Artículos

Minería de datos para identificación de patrones de deserción estudiantil mediante técnicas de agrupamiento en una institución de educación superior

Data Mining to Identify Student Dropout Patterns Using Clustering Techniques at a Higher Education Institution

Oswaldo Villacrés Cáceres
Ingeniero en Sistemas Informáticos. Magíster en Gestión de bases de datos.
Escuela Superior Politécnica de Chimborazo
https://orcid.org/0000-0002-5894-5248
ovillacres@espoch.edu.ec

Byron Paúl Huera Paltán
Ingeniero en Sistemas Informáticos. Magíster en Sistemas de Telecomunicaciones
Escuela Superior Politécnica de Chimborazo
https://orcid.org/0000-0002-1721-2574
bhuera@espoch.edu.ec

Bryan Fabricio Flores Martínez
Ingeniero en Estadística Informática. Máster universitario en inteligencia de negocios
Escuela Superior Politécnica de Chimborazo
https://orcid.org/0009-0009-6644-3131
bryan.flores@espoch.edu.ec

Katherine Maribel Gallegos-Carrillo
Ingeniero en Sistemas Informáticos. Máster Universitario en Ingeniería de Software y Sistemas Informáticos
Escuela Superior Politécnica de Chimborazo
https://orcid.org/0000-0001-6882-8478
katherine.gallegos@espoch.edu.ec

Imaginario Social
ISSN: 2737-6362

Imaginario Social
Entidad editora
REDICME (reg-red-18-0061)

e-ISSN: 2737-6362
julio-diciembre Vol. 8-3-2025
http://revista-imaginariosocial.com/index.php/es/index

Recepción: 26 de febrero de 2025
Aceptación: 08 de junio de 2025

255-270

Resumen

Uno de los problemas que más preocupa a las instituciones de educación superior es la deserción estudiantil debido a sus gravísimas consecuencias. El estudio busca diseñar e implementar una arquitectura tecnológica conforme a un almacén de datos (Data Warehouse), procesos ETL y técnicas de agrupamiento que servirán para la detección de patrones de abandono de la Escuela Superior Politécnica de Chimborazo (ESPOCH). A nivel metodológico, se integró información académica y sociodemográfica, usando sql server e Integration Services, para el conjunto 106 006 instancias. Con esta base se aplicó el algoritmo K-means con el software WEKA y posteriormente se realizó el agrupamiento en cinco clústeres. Los resultados muestran que el primer clúster aglutinó la mayor cantidad registra con 29 232 casos (27,58 %), seguido por los clústeres 2 y 4, permitiendo identificar patrones de deserción asociados a la combinación específica de diferentes variables, tales como: edad, periodo, semestre, asistencia, género, etnia, estado civil, colegio de procedencia, promedio de bachillerato, carrera y zona geográfica. La minería de datos es una herramienta muy útil para convertir grandes volúmenes de información institucional en conocimiento. Esto, a su vez, permite la identificación oportuna de grupos vulnerables. Asimismo, puede ayudar en la toma de decisiones para reducir el abandono académico.

Palabras clave: Arquitectura federada, Bases de datos federadas, Desarrollo, Modelo evaluación, Indicadores.

Abstract

One of the most concerning problems for higher education institutions is student dropout due to its severe consequences. This study aims to design and implement a technological architecture based on a Data Warehouse, ETL processes, and clustering techniques to detect dropout patterns at the Escuela Superior Politécnica de Chimborazo (ESPOCH). Methodologically, academic and sociodemographic information was integrated using SQL Server and Integration Services, consolidating a dataset of 106,006 instances. Based on this, the K-means algorithm was applied using WEKA software, segmenting the data into five clusters. The results show that

the first cluster grouped the highest number of records with 29,232 cases (27.58%), followed by clusters 2 and 4. This allowed for the identification of dropout patterns associated with specific combinations of variables such as: age, academic period, semester, attendance, gender, ethnicity, marital status, high school of origin, high school grade point average (GPA), major, and geographic location. Data mining is a highly useful tool for converting large volumes of institutional information into knowledge. This, in turn, allows for the timely identification of vulnerable groups. Furthermore, it can support decision-making to reduce academic dropout.

Keywords: Federated architecture, Federated databases, Development, Evaluation model, Indicators.

La deserción estudiantil en la educación superior es un fenómeno complejo que no puede reducirse a una sola variable académica. Más bien, la deserción responde a la interacción de múltiples factores: características individuales, desempeño académico, nivel de integración institucional, condiciones sociales y particularidades del entorno educativo en el que se desenvuelve el estudiante.

