TiES Revista de Tecnología e Innovación en Educación Superior

MINERÍA DE DATOS: IDENTIFICANDO CAUSAS DE DESERCIÓN EN LAS INSTITUCIONES PÚBLICAS DE EDUCACIÓN SUPERIOR DE MÉXICO

https://doi.org/10.22201dgtic.26832968e.2019.2.4

Fredy Jesús López Pedraza
Ma. del Consuelo Macías González Edgar R. Sandoval García
https://www.ties.unam.mx/

Fecha de recepción: 29 de junio de 2019 • Fecha de publicación: octubre de 2019 Octubre de 2019 | número de revista 2 • ISSN 2683-2968

Fredy Jesús López Pedraza, Ma. del Consuelo Macías González, Edgar R. Sandoval García
https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

MINERÍA DE DATOS: IDENTIFICANDO CAUSAS DE DESERCIÓN EN LAS INSTITUCIONES PÚBLICAS DE EDUCACIÓN SUPERIOR DE MÉXICO

Resumen

La deserción escolar es un grave problema al que tienen que hacer frente las Instituciones Públicas de Educación Superior, lograr que todos los alumnos concluyan sus estudios es una tarea compleja; cuando se aplica de manera adecuada el proceso de minería de datos a la información de las Instituciones Educativas, permite determinar patrones e identificar las causas de deserción del alumnado, para evitarlas y conseguir que un mayor número de estudiantes termine su formación; por lo anterior es necesario entender las diferentes metodologías que existen, qué técnicas son las que mejor se adaptan al problema, así como utilizar herramientas informáticas que brinden resultados acordes al objetivo de los proyectos.

Conocer las investigaciones realizadas en Instituciones públicas de México, la forma en que fueron implementadas y los logros alcanzados, proporciona un panorama de la situación del problema; no obstante, el sector educativo es un campo muy extenso por explorar, encontrar nuevas variables que incidan en la deserción escolar permitirá implementar acciones concretas para evitarla.

Palabras clave:

Minería de datos, deserción escolar, educación superior.

DATA MINING: IDENTIFYING CAUSES OF STUDENT DROPOUT IN THE PUBLIC HIGHER EDUCATION INSTITUTIONS OF MEXICO

Abstract

Student dropout is a serious problem faced by public higher education institutions, to make that all students complete their studies is a complex task; when the data mining process is properly applied to the information of the Educational Institutions, allows to determine patterns and identify the causes of student dropout, to avoid them and get a greater number of students to finish their studies; so it is necessary to know the different methodologies that exist, which techniques are best adapt to the problem, as well as to use computer tools that provide results according to the objective of the project.

To know the research made in public institutions in Mexico, the way in which they were implemented and the results achieved, gives an overview of the current situation of the problem; however, the educational sector is a very extensive field to explore, to find new variables that affect the student dropout will allow institutions implement concrete actions to avoid it.

Keywords:

Data Mining, Student Dropout, Higher Education.

TiES Revista de Tecnología e Innovación en Educación Superior

Minería de datos: identificando causas de deserción en las instituciones públicas de educación superior de México
Fredy Jesús López Pedraza, Ma. del Consuelo Macías González, Edgar R. Sandoval García
https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

MINERÍA DE DATOS: IDENTIFICANDO CAUSAS DE DESERCIÓN EN LAS INSTITUCIONES PÚBLICAS DE EDUCACIÓN SUPERIOR DE MÉXICO

Introducción

LOS ESTUDIOS REALIZADOS, ENFOCADOS AL ABANDONO ESCOLAR SON DIVERSOS, LA MAYORÍA TRATA DE identificar las causas que lo originan, implementando estrategias para abatir el fenómeno, algunas con mejores resultados que otras, pero al final ninguna ha sido suficiente.

En los últimos 20 años, el sector educativo mexicano se ha vinculado de manera más estrecha con otras ramas del conocimiento, particularmente con las tecnologías computacionales, aprovechando las ventajas que estas representan en cuanto a procesamiento masivo de información, minería de datos y aprendizaje automatizado, a fin de encontrar información inédita que ayude a comprender mejor las causas por las que los estudiantes no concluyen sus estudios. Los resultados obtenidos dan mayor claridad sobre el problema y generan conocimiento para ser empleado en el establecimiento de nuevas acciones con miras a erradicarlo.

Desarrollo

LA DESERCIÓN ESCOLAR HA SIDO Y ES UNO DE LOS GRANDES PROBLEMAS QUE AFECTAN A LAS INSTITUCIONES de educación superior, con el objeto de encontrar una solución que la disminuya, las universidades han explorado nuevas opciones, principalmente en el área de minería de datos; en el desarrollo de este artículo se presenta la información relacionada con el tema, iniciando con los conceptos básicos, sus metodologías, técnicas y algoritmos, así como las herramientas de software más empleadas por los profesionales del sector, para concluir con el análisis de las investigaciones realizadas en las Instituciones Públicas de Educación Superior de México y los resultados alcanzados por éstas.

Educación Superior en México

LA *LEY GENERAL DE EDUCACIÓN DE MÉXICO*, EN SU ARTÍCULO 37, ESTABLECE TRES TIPOS DE EDUCACIÓN: básica, media superior y superior; el tipo superior está integrado por el técnico superior universitario, licenciatura y posgrado [1]; a continuación se muestra la matrícula del ciclo 2017-2018 para conocer los indicadores registrados.

