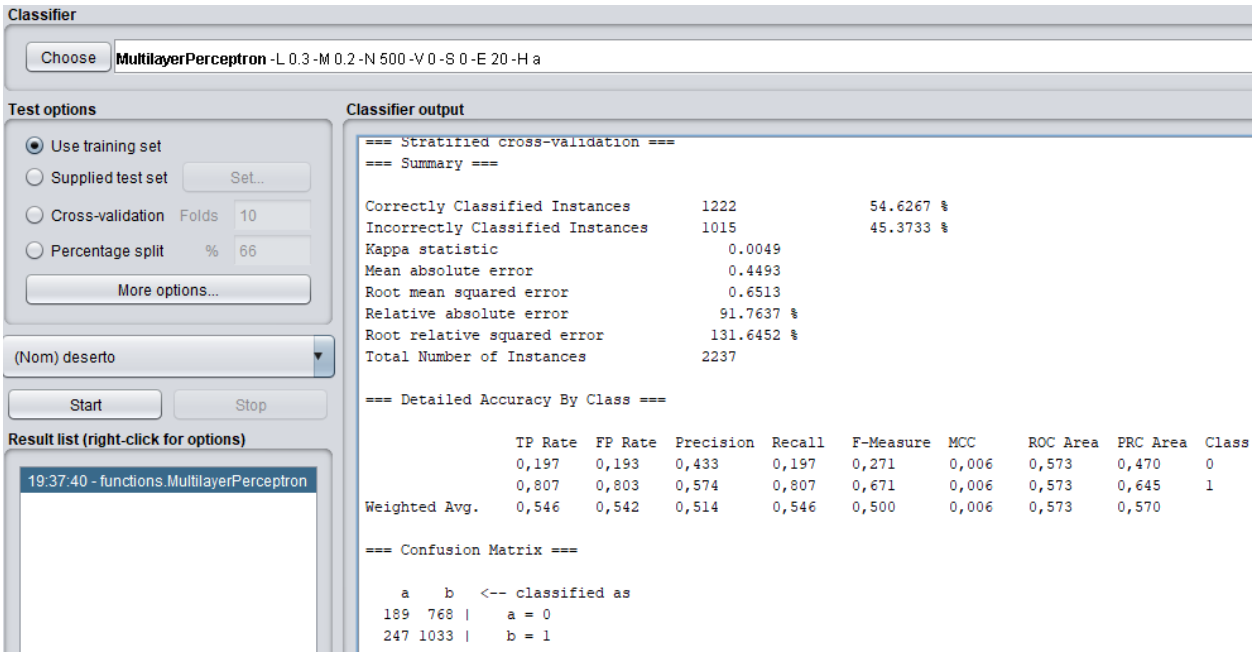
Pruebas Con Una Red Neuronal

Se intento realizar experimentos ejecutando el algoritmo MultiLayerPerceptron que es la red neuronal disponible en WEKA, sin embargo con el conjunto de datos de del periodo 16-02, que es un conjunto de 2237 registros, utilizando una sola capa oculta en la red neuronal, el algoritmo duró en ejecución 26 horas, por lo que el costo computacional de este algoritmo de red neuronal lo hace inviable para este tipo de experimentos en los que se cuenta con una potencia de cálculo limitada, no obstante la maquina donde se ejecutó cuenta con 8 GB de RAM y un procesador raizen 3. Sin embargo, las interacciones del modelo consumen mucho tiempo de máquina.

Como se puede apreciar en la siguiente figura los resultados fueron muy parecidos a los obtenidos en los modelos anteriores en términos de precisión de 0,514 y Recall de 0,546. Y aciertos en clasificaciones de 54.6%.

Figura 39

Ejecución de red Neural, set de datos 16-02



Nota. La figura muestra la ejecución de red Neural, con el set de datos 16-02.

Discusión

En esta etapa final se hizo el análisis de los modelos descubiertos en la etapa de ejecución de los algoritmos de Machine learning. Se analizan los resultados arrojados como salida de la ejecución de los algoritmos, en esta etapa se intentó relacionar los diferentes atributos, los factores que aparecieron en las reglas y arboles de decisiones y a partir de allí interpretar la magnitud del problema de la deserción con la finalidad de facilitar la toma de decisiones.

Teniendo en cuenta que este es un problema de clasificación en el cual están involucrados muchos atributos en el set de datos, se logró dimensionar la complejidad del problema, sobre todo cuando se trata de, cómo se pretendió en este caso, predecir la detección de la deserción estudiantil en una etapa muy temprana del proceso de aprendizaje del estudiante. Esta aseveración se hace teniendo en cuenta que la encuesta de caracterización es una de las primeras actividades que los estudiantes realizan al ingresar a la UNAD. La actividad de diligenciamiento de la caracterización generalmente se finaliza pasadas unas dos semanas a partir del inicio de las actividades del periodo, por lo que predecir la deserción en este punto es un objetivo muy ambicioso, debido a que, en ese momento del ingreso del estudiante a la institución, solo se conoce la información que el estudiante deja consignada mediante la aplicación del instrumento Encuesta de Caracterización.

Los experimentos arrojaron resultados interesantes como por ejemplo que la ciudad donde viven los estudiantes es un atributo que resulta determinante para deserción en esta etapa, sin embargo, por sí sola no es una variable que explica en gran medida el que un estudiante mantenga o no su matrícula activa.

A priori se pudiera pensar que algunos atributos como el número de hijos, o factores como la financiación de la matricula o la vivienda pudieran ser determinantes en la deserción de

los estudiantes, sin embargo una vez configurado el set de datos y obtenidos los modelos , se evidencia el galimatías que representa el reto de predecir la deserción estudiantil, encontrándose que los datos no evidenciaron una correlación significativa con el abandono de los estudios en estas variables, los modelos obtenidos durante los experimentos evidencian que factores como la discapacidad entre otros, no son determinantes para que un estudiante abandone sus estudios. Cuando se inician estudios padeciendo estas condiciones. Al menos en estos experimentos no se evidencia una gran influencia en términos de abandono del proceso de formación.

Los datos procesados por los modelos de clasificación evidencian que el problema de la deserción en la UNAD coincide o se acercan en gran medida con las cifras de retención y permanencia que arrojan las estimaciones de consejería, poniéndose de manifiesto que de los 7881 estudiantes que ingresaron en el 2018 16-01 aproximadamente la mitad en este momento no se encuentran con matrícula activa. Con este estudio se evidencia la magnitud del problema de deserción, siendo plausible que, para los estudiantes nuevos seleccionados en el set de datos a partir del año 2018, el 49% de ellos en este momento (febrero 28 de 2021) no tienen una matrícula activa en la universidad, pero además no matricularon en ningún periodo del año 2020, lo que indica que muy seguramente son estudiantes que abandonaron su proceso de formación académica en esta institución. Cabe aclarar que algunos de estos estudiantes es posible que reingresen en periodos posteriores a este estudio, con lo que se convierten en estudiantes con intermitencias, por lo que es muy complicado dar cifras de deserción exactas, como ya se había encontrado en la literatura revisada.

Los datos analizados por los algoritmos de aprendizaje no evidencian patrones significativos de correlación entre los atributos presentes en la prueba de caracterización de 2018- 16-01, de los estudiantes nuevos que ingresaron en ese periodo. No se pudo establecer de

manera temprana en los atributos con los que consta el set de datos una correlación significativa que defina de manera contundentemente certera si un estudiante desertará o no, sin embargo, con los modelos obtenidos a partir del set de datos de lograron precisiones del orden del 58% hasta 61% de aciertos, cifra que en futuros estudios se aspira a mejorar con la incorporación de atributos como por ejemplo si el estudiante ingresa por homologación externa o en los casos de homologaciones por convenios, datos de rendimiento académico y periodos matriculados y si presenta o no intermitencias.

Con estos experimentos se pudo comprobar que es posible la identificación de los factores que más inciden sobre la deserción estudiantil. Aplicando sistemáticamente técnicas de ranqueo de atributos se logró identificar los más incidentes en la variable de salida, recordando que en la clase a predecir se colocó 1 para indicar que el estudiante abandonó su proceso y 0 si el estudiante tuvo al menos una matrícula desde el periodo 2020 16-01.

En este mismo sentido, el balance de la clase o variable a predecir es muy importante en estos experimentos de predicción o clasificación, toda vez que los algoritmos de clasificación suelen priorizar la clase con mayor número de registros en detrimento de la clase con menor número de registros, lo que podría arrojar resultados poco confiables o con sesgos, así que el balance de los datos debe despertar el interés a la hora de abordar este tipo de estudios en los que se pretende predecir una clase a partir de un conjunto de atributos.