Introducción

Desde la propuesta clásica de Tinto (1975), la permanencia ha sido entendida como un proceso dinámico, en el que la forma en que el estudiante se integra —tanto académica como socialmente— termina influyendo de manera decisiva en su decisión de continuar o abandonar sus estudios. El Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES), define la deserción o abandono estudiantil como la interrupción voluntaria o involuntaria de los estudios universitarios por parte del alumno antes de alcanzar la titulación (CACES, 2020). Además, el CACES en su Modelo de Evaluación Externa con Fines de Acreditación (2023), incorpora la deserción estudiantil como uno de los indicadores cuantitativos del criterio "Condiciones del personal académico, apoyo académico y estudiantes", reconociéndolo como un indicador crítico en la evaluación de la calidad institucional. Esta inclusión no solo tiene un valor técnico, sino también político y ético, pues obliga a las universidades a ir más allá de la simple cuantificación del fenómeno, exigiendo

una comprensión contextualizada y la implementación de estrategias concretas de retención, acompañamiento y equidad (CACES, 2023).

Con la informatización de las diferentes dependencias de una institución de educación superior (IES), gran parte de los procesos son gestionados mediante sistemas informáticos en los cuales se almacenan todas las interacciones. Los sistemas académicos, seguimiento a graduados, talento humano, financieros, sistemas para gestión datos sociodemográficos y plataformas educativas generan grandes cantidades de información susceptible de análisis. Adicionalmente, dado que Romero y Ventura (2013) consideran como un desafío tecnológico importante la integración y explotación de los datos almacenados en sistemas informáticos educativos y que proceden de diversas fuentes, formatos y niveles de granularidad; para Porras et al. (2023) el desarrollo de métodos computacionales para investigar datos relacionandos con el contexto educativo es la minería de datos educativa.
Una característica importante al utilizar minería de datos es que permite ir más allá del análisis manual tradicional, el cual suele quedarse corto frente a volúmenes masivos de información. García Sánchez y Gil Mateos (2023) consideran que para identificar y descubrir comportamientos y relaciones complejas, que a simple vista pasarían desapercibidos, en los grandes volúmenes de datos que generan las instituciones se lo debe realizar mediante la aplicación de procedimientos computacionales, optimizando el tiempo y el esfuerzo necesarios para identificar patrones y relaciones implícitas.

En términos metodológicos, el descubrimiento de conocimiento en bases de datos (KDD) comprende un conjunto de actividades que permiten transformar datos en conocimiento interpretable, en ese sentido, Fayyad et al. (1996) diferenciaron el proceso general de descubrimiento de conocimiento en bases de datos de la minería de datos llegando a ser una de las formulaciones fundamentales del proceso KDD.

Así mismo y puesto que la selección, limpieza y transformación de los datos pueden representar una parte sustancial del esfuerzo total de un proyecto analítico hacen que la selección de datos y el preprocesamiento se constituyan en una etapa particularmente importante como lo han mostrado Zhang et al. (2003) y estudios

posteriores sobre minería de datos. En este sentido, la calidad de los datos de entrada condiciona la utilidad de los patrones posteriormente descubiertos. La literatura metodológica también advierte que selección y preprocesamiento suelen consumir una parte importante del trabajo de minería de datos.

En el ámbito de la educación superior, la analítica de aprendizaje ha adquirido importancia como mecanismo para apoyar el éxito académico. Ifenthaler & Yau (2020) analizaron inicialmente 6.220 publicaciones y seleccionaron 46 publicaciones clave; encontraron evidencia de que distintas aproximaciones de analítica de aprendizaje que pueden apoyar la continuidad y finalización de los estudios, pero señalaron que todavía falta evidencia rigurosa y de gran escala sobre su efectividad.

En la misma línea, Asto-Lázaro & Bermejo-Terrones (2023) analizaron diversos estudios sobre machine learning enfocado en predecir la deserción escolar, destacando cómo ha crecido el interés por aprovechar estas herramientas computacionales para detectar a tiempo a los estudiantes que se encuentran en situación de vulnerabilidad.

Los trabajos más recientes refuerzan esta tendencia. Sandoval-Benavides y López-Ornelas (2024) realizaron una revisión sistemática sobre machine learning para predecir deserción universitaria y encontraron que los modelos avanzados pueden alcanzar altos niveles de precisión, aunque advirtieron problemas relacionados con la heterogeneidad y falta de estandarización de los estudios.