La SEP[1] determina que los indicadores a medir en Educación Superior son la *absorción y cobertura*. El primero se define como el número de alumnos de nuevo ingreso al grado inicial de un nivel educativo, por cada cien egresados del nivel y ciclos inmediatos anteriores [3]; para el ciclo 2017-2018, se alcanzó el 74% [2]. La *cobertura* es el número total de alumnos inscritos en un nivel educativo al inicio del ciclo escolar, por cada cien del grupo de población con la edad reglamentaria para cursar ese nivel [3], es decir, ¿qué porcentaje de

1 Secretaría de Educación Pública

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

Figura 1.
Matrícula del ciclo 2017-2018.
Fuente: Adaptada de [2].

personas con la edad requerida para estudiar determinado nivel realmente lo están haciendo?; en el ciclo de referencia, la matrícula escolarizada y mixta, incluyendo posgrado, fue de 3,864,995 estudiantes, lo cual equivale al 29.5% del total de la población de 18 a 23 años, dicho porcentaje representa la cobertura del nivel [2].

Deserción en Educación Superior

POR OTRA PARTE, LA *DESERCIÓN* ES EL NÚMERO DE ALUMNOS QUE ABANDONAN SUS ESTUDIOS ANTES DE TERMINAR algún grado o nivel educativo [3]. Si bien se ha definido el término, no es indicador del tipo superior. Según [4], hay dificultades para explicar la deserción, ya que presenta variantes particulares que la complican por su carácter longitudinal. Para reforzar el dicho, el autor plantea *¿A partir de qué momento se considerará desertor?, ¿Son desertores los alumnos que lo hacen desde que se inscriben, o debe ajustarse el dato de inscripción?*, las cuales a casi dos décadas de su planteamiento no han sido resueltas.

Aunque no hay una información oficial que permita confrontar la deserción con otros países, estudios de la OECD[2] señalan que México tiene la proporción más baja de adultos con un título de educación superior, logrando apenas el 17%, cifra inferior al promedio que es del 37%, y debajo de otros países de la región donde la media es del 21%; también se reconocen avances en el nivel superior, señalando que en los últimos 16 años el porcentaje de adultos jóvenes que finalizaron educación superior pasó del 17% al 23% y prevé que en el futuro el 26% de los jóvenes mexicanos cuenten con título de nivel superior [5].

La ANUIES[3] , en [6], propone como meta para 2024 reducir al 6%[4] la tasa nacional de abandono en licenciatura y técnico superior, aunque no precisa cómo se determinan dichas cifras ni cómo alcanzarlas.

Sin duda, la *deserción* se mantiene como una situación de alarma por resolver para las instituciones de educación superior. Para atenderla, es preciso conocer

2 Organización para la Cooperación y el Desarrollo Económicos.

3 Asociación Nacional de Universidades e Instituciones de Educación Superior.

4 En el ciclo 2017-2018 la línea base fue del 8.3%

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

las causas que la originan e implementar acciones que la minimicen. Conseguir lo anterior no resulta sencillo, a pesar de que se han implementado variadas estrategias; por ello, en fechas relativamente recientes, las Instituciones han considerado el conocimiento que se extrae de los grandes volúmenes de datos y la información de valor que de ella se genera.

Actualmente, las instituciones exploran alternativas para atender la deserción, buscando respuestas en grandes volúmenes de datos, sin embargo es tal la cantidad de información que el proceso de analizarla, e interpretarla de manera manual, resulta muy tardado y costoso. A continuación se presentan los procesos para apoyar e identificar patrones o causas de deserción.

Big Data

EL TÉRMINO SE TRADUCE COMO DATOS MASIVOS, Y SE REFIERE A TAL CANTIDAD DE INFORMACIÓN QUE, el proceso de analizarla e interpretarla de manera manual, resulta muy tardado y costoso [7]. En [8] se señala respecto a Big Data, que en un principio los datos habían aumentado tanto, que aquellos que se examinaban ya no cabían en la memoria de los ordenadores para poder procesarlos, por lo que hubo que modernizarlos.

En [9] precisan que no solo se refiere al volumen de la información, sino también a la *variedad* del contenido y a la *velocidad* con la que se genera, almacena y analiza, lo cual se conoce como las 3V. En [10] refieren que varían según las características de las organizaciones, para unas, prima el *volumen*; para otras es la *velocidad*; y otras consideran mejor la *variabilidad* de las fuentes. Es claro, que representa un activo de valor, por lo que es más frecuente almacenar datos. Sin embargo, el beneficio se obtiene cuando se procesan adecuadamente, se identifican patrones, tendencias y limitantes de la información. SAS Institute señala que los datos fluyen de todas partes a velocidades y volúmenes nunca vistos, pero tomar decisiones eficientes no depende de la cantidad, de hecho, tener tantos, puede ser un obstáculo [11].

Data Mining[5]

ES UN PROCESO DE SELECCIÓN, EXPLORACIÓN, MODIFICACIÓN, MODELIZACIÓN Y VALORACIÓN DE LOS DATOS con el objetivo de descubrir patrones desconocidos o no detectados a través de procesos manuales, incluso se pueden utilizar para predecir comportamientos futuros [11]. Por su parte en [12] refiere que descubre relaciones, tendencias, desviaciones, comportamientos atípicos, patrones y trayectorias ocultas, con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento; [10] indica que es un proceso que utiliza técnicas estadísticas, matemáticas, inteligencia artificial y de aprendizaje automático para extraer e identificar información útil que se convierte en conocimiento a partir de grandes bases de datos; además de que puede realizar dos operaciones básicas: predecir tendencias y comportamientos y/o identificar patrones desconocidos.

Metodologías

ES NECESARIO IMPLEMENTAR METODOLOGÍAS APROPIADAS AL OBJETIVO DEL PROYECTO Y ACORDES CON los datos a manipular para minarlos. Su utilización permite realizar el proceso en forma sistemática y no trivial, al proveer una guía para la planificación y ejecución del proyecto, estableciendo fases, tareas a realizar y cómo llevarlas a cabo [13]. En 2014, KDnuggets[6] realizó una encuesta [14] donde preguntó a profesionales de datos ¿qué metodología principal utilizaron en el último año para realizar sus proyectos de análisis, minería de datos o ciencia de datos?, los resultados se muestran en la siguiente imagen, posteriormente se describen las tres más elegidas.

