Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Análisis y Comparación de Algoritmos No Supervisados para Detectar Riesgo de Deserción Estudiantil
Analysis and Comparison of Unsupervised Algorithms to Detect Student Dropout Risk

Romero-Rodríguez, W.J.G.*, García-De-La Rosa, L.A., González-Páramo, A.
Tecnológico Nacional de México / Instituto Tecnológico Superior de Guanajuato; C.P. 36262 Guanajuato, Guanajuato, México (https://orcid.org/0000-0002-9256-9784, https://orcid.org/0000-0002-0866-9783, https://orcid.org/0000-0002-1363-5772).
wendolyjgrr@gmail.com*; lgarcia@itesg.edu.mx; agonzalez@itesg.edu.mx

Innovación tecnológica: Se presenta en análisis de factores y comparación de algoritmos no supervisados para detectar casos de riesgo de deserción estudiantil.

Área de aplicación industrial: En el área de educación, para apoyar al área de tutoría a detectar los estudiantes posibles a desertar de sus estudios.

Recibido: 09 enero 2025
Aceptado: 08 diciembre 2025

Abstract

The COVID-19 pandemic had significant repercussions across various sectors, with the education sector being one of the most affected. The inclusion, continuity, and timely graduation of students enrolled in higher education are among the priorities established in the General Education Law. To achieve this, strategies and measures must be implemented to promote student retention in higher education institutions. For this reason, one of the primary educational challenges is providing teachers with the necessary tools and resources to identify and refer cases involving violence, mental health concerns, and potential school dropout.

Through the application of data mining techniques in education, it has been possible to predict academic performance, create predictive models for student retention, and define behavioral profiles. A review of the literature has shown that student dropout is influenced by multiple factors, such as academic, economic, and social variables. By utilizing clustering algorithms, more detailed information and insights into dropout patterns can be obtained, supporting informed decision-making in higher education.

In this project, the indicators will be analyzed and defined to monitor the academic performance of students at the ITESG campus in two phases: first using all student features and then using only the relevant features to compare results. This dataset will be managed by an intelligent system that will

use unsupervised artificial intelligence algorithms, such as k-means and fuzzy c-means (FCM), to cluster students and detect potential dropout cases, which can then be referred to the tutoring area. The results suggest that the FCM algorithm performs best for detecting students at risk of dropping out.
The objective of this work is to analyze and compare the performance of the unsupervised algorithms for detecting students at risk of dropping out, using a dataset from the Instituto Tecnológico Superior de Guanajuato and considering both the full set of features and only the most relevant ones identified through statistical tests.

Keywords: k-means, fuzzy c-means, education, school dropout, classification, artificial intelligence.

Resumen

La pandemia del virus COVID-19 tuvo repercusiones significativas en diversos sectores, siendo el sector educativo uno de los más afectados. La inclusión, continuidad y graduación oportuna de los estudiantes inscritos en la educación superior es una de las prioridades de la Ley General de Educación. Para lograr esto, deben implementarse estrategias y medidas que promuevan la retención estudiantil en las instituciones de educación superior. Por esta razón, uno de los principales desafíos educativos es proporcionar a los docentes las herramientas y recursos necesarios para identificar y canalizar casos como la violencia, la salud mental y posibles casos de abandono escolar.
Mediante la aplicación de técnicas de minería de datos en educación, se ha logrado predecir el rendimiento académico, la creación de modelos predictivos para la retención estudiantil y definir perfiles de comportamiento. Una revisión del estado del arte ha concluido que el abandono escolar está influenciado por múltiples factores, como variables académicas, económicas y sociales. Al utilizar algoritmos de agrupamiento, se puede obtener información más detallada y conocimientos sobre los patrones de abandono, lo que permite tomar decisiones informadas en los niveles de educación superior.
En este proyecto, se analizarán y definirán los indicadores para monitorear el rendimiento escolar de los estudiantes del campus universitario ITESG en dos fases: primero con todas las características de los estudiantes y luego solo con las características relevantes para comparar sus resultados. Este conjunto de datos será gestionado por un Sistema Inteligente que hará uso de algoritmos de clasificación no supervisados de Inteligencia Artificial, como K-Means y Fuzzy C-Means (FCM), para agrupar y detectar posibles casos de abandono escolar y poder canalizarlos al área de tutoría. Concluyendo que el algoritmo FCM es el mejor para este trabajo al detectar casos en riesgo de deserción escolar.
El objetivo de este trabajo fué analizar y comparar el desempeño de los algoritmos no supervisados anteriormente mencionados, para detector estudiantes en riesgo de deserción escolar, empleando un conjunto de datos del Instituto Tecnológico Superior de Guanajuato y considerando tanto todas sus características como solo las más relevantes identificadas mediante pruebas estadísticas.

Palabras clave: K-means, Fuzzy c-means, Educación, Deserción escolar, Clasificación, Inteligencia artificial.

1. Introducción

El impacto que tuvo la pandemia por COVID-19 repercutió en diferentes sectores, particularmente al sector educativo. La inclusión, continuidad y egreso oportuno de los estudiantes inscritos en Educación Superior, es una de las prioridades en la Ley General de Educación, por lo que se deben establecer estrategias y medidas para promover su permanencia en instituciones de educación superior [1].

La deserción estudiantil a nivel superior es un desafío global para las instituciones educativas, en el cual sería deseable identificar de manera temprana los casos de estudiantes que se encuentran en riesgo de abandonar sus estudios para poder implementar estrategias para intervenir de manera más efectiva, de tal manera que se puedan aumentar las tasas de retención y así mejorar los índices académicos.

Es de suma importancia el apoyar proyectos académicos que atiendan las causas del abandono escolar, las cuales incrementaron a causa de la pandemia por COVID-19, el cual es uno de los objetivos principales del Programa de Expansión de la Educación Media Superior y Superior, el cual se justifica de acuerdo con el Programa Sectorial de Educación 2020-2024 [2].

De acuerdo con las estadísticas que reporta la Secretaría de Educación Pública (SEP) para Guanajuato, de cada 100 estudiantes que iniciaron el ciclo escolar 2001-2022, solo 19 egresaron de educación de nivel superior. Siendo 5 menos que en el resto del país y 27 menos que en la Ciudad de México [2]. Es por ello, que uno de los retos educativos es que los docentes cuenten con las herramientas y recursos para detectar y canalizar casos de violencia, salud mental y posibles casos de deserción escolar.

Por medio de la aplicación de técnicas de minería de datos en la educación, se ha podido predecir el desempeño, creación de modelos predictivos para la permanencia escolar y definir perfiles de comportamiento. De acuerdo con el estado del arte, se ha concluido que la deserción escolar no depende de un solo factor y se pueden aplicar diversos algoritmos para obtener más información respecto deserción a nivel superior [3]. La minería de datos educativos (Educational Data Mining) es un campo emergente, en el que se aplican técnicas de inteligencia artificial para explorar y analizar el rendimiento escolar en estudiantes, teniendo como objetivo principal el prevenir la deserción escolar, retroalimentar a docentes y tutores para la búsqueda de mejoras en el proceso de aprendizaje [4].