Por otra parte, la selección del algoritmo constituye otro elemento relevante. Los métodos de agrupamiento, particularmente K-means, permiten identificar grupos de objetos con características similares sin requerir necesariamente una variable objetivo previamente definida. Esto resulta apropiado cuando el propósito inicial consiste en descubrir estructuras o perfiles de estudiantes antes que construir un modelo de clasificación.

El agrupamiento es una alternativa exploratoria adecuada cuando la institución no dispone de una variable objetivo suficientemente consolidada o busca descubrir perfiles desconocidos y es un enfoque pertinente en IES que no necesariamente cuentan con sistemas analíticos consolidados pero que poseen información histórica

acumulada. La integración de información académica, sociodemográfica y administrativa puede permitir identificar combinaciones de características que merecen atención institucional.

En este contexto, el presente trabajo investigativo analiza la solución tecnológica implementada para identificar patrones de deserción estudiantil mediante minería de datos, enfatizando la arquitectura del almacén de datos, el proceso ETL, la preparación del conjunto de datos y la aplicación del algoritmo K-means, así como interpretar los patrones obtenidos desde una perspectiva de analítica institucional.

Metodología

Para la presente contribución científica se enfatiza la dimensión aplicada y tecnológica, específicamente en la implementación y evaluación del artefacto informático.

La implementación de la solución se realiza bajo la metodología Descubrimiento de Conocimiento en Base de Datos - Knowledge Discovery in Databases (por sus siglas en inglés KDD), que se estructura en las siguientes etapas:

1. Selección de datos
2. Preprocesamiento
3. Transformación
4. Minería de datos
5. Interpretación y evaluación

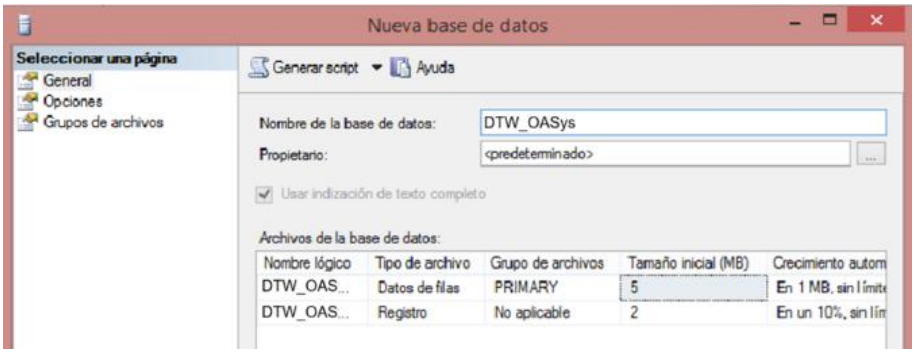
Para la selección de datos se consideraron los datos almacenados en el sistema informático OASys cuyo motor de bases de datos es SQL-SERVER 2008, esta decisión se la realizó pues es el sistema que almacena datos académicos y personales de los estudiantes.

Resultados

En la seleccionan de datos se identificaron 7 vistas de la base de datos: vEstudiante, vColegios, vEspecialidad, vMateria, vProfesor, vPeriodo y vDeserción.

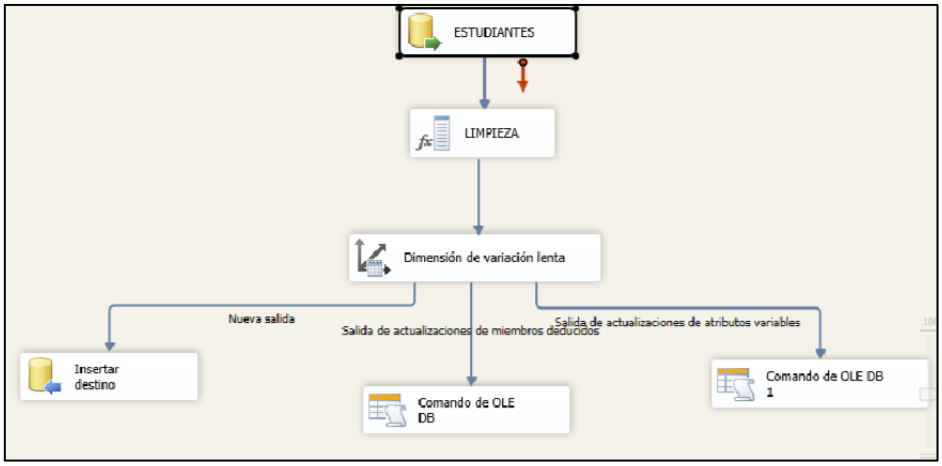
En el preprocesamiento se crea el base de datos para el data warehouse, como se observa en la **Error! Reference source not found.**.