- *CRISP-DM*[7] , tiene un proceso de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación. La sucesión de fases no es necesariamente rígida, cada una de ellas es descompuesta en varias tareas generales de segundo nivel que se proyectan a tareas específicas [13]. Esta metodología no solo garantiza la adecuada planeación sino una mayor efectividad de los resultados [15].
- *SEMMA*[8] , creada por SAS Institute, se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos, se enfoca especialmente en aspectos técnicos, excluyendo actividades

5 Minería de Datos

6 Sitio web líder en inteligencia artificial, análisis de datos, minería de datos, ciencia de datos y aprendizaje automatizado

7 Cross Industry Standard Process for Data Mining

8 Sample, Explore, Modify, Model and Assess

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

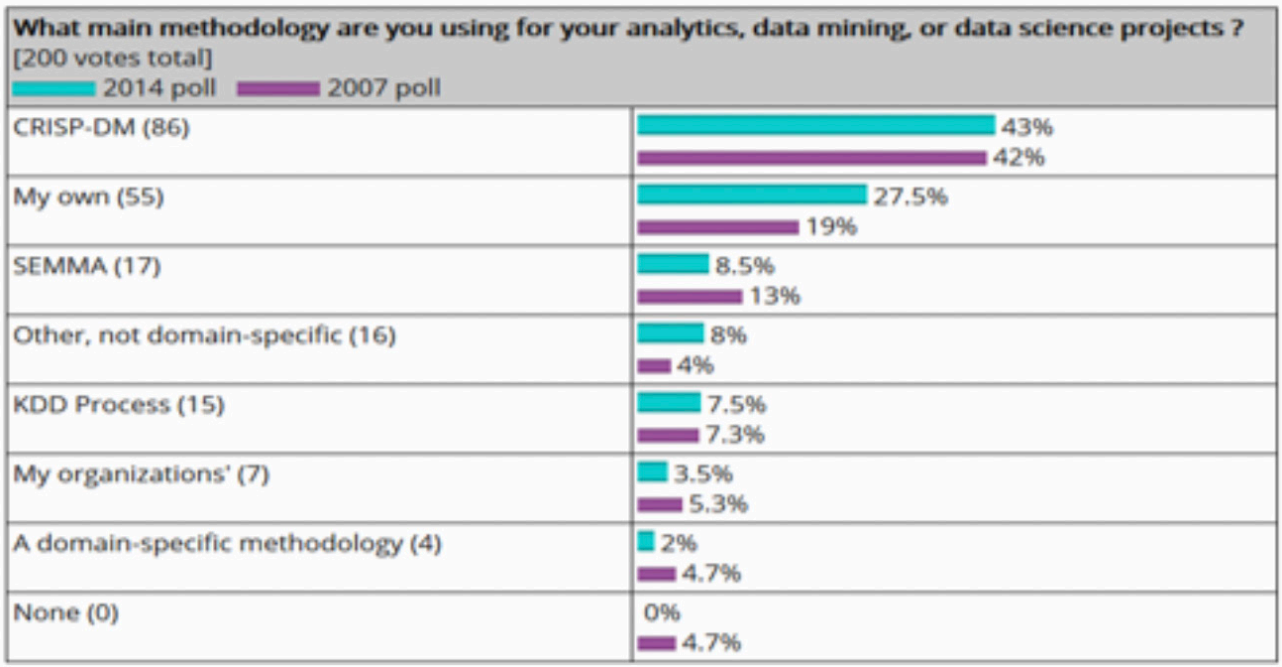

Figura 2.
Metodologías más utilizadas en 2014 para realizar proyectos de análisis, minería de datos o ciencia de datos.
Fuente: Tomada de [4].

de análisis y comprensión del problema que se está abordando [13]. Sus fases son selección, exploración, limpieza, transformación, minería de datos, evaluación y difusión [16].

- *KDD*[9] , tiene sus orígenes hace más de tres décadas. Ya que fue la primera en aparecer, es común que se utilice como sinónimo de minería de datos por lo que se utiliza mayormente para hacer referencia al proceso completo de descubrimiento de conocimiento [13]. El proceso de *KDD* consiste en transformar información de bajo nivel en conocimiento de alto nivel, es interactivo e iterativo, considera las etapas de comprensión del dominio de aplicación, extracción de los datos objetivo, preparar los datos, minería de datos, interpretación y utilización del conocimiento descubierto [17].

Técnicas y Algoritmos

LAS TÉCNICAS Y ALGORITMOS INTENTAN OBTENER MODELOS O PATRONES A PARTIR DE LOS DATOS recopilados, constituyendo el enfoque conceptual para extraer la información y ser implementadas por algoritmos [9]; en [18] señala de que establecen modelos utilizando datos de ejemplo o experiencias pasadas. De este modo identifican patrones o regularidades y se construyen buenas aproximaciones al problema.

La elección de la técnica viene determinada por dos condicionantes: el tipo de datos y el objetivo que se quiera lograr [17]. En [9] se clasifica en dos categorías[10] : supervisadas o predictivas y no supervisadas o descriptivas. Las predictivas se utilizan para prever el valor de un atributo de un conjunto de datos, denominado etiqueta, conocidos otros atributos, a partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos, con esta relación se predicen datos donde la etiqueta es desconocida. En contraparte, las descriptivas ayudan a la comprensión, tratando de ordenar los ejemplos en determinado orden, según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial

9 Knowledge Discovery in Databases

10 Considerando que hay modelos predictivos que también pueden ser descriptivos y los modelos descriptivos también pueden emplearse para realizar predicciones, esta clasificación principalmente señala el propósito para el que son más utilizadas estas técnicas

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

clase. Este es el proceder de sistemas que realizan *clustering* conceptual y de los que adquieren nuevos conceptos como los de asociación.

