Educational data mining for the analysis of student desertion

Emilcy J. Hernández-Leal[1], Diana Patricia Quintero-Lorza[1], Juan Camilo Escobar-Naranjo[1], Juan Sebastián Ramírez-Gómez[1], Néstor D. Duque-Méndez[1]

[1] Universidad Nacional de Colombia, Sede Manizales
ejhernandezl@unal.edu.co, dpquinterol@unal.edu.co,
jcescobarn@unal.edu.co, jsramirezgo@unal.edu.co,
ndduqueme@unal.edu.co

Abstract. Student desertion is one of the phenomena that concerns the most to educational institution's directives in the different academic levels. In the particular case of the higher education, such a phenomenon is relevant, to identify the characteristics or factors most relevant that influence the problem, have a high value. This current paper shows the results of the analysis on student desertion in the Universidad Nacional de Colombia – headquarters Manizales, applying educational data mining techniques. To execute the sample, a dataset with 655 records was used in the test, gathered from first academic semester (2009) to the second academic semester (2014), and an open source software named Weka whose version is 3.8.2. Moreover, classification techniques such as decision trees and induction rules were used; In addition, attributes selection techniques were applied in order to reduce the dimensionality of the set of initial data. The results of the classification algorithms with the initial data set are shown, and also, their modifications are compared to throw the best models obtained.

Keywords: Student desertion, Educational data mining, Classification algorithms.

1 Introducción

En el campo académico, se puede definir la deserción como una situación a la que se enfrenta un estudiante cuando no logra cumplir o completar un proyecto o proceso educativo [1]. La deserción estudiantil es uno de los problemas que más preocupa a los sistemas educativos y en particular a las directivas de las instituciones de educación. Este fenómeno, ha sido estudiado desde décadas atrás [2], es una problemática que se presenta en los diferentes niveles educativos, desde la básica primaria, pasando por la secundaria, media y en la educación superior [3], [4], [5].

Existen diferentes categorizaciones para la deserción estudiantil, una clasificación para esta situación en educación superior, propone dos tipos, la deserción respecto al factor tiempo y la deserción respecto al factor espacio. Para el primer tipo, se hace a su vez una sub-división en: deserción precoz, esta se refiere a los estudiantes que son admitidos, pero no llegan a realizar la matrícula; deserción temprana, se entiende cuando el

estudiante abandona los estudios en los primeros cuatro semestres y deserción tardía, se da cuando el estudiante abandona el programa a partir del quinto semestre. Para el segundo grupo, respecto al factor espacio, se dividen en: deserción interna, se considera cuando el estudiante abandona el programa académico, pero para cambiar por otro programa que ofrece la misma institución, es decir, hace un traslado de programa; deserción institucional, se da cuando el estudiante abandona una institución para iniciar estudios en otra y finalmente, la total deserción del sistema educativo, como se indica, el estudiante no retoma sus estudios en dicho nivel académico [6].

El fenómeno de la deserción universitaria afecta a instituciones en todo el mundo, en [7] se reporta que según la Oficina Regional de Educación para América Latina y el caribe – OREALC, para el año 2010, se tenía que solo uno de cada diez jóvenes entre los 25 a 29 años presentaba más de cinco años en educación superior; en el caso de la comunidad europea, particularmente para España en el año 2015, se comunicaba, según el Ministerio de Educación, Cultura y Deporte, que el 19% de los estudiantes desertan de la universidad en el primer año de estudio. Más en detalle, en [8] se presenta que según la OCDE (Organización para la Cooperación y el Desarrollo Económico) para el año 2012, los países que presentaban mayor abandono universitario eran México y Turquía con un 38%, luego estaban Suecia (36%) y Portugal (31%); mientras que los de menos deserción para ese año eran Alemania (4,03%), Finlandia (0,45%) y Países Bajos (0,7%).

En cuanto al análisis de los factores o causas asociados a la deserción estudiantil, existen aportes de múltiples autores que han creado modelos completos en torno al tema. Sin embargo, se puede encontrar que la mayoría coinciden en agrupar los factores frente a variables de tipo personal, familiar e institucional [6]. Dado lo anterior, los estudios sobre análisis y propuestas de modelos para predecir la deserción suelen usar atributos que se recolectan dentro del sistema educativo y que corresponden a estas características. En este sentido, las instituciones educativas cuentan con muchos de los datos necesarios y usados para estudiar esta problemática y poder hacer frente a ella.

