Análisis de técnicas de agrupamiento para detección de valores anómalos en rendimiento estudiantil en una institución de educación superior

Analysis of clustering techniques for the detection of outliers in student performance in higher education institutions

https://doi.org/10.5281/zenodo.14680639

AUTORES: Angelo Isaac Iturralde Borja[1]*

Patricia Jimbo Santana[2]

DIRECCIÓN PARA CORRESPONDENCIA: aiiturralde@uce.edu.ec

Fecha de recepción: 15 / 10 / 2024

Fecha de aceptación: 12 / 12 / 2024

RESUMEN

En las universidades, la gestión de datos enfrenta riesgos significativos debido al desconocimiento generalizado. Sin embargo, la recolección y el análisis de estos datos son fundamentales para garantizar una enseñanza de calidad y la transparencia en los procesos académicos. A menudo, se presentan irregularidades en el registro de las calificaciones de los estudiantes, ya sea por errores involuntarios, intentos de fraude o actos de corrupción. En este artículo, aplicamos técnicas de clusterización, específicamente el método "K-means", para detectar estos posibles casos de manera oportuna. Esta metodología se basa en la capacidad de identificar valores o datos anómalos, lo que permite mejorar la detección de incidentes y proporciona una herramienta eficaz para el control y la supervisión de los registros académicos.Además, estas técnicas pueden utilizarse para optimizar la precisión de los procesos de evaluación, mejorando así la calidad de la información disponible para la toma de decisiones. Esto facilita la implementación de medidas correctivas oportunas,

[1]* https://orcid.org/0009-0009-8959-1305, Universidad Central del Ecuador, aiiturralde@uce.edu.ec
[2] https://orcid.org/0000-0001-7432-1622, Universidad Central del Ecuador, prjimbo@uce.edu.ec

reduciendo el riesgo de que errores o malas prácticas afecten tanto el rendimiento académico de los estudiantes como la reputación de la institución.

Palabras clave: *k-means, clusterización, educación superior, valores anómalos*

ABSTRACT

In universities, data management faces significant risks due to widespread ignorance. However, the collection and analysis of this data is essential to ensure quality teaching and transparency in academic processes. Irregularities often occur in the recording of students' grades, whether due to involuntary errors, attempted fraud or acts of corruption. In this article, we apply clustering techniques, specifically the "K-means" method, to detect these potential cases in a timely manner. This methodology is based on the ability to identify anomalous values or data, which allows for improved incident detection and provides an effective tool for the control and supervision of academic records. Furthermore, these techniques can be used to optimize the accuracy of assessment processes, thus improving the quality of information available for decision-making. This facilitates the implementation of timely corrective measures, reducing the risk of errors or bad practices affecting both the academic performance of students and the reputation of the institution.

Keywords: *k-means, clustering, higher education, outliers*

INTRODUCCIÓN

El volumen de datos que se extrae en las instituciones de educación superior es muy grande pero estas no aprovechan los datos para obtener información relevante para su toma de decisiones, el principal problema es la dificultad para realizar este análisis por el humano pero esto ha disminuido con la mejora de las computadoras que pueden realizar cálculos de forma eficiente sin importar el tipo de dato, en la actualidad se puede analizar textos, imágenes y más formatos con técnicas basadas en la identificación de patrones o características similares que ayuden al entendimiento de los datos.(Peng' et al., n.d.) (Devasia et al., n.d.)

La aplicación de minería de datos educacional como la llaman algunos autores,(Dol & Jawandhiya, 2022), (Venkatachalapathy et al., 2017) es importante para poder encontrar patrones escondidos mediante la selección, pre procesamiento, transformación de los datos,

aplicación de los modelos obteniendo así información relevante que permita realizar un seguimiento a los estudiantes y docentes para poder conocer el rendimiento de estos con el fin de mejorar la calidad de educación que se brinda con los avances del conocimiento en análisis de datos en técnicas de minería como clasificación, agrupamiento, reglas de asociación, regresión, detección de anomalías, (Muhammed, 2021) se ha podido obtener predicciones sobre el abandono de estudiantes de las carreras, rendimiento académico, comportamiento de profesores (Bae et al., 2020) identificando en qué áreas existen conflictos logrando aplicar mejoras en los procesos.

Realizamos una investigación de diferentes autores que han aplicado las técnicas de agrupamiento para entender cómo podemos aprovechar esos conocimientos en nuestro trabajo.

La detección de fraudes financieros (Min & Lin, 2018) es muy importante y se ha convertido en una necesidad para las entidades se propone un modelo optimizado de detección del fraude financiero que integra técnicas de selección de características y de clasificación de aprendizaje automático. El estudio combina varios métodos de minería de datos, como árboles de decisión, Naive Bayes, máquinas de vectores soporte y redes neuronales artificiales, para mejorar la precisión de la detección del fraude. Utilizando un enfoque híbrido, la investigación pretende proporcionar una metodología sólida para detectar actividades fraudulentas en los estados financieros (Kültür & Ufuk Çağlayan, n.d.). El estudio contribuye a este campo demostrando la eficacia de combinar múltiples técnicas de minería de datos para mejorar el rendimiento global de los modelos de detección de fraudes en los estados financieros. (J. Yao, 2018).