Los algoritmos permiten desarrollar las técnicas paso a paso, por esto es indispensable un entendimiento de alto nivel, para comprender sus parámetros y características para preparar los datos a analizar [9]. En [19] precisan que un algoritmo es un conjunto de heurísticas, cálculos y operaciones que permiten crear un modelo a partir de determinados datos, a través de un gran número de iteraciones se determinan los parámetros óptimos para crear el modelo.

En la encuesta 2016 de [20], se preguntó ¿Qué algoritmos utilizaron en los últimos 12 meses para una aplicación real relacionada con ciencia de datos? Los resultados se muestran en la siguiente imagen.

Los algoritmos de regresión logística son un tipo de análisis estadístico orientado a la predicción de una variable categórica en función de otras variables consideradas como parámetros predictores [21]; destaca que el valor a predecir es numérico de tipo dicotómico; la regresión lineal, se utiliza frecuentemente con variables cuantitativas, y son considerados algoritmos de predicción.

Los algoritmos de *clustering*, agrupamiento o segmentación, parten de una medida de proximidad entre individuos y a partir de ahí, buscan los grupos más parecidos entre sí, según una serie de variables mesuradas [21]. Los árboles de decisión, considerados de clasificación, permiten dividir datos en grupos basados en los valores de las variables; determinan las variables más significativas para un elemento dado, el mecanismo base consiste en elegir un atributo como raíz y desarrollar el árbol según esas variables [16].

Software

En 2018, KDnuggets realizó la 19ª encuesta anual para conocer que software de análisis, minería de datos y ciencia de datos usaron los expertos en el desarrollo de sus proyectos, participaron más de 2,300 votantes de todo el mundo, quienes eligieron de una lista de más de 90 programas. El reporte en [22] destaca los 11 programas más utilizados que se muestran en la siguiente imagen.

- *Python* es un lenguaje de programación de alto nivel, interpretado y multipropósito [23]; cuenta con facilidades para la programación orientada a obje-

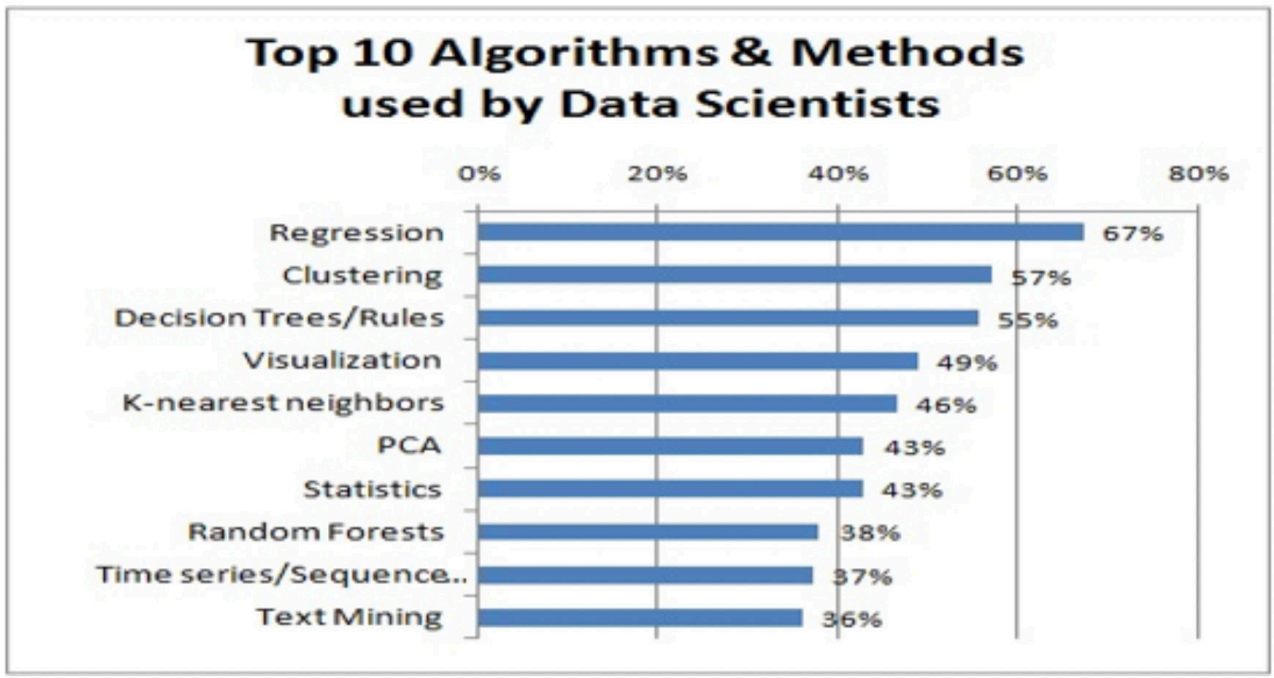

Figura 3.
Algoritmos más utilizados en 2015 para una aplicación real relacionada con ciencia de datos.
Fuente: Tomada de [20].

Fredy Jesús López Pedraza, Ma. del Consuelo Macías González, Edgar R. Sandoval García
https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