Figura 1: Data Warehouse



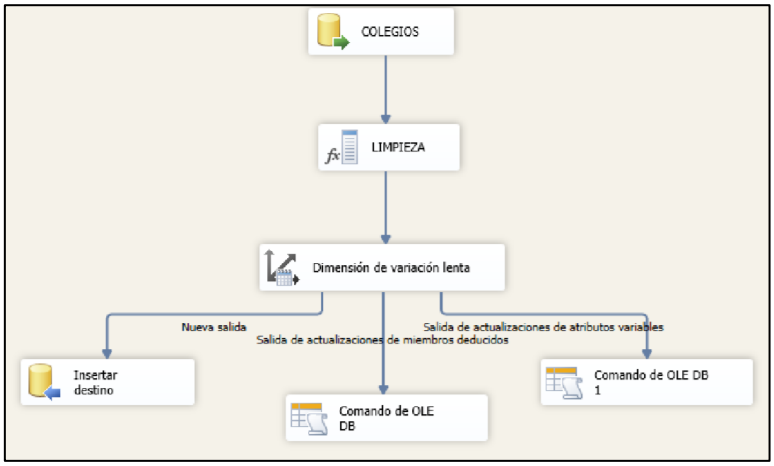
La **Error! Reference source not found.** presenta el ETL para la vista estudiantes y recuperar los datos: cedula, genero, etnia, discapacidad, fecha_nacimiento, estado_civil, titulo_colegio, especialidad_colegio, promedio_colegio y código_colegio:

Figura 2: ETL estudiante



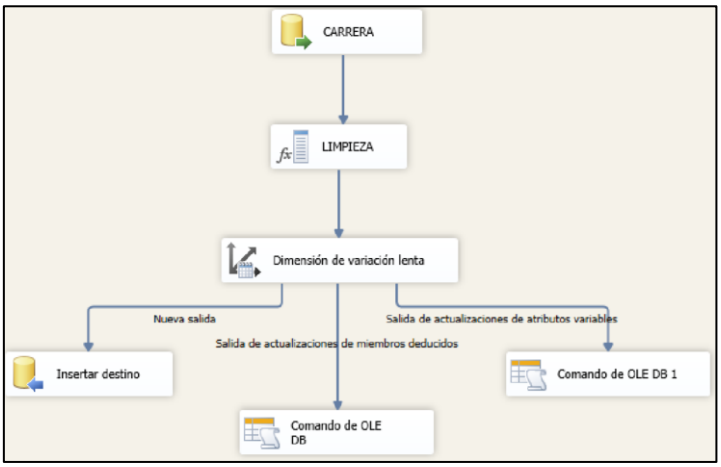
La **Error! Reference source not found.** representa el ETL para la vista colegio y recuperar los datos: codigo_colegio, colegio, ciudad, cantón y provincia.

Figura 3: ETL colegio



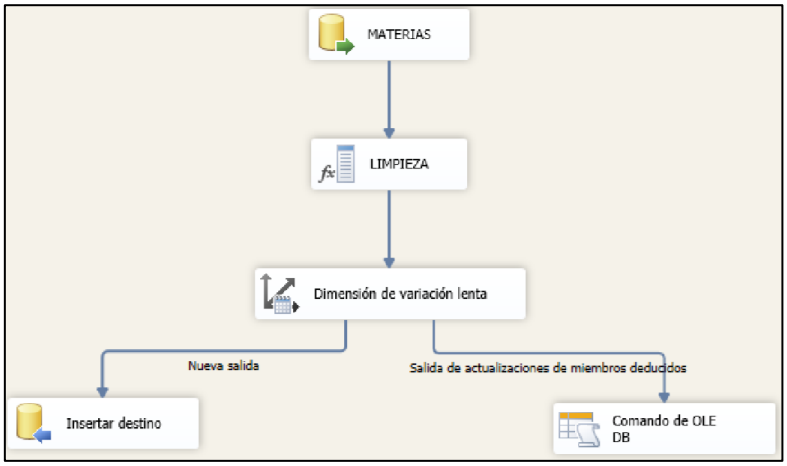
La **Error! Reference source not found.** representa el ETL para la vista carrera y recuperar los datos: código_carrera, carrera, facultad.