La aplicación de la metodología CRISP-DM aportó al desarrollo de este proyecto las siguientes ventajas: Es una metodología No–propietaria, que ofrece mucha libertad o independencia con respecto a las herramientas que se utilicen. CRISP-DM es imparcial con relación a las herramientas, y está encaminado al análisis técnico, así como a 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 WEKA representan un soporte importante para este tipo de investigaciones, toda vez que al ser herramientas con licencia open source abaratan los costos en el desarrollo de proyectos de esta naturaleza, con mayor razón 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, así 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 mining con pocos recursos económicos.

La información es uno de los activos más importantes para las organizaciones, y el hecho de poseer o almacenar datos de los 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 máximo capacidad es necesario realizar procesos que los organice con la finalidad de convertirlos en información, una vez organizados estos datos, 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 elemento que permite una ventaja competitiva para la organización.

Limitaciones

En muchas ocasiones se habla de la minería de datos como la panacea, el mundo ideal donde solo se necesita tener datos para predecir cualquier comportamiento de una población y en efecto es así en ocasiones, pero no es menos real que existen fenómenos con un componente aleatoriedad lo cual se incorpora a las decisiones que toman las personas como por ejemplo, ¿por qué una persona se cambia de operador móvil?, que no necesariamente responden a patrones sino, por ejemplo a un altercado con el conyugue. Se puede afirmar que este es un tipo de limitantes en los proyectos de Data mining, en los que los pronósticos suelen incluir demasiada incertidumbre y por más datos que se tengan del cliente los pronósticos no son los mejores.

Es así como en el desarrollo de este proyecto se tuvieron situaciones o vicisitudes que condicionaron algunas áreas del desarrollo con ciertas limitaciones las cuales impidieron obtener mejores resultados. A continuación, se presenta un resumen de las limitaciones más importantes que se tuvieron durante el desarrollo del proyecto, las cuales podrían considerarse como oportunidades de mejoras hacia el futuro. Estas limitaciones tuvieron que ver especialmente con el acceso a la data de los estudiantes, La universidad cuida de manera celosa los datos de los estudiantes que tiene bajo su custodia. Esto es comprensible toda vez que esta es información sensible y que está regulada por la normativa de habeas data, sin embargo, esto se convirtió en una limitante para el desarrollo de del proyecto. Un ejemplo de ello es que no se pudo incluir en el set de datos la información correspondiente a los estudiantes que ingresan por homologación, bien sea externa o por convenio con otras instituciones como el SENA, algunos datos de experimentos que se vienen realizando con set de datos más pequeños revelan que es posible mejorar las métricas de precisión, exactitud y sensibilidad, si se incluye el atributo de

homologación. Así mismo hubo también limitantes de tiempo. El desarrollo de este proyecto coincidió con la pandemia mundial del covid-19, lo cual también constituyó una limitante para obtener mejores resultados, Por ejemplo, el hecho de no estar presencialmente en el centro, propició que algunos trámites para la consecución de un mejor set de datos no se llevaran a cabo de la mejor manera, además del el stress que genera esta situación de pandemia en los stakeholders también pudo haber afectado de manera indirecta la obtención de mejores resultados

Otro factor limitante que no se tuvo en cuenta, por no tener los datos disponibles fue la información de egresados, no obstante, al ser estudiantes que ingresaron en el periodo 2018-I 16-01 es muy poco probable que algunos de estos estudiantes hayan podido graduarse al corte 2020 16-06.

La intermitencia en la matricula que algunos estudiantes suelen presentar en el desarrollo de los programas académicos es una limitante que dificulta el medir con absoluta precisión si un estudiante deserta o no de su proceso de formación académica superior, aunque para este estudio se tomó una ventana de matrícula de 6 periodos académicos consecutivo, los indicadores de deserción podrían no ser muy exactos en ese sentido y esto concuerda con el autor (Torres et al., 2015) que en este sentido afirma que no es posible medir de manera absoluta los índices de deserción debido a que a que hay estudiantes que salen del sistema por un tiempo y luego regresan, presentado intermitencias en su proceso de formación.

Conclusiones

Se realizó el análisis diagnostico a través del cual se pudo tener una mejor comprensión del problema de la deserción, identificando los elementos necesarios para abordarlo desde el campo de Data Mining, además de posibilitar la revisión de los estudios que se vienen desarrollando en este campo del conocimiento en Latinoamérica y Colombia. La revisión bibliográfica para este proyecto, permitió encontrar gran cantidad de información relacionada con este tipo de estudios por lo que se puede concluir que el problema de la deserción es un tema que está vigente, y en el cual las instituciones vienen haciendo esfuerzos para mitigarlo, esta revisión permitió el abordaje de la problemática desde la aplicación de técnicas de machine learning intentando obtener modelos de predicción que facilitan la toma de decisiones en las instituciones , en este sentido este estudio no pretende darle solución al fenómeno de la deserción sino que más bien propende por concebir una de las herramientas que resultan muy útiles, sobre todo si se combina con otras estrategias que coadyuvan a la mitigación del problema.

Se hizo un diagnóstico de la percepción de la deserción en el cuerpo docente mediante la aplicación del instrumento tipo cuestionario lo cual posibilitó obtener datos muy interesantes acerca de cómo se percibe la problemática de deserción en el cuerpo docente, con resultados que permitieron concluir que el cuerpo docente considera que un modelo de predicción de la deserción estudiantil en la UNAD contribuye en gran medida al mejoramiento de la retención y permanencia de los estudiantes. Observándose una buena percepción acerca de cómo la utilización de tecnologías de data mining contribuye a mejorar problemáticas del contexto como el abandono de los estudios, lo que aporta a la solución de la pregunta problema del estudio, alineándose con el objetivo general propuesto. Esto coincide con (Heredia et al., 2015), en su propuesta de un modelo predictivo de deserción estudiantil basado en arboles de decisión.

También se evidenció que la percepción que tienen los docentes con respecto al comportamiento de la deserción en sus cursos revela que la mayoría de los tutores considera que la deserción se ha mantenido y un menor porcentaje de tutores considera que ha aumentado. Esto indica la relevancia que tiene este estudio toda vez que la deserción es un problema presente en la institución y afecta negativamente los indicadores de esta. También se identificaron los factores considerados por los docentes como más incidentes: la disponibilidad de tiempo, factores económicos, factores sociales, y acceso a las herramientas digitales, considerando la disponibilidad de tiempo como el factor más incidente, lo que coincide con autores como (Torres et al., 2015) quien encontró que el nivel socioeconómico es el principal factor asociado a la deserción. Otro elemento encontrado fue que el 53% consideró que los índices de aprobación influyen en la deserción, lo cual indica que el abandono de los estudios comienza con este síntoma que luego desemboca en el abandono mismo, coincidiendo con lo planteado por (Ángel & Facundo, 2009) que relaciona el rendimiento académico con retención y por ende el abandono de los estudios. Los datos obtenidos en el diagnostico apuntan a que la percepción de los docentes es positiva en cuanto al uso de la IA para la mitigación del fenómeno de la deserción, con el 84,2% de favorabilidad. Estos resultados son pertinentes toda vez que, según (Berlanga, 2016) las técnicas Aprendizaje Automático (campo de investigación de la IA) y la Minería de datos comparten gran cantidad de técnicas, apuntando al uso de algoritmos que pueden extraer conocimiento a partir de datos. Estos resultados dan cuenta de la razón para el desarrollo del modelo de deserción mediante las técnicas de Data Mining, propuesto en objetivo general.

La realización de este estudio permitió evidenciar como el desarrollo de un modelo de predicción de la deserción, contribuye en gran medida a mitigar el problema del abandono de los estudios, facilitando a la organización un soporte para proponer medidas más efectivas para

mitigar el problema de la deserción. Permitiendo aprender mediante estos modelos de aprendizaje supervisado desde los datos históricos de los estudiantes, revisando como ha sido su comportamiento en cuanto a matriculas obteniendo modelos que pueden ser aplicados o extrapolados a estudiantes que ingresan a cualquier periodo y de esta manera identificar aquellos estudiantes con mayores probabilidades de abandono de los procesos de aprendizaje.