Los algoritmos de clasificación no supervisada forman parte de las técnicas de Inteligencia Artificial que están siendo ampliamente usados como herramientas innovadoras para el monitoreo de casos de riesgo. A diferencia de los algoritmos de clasificación supervisados, estos no requieren de etiquetas previamente asignadas, lo que los convierte en ideales para explorar y detectar patrones ocultos que permitan segmentar a los estudiantes en grupos según sus características. Algunas escuelas han aplicado clusterización y técnicas de minería de datos para analizar el rendimiento académico de estudiantes universitarios, tal como [5], en el cual se usó análisis de clusterización para analizar el rendimiento de los estudiantes y poder distinguir las categorías de cada alumno. Al cual agregaron el uso de un algoritmo K-means, combinado con un modelo determinista para analizar el desempeño de los estudiantes [6].

En el trabajo [7], se han aplicado algoritmos de clusterización para detectar seis grupos de estudiantes, de acuerdo con el análisis de los patrones de interacción.

En el estado del arte se han identificado que para el análisis de rendimiento escolar de estudiantes se han usado algoritmos de clasificación supervisada [8] [9] y no supervisada [10] [11] [5] [12], con buenos resultados en la minería de datos educativos, con el objetivo de analizar los datos educativos y usa los datos existentes, mejorando la calidad de la educación y proceso de aprendizaje.

En este estudio se tiene como objetivo analizar y definir los indicadores clave para monitorear el desempeño escolar de los estudiantes del Instituto Tecnológico Superior de Guanajuato (ITESG), los cuales posteriormente serán gestionados mediante un Sistema Inteligente que implementará algoritmos de clasificación no supervisada para identificar casos potenciales de deserción escolar y canalizarlos al área de tutoría correspondiente. Se propuso el uso de los algoritmos K-Means y Fuzzy C-Means (FCM), los cuales serán comparados en términos de las métricas de evaluación para determinar cuál resulta más adecuado en este análisis. Los resultados de este análisis son importantes para la prevención temprano de estudios en riesgo de deserción y al diseño de estrategias de intervención enfocadas en la retención estudiantil a nivel superior. Por tanto, el objetivo principal de esta investigación es comparar el rendimiento de los algoritmos K-Means y Fuzzy C-Means en la detección de patrones de deserción estudiantil en dos fases de análisis, una con todas las variables y otra con las características más significativas, con el fin de determinar cuál ofrece una clasificación más precisa y útil para la toma de decisiones institucionales.

Marco Teórico

a. Factores asociados a la deserción escolar

La deserción escolar ha sido identificada por un conjunto de diversos factores que pueden afectar el rendimiento y continuidad académica, los cuales se pueden agrupar en categorías como académicos, económicos, sociales e incluso psicológicos. A partir del estado del arte se pudieron encontrar diversos trabajos en los cuales se han aplicado diversas técnicas de clasificación supervisada y no supervisada, con el objetivo de categorizar a los estudiantes en base a su rendimiento.

Como se ha mencionado anteriormente, algunas escuelas han usado técnicas de agrupamiento y minería de datos para categorizar el rendimiento académico de los estudiantes. En algunos de ellos se han aplicado Arboles de decisión, para clasificar a los estudiantes en riesgo de deserción, tal como en [13] en el cual se obtuvo un modelo en el cual la cantidad de asignaturas aprobadas fue una variable significativa por encima de las demás, así como se concluyó que la calificación del examen de admisión fue la menos significativa de todos los demás factores.

En [14] se hace uso de árboles de inferencia condicional para realizar una clasificación binaria con el objetivo de predecir si un alumno se graduará o terminará por desertar de sus estudios, cuyos factores fueron las calificaciones de escuela secundaria y satisfacción del estudiante. En [15] también se hace uso de árboles de decisión en conjunto con regresiones logísticas, donde los factores a tomar en cuenta fueron el género, origen (alemán o no alemán), número de intentos para examen, semestre, resultado de examen, promedio de exámenes reprobados y promedio de exámenes aprobados.

[16] estudió qué características podrían ayudar a predecir el abandono escolar de los estudiantes, usando técnicas como redes neuronales y regresión logística, concluyendo que los factores más significativos fueron la

cantidad de créditos acumulados, cantidad de cursos reprobados y número de actividades realizadas en la plataforma Moodle.

De acuerdo con [17], se pudo predecir de manera correcta el 91% de los estudiantes desertores y no desertores, tomando en cuenta 10 variables y teniendo información de estudiantes de todos los semestres. En este estudio se aplicaron algoritmos como Bosques Aleatorios y Máquinas de Soporte Vectorial, en el cual los factores usados fueron género, residencia actual, apoyos económicos, cursos aprobados y cursos que necesita para graduarse.

A continuación, en la Tabla 1 se presenta un resumen de los factores más comunes que pueden ayudar a predecir la deserción escolar de acuerdo con las referencias consultadas, en la cual se puede ver el número de coincidencias que tiene cada factor y el nivel de impacto para poder tener una idea de qué factores pueden ser lo más significativos y así poder elegir estos para nuestras pruebas.

Tabla 1. Factores más comunes de deserción escolar.

Clasificación	Factor	Número de coincidencias	Nivel de impacto
Académicos	Resultados en prueba de admisión	7	Alto
	Promedio de notas	12	Alto
	Cantidad de asignaturas aprobadas	5	Medio
	Porcentaje de asistencia a clases	1	Bajo
	Grado de satisfacción con la carrera elegida	3	Medio
Económicos	Apoyos económicos por Institución	1	Bajo
	Ingreso familiar	2	Bajo
Sociales	Zona de residencia	6	Medio
	Estado Civil	1	Bajo
	Género	1	Bajo
	Sexo	2	Bajo
	Edad	4	Medio
	Máximo nivel educativo alcanzado por los padres	3	Medio

Aunque los factores psicológicos presentan menor incidencia cuantitativa en las referencias analizadas, se incluyeron variables relacionadas como "problemas personales" y "salud mental", dado su impacto comprobado en la retención estudiantil [3] [15], lo cual permite una visión más integral del fenómeno. Cabe destacar que, dentro de la Coordinación de Tutorías del ITESG, se aplican encuestas a los alumnos que han desertado, y estos factores han sido de los más señalados como causas asociadas a la deserción escolar.

b. Algoritmo K-Means

El algoritmo K-Means es uno de los más populares para resolver problemas de clusterización. Este algoritmo intenta agrupar n elementos en subgrupos k definidos por el usuario, donde k debe ser menor o igual que n. La agrupación se realiza de manera iterativa, minimizando la suma de las distancias al cuadrado y los centroides de los elementos hasta que ya no haya cambios en la estructura o se alcance un umbral [18].

Este algoritmo se puede resumir en el siguiente pseudocódigo:

1. Se define el número de clústeres o grupos K en los cuales desea clasificar los datos.
2. Se eligen puntos iniciales para los centroides K, los cuales pueden ser elegidos de manera aleatoria o mediante métodos como K-means++.
3. Se mide la distancia de cada punto en el conjunto de datos hacia cada centroide y se le asigna el clúster más cercano.
4. Se recalcula el centroide de cada clúster como el promedio de las coordenadas de los puntos asignados a ese clúster.
5. Repetir pasos 3 y 4 hasta que los centroides ya no cambien significativamente o se haya alcanzado el máximo número de iteraciones.

c. Algoritmo Fuzzy C-Means

El algoritmo Fuzzy C-Means (FCM), es una técnica de clusterización usada en problemas en los cuales los datos no tienen etiquetas asignadas [19]. Su principal diferencia a comparación del algoritmo K-Means es que, para este algoritmo, un dato puede pertenecer a más de un clúster de manera simultánea, teniendo diferentes grados de pertenencia que se representan por valores en un rango entre 0 y 1. Esto lo convierte en un algoritmo popular para modelar incertidumbre y manejar datos difusos.