Figura 4: ETL carrera



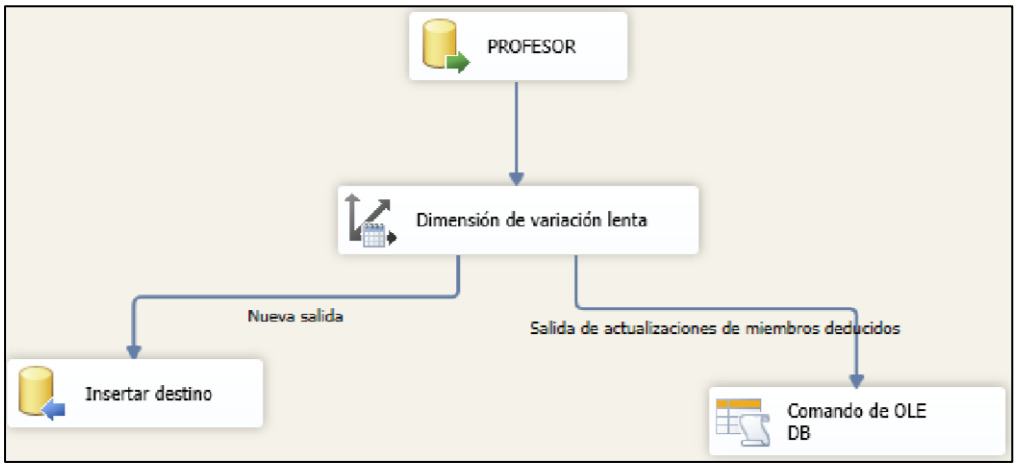
La **Error! Reference source not found.** representa el ETL para la vista materia y recuperar los datos: código_materia, materia y horas.

Figura 5: ETL materias



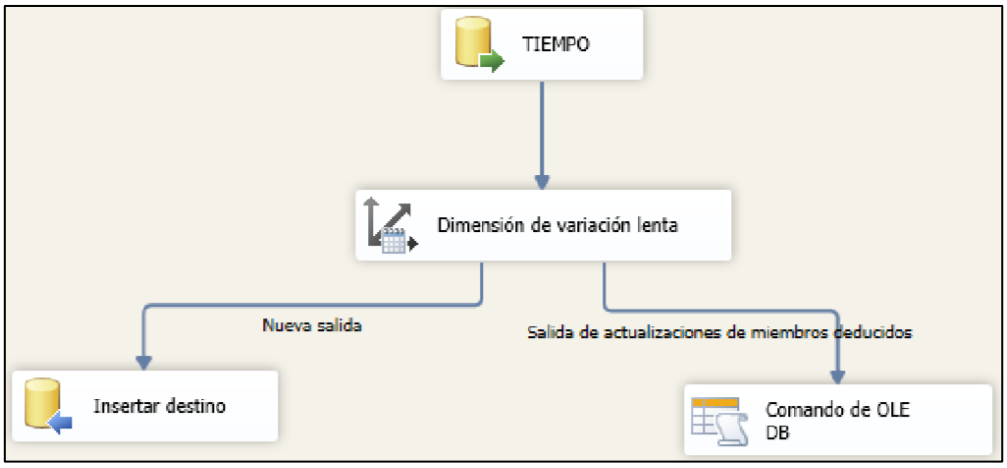
La **Error! Reference source not found.** representa el ETL para la vista profesor y recuperar los datos: cedula_profesor, profesor.

Figura 6: ETL profesor



La **Error! Reference source not found.** representa el ETL para la vista periodo y recuperar los datos: código_periodo, periodo, año.