La aplicación de técnicas de minería de reglas de asociación y clasificación en el análisis de datos educativos permite identificar con éxito patrones y relaciones significativas en conjuntos de datos educativos mediante la minería de reglas de asociación.(Poonam & Dutta, 2012) Además, los algoritmos de clasificación empleados en esta investigación predijeron eficazmente el rendimiento y el comportamiento de los estudiantes basándose en datos históricos. Los resultados demostraron el potencial de estos métodos de análisis de datos para mejorar las estrategias de enseñanza, personalizar las experiencias de aprendizaje y optimizar los resultados educativos. En general, un estudio realizado muestra los valiosos conocimientos obtenidos de la utilización de la minería de reglas de asociación

y la clasificación en el análisis de datos educativos, destacando su impacto en la mejora de los procesos de toma de decisiones y el éxito de los estudiantes en los entornos educativos. (Pornthep Rojanavasu, 2019)

En (Abe, 2019) se utilizan técnicas de minería de datos "clasificación, agrupamiento" y aprendizaje automático para analizar macrodatos educativos en entornos universitarios con la aplicación de estos métodos, el estudio descubre con éxito tendencias, patrones y relaciones ocultas en los conjuntos de datos educativos. Con el modelado predictivo (Pramanik et al., 2020), realizan predicciones precisas con datos históricos lo que permite mejorar los procesos de toma de decisiones y los resultados demuestran la eficacia de estas herramientas analíticas avanzadas para optimizar las estrategias de retención de estudiantes y mejorar las prácticas educativas generales las universidades.

Para el presente artículo nos enfocamos en la detección de valores anómalos en instituciones de educación superior y otros autores en (Hidayat et al., n.d.) nos dan una guía y se centra en la utilización de técnicas de minería de datos educativos (EDM Educational Data Mining) para evaluar el rendimiento de los estudiantes y mejorar el entorno de aprendizaje. El arículo emplea reglas de asociación y técnicas de clasificación para identificar patrones en los datos de los estudiantes y reorganizar los cursos virtuales en función de dichos patrones. Los resultados de la investigación demuestran la eficacia de EDM para predecir el rendimiento académico de los estudiantes y mejorar los resultados educativos. Las técnicas utilizadas en este estudio fueron:

Reglas de asociación: Se emplearon para identificar patrones de conocimiento dentro de los datos educativos, ayudando en la reestructuración de cursos virtuales basados en estos patrones (Pamula et al., 2011).

Técnicas de clasificación: Utilizadas para predecir el rendimiento y el comportamiento de los estudiantes, permitiendo la reorganización de las estrategias educativas para mejorar los resultados de los estudiantes.

Al aprovechar estas técnicas de EDM, la investigación pretende optimizar los procesos de evaluación de los estudiantes, predecir el éxito académico y mejorar el entorno general de aprendizaje de los estudiantes. Para complementar el trabajo anterior (Dol, 2021) explora la aplicación de técnicas de clasificación en la minería de datos educativos, centrándose específicamente en los algoritmos Perceptrón multicapa, Bosque aleatorio y Máquina de

vectores de apoyo. Entre estas técnicas, el estudio identifica Random Forest para analizar datos educativos y predecir el rendimiento de los estudiantes. Los resultados ponen de relieve el rendimiento superior de Random Forest a la hora de potenciar los procesos de evaluación de los estudiantes y mejorar los resultados educativos y lograr predicciones más precisas optimizando el entorno de aprendizaje para los estudiantes.

(Sagardeep Roy & Shailendra Narayan Singh, 2017) identifica los algoritmos de clasificación en la minería de datos educativos. Esta técnica ha demostrado su eficacia a la hora de analizar datos educativos y predecir el rendimiento de los estudiantes, mejorando en última instancia los resultados educativos y mejorando el entorno general de aprendizaje. El artículo destaca la importancia de aprovechar los big data con fines educativos y el potencial de los algoritmos de clasificación en la minería de datos para lograr predicciones más precisas y optimizar el entorno de aprendizaje de los estudiantes (Min & Lin, 2018) (Kanjanawattana, 2019).