Este algoritmo se puede resumir de la siguiente manera [20]:

Entradas:

- u_{ij} es el grado de pertenencia de x_i al clúster j.
- x_i es el i-ésimo de los datos medidos de la d-dimensión.
- c_j es el centro de la dimensión d del clúster.

1. Inicializar aleatoriamente u_{ij}
2. Calcular los centros de los clústeres, usando la ecuación:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

3. Actualizar la nueva matriz de partición difusa, usando la ecuación:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. Repetir los pasos 2 y 3 hasta que se cumpla el criterio de parada.

Aunque los algoritmos supervisados como árboles de decisión o redes neuronales han demostrado dar buenos resultados en la predicción para datos, requieren conjuntos de datos etiquetados. En contraste, los algoritmos no supervisados como K-Means y Fuzzy C-Means permiten identificar patrones

y agrupaciones sin conocimiento previo, resultando ideal en contextos donde la información de deserción escolar no está completamente definida o se requiere descubrir grupos potenciales de riesgo.

2. Metodología

La metodología se basa en aplicar técnicas de minería de datos y análisis mediante algoritmos de clasificación no supervisada. El procesamiento y análisis de datos se realizaron con Python versión 3.11.3 utilizando las librerías pandas, scikit-learn, numpy y matplotlib para la manipulación, normalización, visualización y evaluación de resultados.

El algoritmo Fuzzy C-Means fue implementado con la librería fcmeans y la selección de características se realizó mediante la función chi^2 del módulo sklearn.feature_selection.

a. Dataset y características

Se realizó la recolección de los datos académicos históricos del 2019 al 2023 de estudiantes del campus Instituto Tecnológico Superior de Guanajuato (ITESG) de la carrera de Ingeniería en Sistemas Computacionales, entre los cuales se tienen datos como sus calificaciones, asistencia, cursos aprobados y no aprobados, datos socioeconómicos, nivel educativo de los padres o tutores, información institucional como si el estudiante cuenta con apoyos financieros o si ha sido canalizado a alguna tutoría psicológica o académica y además los diversos motivos como problemas personales, salud mental, falta de apoyo económico, administrativos, etc. que les hicieron desertar. Esto se hizo con la finalidad de poder analizar los datos necesarios que pudieran ayudar a medir el desempeño escolar y su posterior categorización de deserción escolar. Para ello se usaron entrevistas y mesas de trabajo en colaboración con el área de tutorías del campus, el cual proporcionó datos de 150 estudiantes entre los cuales 50 desertaron, 50 fueron graduados y 50 actualmente están inscritos, pero que se consideran irregulares.

En base al estado del arte y a las bases de datos institucionales del campus consultado, se identificaron los atributos relevantes que pudieran ayudar a generar indicadores para detectar posibles casos de riesgo de deserción en alumnos del ITESG, entre los cuales se encuentran factores externos e internos como:

- Residencia actual (distancia en kilómetros a la ciudad del campus).
- Nivel académico máximo alcanzado por padres (No cuenta o Primaria, Secundaria o Preparatoria, Estudios Superiores o Estudios Posgrado).
- Problemas personales (Si / No)
- Hijos (Si / No)
- Trabaja (Si / No)
- Estado civil (Libre / En compromiso)
- Embarazo (Si / No)
- Problemas de salud mental (Si / No)
- Adicciones (Si / No)
- Enfermedad diagnosticada (Si / No)
- Apoyo económico institucional (Si / No)
- Asignaturas en repetición (Número)
- Asignaturas en especial (Número)
- Asistencia a clases (0 – 10, donde 0 significa poca asistencia y 10 es buena asistencia)
- Promedio actual (Número)
- Semestre actual (Número)

b) Preprocesamiento de los datos

Fue necesario transformar las variables categóricas, como estado civil y problemas personales, entre otras, en valores numéricos. Para ello se asignaron valores binarios para representar la presencia o ausencia de cada categoría, donde el 0 indica ausencia o "No" y el 1 representa presencia o "Si". En el caso del nivel académico máximo alcanzado por los padres, se asignó 0 cuando el valor correspondía a "No cuenta o Primaria", 1 para "Secundaria o Preparatoria" y 2 para "Estudios Superiores o Posgrado". Para el factor de Estado civil, se asignó un 0 para el valor de "Libre" y un 1 en caso de "En compromiso". Esta transformación permitió que tanto K-Means como Fuzzy C-Means pudieran procesar correctamente los datos sin sesgos asociados a la naturaleza cualitativa de algunas variables.

Además, durante el preprocesamiento se revisó la base de datos para identificar los valores nulos o inconsistentes. Los registros con valores faltantes en variables clave fueron eliminados y se verificó la presencia de valores atípicos mediante el método de rango intercuartílico. Los valores que distorsionaban significativamente las distribuciones se ajustaron o eliminaron. Esto garantizó la consistencia y calidad del conjunto de datos antes de aplicar los algoritmos.

Para garantizar que los algoritmos de clasificación no supervisada traten todas las características con la misma importancia, sin importar las unidades o rangos en que se encuentran, es importante el escalado o normalización. Esto mejorará la convergencia de K-Means y Fuzzy C-Means. En este trabajo se aplicó la técnica de Normalización StandardScaler de Python [21] [22], la cual convierte los datos para tener:

$$z = \frac{x - \mu}{\sigma}$$

Donde:

- z es el valor escalado o estandarizado de la característica x.
- x es el valor actual de la característica x.
- μ es la media de la característica x en el conjunto de datos.
- σ es la desviación estándar de la característica estandarizada en el conjunto de datos.

Para aplicar los algoritmos de clasificación no supervisada como K-Means y Fuzzy C-Means (FCM), se cuenta con un conjunto de datos que contienen 16 características que representan a cada estudiante y su etiqueta de riesgo alto, medio o bajo de deserción. Cabe mencionar que, aunque en este análisis se están implementando algoritmos de clasificación no supervisada, las etiquetas sólo se tomarán en cuenta para la generación de las matrices de confusión y así poder evaluar el desempeño del algoritmo para agrupar.

En la primera fase del análisis, con el fin de identificar los indicadores o factores que permitan detectar casos de riesgo de deserción en los estudiantes de nivel superior, se usaron las 16 características completas para aplicar los algoritmos de clasificación. De esta forma se pudo evaluar el desempeño de ambos algoritmos para agrupar a los estudiantes en 3 clústeres. Estos 3 grupos representan a los estudiantes que se encuentran en un riesgo alto de desertar, los que están en peligro medio de deserción y los que se predice que serán graduados sin inconvenientes. Las características elegidas para iniciar la primera fase fueron seleccionadas de acuerdo a su impacto en el estado del arte y en los datos recopilados por

ITESG, algunos de los originales como Resultados en prueba de admisión.