La minería de datos, es un enfoque de análisis de datos bajo el cual se agrupan una serie de técnicas que se rigen por el aprendizaje supervisado y no supervisado y que son usadas para afrontar diferentes problemas que involucran el tratamiento de datos a partir de los cuales se pueden descubrir patrones, tendencias y en general extraer conocimiento [9]. La minería de datos puede ser aplicada a diferentes dominios, en el caso particular de los ambientes educativos se ha consolidado una corriente denominada minería de datos educativa, la cual se concentra en el estudio y análisis de los datos generados en este campo con técnicas de tipo descriptivo o predictivo [10]. La minería de datos educativa está fuertemente asociada con las analíticas de aprendizaje, aunque se pueden desarrollar por separado, se suelen encontrar algunos estudios en los que se complementan [11].

Teniendo en cuenta lo anterior, este trabajo se orientó a detectar los factores que influyen en el fenómeno de la deserción universitaria haciendo uso de la minería de datos

educativa en un caso de estudio de una institución de educación superior pública de la ciudad de Manizales en Colombia. En lo que sigue del documento, se presentan las siguientes secciones; en la sección 2 se describen algunos trabajos relacionados, en los cuales se presentan estudios sobre deserción estudiantil y que emplean minería de datos; en la sección 3 se describe el conjunto de datos y el pre-procesamiento que se realizó a estos; en la sección 4 se presentan las pruebas realizadas con la aplicación de las técnicas de minería de datos; en la sección 5 se reportan los resultados y finalmente en la sección 6 se concluye y presentan algunos planteamientos para trabajos futuros.

2 Trabajos relacionados

Como se mencionó anteriormente, la deserción estudiantil es un fenómeno que genera alta preocupación en las instituciones educativas de diferentes niveles, dado lo anterior, se encuentran un buen número de trabajos que incluyen el estudio de aspectos asociados a este fenómeno. A continuación, se traen a colación algunos trabajos relacionados.

Uno de los aspectos clave a la hora de estudiar la deserción o abandono estudiantil son las variables que se asocian y que pueden ayudar a explicar o incluso predecir dicha deserción. En [12] proponen agrupar las causas de la deserción en cinco variables: pérdida de semestre, dificultad financiera, ingreso al mercado laboral, otros intereses atraen al estudiante e indeterminado; el estudio se realiza en una facultad de Ingeniería de Sistemas de una institución de educación superior con una muestra de 707 sujetos; los resultados dejan ver que la causa predominante de la deserción para estos estudiantes son los factores agrupados bajo el apelativo de indeterminado; sin embargo, seguida de esta se encuentra la dificultad financiera, otros intereses en los estudiantes, la pérdida del semestre y por último el ingreso al mercado laboral. A partir de la investigación surgieron estrategias para mitigar este fenómeno, como el otorgamiento de bonos de matrícula, monitoria social, reliquidación de matrícula y acompañamiento individual por sicología.

Por otra parte, en [13] toman los datos de ingreso de los estudiantes a la institución de educación superior (personales y antecedentes educativos) y los datos que se generan durante el periodo de estudio y mediante la aplicación de técnicas de minería de datos, identifican los factores que influyen en la deserción. Los autores aplicaron algoritmos de clasificación (reglas y árboles de decisión) y encontraron que uno de los principales factores asociado a la deserción, en su caso de estudio, es la cantidad de asignaturas que aprueban los estudiantes en el primer año de carrera universitaria, también se destacan la edad de ingreso y la procedencia.