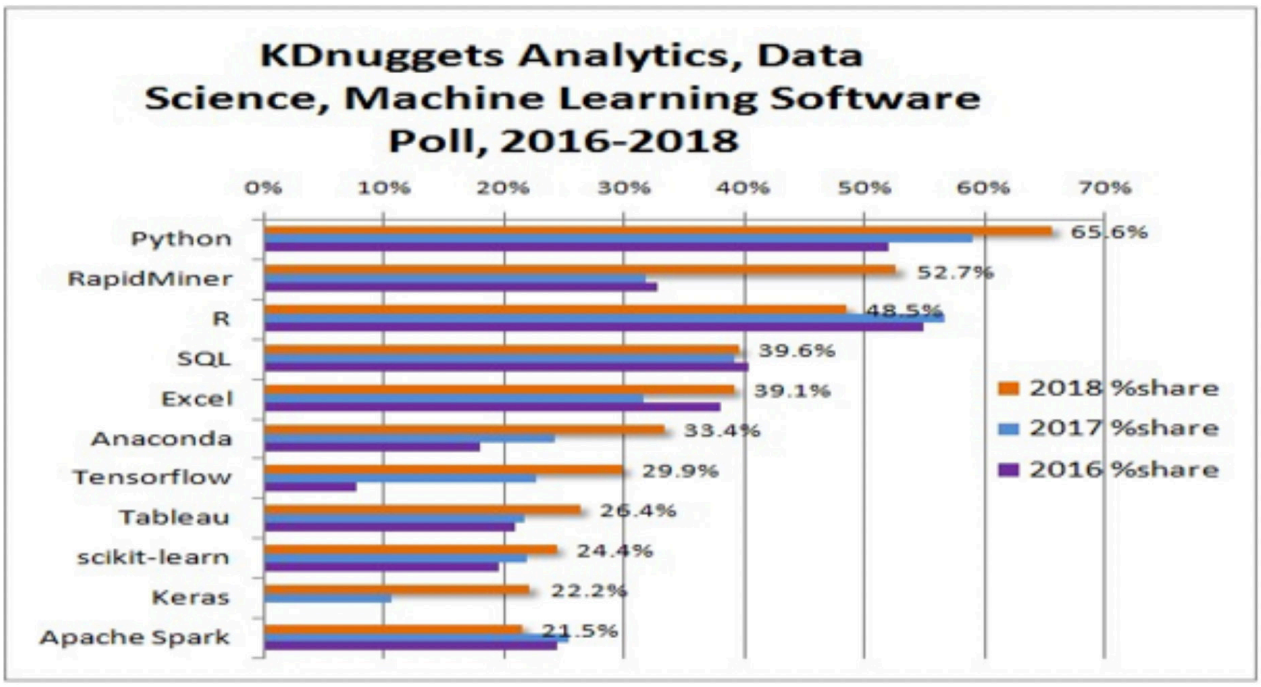

Figura 4.
Principales programas de análisis, minería de datos y ciencia de datos que usaron los expertos en el desarrollo de sus proyectos en 2018.
Fuente: Tomada de [22].

tos, imperativa y funcional, por lo que se considera un lenguaje multiparadigmas [24]. Los motivos principales del creciente uso de *Python* son las numerosas librerías con las que cuenta y su integración con aplicaciones como *MongoDB, Hadoop* o *Pentaho* [25].

- *RapidMiner* es un entorno que contiene procedimientos de *data mining* y aprendizaje automático, el proceso puede hacerse mediante operadores arbitrariamente anidados, descritos en ficheros XML y creados con la interfaz gráfica de usuario, también integra esquemas de aprendizaje y evaluadores de atributos del entorno Weka[11] y esquemas de modelización estadística de *R-Project* [9].
- *R* es un lenguaje de programación de código abierto con un entorno de programación apto para cálculos estadísticos y gráficos. Es muy popular para manipular algoritmos con datos no estructurados, ofrece gran variedad de estadísticas y técnicas gráficas [10]. Dispone de muchos paquetes para la creación de gráficos que le aportan capacidades avanzadas en la visualización de datos y resultados del análisis [25]. Abarca multitud de campos y permite combinar diferentes funciones para análisis más complejos [26].

Estudios realizados

La minería de datos ha tenido presencia en diferentes estudios con la finalidad de encontrar causas de la deserción escolar. En el caso particular de las Instituciones de Educación Superior públicas de México son distintas las investigaciones y muy variados los resultados alcanzados. A continuación, se abordan algunos de los más representativos.

- El primero a analizar es el caso del Instituto Tecnológico de Roque, publicado en el documento *Prototipo de Minería de Datos en la detección oportuna de Estudiantes en riesgo de Abandono Escolar GUÍA (Gestión Universitaria Integral del Abandono)* [27]. Su objeto es el desarrollo de un prototipo web que permita predecir de manera temprana que estudiantes se en-

11 Waikato Environment for Knowledge Analysis, programa computacional

Fredy Jesús López Pedraza, Ma. del Consuelo Macías González, Edgar R. Sandoval García
https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

cuentran en mayor riesgo de abandonar sus estudios. Cabe precisar que esta investigación aún se encuentra en desarrollo y lo referido en este documento son los avances y las acciones a realizar en el futuro. Para identificar las causas de abandono escolar, los autores utilizan el análisis estadístico desarrollado en el proyecto Alfa GUÍA[12] , y lo sistematizaron por medio del prototipo web para aplicarlo a los estudiantes del primer año de la carrera de Ingeniería en Tecnologías de la Información y Comunicaciones en la Institución antes referida. La encuesta consta de 34 preguntas que miden 82 variables las cuales se clasifican en cinco categorías, denominadas factores, individual, académico, sociocultural, económico e institucional. A pesar de que el estudio no está concluido, presentan los resultados obtenidos del análisis estadístico, y determinan que dentro de la muestra (no se especifica el tamaño), existe una probabilidad de deserción escolar del 24%. También señala que los factores que más inciden en la deserción son el académico, el económico y el individual. Por último, los investigadores precisan que la siguiente fase de su investigación considera la utilización de técnicas de minería de datos para determinar patrones que pudieran no haber sido identificados mediante el análisis estadístico.