Figura 7: ETL periodo

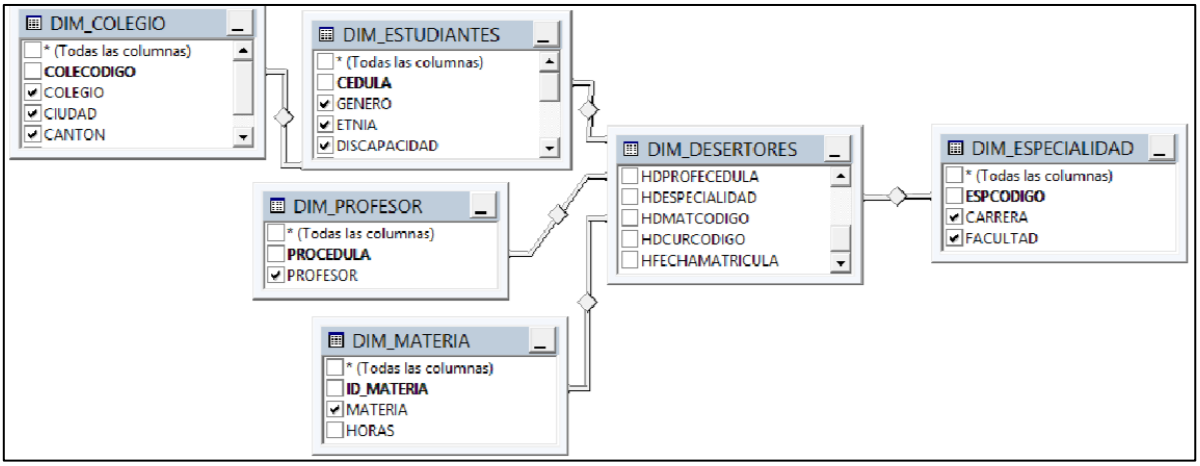


La **Error! Reference source not found.** representa el ETL para la vista deserción y recuperar los datos: código_desertor, periodo, cedula_estudiante, num_veces, asistencia, arrastre, cedula_profesor, carrera, código_materia, código_curso y fecha_matricula.

Figura 8: ETL desertores

La transformación se realiza mediante la consulta sql (realizada en modo gráfico) desde el data warehouse, como se observa en la **Error! Reference source not found.**:

Figura 9: Consulta SQL en modo gráfico desde el Data Warehouse



En la etapa de minería de datos y utilizando la herramienta WEKA para los datos generados con la consulta SQL desde el data warehouse, se aplicó el algoritmo SimpleKMeans con un número de clusters igual a 5, obteniéndose los siguientes resultados:

Figura 10: Número de instancias agrupadas

```
Clustered Instances

0       10300 ( 10%)
1       29995 ( 28%)
2       28387 ( 27%)
3        8542 (  8%)
4       28782 ( 27%)
```

Figura 11: Resultados Full Data y clusters

FULL DATA

Attribute	Full Data (106006.0)
1»¿EDAD_DESERCION	21.2689
PERIODO	MAR/12-AGO/12
HDASISTENCIA	87.7517
CURSO	3.5061
GENERO	MASCULINO
ETNIA	MESTIZA
DISCAPACIDAD	NINGUNA
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	CIENCIAS
PROMEDIO_CELEGIO	16.0173
CARRERA	INGENIERIA CIVIL
FACULTAD	INGENIERIA EN SISTEMAS ELECTRONICA E INDUSTRIAL
COLEGIO	INS TEC SUP BOLIVAR
CIUDAD	AMBATO
CANTON	AMBATO
PROVINCIA	TUNGURAHUA
MATERIA	FISICA I

CLOUSTER 0

Attribute	Cluster# 0 (12015.0)
1»¿EDAD_DESERCION	20.6553
PERIODO	OCT/15-MAR/16
HDASISTENCIA	89.9958
CURSO	3.3608
GENERO	MASCULINO
ETNIA	MESTIZA
DISCAPACIDAD	NINGUNA
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	CIENCIAS
PROMEDIO_CELEGIO	16.4957
CARRERA	INGENIERIA MECANICA
FACULTAD	INGENIERIA CIVIL Y MECANICA
COLEGIO	INS TEC IND RAMON BARBA NARANJ
CIUDAD	LATACUNGA
CANTON	LATACUNGA
PROVINCIA	COTOPAXI
MATERIA	LENGUAJE Y COMUNICACION

CLOUSTER 1

Attribute	1 (29232.0)
1»¿EDAD_DESERCION	21.9222
PERIODO	SEP/11-FEB/12
HDASISTENCIA	87.5962
CURSO	4.5356
GENERO	MASCULINO
ETNIA	MESTIZA
DISCAPACIDAD	NINGUNA
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	CIENCIAS
PROMEDIO_CELEGIO	16.4301
CARRERA	INGENIERIA CIVIL
FACULTAD	INGENIERIA CIVIL Y MECANICA
COLEGIO	INS TEC SUP BOLIVAR
CIUDAD	AMBATO
CANTON	AMBATO
PROVINCIA	TUNGURAHUA
MATERIA	EMPLEO DE NTICS II