Con el desarrollo de este proyecto fue posible proveer mecanismos de predicción de la deserción mediante la realización de experimentos aplicando algoritmos de machine learning al set de datos de los estudiantes de primera matricula del curso Herramientas digitales para la gestión del conocimiento del año 2018 16-01, obteniendo así modelos de aprendizaje supervisado a partir de esos datos. En los modelos concebidos a partir del aprendizaje supervisado, se lograron medidas de precisión entre 58% y 61 %. de efectividad para saber si un estudiante desertará del proceso. Este porcentaje de precisión obtenido en el modelo de predicción de la deserción estudiantil en la UNAD, puede que el lector se vea tentado a considerar que este no es un resultado prometedor o significativo, sin embargo, si se hace un análisis a profundidad es posible percatarse que la deserción es un problema multifactorial complejo, y estos porcentaje de predicción logrados por encima del 58% son bastante significativos, teniendo en cuenta que se puede obtener a solo 2 semanas de haberse matriculado en la UNAD el estudiante, lo cual es muy significativo si se tienen en cuenta que en esos momentos solo se cuenta con esos datos del estudiante, para detectar de manera temprana quienes están en riesgo.

Mediante la utilización de datos históricos del 2018 16-02 se pudieron realizar pruebas sobre los modelos obtenidos y observar el comportamiento de esos estudiantes en términos de matrículas, logrando resultados prometedores en el desempeño del modelo. Con lo cual se

concluye que apuntando a la consecución de un set de datos optimo es posible obtener resultados muy confiables en el modelo de predicción, concibiendo de esta manera una herramienta cuyo uso contribuye en gran medida a mitigar el problema de deserción.

El desarrollo de este proyecto tiene un impacto significativo sobre el abordaje del problema de la deserción, en el sentido en que el problema de la deserción representa unos elevados costos sociales para las instituciones y para el país. Así quedó evidenciado en el diagnóstico de la percepción del cuerpo docente donde el 42,1% coincidieron en que la deserción se asocia con bajas tasas de crecimiento económico y el aumento de la brecha de pobreza. Por lo que mitigando el problema se está contribuyendo a un mundo más justo y equitativo. Este impacto tiene que ver con el avance hacía una educación en la que se tomen en cuenta las particularidades de cada estudiante para lo cual las herramientas propuestas en este proyecto son fundamentales en la extracción de los factores particulares de cada individuo. Aquí se coincide con (Hai-ling et al., 2018) en que es posible educar de una forma más personalizada y esta experiencia de aprendizaje causa motivación en el estudiante.

Este estudio abre el camino para nuevas investigaciones sobre predicción de la deserción estudiantil de manera temprana, aplicando técnicas de minería de datos basada en machine learning. La analítica de datos con minería de datos puede ser de gran significancia debido al enorme impulso que pueden proporcionar en la solución de estos problemas complejos, como la deserción estudiantil.

Es posible obtener mejores resultados de precisión a medida que el estudiante avance en el desarrollo de sus procesos de aprendizaje, específicamente si se incluye en el modelo atributos como el promedio de calificación del estudiante, el número de créditos matriculados, El número de créditos aprobados. Estos atributos pueden mejorar enormemente la precisión del modelo,

como se comprobó en algunos experimentos realizados de manera paralela a este estudio. Al igual que el atributo de homologación, que hasta el momento de presentar este informe no fue posible obtenerlos desde RyC.

La consecución del set de datos es una tarea costosa la cual debe hacerse con mucha dedicación si se quiere que los resultados a partir de la aplicación de técnicas de Machine learning sean los mejores. En este tipo de estudios es ideal que se involucre a la alta gerencia para poder tener acceso privilegiado a las fuentes de datos de las organizaciones de tal manera que las tareas de consecución de la información no se conviertan en un problema en sí mismos, que consuma toda la energía del investigador en lugar de emplearla en el logro del mejor modelo posible.

El aprendizaje obtenido como producto del desarrollo de este proyecto es invaluable. El hecho de poder aplicar los conocimientos aprendidos durante la Maestría en Gestión de TI es una experiencia bastante significativa, en la que se aplicaron esas teorías a problemas del entorno. Lo cual deja las bases sentadas para el desarrollo futuro de este tipo de proyectos que involucra la aplicación de analítica de datos mediante Inteligencia Artificial para solucionar problemas del contexto.

Como resultado esta investigación se ha divulgado los avances obtenidos durante el desarrollo de la misma, de esta manera se participó en dos ocasiones en la feria de proyectos de la maestría en gestión de TI con 2 ponencias, además se participó en la convocatoria para publicación de Working Paper realizada al interior de la UNAD, y en la cual se obtuvo la publicación de articulo denominado “Minería de datos para la predicción de la deserción estudiantil en la Universidad Nacional Abierta Y A Distancia” el cual tiene como propósito divulgar una propuesta de investigación en el marco de la maestría en gestión de TI de la

universidad abierta 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. (Ávila Pérez & Medina Cruz, 2021)

Referencias

Abadía, C., Guerrero, H., Rodríguez, J., Vela González, P. A., Martínez, H., Villamizar, A. N., Sánchez, D. A. A., & Aguilar, L. A. P. (2018). *Informes de Gestión de Procesos -IGP- vigencia 2018 - Ciclo Vida Del Estudiante*. https://sig.unad.edu.co/documentos/sgc/informes_gestion/2018/periodo_1/1er_IGP_2018_ciclo_de_vida_del_estudiante.pdf

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

Apache Hadoop. (2018). Apache Software Fundation. https://hadoop.apache.org/docs/stable/

Ávila Pérez, M., & Medina Cruz, J. (2021). Minería de datos para la predicción de la deserción estudiantil en la Universidad Nacional Abierta y a Distancia. In *Documentos de trabajo ECBTI*. https://hemeroteca.unad.edu.co/index.php/wpecbti/article/view/3887/4317

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

Barrera Suárez, C. M., Reyes Murillo, I. G., & Silva Santander, C. H. (2010). Deserción académica y agente inteligente para su control y seguimiento: un aporte al sistema de consejería de la UNAD. *Revista de Investigaciones UNAD*, *9*(2), 255. https://doi.org/10.22490/25391887.687

Berlanga, A. (2016). El camino desde la inteligencia artificial al Big Data. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *2*(9), 9–11. http://www.revistaindice.com/numero68/p9.pdf

Casas, J., Repullo, J. R., & Donado, J. (2003). La encuesta como tecnica de investigacion. *Atención Primaria*, *31*(8), 527–538. http://www.unidaddocentemfyclaspalmas.org.es/resources/9+Aten+Primaria+2003.+La+Encuesta+I.+Custionario+y+Estadistica.pdf

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)*. Retrieved April 30, 2020, from http://recercat.cat/handle/2072/335833

Cuestas, A. (2020). *Informes de Gestión de Procesos - vigencia 2020 GESTIÓN DEL TALENTO HUMANO*. https://sig.unad.edu.co/documentos/sgc/informes_gestion/2020/periodo_2/2do_IGP_2020_gestion_talento_humano.pdf

Cuji, B., Gavilanes, W., & Sanchez, R. (2017). Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Espacios*, *38*(55), 17. http://ww.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

Del Toro Díaz, W. (2013). Factores que contribuyen en la deserción de los estudiantes de la Escuela de Ciencias Administrativas en la UNAD – CEAD Simón Bolívar - Cartagena. *Revista de Investigaciones UNAD*, *12*(1), 183. https://doi.org/10.22490/25391887.1167

Díaz, F., & Osorio, M. (2013). Aplicando estrategias y tecnologías de Inteligencia de Negocio en sistemas de gestión académica. *Sedici.Unlp.Edu.Ar*. http://sedici.unlp.edu.ar/handle/10915/27157

Disla, R., & Llaugel, F. (2015). Un Modelo Predictivo de Deserción Escolar para la República Dominicana. *Researchgate.Net*. 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

Dominguez, D. (2018). *Big Data, analítica del aprendizaje y educación basada en datos (Big & Data-driven Education). Recuperado de:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3124369.*

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

Fayyad, U., & Piatetsky-Shapiro, Gregory Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*, 37. https://wvvw.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131

González-Loaiza, R. (2018). Proyecto de graduación genera herramienta para pronosticar deserción de estudiantes del TEC. *Investiga.TEC*, *33*, 3. https://revistas.tec.ac.cr/index.php/investiga_tec/article/download/3872/3448

Hai-ling, L., Jun-huai, L., Jun, P., & Troisi, O. (2018). June). Big Data Technology Applied to Learning Behavior Evaluation System. *In International Conference on Big Data and Artificial Intelligence BDAI Pp IEEE Otroisiunisa I, 2018 SRC*, 10–17.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Morgan Kaufmann. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-

Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. https://doi.org/10.1109/TLA.2015.7350068

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

Kuna, H., García, R., & Villatoro, F. R. (2010). Pattern Discovery In University Students Desertion Based On Data Mining. *Proceedings of The IV Meeting on Dynamics of Social and Economic Systems*, *2*(2), 11. https://www.researchgate.net/profile/Ramon_Garcia-Martinez/publication/266939899_Pattern_discovery_in_university_students_desertion_based_on_data_mining/links/54b6bca30cf2e68eb27efce1.pdf