En otra aplicación importante de estas técnicas tenemos la relación entre el comportamiento del profesor y el rendimiento del alumno el estudio revela que ciertos comportamientos, como proporcionar retroalimentación oportuna, asistencia personalizada y crear un entorno de aprendizaje atractivo, influyen positivamente en el rendimiento de los alumnos. Con el análisis de patrones en las interacciones profesor-alumno y los resultados académicos, la investigación identifica factores clave que contribuyen a mejorar el rendimiento de los estudiantes. Los resultados sugieren que optimizar el comportamiento de los profesores basándose en datos puede mejorar los resultados de los alumnos y aumentar la eficacia de las prácticas educativas. (Devasia et al., n.d.)

Por otro lado, Anoopkumar y Rahman explora diversas técnicas de minería de datos, como algoritmos de clasificación y métodos de agrupación, para analizar datos y predecir el progreso de los estudiantes. La investigación pretende contribuir al avance de la minería de datos mediante la identificación de los factores clave y las técnicas que pueden conducir a la mejora de los resultados de los estudiantes y las estrategias de educación personalizada.

Para la detección de valores atípicos con minería de datos (Poonam & Dutta, 2012) destaca la importancia de identificar patrones en los datos que se desvían del comportamiento esperado, (Amit. Agarwal, 2015) se explora diferentes metodologías para detectar valores atípicos y su aplicación en la resolución de problemas de la vida real, el análisis de redes

basado en comunidades y la caracterización de valores atípicos temporales en redes dinámicas, la investigación pretende ofrecer valiosas perspectivas a los investigadores. La detección de valores atípicos desempeña un papel vital en diversas aplicaciones, como la detección de fraudes en transacciones de comercio electrónico y el descubrimiento de anomalías en datos de redes, lo que la convierte en una tarea fundamental en las prácticas contemporáneas de minería de datos.(Pramanik et al., 2020) Con este precedente podemos evidenciar lo que (Lakshmi Sreenivasa Reddy.D et al., 2014) nos cuenta en su artículo donde utilizó la detección de valores atípicos para identificar estudiantes peculiares a partir de bases de datos de estudiantes. El estudio tomó estudiantes con comportamientos diferentes en los centros educativos, que pueden dar lugar a diversos problemas en clase. Empleando el análisis de valores atípicos, la investigación pretende reconducir a estos estudiantes peculiares. El artículo presenta técnicas relacionadas con datos de atributos categóricos, que se utilizan para recopilar datos de estudiantes de B. Tech de diferentes facultades para realizar experimentos. (J Md Zubair Rahman Principal, n.d.)

Este artículo está elaborado con la siguiente estructura: la sección 2 describe los trabajos relacionados, la sección 3 presenta los métodos aplicados, la sección 4 expone los resultados obtenidos y la sección 5 realiza una discusión del tema y por último la sección 6 tenemos las conclusiones.

METODOLOGÍA

El descubrimiento de conocimientos en bases de datos (KDD) es esencial para detectar eficazmente los valores atípicos en los sistemas de calificación de estudiantes nos proporciona una metodología estructurada para extraer información significativa de conjuntos de datos complejos (S. Agarwal, 2014). Aprovechando el KDD, realizamos este proceso para asegurar la calidad de datos con los que se trabajó en la investigación, este abarca varias etapas, como la selección de datos, el preprocesamiento, la transformación, la extracción, la evaluación y la presentación de conocimientos.

En el documento se trabajó con 4 archivos obtenidos de una institución de educación superior para su respectivo análisis y después de aplicar KDD para realizar la limpieza de datos se procede a aplicar la metodología K-means que al ser una de las herramientas más utilizadas para análisis de valores anómalos en educación como detección de fraudes,

corrupción o errores en calificaciones de los estudiantes nos permitirá detectar patrones y anomalías los datos, al dividir los datos en grupos que tienen características similares nos permiten identificar los valores que se desvían de forma evidente de nuestro proceso.

La eficacia de esta herramienta es por su naturaleza de aprendizaje no supervisado que no requiere datos etiquetados y es crucial en contextos educativos en los que los datos históricos de calificación pueden estar incompletos o sesgados al agrupar a los estudiantes en función de sus métricas de rendimiento podemos poner de relieve patrones de puntuación inusuales que justifiquen una investigación más profunda.

- Datos de Estudiantes: Archivo que contiene datos personales, académicos y socioeconómicos de los estudiantes.
- Proceso de Calificación: Archivo que documenta los procesos de calificación realizados por los docentes.
- Notas: Información consolidada sobre calificaciones y rendimiento estudiantil en el periodo 2017-2022.
- Datos de Docentes: Detalles sobre los profesores, asignaturas impartidas y resultados relacionados con el rendimiento de los estudiantes.

1. **Datos de Estudiantes**

En este documento tenemos un total de 491627 filas al cual se realizó un análisis identificar columnas y registros con valores nulos donde se encuentran 566,761 valores nulos distribuidos en varias columnas del archivo y se eliminan los registros que contenían valores nulos, procedemos a identificar y eliminar registros 340,617 duplicados en el archivo.