Para la segunda fase del análisis, se ha aplicado la Prueba de Chi-cuadrado, la cual es una técnica de selección de las características más relevantes de un conjunto de datos que influyen en su clusterización [21]. Esta herramienta es importante para reducir el número de dimensiones de los datos, la cual hace un filtrado de las características o factores que obtuvieron un valor $p < 0.05$ y se muestran en la Tabla 2, las cuales se consideran como las características más relevantes y las cuales se usarán para la aplicación de los algoritmos de clusterización.

Tabla 2. Características más relevantes aplicando Prueba de Chi-cuadrado con base en datos de ITESG.

Característica	Valor Chi^2	P-Valor
Residencia Actual	9.363636	0.009262158
Nivel Académicos padres	52.354839	0
Trabaja	17.705882	0.000142961
Casado	29.485149	0.000000396
Asignaturas en repetición	136.721429	0
Asignaturas en especial	137.550173	0
Asistencia	1538.59214	0
Promedio Actual	3059.14825	0

De acuerdo con la prueba Chi-cuadrado, se filtraron 8 características identificadas como relevantes y se usaron para reentrenar los algoritmos K-Means y Fuzzy C-Means (FCM) y poder realizar su posterior comparación con los resultados obtenidos entre la Fase 1 y 2 del análisis, con esto se puede concluir si se mejora la agrupación de los datos con las características relevantes. En la Tabla 3 se puede observar la comparación de los factores de deserción identificados en el estado del arte consultado y los factores que se tienen recopilados en los datos recopilados por los estudiantes del Instituto Tecnológico Superior de Guanajuato, a los cuales se les aplicó la Prueba Chi-Cuadrada para categorizar el nivel de impacto que pueden tener en la clusterización si se toman en cuenta. Hay algunos factores que no se encontraron en las referencias consultadas, pero que en las pruebas realizadas denotan tener un impacto alto para agrupar los datos y con los cuales, si se cuenta en el conjunto de datos recopilado, tal como la cantidad de asignaturas en repetición, cantidad de asignaturas en especial, cuenta con trabajo o no. Existen otros factores como semestre actual y cuenta con hijos o no, que se vieron como factores de nivel de impacto medio y que no se habían tomado en cuenta en el estado del arte, lo cual podría ser una aportación como factores en la detección de estudiantes en riesgo de deserción.

El análisis se dividió en dos fases complementarias:

- Fase 1: Se incluyen las 16 características para observar el comportamiento general del modelo.
- Fase 2: Se aplicó la prueba de Chi-cuadrado para seleccionar las variables más relevantes en la agrupación de clústeres.

De esta manera, el objetivo fue determinar si un menor número de características relevantes puede mejorar la precisión y coherencia de los grupos formados.

Tabla 3. Comparación de factores de deserción estado del arte y Datos recopilados con base ITESG.

Clasificación	Factor	Nivel de impacto	
		Estado del arte	Datos recopilados
Académicos	Resultados en prueba de admisión	Alto	N/A
	Promedio de notas	Alto	Alto
	Cantidad de asignaturas aprobadas	Medio	N/A
	Cantidad de asignaturas en repetición	N/A	Alto
	Cantidad de asignaturas en especial	N/A	Alto
	Porcentaje de asistencia a clases	Bajo	Alto
	Grado de satisfacción con la carrera elegida	Medio	N/A
	Semestre Actual	N/A	Medio
Económicos	Apoyos económicos por Institución	Bajo	Medio
	Ingreso familiar	Bajo	N/A
	Trabaja	N/A	Alto
Sociales	Zona de residencia	Medio	Alto
	Estado Civil	Bajo	Alto
	Género	Bajo	N/A
	Edad	Medio	N/A
	Máximo nivel educativo alcanzado por los padres	Medio	Alto
	Hijos	N/A	Medio
	Embarazo	N/A	Bajo
	Problemas de Salud mental	N/A	Bajo
	Enfermedad Diagnosticada	N/A	Bajo
	Problemas personales	N/A	Bajo
	Adicciones	N/A	Bajo

Cabe mencionar que algunas variables utilizadas en el estado del arte, como el grado de satisfacción con la carrera, ingreso familiar, género o edad, no fueron incluidas en el presente análisis debido a que dichos datos no se encontraban disponibles en la base institucional proporcionada por el ITESG. Asimismo, la variable cantidad de asignaturas aprobadas fue omitida por presentar alta correlación con los factores asignaturas en repetición y asignaturas en especial, los cuales mostraron un impacto significativo en las pruebas exploratorias. Estos factores, aunque no aparecen ampliamente documentados en el estado del arte, emergieron como variables de alta influencia en el contexto local, aportando una perspectiva complementaria a los estudios previos sobre deserción escolar.

c) Agrupación de datos

Para evaluar la calidad y consistencia de los clústeres generados por los algoritmos, se aplicaron diversas métricas. En primer lugar, se utilizó el coeficiente de Silhouette [24], que mide simultáneamente la cohesión interna y la separación entre grupos, indicando qué tan bien se encuentran definidos los clústeres. Además, se analizaron métricas derivadas de las matrices de confusión (exactitud, precisión, sensibilidad y F1-Score), con las cuales se pudo comparar el rendimiento de los algoritmos K-Means y Fuzzy C-Means en ambas fases del estudio.

Estas métricas ofrecen una valoración objetiva sobre la capacidad de cada modelo para clasificar correctamente los estudiantes en los diferentes niveles de riesgo de deserción.

Como se mencionó anteriormente, para elegir el número óptimo de clústeres k que pueden agrupar los datos, se aplicó la métrica de coeficiente de Silhouette. En la Figura 1, se muestra el coeficiente de Silhouette promedio para diferentes números de clústeres k, en el cual se puede ver que el valor máximo se encuentra cuando $k = 2$, lo cual indica que los datos agrupados en 2 clústeres pueden tener mayor cohesión y su separación es más significativa. Cuando $k > 5$, los clústeres no están tan definidos y algunos datos pueden ser agrupados erróneamente. Con $k = 3$, se considera un valor aceptable de clústeres para tener una segmentación suficiente sin añadir complejidad innecesaria.

Es por ello por lo que el modelo se ha reducido a 3 categorías, las cuales se pueden interpretar como:

- Riesgo bajo: estudiantes con buen desempeño académico y que no se consideran que pudieran desertar de sus estudios.
- Riesgo medio: estudiantes que tienen indicadores de alerta que les pudiera hacer desertar de sus estudios.
- Riesgo alto: estudiantes con alto índice de desertar de sus estudios, debido a problemas académicos y/o personales.