- Por otra parte, en el Centro Universitario UAEM[13] Valle de México, realizaron un *Análisis Comparativo de Algoritmos de Minería de Datos para Predecir la Deserción Escolar* [28] a fin de determinar qué algoritmo de clasificación obtiene mejores resultados para predecir la deserción escolar. En esta investigación, mediante la metodología KDD, se analizaron las calificaciones por asignatura obtenidas durante el primer año de estudios de los alumnos de Ingeniería en Sistemas y Comunicaciones, de las generaciones 2008, 2009 y 2010. Los algoritmos comparados fueron árboles de decisión y bayesianos, particularmente, ID3[14] , C4.5[15] (J48), Naive Bayes Tree, Naive Bayes, y Redes Bayesianas[16] , con apoyo del software Weka; las pruebas se realizaron en dos bloques, uno para los datos de tipo nominal y otro para los datos de tipo numérico. En ambos casos se hicieron 3 experimentos variando los datos, todos los modelos fueron evaluados mediante validación cruzada de 10 pliegues, que divide el conjunto de datos en diez partes, utilizando nueve partes para entrenamiento y una para prueba. Los resultados obtenidos concluyen que es posible obtener un modelo de predicción de la deserción escolar fiable, utilizando las calificaciones obtenidas por los alumnos en el primer año de sus estudios. Los mejores algoritmos son Naive Bayes Tree y J48, ya que el primero es, desde el punto de vista cuantitativo, el que tiene menor error al momento de clasificar; si se considera aspecto cualitativo, el árbol generado por el algoritmo J48 provee mejor información y de mayor utilidad pues precisa las asignaturas que influyen en mayor medida al abandono de los estudios por parte del alumno.
- Otro caso a considerar es el realizado en una Institución de Educación Superior del Estado de México, no se precisa el nombre de la misma, donde un grupo de investigadores realizó el *Diseño de un Modelo predictivo aplicando Minería de Datos para identificar causas de Deserción Estudiantil Universitaria* [29], llamado PredATIS, el cual se basa en reglas de clasificación y selección de atributos, y cuyo objetivo fue identificar patrones relacionados con los aspectos de mayor influencia en la deserción estudiantil. La investigación se realizó mediante un análisis exploratorio, correlacional y explicativo, a partir del cual se creó un modelo de minería de datos; los datos empleados para el entrenamiento es una muestra de 170 estudiantes de un programa educativo (no se especifica cual), del tercer período escolar del año 2017, en la cual consideraron 39 variables agrupadas en 6 factores: personales, vocacionales, académicos, socioeconómicos, de salud y otros. El entrenamiento del modelo se realizó con el software Weka, empleando algoritmos de árboles de decisión J48 y REPTree[17] , así como reglas de clasi-

12 Proyecto de la UNESCO para la mejora de los índices de permanencia de los estudiantes de Enseñanza Superior

13 Universidad Autónoma del Estado de México

14 Es utilizado en la construcción de árboles de decisión, principalmente para aspectos de inteligencia artificial

15 Es empleado para generar un árbol de decisión de forma recursiva, su uso primordial es en técnicas de clasificación

16 Son clasificadores estadísticos que determinan la probabilidad de que una instancia pertenezca a una clase determinada

17 Permite construir un árbol de decisión, considerado de aprendizaje de decisión rápida

Fredy Jesús López Pedraza, Ma. del Consuelo Macías González, Edgar R. Sandoval García
https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

ficación basadas en JRIP[18] , OneR[19] y ZeroR[20] ; los resultados evidencian que el algoritmo J48 permitió identificar de mejor manera las causas que más influyen en la deserción, precisando que un estudiante está en riesgo de baja si tiene planes de matrimonio, ha presentado exámenes extraordinarios de las asignaturas técnicas del perfil, su mayor tiempo libre lo ocupa en realizar actividades culturales, deportivas o de entretenimiento, así como dedicar poco tiempo en estudiar o visitar bibliotecas, adicional a que la carrera elegida no fue su primera opción y es obligado por sus padres o por la cercanía a la universidad. Finalmente, cuando se compararon los resultados con el reporte real de la Institución se obtuvo un porcentaje de 90% de clasificaciones correctas.

- En el Tecnológico de Estudios Superiores de Jocotitlán se realizó una investigación presentada en el documento *Minería de datos aplicada para la identificación de factores de riesgo en alumnos* [30], cuyo objetivo fue implementar un sistema que realiza más eficiente el proceso de tutorías. Para esto los investigadores analizaron información de 831 estudiantes de las generaciones de 2008 a 2013 de la carrera de Ingeniería en Sistemas Computacionales. Tomando como base la metodología KDD, emplearon técnicas de *clustering* y reglas de asociación, particularmente los algoritmos K-Means[21] y A priori[22] . Para determinar las reglas de asociación utilizaron datos sobre práctica de deporte, problemas económicos, si trabajan o no, promedio de bachillerato, interrupciones en sus estudios y el campo de acción de la carrera; para el agrupamiento emplearon datos como la carrera elegida, tratamientos médicos, dependencias económicas, si están casados o tienen hijos y conocimiento de programas de becas. Posteriormente realizaron el preprocesamiento de datos, la aplicación de técnicas de minería de datos hasta llegar a la interpretación y evaluación de los resultados. La investigación indica que con las reglas de asociación obtuvieron un panorama general de la situación de los alumnos estudiados y las causas probables por las que abandonan sus estudios, destacando que: los alumnos que practican algún deporte tienen problemas económicos; la mayor parte de los estudiantes que desertan tienen problemas económicos; los desertores obtuvieron buenos resultados en niveles de estudio previos, remarcando que la mayoría de los que desertan nunca interrumpieron sus estudios en los niveles básico y medio superior. Con la aplicación del algoritmo de K-Means señalan algunas premisas de los resultados, mas no concluyen si las mismas son factores en el abandono de los estudios.
- Por último, está la investigación realizada en la Universidad Tecnológica de Izúcar de Matamoros (UTIM) donde en el documento *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos* [31]exponen los resultados obtenidos, que concluyen en una herramienta que permite calcular la probabilidad de deserción de cada uno de los estudiantes de la UTIM, apoyando de esta manera el proceso de tutorías de dicha institución. Si bien los autores no especifican explícitamente la metodología que siguen, señalan las fases implementadas, mismas que corresponden a la metodología KDD; utilizando como fuente de datos los resultados del EXANI–II[23] de los alumnos inscritos y de los alumnos que causaron baja en la universidad, llevaron a cabo la selección, limpieza y transformación de los datos. Posteriormente realizaron la clasificación de los mismos, donde emplearon un árbol de decisión mediante el algoritmo C4.5 y el método de aprendizaje de los *k* vecinos más cercanos. A continuación generaron los modelos para los dos algoritmos y con la ayuda del software Weka se probaron ambos con diferentes variables, obteniendo que el modelo del árbol de clasificación tuvo una mejor precisión con un porcentaje del 98.98%, mientras que los *k* vecinos más cercanos apenas superó el 70%. Los autores refieren en los resultados que los alumnos desertan por tres causas: la edad, ya que tiene que ver con la madurez y perspectiva de futuro de los estudiantes; los ingresos familiares, para alum-