Después de completar la limpieza de los datos, el archivo quedó con un total de 151,010 registros, consolidando toda la información relevante para el análisis.

De las 36 columnas originales, se identificó las que tenían más del 60% de vacíos o nulos y se eliminaron, de igual forma se eliminan las columnas que incluyen detalles como correos electrónicos, direcciones, teléfonos y características de conectividad por anonimización de los datos.

En la tabla 1 podemos ver las columnas que se mantienen para el análisis del archivo Datos Estudiantes y el significado de los encabezados a trabajar.

Columna	Detalle
CEDULA	Número de cédula
PAIS_ORIGEN	País de origen
PROVINCIA_NACIMIENTO	Provincia de nacimiento
CANTON_NACIMIENTO	Cantón de nacimiento
DISPONIBILIDAD_INTERNET	Disponibilidad de Internet
POSEE_MULTIMEDIA	Disponibilidad de multimedia

Tabla 1. Columnas en los Datos de Notas

2. **Proceso de Calificación**

En este documento tenemos un total de 4’875.970 filas al cual se realizó un análisis identificar columnas y registros con valores nulos donde se encuentran 1,157,017 valores nulos distribuidos en varias columnas del archivo y se eliminan los registros que contenían valores nulos, procedemos a identificar y eliminar registros 942.646 duplicados en el archivo.

Después de completar la limpieza de los datos, el archivo quedó con un total de 151,010 registros, consolidando toda la información relevante para el análisis.

La columna PRCL_FECHA fue procesada para generar la columna derivada PERIODO, estableciendo el formato de año académico (AAAA-AAAA). Esto fue necesario ya que los periodos académicos son semestrales y la gran cantidad de fechas generaba errores en el análisis.

Se creó la columna Descripcion_Proceso, generando un diccionario que identifica los procesos existentes.

Diccionario de Procesos

a) Registro de notas
b) No tiene datos en PRCL_OBSERVACION
c) Recalificación por director
d) Evaluación cualitativa por docente
e) Registro de notas por nuevo docente
f) No tiene datos en PRCL_OBSERVACION

g) Registro de notas de suficiencia

h) Rectificación automática de notas

Para poder hacer la relación entre tablas, necesitamos que el valor en la columna CLF_ID caso contrario no se podrá unir con las otras tablas, por tanto, las filas que no poseen este valor serán eliminadas.

En la tabla 2 tenemos las columnas que se mantienen para el análisis del archivo Proceso Calificación y el significado de los encabezados a trabajar.

Columna	Detalle
CLF_ID	Identificador único de calificaciones
PRCL_FECHA	Fecha asociada al proceso
PRCL_TIPO_PROCESO	Tipo de proceso relacionado con las notas
PERIODO	Período académico derivado de PRCL_FECHA
Descripcion_Proceso	Descripción categorizada del tipo de proceso

Tabla 2. Columnas en Proceso Calificación

3. **Datos de Docentes**

En el análisis inicial del archivo de datos docentes se tiene 104619 registros, no se encontraron valores nulos. El archivo contenía 18 columnas originales, de las cuales se eliminaron aquellas que contenían datos personales o porcentuales de rendimiento del docente, ya que esta información también se encontraba en el archivo de notas de los estudiantes. Solo se requería identificar al docente para vincular la nota ingresada.

Además, se identificaron y eliminaron 3,689 registros duplicados, finalizando con un total de 100,930 filas después de la limpieza.

En la tabla 3 tenemos las columnas que se mantienen para el análisis del archivo Datos docentes y el significado de los encabezados a trabajar

Columna	Detalle
PERIODO	Período académico asociado al docente.
FACULTAD	Facultad a la que pertenece el docente.
CARRERA	Carrera asociada al docente.

MATERIA	Nombre de la materia impartida.
TOTAL_APROBADOS	Número total de estudiantes aprobados en la materia.
TOTAL_REPROBADOS	Número total de estudiantes reprobados.
TOTAL_RETIRADOS	Número total de estudiantes que se retiraron de la materia.
CEDULA_DOCENTE	Cédula de identidad del docente.
DOCENTE	Nombre completo del docente.

Tabla 3. Columnas Datos Docentes

4. Notas Estudiantes

En la revisión inicial del archivo de Notas Estudiantes, se identificaron 516512 registros, 1,551,847 valores nulos distribuidos en varias columnas que fueron eliminados para garantizar la calidad del análisis. De las 32 columnas originales, se realizó un análisis exhaustivo para determinar la relevancia de cada una, priorizando aquellas con menos del 60% de valores vacíos o nulos. Asimismo, se descartaron columnas que contenían información redundante o confidencial, ya que no aportaban valor directo al estudio. Como resultado, se conservaron 12 columnas clave, que incluían datos esenciales como el período académico, la facultad, la carrera, la cédula y otros relacionados directamente con el desempeño académico de los estudiantes.