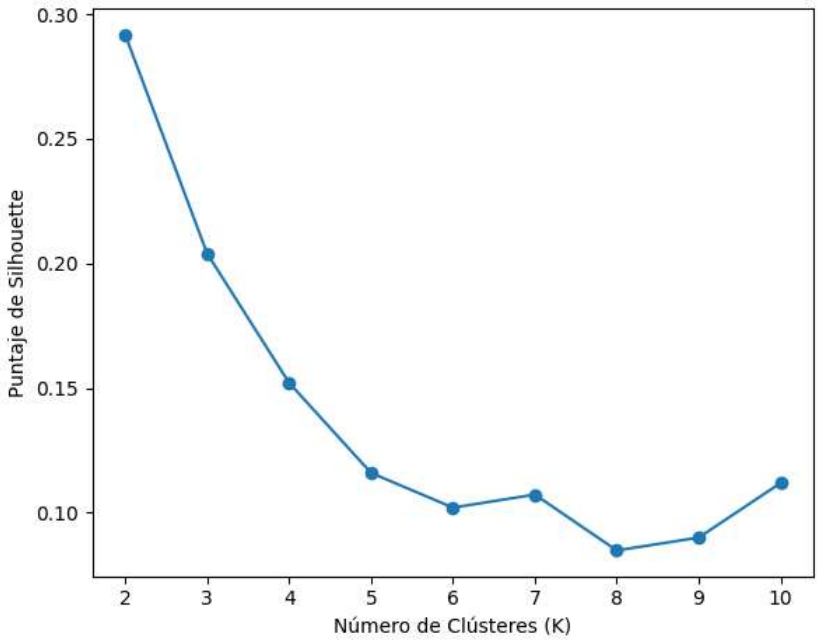

Figura 1. Coeficiente de Silhouette.

Todas las figuras y tablas son autoría propia, generadas con datos proporcionados por el área de tutoría del ITESG. El conjunto de datos utilizado tiene carácter confidencial, ya que contiene información sensible de los estudiantes y no puede ser publicado de manera abierta.

En el caso del algoritmo Fuzzy C-Means, los parámetros se definieron inicialmente conforme a los valores más reportados en el estado del arte, con el fin de mantener la comparabilidad de los resultados. Posteriormente, dichos parámetros se

ajustaron en función del rendimiento observado en las matrices de confusión, seleccionando la combinación que mostró mayor precisión y estabilidad en la clasificación. Los valores finales se reportan en la sección de Resultados.

3. Resultados

Es importante el mostrar una visualización de los clústeres generados por los algoritmos de clasificación no supervisada, esto debido a que se debe evaluar qué tan bien están agrupando los datos. Sin embargo, en este caso se están manejando datos con 16 dimensiones en la primera fase de las pruebas y 8 dimensiones en la segunda fase, por lo que es difícil interpretar gráficamente los clústeres en su espacio inicial. Es por esto que se usa el Análisis de Componentes Principales (PCA), para reducir el número de dimensiones a 2 principales. Estas dos dimensiones permiten representar los datos en un espacio bidimensional sin que se pierda información importante para los clústeres.

En la Figura 2 se presenta la visualización de los tres clústeres generados en la primera fase mediante el algoritmo K-Means, utilizando las 16 características consideradas. En esta representación, el clúster 2 agrupa a los 77 estudiantes clasificados con bajo riesgo de deserción, el clúster 1 corresponde a los 23 estudiantes con riesgo medio de deserción, y el clúster 0 incluye a los 50 estudiantes con alto riesgo de deserción.

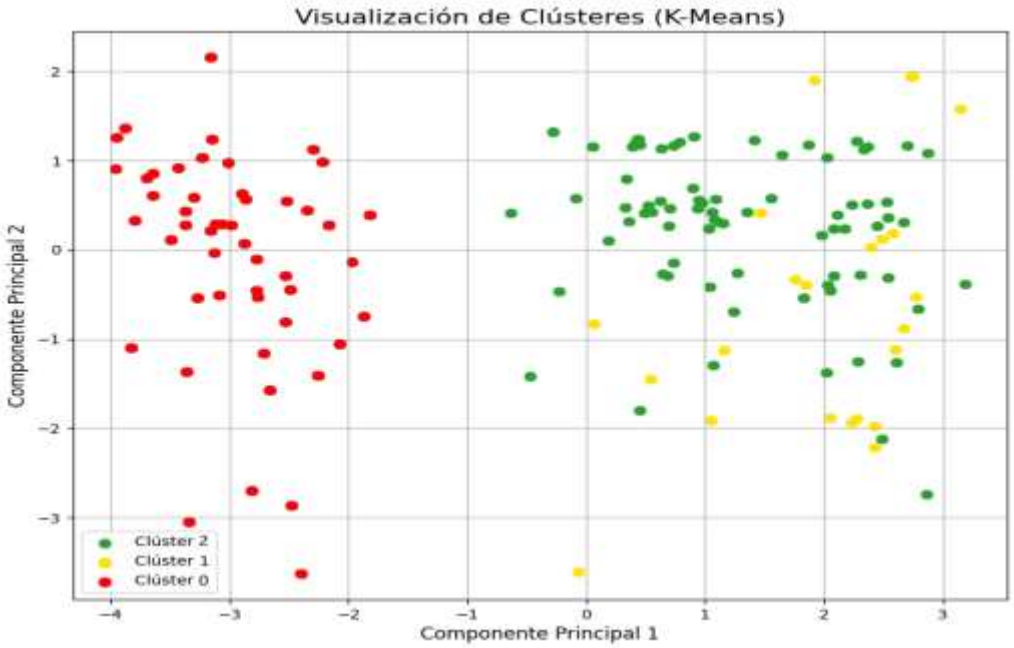

Figura 2. Visualización de Clústeres (K-Means) usando 16 características.

La Figura 3 presenta la matriz de confusión generada por la primera fase mediante el algoritmo K-Means, utilizando las 16 características consideradas, en la cual se puede observar que se clasifican de manera correcta por completo a los 50 estudiantes que estaban clasificados en riesgo alto de deserción, 7 estudiantes correctamente en riesgo medio y 34 estudiantes correctamente clasificados en riesgo bajo de deserción. Sin embargo, los restantes fueron clasificados de manera incorrecta, 43 estudiantes de riesgo medio fueron clasificados como riesgo bajo, mientras que 16 estudiantes de riesgo bajo fueron clasificados como riesgo medio.

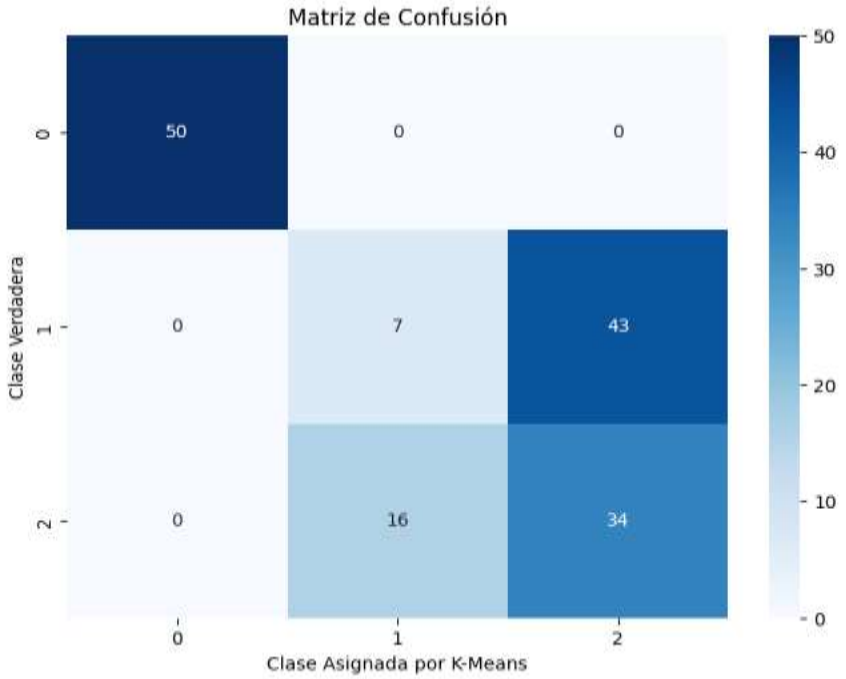

Figura 3. Matriz de confusión (K-Means) usando 16 características.