18 Admite la poda incremental reducida, para producir reducción de errores

19 Basado en el algoritmo ID3, clasifica desde el conjunto de datos de entrenamiento

20 Predice la clase mayoritaria, si es nominal, o el valor promedio, si es numérico

21 Algoritmo de clasificación no supervisada, que agrupa objetos en k grupos basándose en sus características

22 Permite encontrar eficientemente conjuntos de ítems frecuentes que sirven de base para generar reglas de asociación

23 Examen Nacional de Ingreso a la Educación Superior

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

nos de 18 años o menos, puesto que dependen de los ingresos familiares; la tercera causa es el nivel de inglés, para alumnos mayores de 18 años. Para minimizar estas causas, se implementó una aplicación web, a través de la cual los tutores determinan el factor de riesgo de manera oportuna, y dan seguimiento a aquellos estudiantes vulnerables. En este punto vale la pena precisar que, si bien en todas las investigaciones analizadas sugieren estrategias para minimizar las causas de la deserción escolar, en su mayoría programas de tutoría o de seguimiento académico, en ninguna se refieren los resultados alcanzados de dichas acciones, tampoco señalan si las acciones sugeridas realmente están logrando disminuir los índices de abandono escolar. Esta falta de conclusiones puede deberse a que tres de los cinco estudios analizados fueron publicados recientemente, en los dos últimos años para ser exactos, otro más en 2013 y el más antiguo es de 2010.

Conclusión

La educación superior en México es amplia y compleja, presenta problemas de distintas magnitudes, unas simples y otras más complicadas cómo la cobertura, la absorción, la reprobación y la deserción. No obstante, es obligación del Estado garantizar el derecho a la educación de los mexicanos. A partir de la recopilación realizada, se concluye que hay muchas posibilidades para implementar minería de datos en el sector educativo; existen diferentes metodologías, técnicas, algoritmos y herramientas de software para ser empleadas en el proceso que en conjunto facilitan el análisis de la información y generan conocimiento que da mayor certeza a la toma de decisiones y reduce los problemas.

Respecto a la deserción escolar, desde siempre se ha sabido que es un problema multifactorial, esta premisa cobra mayor fuerza cuando se comparan los resultados de los estudios y se observa que los mismos determinan los factores que lo originan, encontrándose que principalmente son de tipo personal, académico y socioeconómico.

Adicionalmente, es importante destacar que ninguno de los casos analizados precisa resultados de las acciones implementadas después de haber identificado las causas de origen de la deserción; esta ausencia de información, aunque no se señala, bien puede deberse a que las investigaciones principalmente son realizadas por docentes, quienes carecen de la autoridad para implementar estrategias institucionales al respecto, otro factor puede ser el alcance de la investigación, y en todos los casos las investigaciones se limitan a identificar causas de deserción y proponer acciones que pueden disminuirla, sin llegar a la implementación y menos aún evaluar los resultados de su aplicación.

Otro punto importante a comentar es referente al enfoque que se da a los estudios, encontrándose que en todos los casos, las variables analizadas estuvieron directamente relacionadas con los estudiantes, lo que induce a pensar que en la mayoría de los casos, los orígenes del problema involucran a los alumnos; sin embargo, habrá que preguntarse si esto siempre es cierto o hay factores no relacionados con el alumno que influyen en su deserción.

Para reafirmar o desmentir lo anterior, los que escriben desarrollan una investigación de grado, cuyo objetivo es determinar, mediante técnicas de minería de datos, si las características, perfil y desempeño de los docentes influyen en la deserción de los estudiantes de educación superior. La investigación antes referida está en la fase de desarrollo, y aun no cuenta con resultados que puedan comentarse en este documento. A referencia se comenta para identificar una arista más del problema de deserción y sentar el precedente de la investigación que se realiza.

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

BIBLIOGRAFÍA

[1] Gobierno de los Estados Unidos Mexicanos, *Ley General de Educación*, México, 2018.

[2] Dirección General de Planeación, Programación y Estadística Educativa, "Sistema Educativo de los Estados Unidos Mexicanos, Principales Cifras 2017-2018," *Secretaría de Educación Pública*, México, 2018

[3] Dirección General de Planeación y Programación, "Glosario. Términos utilizados en la Dirección General de Planeación y Programación," *Secretaría de Educación Pública*, México, 2008.

[4] F. Martínez Rizo, "Estudio de la eficiencia en cohortes aparentes," *Deserción, rezago y eficiencia terminal en las IES. Propuesta metodológica para su estudio*, Serie Investigaciones, México, ANUIES, 2001.

[5] OECD, "Higher Education in Mexico: Labour Market Relevance and Outcomes," *OECD* Publishing, Paris, 2019.

[6] ANUIES, "Visión y Acción 2030, Propuesta de la ANUIES para renovarla educación superior en México," *Publicaciones ANUIES*, México, 2018.

[7] IBM, "¿Qué es Big Data? Todos formamos parte de ese gran crecimiento de datos," junio 18, 2012. [En línea]. Disponible en: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html. [Consultado en abril 9, 2019].