En la tabla 4 tenemos las columnas que se mantienen para el análisis del archivo Notas Estudiantes y el significado de los encabezados a trabajar.

Columna	**Detalle**
CLF_ID	ID único del curso
PERIODO	Período académico
FACULTAD	Facultad correspondiente
CARRERA	Nombre de la carrera
CEDULA	Cédula del estudiante
SEXO	Género del estudiante
REGIMEN	Régimen académico
SEMESTRE	Semestre del estudiante
MATERIA	Nombre de la materia

NOTA_FINAL	Nota final del estudiante
ESTADO	Estado académico del estudiante
CI_DOCENTE	Cédula del docente

Tabla 4. Columnas en Notas Estudiantes

a) Preparación de Datos

Se creó un archivo maestro consolidando las bases de datos limpias para realizar un análisis integral, en la primera etapa fue la unión de Notas Limpias y Proceso de Calificación para ello, los datos de las notas de los estudiantes (df_notas_limpio) se combinaron con los datos del proceso de calificación (df_proceso_completo) utilizando la columna CLF_ID como llave, empleando un método de unión left join que permitió preservar la información principal de las notas.

Posteriormente, se realizó la unión con los Datos de Docentes Limpios, enlazando la columna CI_DOCENTE en el conjunto de notas con CEDULA_DOCENTE en el conjunto de docentes. Este paso incorporó información detallada sobre los docentes asociados a las calificaciones de los estudiantes.

Para garantizar la calidad del archivo maestro, se procedió a revisar valores nulos y duplicados. Inicialmente, el archivo contenía 1,007,130 filas. Tras la eliminación de duplicados, el conjunto de datos se redujo a 510,096 filas, mejorando su consistencia.

Después se analizaron las columnas con más del 50% de valores nulos, las cuales fueron eliminadas para evitar sesgos en el análisis. Tras estas modificaciones, el archivo quedó reducido a 13 columnas clave. Se realizó una revisión específica de las columnas PRCL_TIPO_PROCESO y Descripcion_Proceso, esenciales por describir el tipo y la naturaleza del proceso académico relacionado con las notas. La eliminación de registros con valores nulos en estas columnas dejó el archivo final con 299,078 filas, listo para el análisis.

b) Separación de Datos por Facultad

Para tener un análisis detallado los datos se segmentaron en subconjuntos individuales por facultad. Este paso permitió adaptar el estudio a las características particulares de cada facultad. A continuación, se presentan las facultades y el número de registros correspondientes:

Filosofía, Letras y Ciencias de la Educación: 131,568 registros.

Ciencias Administrativas: 90,479 registros.

Ciencias Médicas: 77,031 registros.

Esta segmentación permitió abordar los análisis de manera específica para cada facultad, destacando las tendencias y patrones únicos.

c) Normalización y Codificación de Datos

La preparación del archivo maestro incluyó un proceso de normalización y codificación de los datos:

Codificación de columnas categóricas:

Las columnas SEXO y ESTADO fueron transformadas en variables binarias:

1 para masculino/aprobado.

0 para femenino/reprobado.

Las columnas FACULTAD_x, CARRERA_x y MATERIA_x fueron convertidas a valores numéricos mediante un codificador de etiquetas (LabelEncoder).

d) Normalización de datos numéricos:

Las variables seleccionadas fueron estandarizadas utilizando el método StandardScaler. Esto garantizó que todas las características tuvieran un peso equitativo, eliminando sesgos derivados de las escalas originales de los datos.

Este proceso aseguró la calidad y relevancia de los datos para los análisis posteriores, sentando las bases para la clusterización y detección de patrones.

RESULTADOS

Con la aplicación de K-menas logramos obtener los outliers divididos por facultad, se realizó de esta manera porque cada facultad tiene características diferentes y al dividirlo se logró obtener mejores resultados.

Para determinar el número de clusters óptimo se trabajó con el método de codo y en todos los casos recomendó 5 clusters.

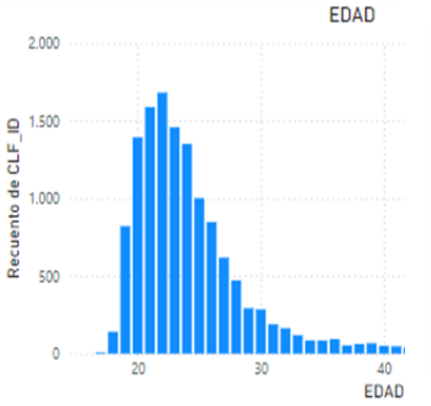
En los siguientes gráficos podremos ver un análisis estadístico de los clusters solo con valores anómalos, no se incluyó los valores normales para poder enfocarnos en la identificación de valores que nos puedan brindar información clave.