En [14] se propone un modelo predictivo para la deserción académica, para ello utilizaron como fuentes de datos: estadísticas de deserción de los últimos años en áreas del saber, zonificación y discriminación por programa, área del conocimiento, condiciones sociales, resultados pruebas ICFES, pensum y acuerdo académico en el que se encuentran los estudiantes. A partir de estas fuentes de información crearon una bodega de

datos para centralizar los registros y extrajeron un datamart para hacer los análisis aplicando un árbol de decisión. Los resultados encontrados indican que, para el caso de estudio, Facultad de Ingeniería de una universidad de la capital colombiana y con datos de los años 2009 a 2015, la cantidad de materias cursadas es un factor influyente en el modelo generado, asimismo, el género es un factor relevante, por otro lado, factores socioeconómicos también intervienen, ya que se encontró que los estudiantes que viven en localidades distantes a la ubicación de la facultad tienen mayor probabilidad de desertar de sus estudios.

Como lo indican en [15], el abandono o deserción universitaria es un problema que genera costos altos para el estudiante y en general para la sociedad; los autores analizaron varias metodologías utilizadas para la creación de modelos predictivos del fenómeno: análisis de correlaciones, análisis de regresión logística, análisis de supervivencia y minería de datos; además, deciden utilizar la primera metodología para llevar a cabo un estudio particular con datos de la cohorte de nuevo ingreso del año 2010/11 a la Universidad de Oviedo. El conjunto de datos estudiado, estaba formado por una muestra de 5215 estudiantes de los que, en septiembre de 2012, 4194 permanecían en la titulación inicialmente matriculada, 363 habían cambiado de titulación a otra dentro de la Universidad y 658 habían abandonado la Universidad de Oviedo. Los resultados obtenidos indican que hay una relación entre el rendimiento académico previo (nota de ingreso al programa de estudios) y la deserción, quienes tienen bajo rendimiento previo, presentan mayor riesgo de abandono de estudios. También encuentran relación entre la asistencia a clase y el fenómeno estudiado, puesto que tienen mayor probabilidad de permanecer en la institución los estudiantes que asisten con mayor frecuencia a clases.

En este orden de ideas, se encuentra que en los últimos años varios autores han coincidido en que las técnicas de minería de datos son una buena alternativa para abordar los análisis en cuanto al problema de la deserción universitaria, tanto para la identificación de las variables o factores que influyen, como para la generación de modelos predictivos que ayuden a prevenir el abandono. Entre las técnicas utilizadas se encuentran los árboles de decisión [16], [17]; los k vecinos cercanos [18]; redes bayesianas y reglas [19]. En [4] se muestra la aplicación de técnicas de minería para la predicción del fracaso y abandono escolar; los autores emplearon métodos de clasificación como reglas de inducción y árboles de decisión. Las pruebas fueron realizadas con un conjunto de datos reales de 670 estudiantes de educación media; es de destacar de este trabajo, que se incluye un paso de balanceo de datos, lo cual no se consideró en los demás trabajos previos revisados, este balanceo se realiza con el ánimo de equilibrar el número de registros de estudiantes desertores con los no desertores.

3 Conjunto de datos y pre-procesamiento

El conjunto de datos con el cual se realizaron las pruebas fue conformado con los registros aportados por la dirección académica de la Universidad Nacional de Colombia – sede Manizales. En total se logró tener una muestra de 655 registros y 71 atributos.

En la Tabla 1 se presentan las variables utilizadas. Los registros que conforman el dataset corresponden a los periodos 2009 – 2015 y provienen de tres fuentes de datos:

- Registros de admisión y matrícula
- Reporte de estudiantes bloqueados por cada periodo (semestre)
- Registros de cursos inscritos por periodo

Tabla 1. Atributos utilizados y su fuente

Fuente	Variable
Información de admisión y matrícula	Periodo, cód. facultad, plan, carrera, código, tipo doc., documento, año, mes, día, sexo, estado civil, creación, inicio estudios, tipo acceso, acceso, tipo subacc, subacceso, estrato, cód. depto., cód. mpio, nacionalidad1, extranjero, lugar residencia, becado, título pregrado, tipcolegio, jorcolegio, carcolegio, calendario, colegio-depto., col depto., tipo vivienda, vivienda, nro. hnos., pbm calculado, rcb pago, matricula, bienestar, sistematización, seguro, A1 pensión, B1 colegio, A2 estrato, B2 lugar res, B3 vivienda, A3 ingresos, B4 nro. hijos, aprobadas, homologadas, inscritas, cred. apro, cred adi, cred homo, cred ins, graduado, papa.
Información de estudiantes bloqueados por periodo	Documento, nombre programa, municipio nacimiento, cód. núm., causa.
Información de cursos inscritos por periodo	Documento, cód. dep, id assignatura, grupo, nom asignatura, créditos, calificación codalf, tipología, periodo materia, tco codalf, estado act.