[8] V. Mayer-Schönberger y K. Cukier, *Big Data. La revolución de los datos masivos*, Madrid: Editor digital Titivillus, 2013.

[9] J. García, J. M. Molina, A. Berlanga, *et al., Ciencia de datos. Técnicas analíticas y aprendizaje estadístico*. Bogotá: Alfaomega Colombiana, 2018.

[10] L. Joyanes Aguilar, *Big Data: Análisis de grandes volúmenes de datos en organizaciones*. México: Alfaomega Grupo Editor, 2013.

[11] SAS Institute, "La Minería de Datos de la A a la Z: Cómo Descubrir Conocimientos y Crear Mejores Oportunidades," 2015. [En línea]. Disponible en: https://www.sas.com/es_mx/whitepapers/data-mining-from-a-z-104937.html. [Consultado en abril 22, 2019].

[12] B. Beltrán Martínez, *Puebla: Benemérita Universidad Autónoma de Puebla*. Facultad de Ciencias de la Computación, 2019.

[13] S. Gordillo, A. S. Haedo y J. M. Moine, "Estudio comparativo de metodologías para minería de datos," *XIII Workshop de Investigadores en Ciencias de la Computación*, Argentina, 2011.

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

[14] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," *KDnuggets*, 2014. [En línea]. Disponible en: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. [Consultado en febrero 3, 2019].

[15] J. L. Cendejas Valdez, M. Á. Acuña López, G. Cortes Morales y G. Bolaños Jiménez, "El uso de modelos y metodologías de minería de datos para la inteligencia de negocios," *Revista de Sistemas Computacionales y TIC's*, vol. 3, no. 8, pp. 54-63, junio 2017.

[16] M. Pérez Marqués, *Minería de datos a través de ejemplos*. Madrid: RC Libros, 2014.

[17] J. C. Riquelme, R. Ruiz y K. Gilbe, "Minería de Datos: Conceptos y Tendencias", *Revista Iberoamericana de Inteligencia Artificial*, vol. 10, no. 29, pp. 11-18, 2006.

[18] J. T. Palma Méndez y R. Marín Morales, *Inteligencia Artificial: Métodos, técnicas y aplicaciones*. España: McGraw-Hill / Interamericana de España, S.A.U., 2008.

[19] Microsoft, "Algoritmos de minería de datos (Analysis Services: Minería de datos)," abril 4, 2018. [En línea]. Disponible en: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017 [Consultado en abril 04, 2019].

[20] G. Piatetsky, "Top Algorithms and Methods Used by Data Scientists," *KDnuggets*, 2016. [En línea]. Disponible en: https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html. [Consultado en febrero 23, 2019].

[21] T. Aluja, "La Minería de Datos entre la Estadística y la Inteligencia Artificial," *QUESTIIO*, vol. 25, no. 3, pp. 479-478, 2001.

[22] G. Piatetsky, "Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis," *KDnuggets*, 2018. [En línea]. Disponible en: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html. [Consultado en abril 3, 2019].

[23] A. Fernández Montoro, *Python 3 al descubierto*. Madrid: RC Libros, 2012.

[24] I. Challenger Pérez, Y. Díaz Ricardo y R. A. Becerra García, "El lenguaje de programación Python," *Ciencias Holguín*, vol. XX, no. 2, pp. 1-13, 2014.

[25] P. Rochina, "Python vs R para el análisis de datos," noviembre 16, 2016. [En línea]. Disponible en: https://revistadigital.inesem.es/informatica-y-tics/python-r-analisis-datos/. [Consultado en mayo 20, 2019].

[26] L. A. Vargas, J. H. Farfán, F. Aramayo, et al., "Comparación de las principales herramientas de Data Mining y Análisis de Sábanas Telefónicas," *II Segunda jornada Argentina de Tecnología, Innovación y Creatividad*, 2016.

[27] Y. D. Guzmán Islas, E. Ramos Ojeda y A. Guzmán Zazueta, "Prototipo de Minería de Datos en la detección oportuna de estudiantes en riesgo de abandono escolar GUÍA (Gestión Universitaria Integral del Abandono)," *Pistas Educativas*, vol. 39, no. 129, pp. 64-79, 2018.

[28] M. Quintana López, J. C. Trinidad Pérez, S. J. Morales Escobar, *et al.*, "Análisis Comparativo de Algoritmos de Minería de Datos para Predecir la Deserción Escolar," *Research in Computing Science*, vol. 67, pp. 13-23, 2013.

[29] P. N. Maya Pérez, J. R. Aguilar C., R. A. Zamora R., *et al.*, "Diseño de un Modelo predictivo aplicando Minería de Datos para identificar causas de Deserción Estudiantil Universitaria," *Strategy, Technology & Society*, vol. 7, no. 2, pp. 11-39, 2018.

https://doi.org/10.22201/dgtic.26832968e.2019.2.4
Octubre 2019 • Número 2 • ISSN 2683-2968
https://www.ties.unam.mx/

[30] A. Reyes-Nava, A. Flores-Fuentes, R. Alejo, *et al.*, "Minería de datos aplicada para la identificación de factores de riesgo en alumnos," *Research in Computing Science*, no. 139, pp. 177-189, 2017.

[31] S. Valero Orea, A. Salvador Vargas y M. García Alonso, "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," *Recursos digitales para la educación y la cultura, volumen Kaambal*, Mérida, Yucatán, Universidad Tecnológica Metropolitana, Mérida, Yucatán, México y Universidad de Cádiz, Andalucía, España: 2010, pp. 33-39.

Cómo se cita

F.J. López, M.C. Macías y E.R. Sandoval, "Minería de datos: identificando causas de deserción en las instituciones públicas de educación superior de México," *TIES, Revista de Tecnología e Innovación en Educación Superior*, n.o. 2, octubre, 2019. [En línea]. Disponible en: https://www.ties.unam.mx/ [Consultado en octubre, 2019].