CLOUSTER 2

Attribute	2 (28020.0)
1»¿EDAD_DESERCION	21.5761
PERIODO	SEP/07-FEB/08
HDASISTENCIA	86.3478
CURSO	3.6192
GENERO	FEMENINO
ETNIA	NINGUNO
DISCAPACIDAD	NO IDENTIFICADO
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	CIENCIAS
PROMEDIO_CELEGIO	15.2484
CARRERA	MEDICINA
FACULTAD	CIENCIAS DE LA SALUD
COLEGIO	COL EXP AMBATO
CIUDAD	AMBATO
CANTON	AMBATO
PROVINCIA	TUNGURAHUA
MATERIA	TECNICAS DE ESTUDIO

CLOUSTER 3

Attribute	3 (7893.0)
1»¿EDAD_DESERCION	20.987
PERIODO	SEP/10-FEB/11
HDASISTENCIA	84.9411
CURSO	2.9982
GENERO	FEMENINO
ETNIA	NINGUNO
DISCAPACIDAD	NO IDENTIFICADO
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	CIENCIAS
PROMEDIO_CELEGIO	16.2539
CARRERA	ING. MARKETING Y GESTION DE NEGOCIOS
FACULTAD	CIENCIAS ADMINISTRATIVAS
COLEGIO	INS TEC VICTORIA VASCONEZ CUVI
CIUDAD	LATACUNGA
CANTON	LATACUNGA
PROVINCIA	COTOPAXI
MATERIA	TECNICAS DE ESTUDIO

CLOUSTER 4

Attribute	4 (28846.0)
1»¿EDAD_DESERCION	20.641
PERIODO	MAR/12-AGO/12
HDASISTENCIA	89.1073
CURSO	2.5635
GENERO	MASCULINO
ETNIA	MESTIZA
DISCAPACIDAD	NINGUNA
ESTADO_CIVIL	SOLTERO
TITULO_CEOLEGIO	COMERCIO Y ADMINISTRACION
PROMEDIO_CELEGIO	16.082
CARRERA	GENERAL
FACULTAD	INGENIERIA EN SISTEMAS ELECTRONICA E INDUSTRIAL
COLEGIO	INS SUP TEC HISPANO AMERICA
CIUDAD	AMBATO
CANTON	AMBATO
PROVINCIA	TUNGURAHUA
MATERIA	FISICA I

Discusión

Los resultados indican que la propuesta no se limita a usar únicamente el algoritmo K-means, sino como una arquitectura de analítica educativa que integra todo el proceso de gestión de los datos, desde su preparación e integración hasta su análisis y actualización. Esta diferencia es importante porque la utilidad de un algoritmo de minería de datos no depende únicamente de su capacidad para encontrar patrones,

sino también de que la institución cuente con sistemas necesarios para preparar, integrar y actualizar información.

La construcción del almacén de datos (data warehouse) constituye una importante contribución tecnológica, especialmente porque permite diferenciar los datos destinados a la operación institucional de los datos necesarios para el análisis. Esta estrategia está vinculada con la evolución de la minería de datos en educación, donde las investigaciones han ido cambiando poco a poco de analizar pequeños conjuntos de datos hacia ecosistemas de datos provenientes de diferentes fuentes. Romero y Ventura (2013) ya citaban que la característica centrl del área educativa son la diversidad de fuentes con las que cuentan.

En todo proyecto de minería de datos, el preprocesamiento es una de las etapas críticas y por ende importante debido a que los datos pueden tener errores, inconsistencias o variables mal estructuradas que pueden afectar la calidad de los resultados. Por ello, una contribución a destacar es la implementación de los procesos específicos para estudiantes, colegios, carreras, materias, profesores, períodos académicos y desertores lo cual se lleva a cabo en el proceso ETL. Esta decisión es metodológicamente importante porque evita enviar directamente los datos de las transacciones al algoritmo.

Así mismo, la integración de información desde diferentes dimensiones es un aporte importante evidenciado mediante el conjunto de datos Full Data que reúne variables académicas, demográficas, geográficas e institucionales, alcanzando un total de 106.006 registros. Esta combinación ofrece una visión más amplia de la realidad estudiantil y permite ir más allá de los análisis tradicionales centrados únicamente en las calificaciones o en los datos de matrícula, resultados que coincide con la literatura sobre analítica de aprendizaje o simplemente conocida por la expresión en inglés learning analytics.