Fuente: Elaboración propia

Una de las tareas que más tiempo y esfuerzo demandan en los procesos de minería de datos es la fase correspondiente al pre-procesamiento de los datos, también conocida como ETL (extraer, transformar y cargar por sus siglas en inglés), esta fase es requerida para poder organizar los datos originales de tal manera que se puedan aplicar los algoritmos particulares. En este orden de ideas, el primer paso que se siguió fue la integración de los datos provenientes de las tres fuentes descritas anteriormente, esta integración consiste en unificar en un solo dataset la información disponible de cada estudiante. La cantidad inicial de registros era de: 5690.

Posteriormente, se hizo un trabajo de limpieza para lograr que en el conjunto de datos solo se mantuvieran los registros de los estudiantes que contaban con el 100% de la información. Por ejemplo, si un estudiante, dentro de su información de admisión y matrícula, tenía un dato faltante, como el número de hermanos, entonces éste se excluía del conjunto de datos. En esta fase también se tuvo una dificultad al no contar con un diccionario de datos consolidado que permitiera identificar con claridad cada uno de los atributos, por ello, se tuvo que retirar del dataset algunos atributos que no se lograron identificar. A su vez, algunos atributos correspondían a la misma información, por ejemplo, se tenía el código del motivo de bloqueo y en otro atributo la descripción de dicho motivo, por lo cual, se eliminó el atributo con el código y se dejó el atributo categórico que contenía la descripción.

Al finalizar esta fase de ETL, se creó un fichero con formato .ARFF, que es el formato propio de Weka [20], software seleccionado para realizar las pruebas, dadas sus bondades en cuanto a algoritmos de minería de datos disponibles y facilidad de uso. Después de realizar estas tareas de pre-procesado, se cuenta con un dataset de 49 atributos de 655 estudiantes. Siguiendo los pasos sugeridos en [4] se aplicaron seis algoritmos de selección y a partir de los resultados arrojados por estos se seleccionaron los atributos con una frecuencia superior o igual a dos, es decir, si un atributo era seleccionado por al menos dos de los algoritmos, se decidía llevarlo al nuevo dataset; con lo cual se llegó a un segundo conjunto de datos con 20 atributos. En la tabla 2 se muestran los resultados de la aplicación de los seis algoritmos de selección por medio del software Weka.

Tabla 2. Mejores atributos según cada algoritmo de selección

Algoritmo	Atributos seleccionados
CfsSubsetEval: Evalúa el valor de un subconjunto de atributos al considerar la capacidad de predicción individual de cada característica junto con el grado de redundancia entre ellas.	Documento, periodo, código, tipo_doc, ano, tipo_acceso, vivienda, nro_hnos, papa, causa, tipología, tco_codalf
OneRAttributeEval: Evalúa el valor de un atributo mediante el uso del clasificador OneR.	Cred_apro, aprobadas, inicio_estudios, causa, código, nro_hnos, ano, periodo, documento, papa, nom_asignatura, cred_adi, vivienda, cod_facultad, carrera, plan, col_depto, nombre_programa
ReliefFAttributeEval: Evalúa el valor de un atributo al muestrear repetidamente una instancia y considerando el valor del atributo dado para la instancia más cercana de la misma clase y diferente. Puede operar en datos de clase discretos y continuos.	Causa, cred_apro, aprobadas, inicio_estudios, periodo, cred_adi, tipología, nom_asignatura, col_depto, ano, jorcolegio, carcolegio, nro_hnos
InfoGainAttributeEval: Evalúa el valor de un atributo midiendo la ganancia de información con respecto a la clase	Cred_apro, aprobadas, inicio_estudios, causa, nom_asignatura, código, documento, ano, periodo, nro_hnos, municipio nacimiento, cred_adi, papa, col_depto, tipología, nombre programa
GainRatioAttributeEval: Evalúa el valor de un atributo midiendo la relación de ganancia con respecto a la clase	Cred_apro, aprobadas, causa, ano, nro_hnos, papa, tipo_doc, documento, código, extranjero, inicio_estudios, periodo, cred_adi, nom_asignatura, estado_civil
SymmetricalUncertAttributeEval: Evalúa el valor de un atributo midiendo la incertidumbre simétrica con respecto a la clase.	Cred_apro, aprobadas, causa, ano, nro_hnos, documento, papa, código, inicio estudios, periodo, nom_asignatura, cred_adi, tipo_doc, municipio nacimiento, estado_civil