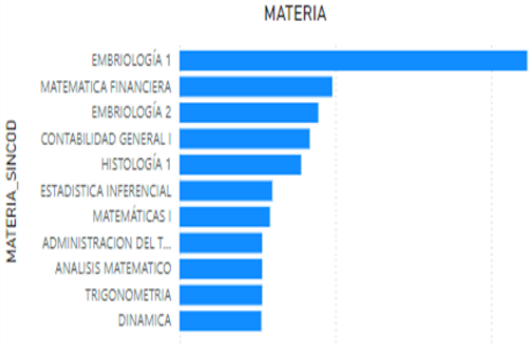
CONTEO VALORES ANOMALOS

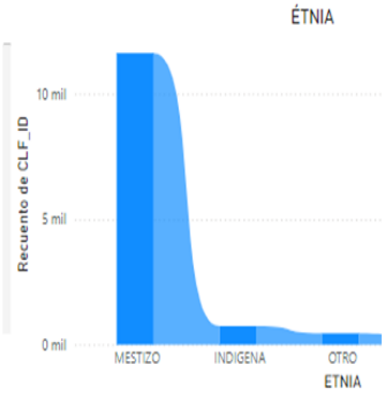
V_ANOMALOS	ID_DOC
210	17098
181	4008
171	10016
167	17110
159	17135
153	17084
151	17036

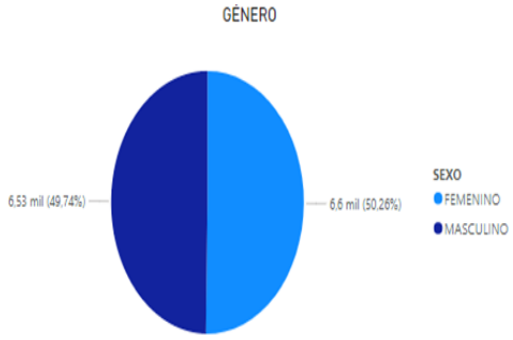
ESTADO MATRÍCULA

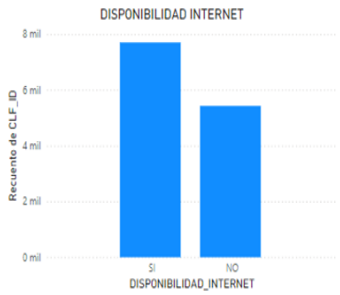
REPROBADO	12473
APROBADO	412
ANULADO	130
RETIRO APROBADO	101
INSCRITO	11
Total	**13127**

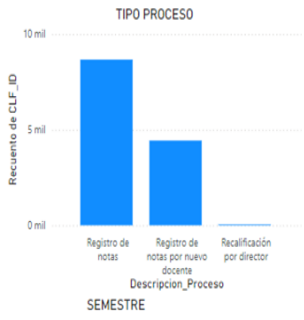


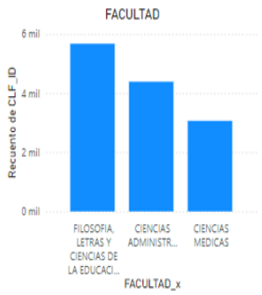












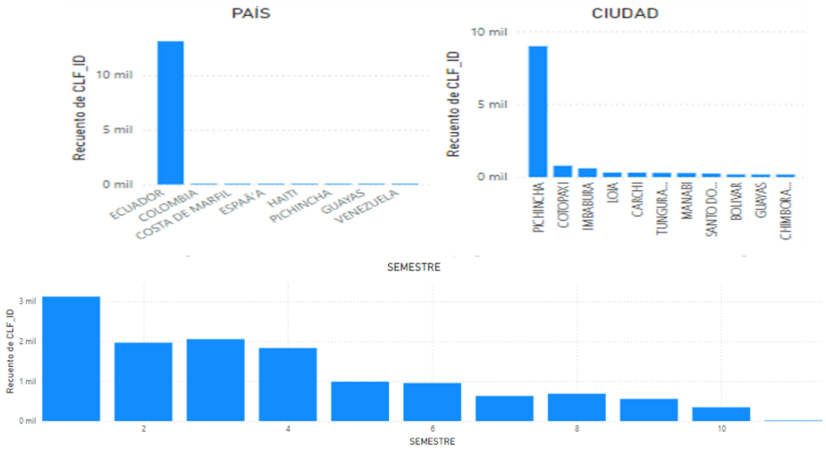


Ilustración 1. Análisis Estadístico Outliers Global

Como podemos apreciar en la ilustración 1 tenemos gráficos estadísticos que resumen la clusterización realizada, tenemos información relevante sobre las características de los valores considerados como anómalos, con esta base procedemos a investigar solo los usuarios que tienen mayor número de anomalías para poder detectar que está sucediendo y permitir a la institución tomar decisiones con el fin de minimizar estos riesgos.