Ifenthaler y Yau (2020) identificaron que los estudios de éxito académico utilizan variables relacionadas tanto con características individuales como con el entorno educativo. De igual forma, de Oliveira et al. (2021) clasificaron las variables empleadas

en investigaciones de deserción en categorías personales, académicas y externas a los estudiantes.

El enfoque utilizado resulta útil en una primera etapa de análisis institucional, a través del cual se conocen las características de los diferentes grupos de estudiantes antes de desarrollar modelos orientados a predecir su comportamiento o nivel de riesgo, partiendo del agrupamiento para identificar de manera más abierta los perfiles y patrones presentes en la información en lugar de partir de una categorización predefinida de estudiantes con alto o bajo riesgo. Esto se logró mediante el algoritmo K-means como herramienta de exploración de los datos y mejorando el entendimiento de la información.

Existe una diferencia metodológica fundamental entre patrón y factor causal. Que un grupo presente una característica que ocurre con mucha frecuencia no significa que dicha característica cause la deserción. Al utilizar el algoritmo K-means se pueden encontrar similitudes y estructuras internas, pero el algoritmo no permite establecer causalidad.

La arquitectura propuesta puede evolucionar hacia un sistema de alerta temprana. Por ejemplo, los clústeres descubiertos podrían utilizarse para definir perfiles de riesgo y luego entrenar modelos supervisados con datos históricos etiquetados.

Conclusiones

Se evidencia que la minería de datos puede ser una herramienta tecnológica útil para aprovechar la información que producen las instituciones de educación superior y convertirla en conocimiento que ayude a detectar patrones vinculados con la deserción estudiantil.

La construcción de una arquitectura que combina almacenes de datos, procesos de extracción, transformación y carga, y a esto se suman las técnicas de análisis de datos, permite integrar información procedente de diferentes fuentes institucionales y la producción de un conjunto de datos para análisis educativo.

El uso de SQL Server e Integration Services permite implementar procesos de extracción, transformación y carga que preparan los datos para su análisis posterior. La existencia de procesos ETL específicos para estudiantes, colegios, carreras, materias, profesores, períodos y desertores constituye un elemento importante de la estructura tecnológica.

Referencias Bibliográficas

Asto-Lázaro, M. S., & Bermejo-Terrones, H. P. (12 de 2023). RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação. *Dialnet, 64*(12), 463-476. Obtenido de https://www.risti.xyz/issues/ristie64.pdf

CACES. (2020). *CACES*. Obtenido de https://www.caces.gob.ec/universidades-y-escuelas-politecnicas-3/

CACES. (2023). *CACES*. Obtenido de https://www.caces.gob.ec/universidades-y-escuelas-politecnicas-3/

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. *Big Data and Cognitive Computing, 5*(4), 64. doi:https://doi.org/10.3390/bdcc5040064

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 13-95. doi:https://doi.org/10.1609%2Faimag.v17i3.1230

García Sánchez, R., & Gil Mateos, J. (12 de 2023). Minería de Datos Educacionales: Descubrir tesoros ocultos durante el aprendizaje. *Revista Científica ECOCIENCIA, 10*, 18-41. doi:https://doi.org/10.21855/ecociencia.100.830

Ifenthaler, D., & Yau, J. Y.-K. (8 de 2020). Utilising learning analytics to support study success in higher education: a systematic review. *Education Tech Research Dev, 68*, 1961–1990. doi:https://doi.org/10.1007/s11423-020-09788-z

Porras, J. M., Lara, J. A., Romero, C., & Ventura, S. (2023). A Case-Study Comparison of Machine Learning Approaches for Predicting Student's Dropout from Multiple Online Educational Entities. *Algorithms, 16*(12). doi:10.3390/a16120554

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery, 3*(1), 12-27. doi:https://doi.org/10.1002/widm.1075

Técnicas de machine Learning para predecir la deserción estudiantil universitaria: Una revisión sistemática de la literatura. (9 de 2024). *Revista Científica Multidisciplinar G-Nerando, 5*(2), 997-1023. doi:https://doi.org/10.60100/rcmg.v5i2.245

Tinto, V. (1975). Dropout From Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research, 45*(1), 89-125. doi:10.2307/1170024

Zhang, S., Zhang, C., & Yang, Q. (30 de 11 de 2003). Data preparation for data mining. *Applied Artificial Intelligence, 17*(5-6), 375-381. doi:https://doi.org/10.1080/713827180