Fuente: Elaboración propia

4 Aplicación de minería de datos educativa

A continuación, se describirán los experimentos realizados. Para la aplicación de los algoritmos, se crearon 20 particiones, 15 de ellas con 33 registros y 5 de ellas con 32

registros, todas de manera aleatoria con los dataset de 20 y de 49 atributos. Con el fin de obtener unos resultados confiables, se realizaron los experimentos con 5 algoritmos de clasificación, en particular, tres de árboles de decisión y 2 de reglas de inducción.

Para conseguir una clasificación óptima, se decidió hacer un primer experimento en el que se empleó el dataset con 49 atributos. En un segundo experimento se emplearon los mismos algoritmos, pero sobre el dataset de los 20 mejores atributos escogidos previamente por los algoritmos de selección. Los algoritmos de clasificación aplicados se tomaron de los disponibles en Weka y se caracterizan por ser fácilmente interpretables, puesto que muestran los resultados en forma de reglas del tipo si – entonces o en ramas de árboles de decisión. En la Tabla 3 se muestran los resultados de la ejecución de los algoritmos con las particiones del dataset de 49 atributos, para llegar a estos resultados se calculó la media de las 20 ejecuciones. Las medidas empleadas son el RAE (Relative Absolute Error), el porcentaje de instancias correctamente clasificadas y el porcentaje de instancias incorrectamente clasificadas, teniendo en cuanta lo trabajado en [4].

Tabla 3. Validación cruzada utilizando los 49 atributos

Algoritmo	RAE (%)	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)
RandomTree	33.54	86.29	13.66
J48	11.47	**95.40**	4.56
REPTree	44.48	85.38	14.57
JRip	11.83	**95.39**	4.56
OneR	28.97	87.02	12.94

Fuente: Elaboración propia

Para el segundo experimento se han usado los ficheros con los mejores 20 atributos, los resultados de la validación cruzada, promedio de las 20 ejecuciones, se muestran en la Tabla 4, utilizando las mismas medidas que en la tabla anterior.

Tabla 4. Validación cruzada utilizando los 20 mejores atributos

Algoritmo	RAE (%)	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)
RandomTree	30.29	89.51	10.46
J48	11.47	**95.43**	4.56
REPTree	44.49	85.42	14.57
JRip	12.27	**95.12**	4.87
OneR	28.98	87.05	12.94

Fuente: Elaboración propia

5 Interpretación de resultados

Al revisar y comparar los resultados de las tablas 3 y 4, se encuentra que el algoritmo de árboles de decisión que más se destaca es el J48, mientras que para los algoritmos de reglas sobresale el JRip. En general, los algoritmos presentan una leve mejora

cuando se utiliza el dataset con los 20 mejores atributos, pero es casi imperceptible. Lo anterior puede indicar que los atributos que fueron desechados por los algoritmos de selección no generaban ruido en el conjunto de datos.

El algoritmo J48 es quien presenta mejores resultados con un porcentaje de instancias correctamente clasificadas del 95,40% (para los 49 atributos) y del 95,43% (para los 20 atributos) y su porcentaje de error absoluto relativo es del 11,47 en los dos casos. Lo sigue el algoritmo RandomTree con un porcentaje de aciertos del 86.29 y 89.51 respectivamente para los dos experimentos, sin embargo, su porcentaje de error absoluto relativo duplica el del J48, al ser de 30.29% en el mejor de los casos. Por su parte, el REPTree es el algoritmo de árboles con el rendimiento más bajo, al solo alcanzar un 85,4% de aciertos y tener un RAE de 44.49% para la ejecución con el dataset de mejores atributos.