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

Lopera, C. (2008). Determinantes de la deserción universitaria en la Facultad de Economía de la Universidad del Rosario. *Economia*, *95*. http://www.urosario.edu.co/FASE1/economia/econ_inve_publicaciones_3.htm

Malvicino, F., & Yoguel, G. (2015). Big Data : Avances Recientes a Nivel Internacional y Perspectivas para el Desarrollo Local. In *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

Márquez, C. (2015). Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos [tesis de doctorado, Universidad de Córdoba]. *Biblioteca Universidad de Córdoba*, 105. https://dialnet.unirioja.es/servlet/tesis?codigo=66343

Mendoza García, D., & Gómez Orduz, M. (2012). Influencia de la virtualidad en la deserción de estudiantes en la Universidad Nacional Abierta y a Distancia. UNAD – CEAD, Yopal (Casanare – Colombia). *Revista de Investigaciones UNAD*, *11*(1), 163. https://doi.org/10.22490/25391887.778

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

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

Mustafa, M. N., Chowdhury, L., & Kamal, M. S. (2012). Students dropout prediction for intelligent system from tertiary level in developing country. *2012 International Conference on Informatics, Electronics and Vision, ICIEV 2012*, 113–118. https://doi.org/10.1109/ICIEV.2012.6317441

Niño, M., & Illarramendi, A. (2015). ENTENDIENDO EL BIG DATA: ANTECEDENTES, ORIGEN Y DESARROLLO POSTERIOR. *DYNA NEW TECHNOLOGIES*, *2*(3), [8 p.]-[8 p.]. https://doi.org/10.6036/nt7835

Noriega, C. O. (2019). Factores personales y académicos que inciden en la deserción temprana

de estudiantes del programa de psicología adscritos a la Unad Cead Ocaña. *Repository.Unad.Edu.Co*. https://repository.unad.edu.co/bitstream/handle/10596/28063/claudia.ochoa.pdf?sequence=1&isAllowed=y

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. *International Journal of Morphology*, *35*(1), 227–232. https://doi.org/10.4067/S0717-95022017000100037

Patiño Garzón, L., & Cardona Pérez, A. M. (2012). REVISIÓN DE ALGUNOS ESTUDIOS SOBRE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA EN COLOMBIA Y LATINOAMÉRICA. *Theoría - Ciencia, Arte y Humanidades*, *21*(1), 9–20. http://revistas.ubiobio.cl/index.php/RT/article/view/1241

Perez, B., Castellanos, C., & Correal, D. (2018, October 5). Applying Data Mining Techniques to Predict Student Dropout: A Case Study. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence, ColCACI 2018 - Proceedings*. https://doi.org/10.1109/ColCACI.2018.8484847

Poveda Velasco, J. C., Poveda Velasco, I. M., & España Irala, I. A. (2020). Análisis de la deserción estudiantil en una universidad pública de Bolivia. *Revista Iberoamericana de Educación*, *82*(2), 151–172. https://doi.org/10.35362/rie8223572

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

Rodríguez Núñez, L., & Londoño Londoño, P. (2011). Estudio sobre deserción estudiantil en los programas de Educación de la Católica del Norte Fundación Universitaria. *Revista Virtual*

Universidad Católica Del Norte, *1*(33), 328–355.

Rodríguez, P., Truffello, R., Suchan, K., Varela, F., Matas, M., Mondaca, J., Céspedes, J., Valenzuela, L., Valenzuela, J. P., & 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 EDUCACION*. http://disde.minedu.gob.pe/handle/123456789/4463

Rodriguez, M., & Vargas E, D. (2013). Diseño no experimental transeccional. *Aprender Haciendo Universidad Yacambú*. https://issuu.com/divargase/docs/dise__o_no_experimental_transeccion

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, Issues 2 SRC-BaiduScholar FG-0). Informativo, C E I.

Salazar, A., Gosálbez, j, Bosch, I., Miralles, R., & Vergara, l. (2004). A case study of knowledge discovery on academic achievement, student desertion and student retention. *TRE 2004. 2nd International Conference Information Technology: Research and Education. IEEE, 2004.* https://www.researchgate.net/profile/Addisson_Salazar/publication/4124229_A_case_study_of_knowledge_discovery_on_academic_achievement_student_desertion_and_student_retention/links/54f8626b0cf28d6deca25e2e/A-case-study-of-knowledge-discovery-on-academic-ac

Sampieri, R., Collado, C., & Lucio, P. (2014). Metodologia de la investigaci6n. *Edici6n McGraw-Hill*. http://observatorio.epacartagena.gov.co/wp-

content/uploads/2017/08/metodologia-de-la-investigacion-sexta-edicion.compressed.pdf

Sánchez-Sánchez, J. C. (2018). Factores asociados a la deserción académica en los programas de las escuelas de la universidad nacional abierta y a distancia – UNAD- CCAV Cartagena*. *Revista Estrategia Organizacional*, *7*(2), 51–66. https://doi.org/10.22490/25392786.2943

Shrivastava, A., & Deshpande, T. (2016). *Hadoop blueprints : use Hadoop to solve business problems by learning from a rich set of real-life case studies*. Packt Publishing Ltd. https://books.google.es/books?hl=es&lr=&id=mn9cDgAAQBAJ&oi=fnd&pg=PP1&dq=Hadoop+Blueprints.+Birmingham,+UK:+Packt+Publishing.+Retrieved+from&ots=jmo98thJEi&sig=fwBLFiyBTyG1QKFmtfPJp2hETCU#v=onepage&q=Hadoop Blueprints. Birmingham%2C UK%3A Packt Publishing

Torres, J., Acevedo, D., & Gallo, L. (2015). Causas y consecuencias de la deserción y repitencia escolar: una visión general en el contexto latinoamericano. *Cultura Educación Y Sociedad*, *6*(2), 157–187. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwixhLSKvPHlAhXLx1kKHeIwCmMQFjAAegQIARAC&url=https%253A%252F%252Frevistascientificas.cuc.edu.co%252Findex.php%252Fculturaeducacionysociedad%252Farticle%252Fdownload%252F904%252Fpdf_127%25

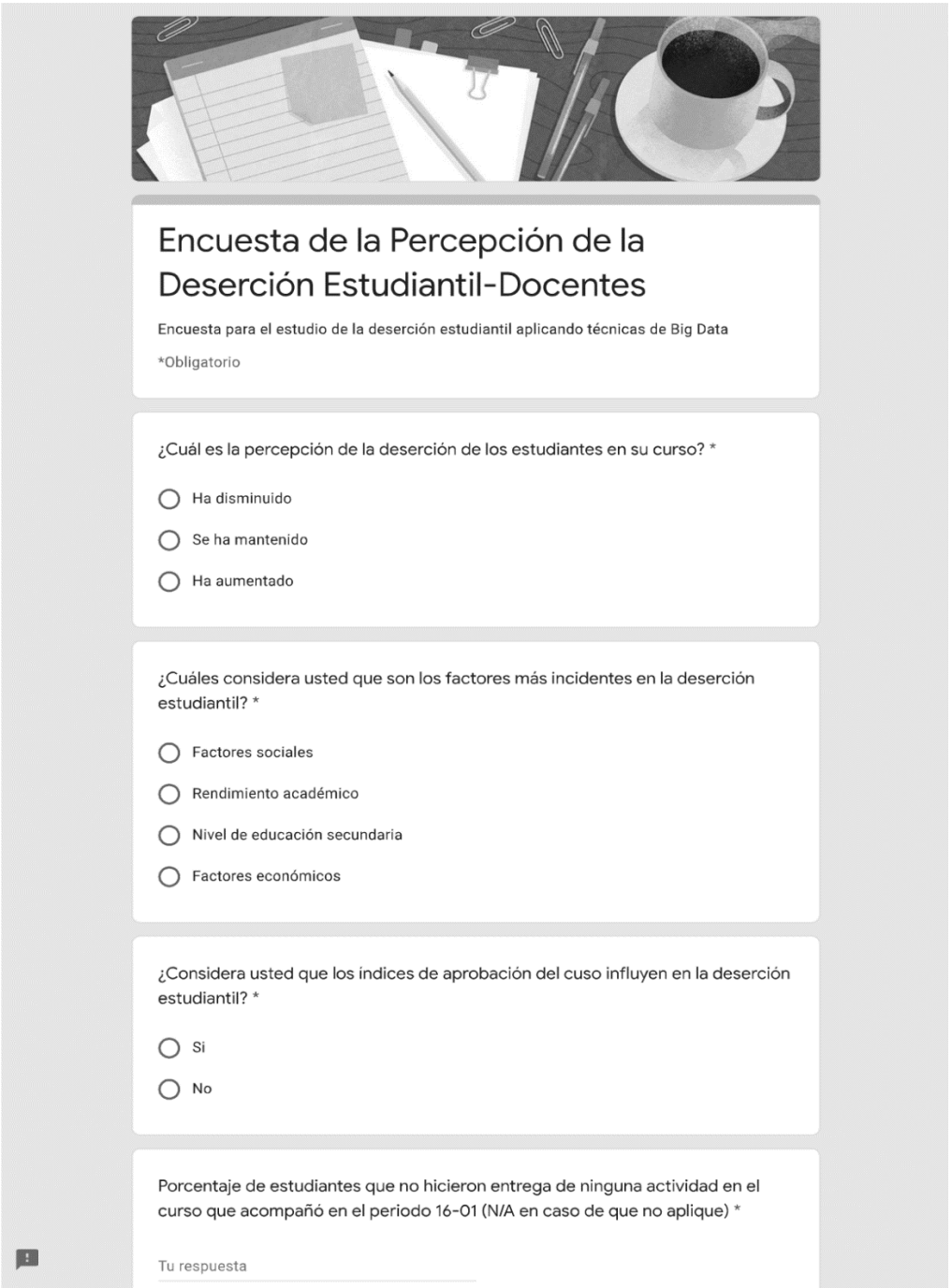
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/c2009-0-19715-5