DISCUSIÓN

Este artículo nos brinda diferentes contribuciones para entender el comportamiento de docentes y estudiantes al momento de realizar los procesos de calificación en el siguiente gráfico podemos ver las anomalías globales que existen en la facultad de medicina, con el análisis estadístico podemos identificar un docente que posee gran cantidad de valores anómalos y procederemos a investigar.

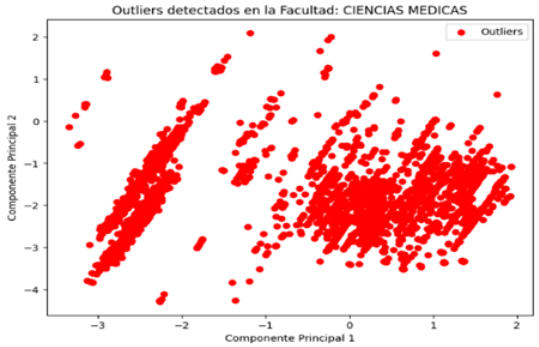


Ilustración 2. Outliers en Ciencias Médicas

En la ilustración 2 tenemos la distribución de datos anómalos encontrados en la facultad de Ciencias médicas con los que vamos a trabajar para la investigación.

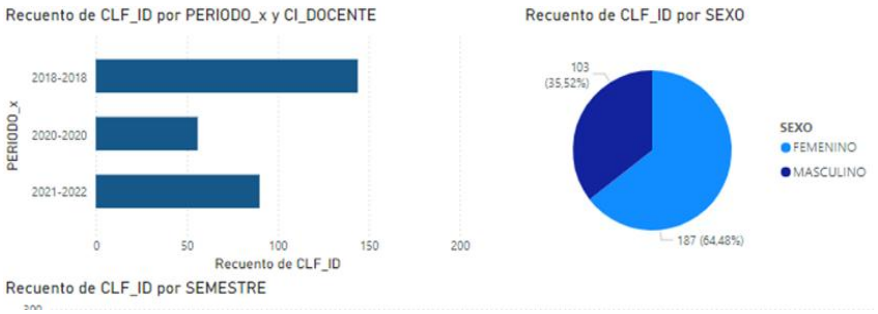


Ilustración 3. Outliers en PERIODOS y GENERO

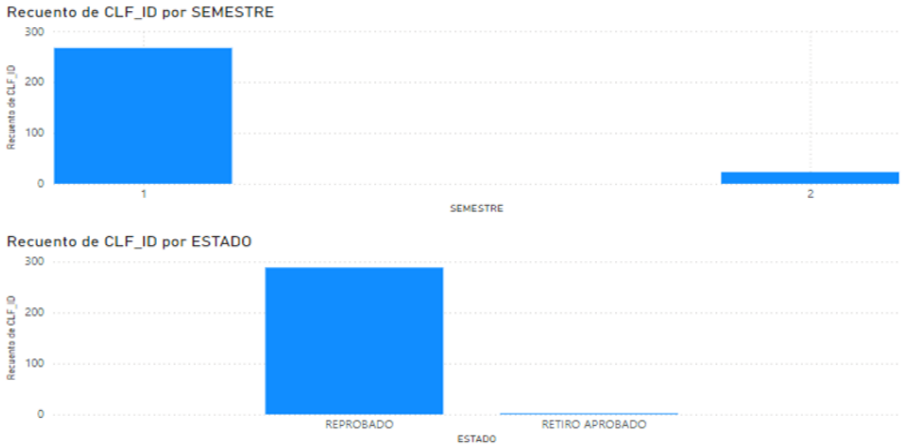


Ilustración 4. Outliers en SEMESTRE y ESTADO

CONCLUSIONES

Uno de los principales objetivos de las instituciones de educación superior es mejorar la calidad de la enseñanza que ofrecen a sus estudiantes. Para lograrlo, es fundamental mantener un control riguroso sobre el rendimiento académico, lo que permite realizar un seguimiento efectivo de aquellos alumnos que enfrentan dificultades. A través del análisis del proceso de calificación y los hallazgos de esta investigación, las instituciones pueden identificar inconsistencias que representen riesgos tanto para docentes como para estudiantes, lo que a su vez mejora la toma de decisiones basada en las características reveladas por el agrupamiento de datos.

Para futuras investigaciones, es recomendable desarrollar modelos predictivos que apoyen a las instituciones en sus operaciones diarias. Utilizando las características obtenidas a partir de los clusters y la información histórica, será posible identificar a los estudiantes que siguen patrones de abandono y promover programas que les ayuden a completar sus estudios con éxito. Asimismo, es crucial detectar las características de aquellos docentes que manipulan datos con el fin de obtener incentivos, lo cual perjudica la integridad institucional y los derechos de los estudiantes.