Por otra parte, de los dos algoritmos de reglas de inducción utilizados, el JRip reportó los mejores resultados y en cuanto al porcentaje de aciertos muy similar al J48; no obstante, presentó una leve desmejora con el dataset de 20 atributos, con un porcentaje de instancias correctamente clasificadas del 95,12% frente a un 95,39% obtenido para la validación con el dataset de 40 atributos. Su porcentaje de RAE también tuvo un leve aumento, pasando del 11.83 al 12.27 del primer experimento al segundo. En cuanto al OneR, a pesar de que tuvo un resultado menos favorable que el JRip, presentó un porcentaje de error absoluto relativo menor que dos de los algoritmos de árboles de decisión, RandomTree y REPTree, con un 28.9%.

Ahora bien, si se revisan los mejores modelos de reglas y árboles encontrados por los algoritmos, se encuentra que para el caso de estudio, los atributos que más influencia presentan en el fenómeno de la deserción representado en la obtención o no del título profesional son el número de créditos aprobados, el PAPA (esta es una calificación utilizada en el interior de la institución que corresponde al Promedio Aritmético Ponderado Acumulado) , el periodo de inicio de estudios, estrato y la causa de finalización de estancia en la universidad.

6 Conclusiones y trabajo futuro

Los algoritmos de clasificación utilizados en este trabajo permitieron conseguir ciertos modelos con el objetivo de predecir la deserción estudiantil, es decir, si se alcanzará a obtener el título universitario o no. Como se comentó en las dos primeras secciones, el problema de la deserción no es fácil de abordar, dado que intervienen múltiples variables en él. Sin embargo, se ratifica que las técnicas de minería de datos educativa son una alternativa viable para el análisis de este fenómeno. Además, en la Universidad Nacional de Colombia sede Manizales, a pesar de tener sistemas de información consolidados, los datos no se habían utilizado para explorar causas o patrones en la deserción estudiantil y este trabajo abre un espacio para promover este enfoque y las tecnologías de análisis de datos.

Se debe resaltar que una de las tareas que exigió mayor esfuerzo y tiempo fue la correspondiente al pre-procesamiento de los datos, ya que de esta depende en gran parte la calidad de los análisis y la fiabilidad de los resultados. En general, se tuvo que recuperar desde tres fuentes diferentes, la información para conformar el conjunto de datos de prueba y también se lograron aplicar algunos algoritmos de selección para buscar los atributos que mejor describieran el modelo, sin embargo, los resultados posteriores llevan a pensar, que para el caso de estudio particular, el dataset inicial y el dataset con los atributos seleccionados no presentaron diferencias mayores, por lo cual, se considera que los atributos no eran redundantes y no generaban ruido.

Con los resultados que generan estos algoritmos de clasificación se puede hacer una selección de los grupos de estudiantes que presentan mayor tendencia a la deserción o que están en una situación de riesgo de abandono, con lo cual, se puede iniciar con ellos la aplicación de algún tipo de medidas preventivas por parte de los docentes o directivas de la institución educativa.

Se comprobó que en los procesos de ingreso y admisión a la institución se suelen tener muchos vacíos de información que hicieron que se tuviese que retirar una buena parte de registros del conjunto de datos. Lo anterior, puede limitar los resultados de esta investigación, por lo cual se recomienda que en la institución se de gran relevancia a la recolección de dicha información para posteriores periodos académicos.

Como trabajo futuro se plantea incluir un nuevo grupo de experimentos realizando balanceo de datos y análisis de costo de la clasificación como lo sugieren en [4]. Así mismo, se socializarán los resultados del estudio con las directivas de la institución, en particular con la dirección académica para iniciar procesos que permitan usar estos análisis en procesos de acompañamiento a estudiantes o planes de mejora y toma de acciones correctivas.

Agradecimientos

Al programa de Formación de Capital Humano de Alto Nivel para el Departamento de Norte de Santander en el marco de la Convocatoria N°753 de Colciencias.