Yue, Y., & Liu, D. (2016). Evaluation of different training programs of innovative education in top international universities using big data analysis. *International Journal of Simulation: Systems, Science and Technology*, *17*(42), 58.1-58.5.

https://doi.org/10.5013/IJSSST.a.17.42.58

Anexos

Anexo 1. Instrumento Encuesta De Percepción



Encuesta de la Percepción de la Deserción Estudiantil-Docentes

Encuesta para el estudio de la deserción estudiantil aplicando técnicas de Big Data

*Obligatorio

¿Cuál es la percepción de la deserción de los estudiantes en su curso? *

- Ha disminuido
- Se ha mantenido
- Ha aumentado

¿Cuáles considera usted que son los factores más incidentes en la deserción estudiantil? *

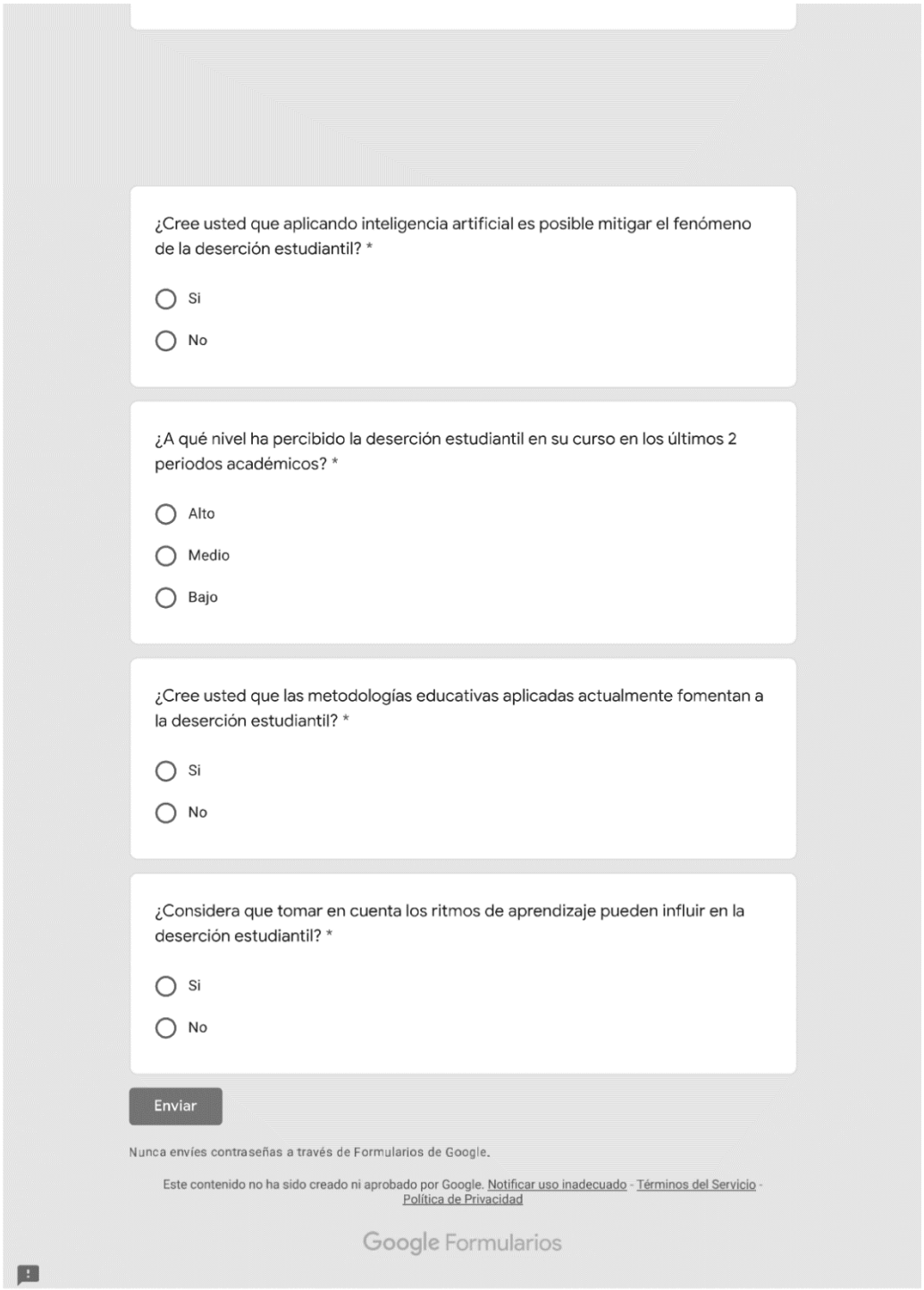
- Factores sociales
- Rendimiento académico
- Nivel de educación secundaria
- Factores económicos

¿Considera usted que los indices de aprobación del cuso influyen en la deserción estudiantil? *

- Si
- No

Porcentaje de estudiantes que no hicieron entrega de ninguna actividad en el curso que acompañó en el periodo 16-01 (N/A en caso de que no aplique) *

Tu respuesta



¿Cree usted que aplicando inteligencia artificial es posible mitigar el fenómeno de la deserción estudiantil? *

- Si
- No

¿A qué nivel ha percibido la deserción estudiantil en su curso en los últimos 2 periodos académicos? *

- Alto
- Medio
- Bajo

¿Cree usted que las metodologías educativas aplicadas actualmente fomentan a la deserción estudiantil? *

- Si
- No

¿Considera que tomar en cuenta los ritmos de aprendizaje pueden influir en la deserción estudiantil? *

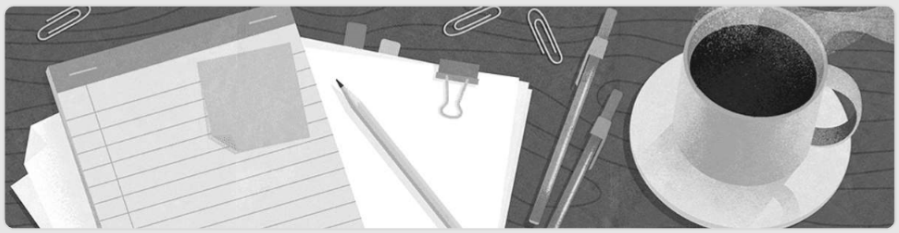
- Si
- No

Enviar

Nunca envíes contraseñas a través de Formularios de Google.