En la Figura 4 presenta la visualización de los tres clústeres generados en la primera fase mediante el algoritmo Fuzzy C-Means, utilizando las 16 características consideradas y los siguientes valores de entrada:

- m=2, peso de fuzzificación[2].
- error=0.005, error de tolerancia.
- 1000 iteraciones.

En esta representación, el clúster 2 agrupa a los 60 estudiantes clasificados con bajo riesgo de deserción, el clúster 1 corresponde a los 40 estudiantes con riesgo medio de deserción, y el clúster 0 incluye a los 50 estudiantes con alto riesgo de deserción.

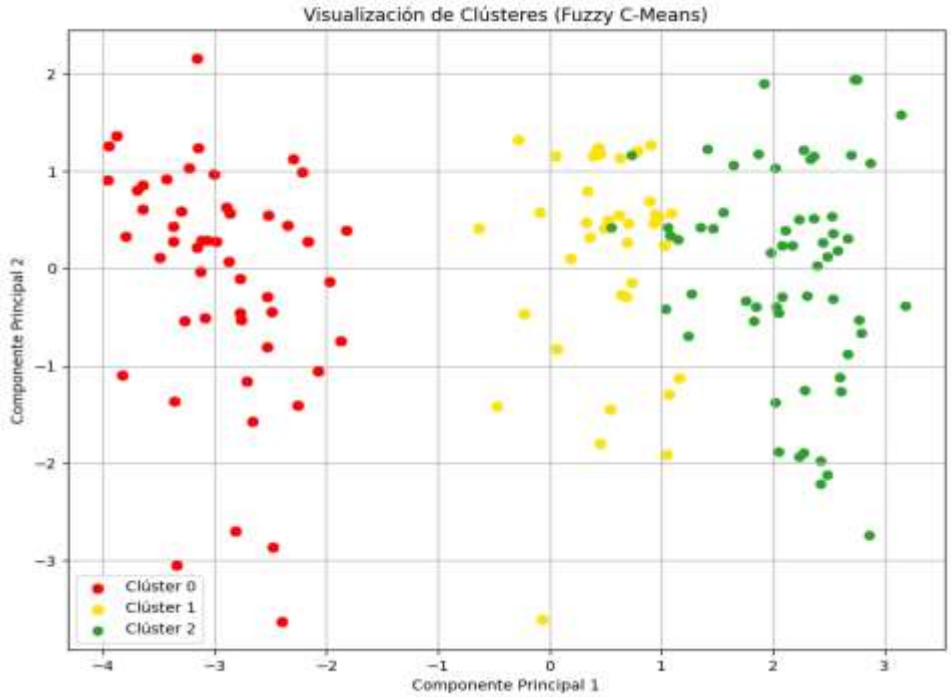

Figura 4. Visualización de Clústeres (Fuzzy C-Means) usando 16 características.

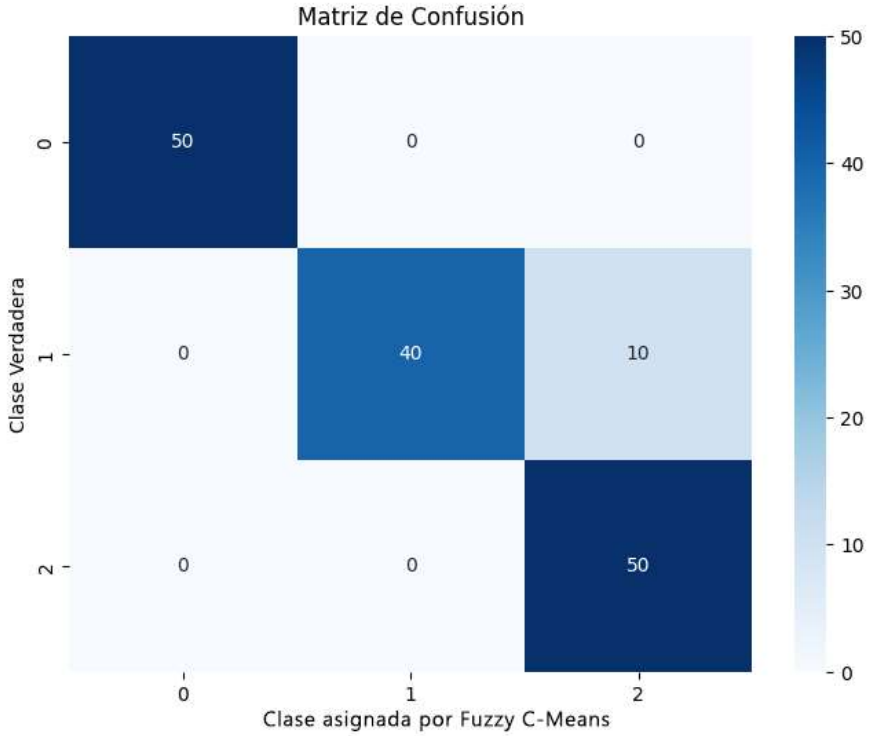

Figura 5. Matriz de confusión (Fuzzy C-Means) usando 16 características.

La Figura 5 presenta la matriz de confusión generada por la primera fase mediante el algoritmo Fuzzy C-Means, utilizando las 16 características consideradas, en la cual se puede observar que se clasifican de manera correcta por completo a los 50 estudiantes que

estaban clasificados en riesgo alto de deserción, 40 estudiantes correctamente en riesgo medio y 50 estudiantes correctamente clasificados en riesgo bajo de deserción. Sin embargo, los restantes fueron clasificados de manera incorrecta, 10 estudiantes de riesgo medio fueron clasificados como riesgo bajo.

En la Tabla 4 se presenta una comparativa de las métricas de evaluación de las matrices de confusión de ambos algoritmos usando 16 características, estas métricas son exactitud, precisión, sensibilidad y F1 Score. Claramente el algoritmo Fuzzy C-Means resultó obtener mejores valores de clasificación para la primera fase en la cual se toman en cuenta las 16 características de los datos, además que su matriz de confusión para el FCM también muestra una mejor agrupación correcta de los estudiantes en los 3 clúster.

Tabla 4. Comparativa de métricas matriz de confusión (16 características).

Métrica	K-Means	Fuzzy C-Means
Exactitud	0.61	0.93
Precisión	0.58	0.94
Sensibilidad	0.61	0.93
F1-Score	0.58	0.93

Para la segunda fase, se tomaron en cuenta las 8 características relevantes tal como se menciona en la Sección de Metodología. En la Figura 6 se presenta la visualización de los tres clústeres generados en esta segunda fase mediante el algoritmo K-Means, utilizando las 8 características relevantes. En esta representación, el clúster 2 agrupa a los 30 estudiantes clasificados con bajo riesgo de deserción, el clúster 1 corresponde a los 21 estudiantes con riesgo medio de deserción, y el clúster 0 incluye a los 99 estudiantes con alto riesgo de deserción.