Finalmente, esta metodología puede extenderse a otras áreas, como la medicina, para identificar patrones en enfermedades peligrosas, o el ámbito financiero, para reconocer características de personas involucradas en fraudes al sistema. La versatilidad del análisis de datos ofrece un amplio espectro de aplicaciones que pueden contribuir significativamente a la mejora en diversos campos.

REFERENCIAS BIBLIOGRÁFICAS

Abe, K. (2019). Data Mining and Machine Learning Applications for Educational Big Data in the University. 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 350–355. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00071

Agarwal, Amit. (2015). Proceedings on 2015 1st International Conference on Next Generation Computing Technologies (NGCT) : September 4th-5th, 2015, Center for Information Technology, University of Petroleum and Energy Studies, Dehradun. IEEE.

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Bae, D. H., Jeong, S., Hong, J., Lee, M., Ivanović, M., Savić, M., & Kim, S. W. (2020). An Effective Approach to Outlier Detection Based on Centrality and Centre-Proximity. Informatica (Netherlands), 31(3), 435–458. https://doi.org/10.15388/20-INFOR413

Devasia, M. T., Vinushree, M., & Hegde, M. V. (n.d.). Prediction of Students Performance using Educational Data Mining.

Dol, S. M. (2021, January 15). Use of Classification Technique in Educational Data Mining. 2021 International Conference on Nascent Technologies in Engineering, ICNET 2021 - Proceedings. https://doi.org/10.1109/ICNTE51185.2021.9487739

Dol, S. M., & Jawandhiya, P. M. (2022). Use of Data mining Tools in Educational Data Mining. Proceedings - 2022 5th International Conference on Computational Intelligence and Communication Technologies, CCICT 2022, 380–387. https://doi.org/10.1109/CCiCT56684.2022.00075

Hidayat, N., Wardoyo, R., & Sn, A. (n.d.). Educational Data Mining (EDM) as a Model for Students' Evaluation in Learning Environment.

J Md Zubair Rahman Principal, A. M. (n.d.). A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration.

J. Yao, J. Z. and L. W. (2018). A financial statement fraud detection model based on hybrid data mining methods. IEEE Press. https://doi.org/10.1109/ICAIBD.2018.8396167

Kanjanawattana, S. (2019). A novel outlier detection applied to an adaptive K-means. International Journal of Machine Learning and Computing, 9(5), 569–574. https://doi.org/10.18178/ijmlc.2019.9.5.841

Kültür, Y., & Ufuk Çağlayan, M. (n.d.). A Novel Cardholder Behavior Model for Detecting Credit Card Fraud.

Lakshmi Sreenivasa Reddy.D, S.Sailaja, Vijaya Bhaskar Velpula, & Dr.B.Raveendrababu2. (2014). Finding Peculiar Students from Student Database using Outlier Analysis: Data Mining Approach. Institute of Electrical and Electronics Engineers (IEEE).

Min, X., & Lin, R. (2018). K-means algorithm: Fraud detection based on signaling data. Proceedings - 2018 IEEE World Congress on Services, SERVICES 2018, 23–24. https://doi.org/10.1109/SERVICES.2018.00024

Muhammed, L. A. N. (2021). Educational Data Mining: Analyzing Teacher Behavior based Student's Performance. Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021, 181–185. https://doi.org/10.1109/IC2IE53219.2021.9649366

Pamula, R., Deka, J. K., & Nandi, S. (2011). An outlier detection method based on clustering. Proceedings - 2nd International Conference on Emerging Applications of Information Technology, EAIT 2011, 253–256. https://doi.org/10.1109/EAIT.2011.25

Peng', Y., Kou, G., Sabatka2, A., Chen', Z., Khazanchil, D., & Shi3, Y. (n.d.). Application of Clustering Methods to Health Insurance Fraud Detection.

Poonam, & Dutta, M. (2012). Performance analysis of clustering methods for outlier detection. Proceedings - 2012 2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012, 89–95. https://doi.org/10.1109/ACCT.2012.84

Pornthep Rojanavasu. (2019). Educational Data Analytics using Association Rule Mining and Classification. IEEE.

Pramanik, A., Sarker, A., Islam, Z., & Hashem, M. M. A. (2020). Public sector corruption analysis with modified K-means algorithm using perception data. Proceedings of 2020 11th International Conference on Electrical and Computer Engineering, ICECE 2020, 198–201. https://doi.org/10.1109/ICECE51571.2020.9393110

Sagardeep Roy, & Shailendra Narayan Singh. (2017). Emerging Trends in Applications of Big Data in Educational Data Mining and Learning Analytics. IEEE.

Venkatachalapathy, K., Vijayalakshmi, V., & Ohmprakash, V. (2017). Educational data mining tools: A survey from 2001 to 2016. Proceedings - 2017 2nd International Conference on Recent Trends and Challenges in Computational Models, ICRTCCM 2017, 67–72. https://doi.org/10.1109/ICRTCCM.2017.53