Este contenido no ha sido creado ni aprobado por Google. Notificar uso inadecuado - Términos del Servicio - Política de Privacidad

Google Formularios



Encuesta de Percepción de la Deserción Estudiantil

*Obligatorio

¿En qué medida considera usted que la implementación de un modelo de predicción de la deserción estudiantil en la UNAD contribuye al mejoramiento de la retención y permanencia de los estudiantes? *

	1	2	3	4	5	
No contribuye	○	○	○	○	○	Contribuye en gran medida

De las siguientes herramientas de Big Data, ¿Cuáles le son familiares o ha interactuado? *

- [ ] Python
- [ ] R
- [ ] Orange
- [ ] Hadoop
- [ ] WEKA
- [ ] Ninguna

¿Cuál es la percepción de la deserción de los estudiantes en el curso que usted es tutor? *

- ○ Ha disminuido
- ○ Se ha mantenido

- ◯ Ha aumentado

¿Cuáles de los siguientes factores considera usted que son los que más inciden en la deserción estudiantil? *

- ☐ Factores sociales
- ☐ Rendimiento académico
- ☐ Calidad de la educación recibida en la secundaria
- ☐ Factores económicos
- ☐ Disponibilidad de tiempo
- ☐ Acceso a las herramientas digitales
- ☐ Otro: _____

¿Considera usted que los índices de aprobación del curso influyen en la deserción estudiantil? *

- ◯ Si
- ◯ No

Porcentaje de estudiantes que no hicieron entrega de ninguna actividad en algún curso que usted haya acompañado. *

Tu respuesta

¿Cree usted que aplicando inteligencia artificial es posible mitigar el fenómeno de la deserción estudiantil? *

- ◯ Si
- ◯ No

¿A qué nivel ha percibido la deserción estudiantil en el curso que usted es tutor en los últimos 2 periodos académicos? *

- ◯ Alto
- ◯ Medio
- ◯ Bajo

¿Cree usted que las metodologías educativas aplicadas actualmente fomentan la retención y permanencia estudiantil? *

- ◯ Si
- ◯ No

¿Considera usted que tomar en cuenta los ritmos de aprendizaje pueden influir en la deserción estudiantil? *

- ◯ Si
- ◯ No

De las siguientes consecuencias sociales provocadas en parte por la deserción estudiantil, ¿Cuál cree usted que es en la que más incide? *

- ◯ Incremento en la tasa de criminalidad
- ◯ Decremento de la tasa de crecimiento económico
- ◯ Aumento de la brecha entre ricos y pobres
- ◯ Incremento de la informalidad en el comercio


Atrás
Enviar
Nunca envíes contraseñas a través de Formularios de Google.
Este contenido no ha sido creado ni aprobado por Google. Notificar uso inadecuado - Términos del Servicio - Política de Privacidad
Google Formularios

Anexo 2 Descripción Del Set De Datos

@relation 'Deserción'
@attribute Programa {'LICENCIATURA EN INGL�S COMO LENGUA EXTRANJERA','INGENIER�A ELECTR�NICA (Resoluci�n 13155)','TECNOLOGIA EN REGENCIA DE FARMACIA (RESOLUCION 08200)','INGENIER�A DE TELECOMUNICACIONES (Resoluci�n 14518)','ADMINISTRACION EN SALUD','TECNOLOGIA EN AUTOMATIZACION ELECTRONICA','INGENIERIA INDUSTRIAL (Resoluci�n 05867)','TECNOLOGIA EN SISTEMAS DE COMUNICACIONES INAL�MBRICAS','ARTES VISUALES','TECNOLOGIA EN SEGURIDAD Y SALUD EN EL TRABAJO','TECNOLOGIA EN SISTEMAS AGROFORESTALES','ADMINISTRACION DE EMPRESAS','TECNOLOGIA EN LOGISTICA INDUSTRIAL','PSICOLOGIA (Resolucion 3443)',AGRONOMIA,'COMUNICACION SOCIAL','INGENIERIA DE SISTEMAS',SOCIOLOGIA,'TECNOLOG�A EN DESARROLLO DE SOFTWARE','TECNOLOGIA EN RADIOLOGIA E IMAGENES DIAGNOSTICAS','FILOSOFIA (Resolucion 10583)','TECNOLOGIA EN GESTION COMERCIAL Y DE NEGOCIOS (Resoluci�n 6544)','TECNOLOGIA EN GESTION INDUSTRIAL (RESOLUCION 16616)',ECONOMIA,'LICENCIATURA EN MATEMATICAS',MUSICA,'TECNOLOGIA EN GESTION DE OBRAS CIVILES Y CONSTRUCCIONES (RESOLUCION 95)','LICENCIATURA EN PEDAGOGIA INFANTIL',ZOOTECNIA,'TECNOLOGIA EN PRODUCCION DE AUDIO','TECNOLOGIA EN GESTION DE TRANSPORTES (RESOLUCION 99)','INGENIERIA DE ALIMENTOS (Resolucion 0575)','LICENCIATURA EN ETNOEDUCACION','LICENCIATURA EN FILOSOFIA','TECNOLOGIA EN PRODUCCION ANIMAL (RESOLUCION 17682)','TECNOLOGIA EN GESTION AGROPECUARIA (RESOLUCION 12019)','INGENIERIA AMBIENTAL','TECNOLOGIA EN PRODUCCION AGRICOLA','CURSOS LIBRES','TECNOLOGIA EN GESTION DE EMPRESAS ASOCIATIVAS Y ORGANIZACIONES COMUNITARIAS (RESOLUCION 12020)'}
@attribute Cead {'(BOGOTA CRA.30)JOSE ACEVEDO Y G�MEZ',VALLEDUPAR,MEDELL�N,SAHAG�N,'LA DORADA',ZIPAQUIR�,'SANTA MARTA',YOPAL,SOCHA,DOSQUEBRADAS,OCA�A,CALI,PALMIRA,PITALITO,FLORENCIA,TUNJA,MARIQUITA,NEIVA,BARRANCABERMEJA,TURBO,BARRANQUILLA,V�LEZ,'CARTAGENA (Roberto de Jes�s Salazar Ramos)',ACACIAS,AGUACHICA,POPAY�N,'SANTANDER DE QUILICHAO',GIRARDOT,FACATATIVA,QUIBD�,'LA PLATA',CURUMAN�,PASTO,BUCARAMANGA,DUITAMA,SOACHA,IBAGU�,'COROZAL (Rub�n del Cristo Mart�nez)','LA GUAJIRA',GACHETA,BOAVITA,GARAGOA,PAMPLONA,FUSAGASUGA,CUMARAL,TUMACO,C�CUTA,'VALLE DEL GUAMUEZ','SAN JOS� DEL GUAVIARE',SOGAMOSO,SOAT�,ARBEL�EZ,CUBAR�,CHIQUINQUIR�,'EL BORDO',PLATO,'SAN VICENTE DEL CAGU�N','PUERTO IN�RIDA',L�BANO,LETICIA,M�LAGA,'PUERTO CARRE�O','PUERTO AS�S','EL BANCO'}
@attribute Zona {'CENTRO BOGOTA Y CUNDINAMARCA',CARIBE,OCCIDENTE,'AMAZONAS Y ORINOQUIA','CENTRO BOYACA','CENTRO ORIENTE','CENTRO SUR',SUR}
@attribute Escuela {'ESCUELA DE CIENCIAS DE LA EDUCACI�N','ESCUELA DE CIENCIAS B�SICAS, TECNOLOG�A E INGENIER�A','ESCUELA DE CIENCIAS DE LA SALUD','ESCUELA DE CIENCIAS SOCIALES, ARTES Y HUMANIDADES','ESCUELA DE CIENCIAS AGR�COLAS, PECUARIAS Y DEL MEDIO AMBIENTE','ESCUELA DE CIENCIAS ADMINISTRATIVAS, CONTABLES, ECON�MICAS Y DE NEGOCIOS'}
@attribute G�nero {F,M}
@attribute Etnia {'No pertenece',Afrocolombianos,Raizales,Wayuu,'No informa','Zen� / sen�',Sikuani,Guambiano,Yanacona,'Wiwa (arzario)',Pastos,'Otro Pueblo Indigena',Pijaos,'Otras comunidades negras','Curripaco o kurripaco',Muisca,'Pueblo RROM','Awa (cuaiker)',Desano,'Embera chami',Kogui,Puinave,'Nasa (pa�z)',Inga,Kankuamo,Bara}
@attribute 'Ciudad Residencia' {'BOGOT� D.C.','LA JAGUA DE IBIRICO',MEDELLIN,MONTERIA,'LA DORADA',TOCANCIPA,NEMOCON,'SAN JERONIMO','SANTA MARTA',YOPAL,'PAZ DE RIO','DOS QUEBRADAS',OCA�A,TULUA,CALI,ITAGUI,GUACARI,PITALITO,FLORENCIA,TUNJA,MOTAVITA,SORACA,MARIQUITA,VALLEDUPAR,NEIVA,BARRANCABERMEJA,TURBO,BARRANQUILLA,'SANTA HELENA DEL OPON',PEREIRA,CARTAGENA,VILLAVICENCIO,AGUACHICA,ARMENIA,'PUERTO BOYACA',POPAYAN,'SANTANDER DE QUILICHAO',OROCUE,GIRARDOT,SUESCA,MADRID,FACATATIVA,SONSON,PACHO,QUIBDO,APARTADO,MOCOA,'LA PLATA',CANDELARIA,PAILITAS,PASTO,CACHIRA,BUCARAMANGA,CIMITARRA,DUITAMA,SOACHA,IBAGUE,PALMIRA,'LA CEJA','SAN ONOFRE',CHINCHINA,VILLAMARIA,ENVIGADO,BUENAVENTURA,JAMUNDI,'PUERTO TEJADA',FLORIDA,ACACIAS,'PUERTO GAITAN','SAN MARTIN','LA GLORIA',RIOHACHA,'SANTA ROSA DE CABAL',VILLETA,'EL COPEY',BOSCONIA,GUACHETA,BELLO,JERICO,SINCELEJO,MONIQUIRA,NUNCHIA,'SAN MATEO',GARAGOA,MANIZALES,AGUAZUL,PAEZ,PAIPA,MUTISCUA,GIRON,FUSAGASUGA,'EL CAIRO',CA�ASGORDAS,RESTREPO,SAHAGUN,'SAN ANDRES SOTAVENTO',BUGALAGRANDE,YARUMAL,'SANTA CATALINA',COMBITA,MANI,GRANADA,TIMANA,CALOTO,SOPO,ITSMINA,TUMACO,SALAZAR,CUCUTA,ABREGO,'SAN ALBERTO',FLORIDABLANCA,CHOCONTA,'LA ESTRELLA',CAICEDONIA,'SAN MIGUEL',ZIPAQUIRA,GUAMAL,'SAN JUAN DE ARAMA',CHIRIGUANA,INIRIDA,FRESNO,CHIGORODO,SOGAMOSO,'LA SALINA',COTA,SAMACA,'SAN JOSE DE FRAGUA',CHIA,'SAN JOSE DEL GUAVIARE','EL TARRA',SUAITA,PAMPLONA,'SABANA DE TORRES','AGUSTIN CODAZZI',CURUMANI,OVEJAS,ESPINAL,'EL CERRITO',BUGA,IPIALES,'LA PINTADA',BARBOSA,COPACABANA,SOLEDAD,ARAUCA,COVARACHIA,CAJICA,PORE,CAREPA,BARICHARA,'PUERTO LIBERTADOR',MELGAR,GUARNE,PIEDECUESTA,MOSQUERA,NOBSA,CABUYARO,RICAURTE,FUNZA,RIONEGRO,VELEZ,FONSECA,CONSACA,FORTUL,HIMARE,PUPIALES,VENECIA,BOLIVAR,VILLANUEVA,MACEO,CHISCAS,'PUENTE NACIONAL',BELALCAZAR,VIRACACHA,'EL TAMBO','SAN MARCOS',MANZANARES,LORICA,GACHETA,YAGUARA,VILLAPINZON,'PAZ DE ARIPORO',PLATO,GUADALUPE,SAMANIEGO,BALBOA,TIBU,HACARI,'CIUDAD BOLIVAR',TAURAMENA,SOATA,'PUEBLO NUEVO',GALAN,BARBACOAS,DOLORES,RAMIRIQUI,MILAN,VALPARAISO,'SAN VICENTE DEL CAGUAN',TAMARA,'PUERTO LLERAS','PUERTO GUZMAN',ORITO,'EL DORADO','PUERTO RICO','VALLE GUAMUEZ',BETULIA,LIBANO,BECERRIL,SOMONDOCO,SOTAQUIRA,BOYACA,VICTORIA,TABIO,CARTAGO,TIMBIQUI,VILLARRICA,TAMINANGO,GACHALA,'SANTA ROSA DE VITERBO',RETIRO,CHIQUINQUIRA,ANAPOIMA,CUMBAL,LETICIA,ANDES,CISNEROS,AMALFI,BOJACA,NILO,SIBATE,FOMEQUE,UBATE,ANTIOQUIA,'CARMEN DE CARUPA','PUERTO TRIUNFO',ENTRERRIOS,ANORI,SANTUARIO,'CARMEN DE VIBORAL','SAN CARLOS','GOMEZ PLATA',SEGOVIA,'SANTA ROSA DE OSOS',REMEDIOS,MARINILLA,NECOCLI,PUEBLORRICO,SABANETA,'PUERTO