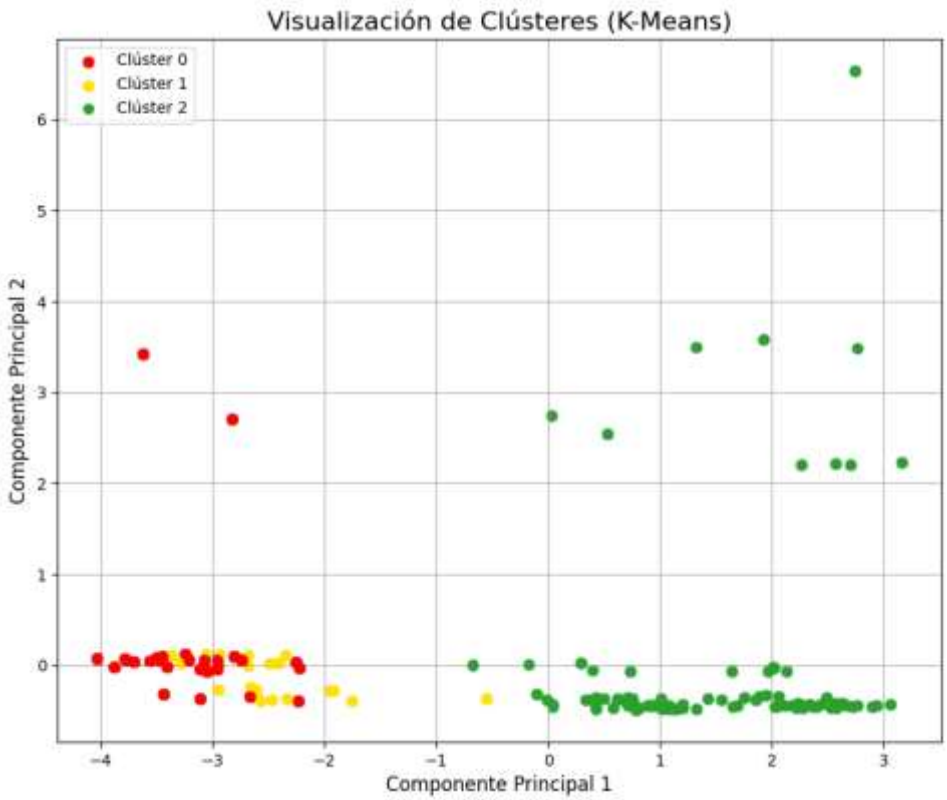

Figura 6. Visualización de Clústeres (K-Means) usando 8 características relevantes.

La Figura 7 presenta la matriz de confusión generada por la segunda fase mediante el algoritmo K-Means, utilizando las 8 características relevantes, en la cual se puede observar que se clasifican de manera correcta por completo a los 30 estudiantes que estaban clasificados en riesgo alto de deserción, 1 estudiante correctamente en riesgo medio y 50 estudiantes correctamente clasificados en riesgo bajo de deserción.

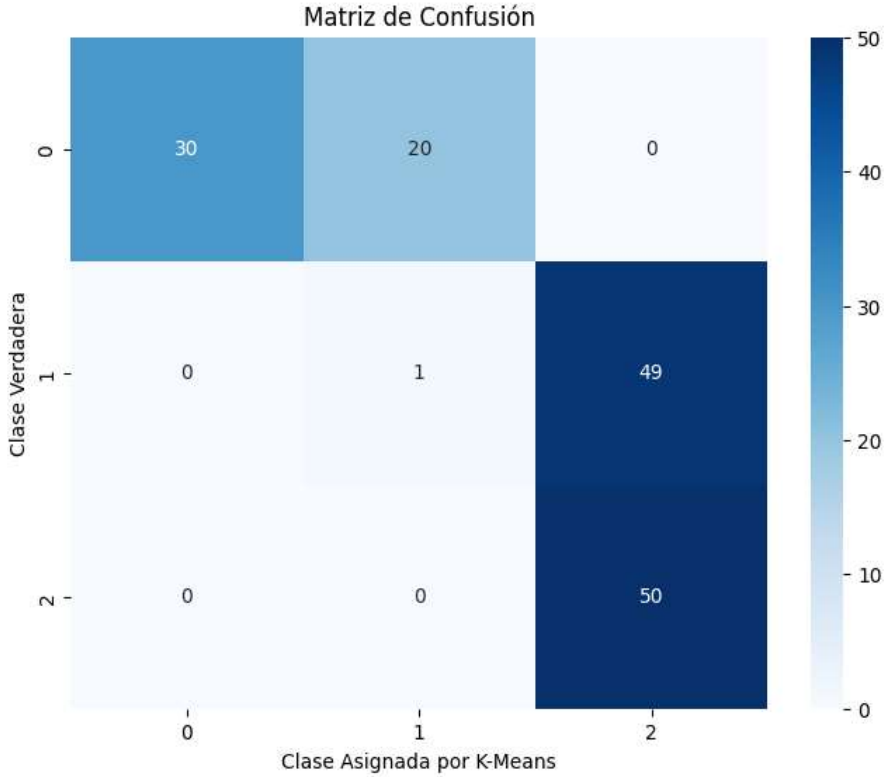

Figura 7. Matriz de confusión (K-Means) usando 8 características relevantes.

En la Figura 8 se presenta la visualización de los tres clústeres generados en la segunda fase mediante el algoritmo Fuzzy C-Means, utilizando las 8 características relevantes. En esta representación, el clúster 2 agrupa a los 51 estudiantes clasificados con bajo riesgo de deserción, el clúster 1 corresponde a los 49 estudiantes con riesgo medio de deserción, y el clúster 0 incluye a los 50 estudiantes con alto riesgo de deserción.

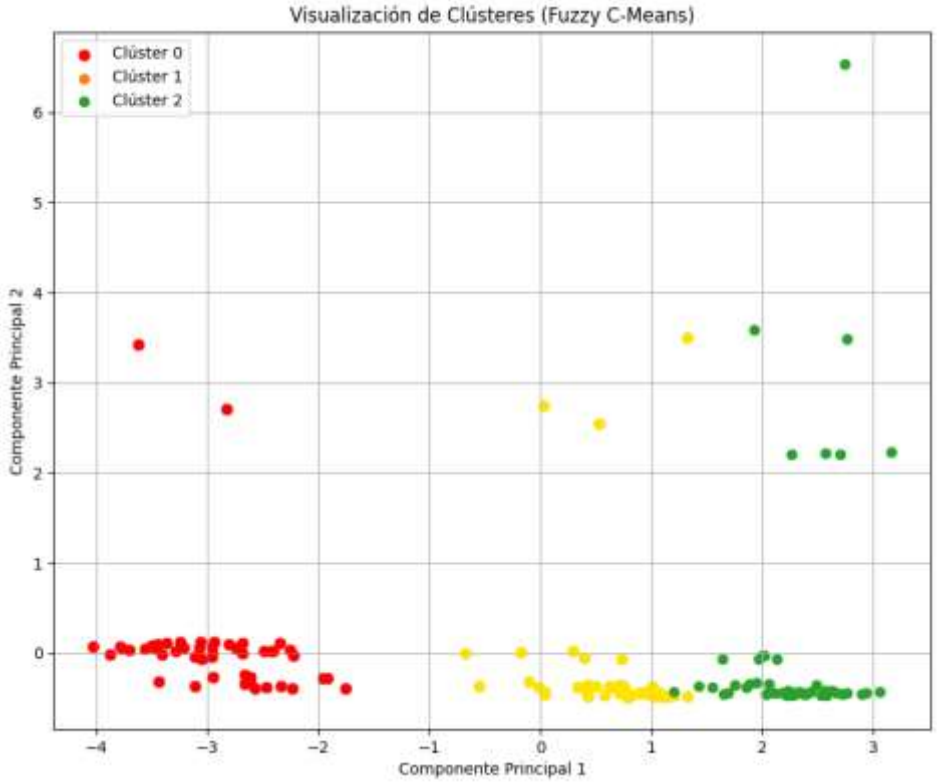

Figura 8. Visualización de Clústeres (Fuzzy C-Means) usando 8 características relevantes.