Referencias

[1] V. Tinto, “Definir la Deserción: Una Cuestión de Perspectiva,” *Rev. Educ. Super.*, no. 71, pp. 33–51, 1982.

[2] V. Tinto, “Dropout from Higher Education: A Theoretical Synthesis of Recent Research,” *Rev. Educ. Res.*, vol. 45, no. 1, pp. 89–125, Mar. 1975.

[3] C. Gómez-Restrepo, A. Padilla Muñoz, and C. J. Rincón, “Deserción escolar de adolescentes a partir de un estudio de corte transversal: Encuesta Nacional de Salud Mental Colombia 2015,” *Rev. Colomb. Psiquiatr.*, vol. 45, pp. 105–112, Dec. 2016.

[4] C. Márquez Vera, C. Romero Morales, and S. Ventura Soto, “Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos,” *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 7, no. 3, pp. 109–117, 2012.

[5] A. Betancur Escobar, “El reto de la permanencia para las IE en Colombia,” *Rev. Reflexiones y Saberes*, vol. 3, no. 5, pp. 1–4, 2016.

[6] H. E. Viale Tudela, “Una aproximación teórica a la deserción estudiantil universitaria,” *Rev. Digit. Investig. en Docencia Univ.*, vol. 8, no. 1, pp. 59–75, 2014.

[7] N. L. Alvarez, Z. Callejas, D. Griol, and M. Durán Benejam, “La deserción estudiantil en educación superior: S.O.S. en carreras de ingeniería informática,” in *Conferencia Latinoamericana sobre el abandono en la educación superior CLABES*, 2017.

[8] G. Izquierdo Cázares and R. C. Mestanza Páez, “Retos de la educación ante la deserción escolar universitaria. Revisión sistemática,” *Rev. Científica Retos la Cienc.*, vol. 1, no. 2, pp. 15–21, Dec. 2017.

[9] M. A. Bramer, *Principles of data mining*. Springer, 2013.

[10] R. Jindal and M. D. Borah, “A Survey on Educational Data Mining and Research Trends,” *Int. J. Database Manag. Syst.*, vol. 5, no. 3, pp. 53–73, Jun. 2013.

[11] Z. Papamitsiou and A. A. Economides, “Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence,” *Educ. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, Oct. 2014.

[12] K. Azoumana, “Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos,” *Rev. Pensam. Am.*, vol. 6, no. 10, pp. 41–51, 2014.

[13] K. B. Eckert and R. Suénaga, “Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos,” *Form. Univ.*, vol. 8, no. 5, pp. 3–12, 2015.

[14] J. E. Sotomonte-Castro, C. C. Rodríguez-Rodríguez, C. E. Montenegro-Marín, P. A. Gaona-García, J. G. Castellanos, and J. G. Castellanos, “Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos,” *Rev. científica*, vol. 3, no. 26, p. 35, Oct. 2016.

[15] A. B. Bernardo Gutiérrez, R. Cerezo Menéndez, L. J. Rodríguez-Muñiz, J. C. Núñez Pérez, E. Tuero Herrero, and M. Esteban García, “Predicción del abandono universitario: variables explicativas y medidas de prevención,” *Rev. Fuentes*, no. 16, pp. 63–84, Jun. 2015.

[16] J. L. Aguirre Mendiola, R. M. Valdovinos Rosas, J. A. Velazquez, R. A. Eleuterio, and J. R. Marcial Romero, “Análisis de deserción escolar con minería de datos,” *Res. Comput. Sci.*, vol. 93, pp. 71–82, 2015.

[17] S. Formia, L. Lanzarini, and W. Hasperué, “Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio,” *Rev. Iberoam. Educ. en Tecnol. y Tecnol. en Educ.*, no. 11, pp. 92–98, 2013.

[18] Y. Marcano and R. Rodríguez, “Minería de datos aplicada a la deserción estudiantil,” *EDUCARE*, vol. 18, no. 2, pp. 31–51, 2014.

[19] K. Eckert and R. Suénaga, “Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos,” *Form. Univ.*, vol. 8, no. 5, pp. 3–12, 2015.

[20] University of Waikato, “Weka.” .