BERRIO',SALGAR,FREDONIA,CALDAS,BETEITIVA,GUATEQUE,JENESANO,FIRAVITOBA,MAGANGUE,'SANTA MARIA','VILLA DE LEYVA',SOCOTA,ANSERMA,QUIPAMA,'SANTA SOFIA',AGUADAS,TOCA,'LA UVITA',MALAGA,CUMARIBO,PENSILVANIA,'PATIA(EL BORDO)',SUBACHOQUE,CAJIBIO,LEIVA,SUAREZ,CALDONO,PELAYA,MONTELIBANO,'RIO DE ORO',CHINU,ASTREA,'MANAURE BALCON DL CESAR',CIENAGA,GUTIERREZ,'PLANETA RICA','LA PAZ',ANOLAIMA,GUASCA,'SAN BERNARDO','EL COCUY',SILVANIA,MEDINA,SABANALARGA,COGUA,FLANDES,'LA CALERA','AGUA DE DIOS',TENA,'LA MESA',GUICAN,'PUERTO SALGAR',GACHANCIPA,PAIME,JUNIN,SUTATAUSA,FUQUENE,SIPI,GARZON,ALTAMIRA,NATAGA,TENJO,ACEVEDO,CAMPOA LEGRE,SALDA&#65533;A,BOJAYA,GIGANTE,RIVERA,'LA ARGENTINA',ALBAN,SALADOBLANCO,'SAN AGUSTIN','SAN JUAN NEPOMUCENO',ARACATACA,'BUENOS AIRES',ISNOS,CUMBITARA,'LA FLORIDA',SANDONA,LINARES,'LA VIRGINIA',POLICARPA,'EL DONCELLO',ECUADOR,GUATICA,BOCHALEMA,SARAVENA,'EL CARMEN',TONA,'SAN CALIXTO',BUENAVISTA,CAUCASIA,PROVIDENCIA,FLORIAN,'VISTA HERMOSA',QUIMBAYA,CALARCA,MONTENEGRO,CAPITANEJO,CIRCASIA,ARATOCA,GIRALDO,'CASTILLA LA NUEVA',AGUADA,SACAMA,SAMPUES,CURITI,'SAN GIL',OIBA,SOCORRO,LANDAZURI,SUCRE,COROZAL,'SAN RAFAEL',TOLU,'CARMEN DE APICALA',FALAN,ICONONZO,CHAPARRAL,ROVIRA,LERIDA,'PUERTO CARRENO',ANZOATEGUI,RIOBLANCO,VILLAHERMOSA,TRUJILLO,'LA UNION','EL AGUILA',CALIMA,GINEBRA,SEVILLA,'PUERTO LOPEZ',PRADERA,ALCALA,RIOFRIO,'CARTAGENA DEL CHAIRA',QUINCHIA,ZARZAL,TRINIDAD,ARAUQUITA,TAME,YUMBO,'SAN PEDRO',MONTERREY,CUMARAL,'PUERTO ASIS',MESETAS,'EL RETORNO','PUERTO CONCORDIA','SAN LUIS DE PALENQUE','LA URIBE',PARATEBUENO,'SAN CARLOS GUAROA','FUENTE DE ORO','LA MACARENA','PUERTO CAICEDO','VILLA DEL ROSARIO','PUERTO COLOMBIA',NEIRA,VIJES,GUAPI,'PUERTO LEGUIZAMO','SAN PABLO','PALMAR DE VARELA',PIENDAMO,SESQUILE,CHARALA,URIBIA,MANATI,CAMPOHERMOSO,BELTRAN,MARMATO,GALAPA,CONVENCION,GIRARDOTA,CACHIPAY,RIOSUCIO,MAICAO,CHINACOTA,PINILLOS,CONTADERO,ZETAQUIRA,PEQUE,SOPLAVIENTO,SIMIJACA,PAUNA,MUZO,SUTAMARCHAN,'LA CRUZ','EL PASO','CIENAGA DE ORO',PASCA,PALESTINA,FUNDACION,GUAVATA,GUEPSA,MACARAVITA,SINCE,CUNDAY,'VALLE DE SAN JUAN',NATAGAIMA,CALAMAR,SOLANO,'SAN JUAN DE BETULIA',SABOYA,TARQUI,AQUITANIA,MACANAL,ULLOA,CERRITO,CHIMA,'PUERTO WILCHES',PURIFICACION,CUBARRAL,'SAN LORENZO',BELMIRA,MITU,'EL CARMEN DE BOLIVAR',BOAVITA,FRONTINO,NECHI,LURUACO,'SANTO TOMAS',URRAO,CUCAITA,TURBACO,BELEN,MARIPI,PAJARITO,TUTA,SOCHA,VENTAQUEMADA,PACORA,MERCADERES,INZA,VITERBO,SILVIA,'SAN DIEGO',MORALES,TAMALAMEQUE,'LA VEGA',GUATAVITA,LENGUAZAQUE,BITUIMA,GUADUAS,SAMANA,CHIPAQUE,ALGECIRAS,AIPE,TADO,CUCUNUBA,TAUSA,UNE,TESALIA,SUAZA,PITAL,MALLAMA,'EL TABLON',CORDOBA,CHACHAGUI,SANTACRUZ,TUQUERRES,ILES,GUACHUCAL,'SAN PEDRO DE CARTAGO',MARSELLA,'VILLA CARO',TEORAMA,'LA PLAYA','LOS PATIOS',GUACA,CACOTA,TOLEDO,COROMORO,HOUSTON,'SAN JOSE DE MIRANDA',MOGOTES,GALERAS,ALBANIA,'LA CUMBRE','SAN ANDRES',LEBRIJA,BARANOA,CAJAMARCA,URUMITA,COELLO,ALVARADO,ANDALUCIA,YOTOCO,'SAN JUAN DEL CESAR',VILLAGARZON,'BELEN DE LOS ANDAQUIES','VIGIA DEL FUERTE','SAN ROQUE','SAN BENITO',PINCHOTE,MACHETA,'LA BELLEZA',VALDIVIA,BRICENO,PESCA,'DON MATIAS','PUERTO PARRA','SAN LUIS DE GACENO',BUESACO,SORA,'EL CASTILLO',ARMERO,CASABIANCA,ATACO,DAGUA,OCAMONTE,ARGELIA,HONDA,'SAN FRANCISCO',TOCAIMA,CONCORDIA,FILANDIA,BARRANCAS,'LOS ANDES','SAN CAYETANO','EL ZULIA',YONDO,SIACHOQUE,CURILLO,'LOS PALMITOS',TIERRALTA,TASCO,CHITA,TENZA,MARQUETALIA,SIBUNDOY,TORIBIO,'SAN SEBASTIAN',VERGARA,QUETAME,NARI&#65533;O,ELIAS,'ALTO BAUDO',BARAYA,YACUANQUER,OPORAPA,APIA,GUAITARILLA,'EL PENON',TARSO,TAMESIS,CARCASI,ZIPACON,'SAN JUAN DE URABA',MORELIA,ROLDANILLO,VIOTA,CAICEDO,HELICONIA,'SAN LUIS',SABANAGRANDE,MALAMBO,MIRAFLORES,RONDON,BERBEO,TURMEQUE,'NUEVO COLON',TUTASA,'SAN PABLO DE BORBUR',SANTANA,TIBASOSA,TOPAGA,MONGUA,SUPIA,'PUEBLO BELLO',CERETE,MANTA,PANDI,UBALA,TELLO,HOBO,'OLAYA HERRERA',GONZALEZ,SARDINATA,'PUEBLO RICO',CALIFORNIA,'SAN VICENTE DE CHUCURI',TOLUVIEJO,MANGOCHI,CORINTO,TORO,SANTIAGO,MAPIRIPAN,'CERRO SAN ANTONIO',NOVITA,'EL BANCO'}
@attribute 'Area Residencia' {Urbana,Rural}
@attribute Desplazado {Si,No}
@attribute 'Disc Auditiva' {'No Aplica',Sordera,Hipoacusia,'Sordoceguera Parcial','Sordoceguera Total'}
@attribute 'Disc Cognitiva' {'No Aplica',Autismo,'Transtorno de aprendizaje',Asperger,'S&#65533;ndrome de Down','Retardo Mental Leve'}
@attribute 'Disc Fisica' {'No Aplica','Dificultades en la Coordinaci&#65533;n',Cuadriplejia,'P&#65533;rdida de extremidades',Hemiplejia}
@attribute 'Disc Mental' {'No Aplica',Psicosis,Esquizofrenia}
@attribute Enfermedad {Ninguna,Diabetes,'Insuficiencia Renal','VIH / Sida',C&#65533;ncer}
@attribute 'Conocio UNAD' {'Pagina Web de la UNAD','Estudiante de la Unad',Internet,Televisi&#65533;n,'Familiar o Amigo no relacionado con la UNAD','Egresado de la Unad',Radio,'Publicidad Impresa','Aviso en prensa','Funcionario de la unad','Instituci&#65533;n en Convenio','Feria acad&#65533;mica',Otro}
@attribute Deporte {'No Aplica','Tenis de mesa','Futbol Once','Ciclo monta&#65533;ismo',Ciclismo,'Artes Marciales',Otro,Voleibol,Baloncesto,Pesas,'Futbol Sala',Rugby,Coleo,Softbol,Patinaje,Gimnasia,Atletismo,Ajedrez,Nataci&#65533;n,Porrismo,Monta&#65533;ismo,'Tenis de campo',Bicicross,Parkour}
@attribute Instrumento {'No Aplica',Bater&#65533;a,Congas,Guitarra,Tromb&#65533;n,Bajo,Caja,Timbales,Piano,Trompeta,Pandero,Guacharaca,Flauta,Otro,Acorde&#65533;n,Quena,Bongo,Viola,Viol&#65533;n,Lira,Clarinete,Tambor,Saxo,Saxof&#65533;n,Arpa,&#65533;rgano,G&#65533;ira,Bandola,Arm&#65533;nica}
@attribute 'Actividad Artistica' {'No Aplica',Canto,Fotograf&#65533;a,Danza,Pintura,'Interpretaci&#65533;n de Instrumento Musical',Teatro,Dibujo,Escultura}
@attribute 'G&#65533;nero Danza' {'No Aplica',Merengue,Salsa,'Folcl&#65533;rica Colombiana',Otro,Tango}

@attribute Edad numeric
@attribute 'Tipo Instituci&#65533;n' {Privado,P&#65533;blico}
@attribute Modalidad {Presencial,Validaci&#65533;n,Distancia,Virtual}
@attribute 'Rendimiento Academico' {Sobresaliente,Aceptable,Deficiente}
@attribute 'Grados Perdidos' {Cero,Uno,'Cinco o m&#65533;s',Cuatro,Dos,Tres}
@attribute '&#65533;ltimo Nivel Alcanzado' {Tecnol&#65533;gico,T&#65533;cnico,'No Aplica',Profesional}
@attribute 'Modalidad Estudios' {'No Aplica',Presencial,'Distancia / Virtual'}
@attribute 'Graduaci&#65533;n Estudios Cursados' {No,Si,'No Aplica'}
@attribute 'Raz&#65533;n Abandono Estudios' {'Insatisfacci&#65533;n con la Universidad','No Aplica','Falta de recursos econ&#65533;micos','Cierre de Instituci&#65533;n o Programa',Enfermedad,'Insatisfacci&#65533;n con el programa cursado','Problemas Familiares','Traslado de ciudad o pa&#65533;s','Falta de tiempo para estudiar','Servicio Militar','Embarazo o parto',Otra,'Bajo Rendimiento Acad&#65533;mico'}
@attribute 'Tomado Cursos Virtuales' {Si,No}
@attribute 'Aprobado Cursos Virtuales' {'Poca motivaci&#65533;n por la modalidad','No Aplica','Dificultades para la comprensi&#65533;n de los materiales'}
@attribute 'Raz&#65533;n Sin Estudiar' {'No Aplica','Falta de tiempo','Estaba satisfecho con los t&#65533;tulos alcanzados','Falta de recursos econ&#65533;micos','No quer&#65533;a estudiar','Condici&#65533;n de desplazamiento','No sab&#65533;a en donde estudiar',Enfermedad,'No sab&#65533;a que estudiar','Situaciones familiares','Servicio Militar',Otra,Interno}
@attribute 'Raz&#65533;n Ingreso al Programa' {'Proyeccion laboral','Actualizaci&#65533;n de conocimientos',Vocacion,'Nivel acad&#65533;mico','Posibilidad de homologaci&#65533;n','Ascenso laboral','Prestigio de la carrera','Posibilidad de crear empresa',Otra,'Duraci&#65533;n de los estudios','Influencia familiar'}
@attribute 'Vivienda Actual' {Arriendo,Familiar,'Lugar de Trabajo',Propia}
@attribute Tics numeric
@attribute nivelIngl {Suficiente,Sobresaliente,Insuficiente}
@attribute deserto {0,1}