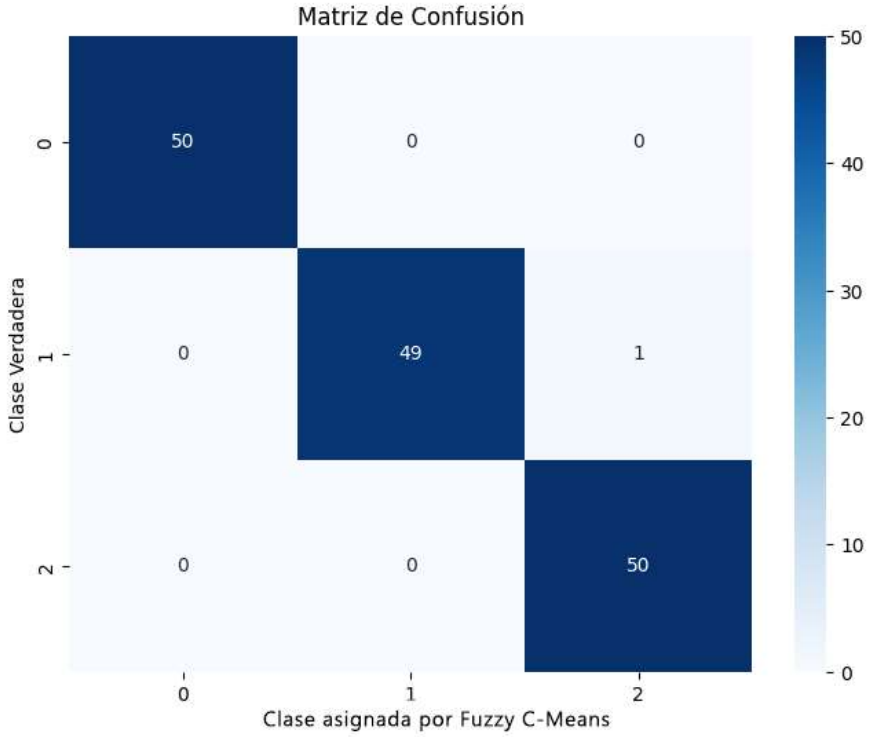

Figura 9. Matriz de confusión (Fuzzy C-Means) usando 8 características relevantes.

La Figura 9 presenta la matriz de confusión generada en la segunda fase mediante el algoritmo Fuzzy C-Means, utilizando las 8 características relevantes, en la cual se puede observar que se clasifican de manera correcta por completo a los 50 estudiantes que estaban clasificados en riesgo alto de deserción, 49 estudiantes correctamente en riesgo medio y 50 estudiantes correctamente clasificados en riesgo bajo de deserción. Esto denota que usando únicamente 8 de las 16 características contempladas inicialmente, se puede obtener una mejor categorización de los estudiantes en los niveles de riesgo de deserción.

En la Tabla 5 se presenta una comparativa de las métricas de evaluación de las matrices de confusión de ambos algoritmos usando 16 características, estas métricas son exactitud, precisión, sensibilidad y F1 Score. El algoritmo Fuzzy C-Means resultó nuevamente obtener mejores valores de clasificación a comparación del K-Means, para esta segunda fase en la cual se toman en cuenta las 8 características relevantes de los datos.

Tabla 5. Comparativa de métricas matriz de confusión (8 características).

Métrica	K-Means	Fuzzy C-Means
Exactitud	0.54	0.99
Precisión	0.52	0.99
Sensibilidad	0.54	0.99
F1-Score	0.48	0.99

Aunque los algoritmos aplicados son de tipo no supervisado, se usaron matrices de confusión únicamente con fines de evaluación externa del desempeño. Las etiquetas originales del conjunto de datos (estudiantes desertores, graduados y en riesgo) no se emplearon durante el proceso de agrupamiento, sino únicamente para comparar los resultados obtenidos y medir el grado de correspondencia entre los grupos formados por los algoritmos y las categorías reales. Este procedimiento permite validar la calidad de los clústeres generados sin violar el principio del aprendizaje no supervisado.

Los resultados obtenidos tanto en primera y segunda fase de los algoritmos K-Means y Fuzzy C-Means (FCM), mostraron que el algoritmo Fuzzy C-Means superó en las métricas evaluadas. FCM mostró una mayor capacidad para representar la pertenencia de las muestras a los clústeres y así poder clasificar a los estudiantes en 3 grupos en función de su riesgo de deserción escolar. Si se comparan los resultados del Fuzzy C-Means usando 16 y 8 características, se concluye que tiene mejor rendimiento para clasificar los estudiantes usando únicamente las 8 características relevantes: residencia actual, nivel académico de padres, cuenta con trabajo, estado civil, cantidad de asignaturas en repetición, cantidad de asignaturas en especial, promedio de asistencia a las clases y promedio actual de calificaciones en todas las asignaturas.

Además, en el análisis del coeficiente de Silhouette confirmó que los clústeres asignados por FCM presentaron una mejor separación respecto a los generados por K-Means, lo cual lo convierte en un algoritmo de clasificación no supervisada adecuado para este análisis de deserción estudiantil para el campus ITESG y así poder generar estrategias para poder evitar el abandono escolar de los alumnos.

Con base en los resultados obtenidos, se propone que la institución fortalezca sus estrategias de intervención temprana mediante un sistema institucional de monitoreo continuo que utilice las ocho características identificadas como más relevantes para la detección de riesgo. La implementación formal del modelo basado en Fuzzy C-Means permitiría canalizar de manera oportuna a los estudiantes con riesgo medio y alto hacia programas de tutoría

académica, apoyo psicológico y acompañamiento socioeconómico. Asimismo, se recomienda integrar estos indicadores en los procesos de seguimiento docente y en la toma de decisiones del área de Tutorías, con el fin de diseñar planes personalizados de atención y reducir los índices de abandono escolar. La adopción de este enfoque sistematizado contribuiría a mejorar la retención estudiantil, optimizar los recursos institucionales y elevar la calidad educativa del ITESG.

4. Conclusiones y Trabajo a Futuro

Esta investigación tuvo como uno de sus objetivos el analizar los factores asociados a la deserción estudiantil en nivel superior en el campus ITESG, mediante la aplicación de algoritmos de clasificación no supervisada. Para ello, se recopilaron datos relevantes de los estudiantes que han desertado y actualmente se encuentran inscritos, además se consultó en el estado del arte cuáles eran los factores clave que influyen en la deserción escolar.

A través de este análisis de factores externos e internos involucrados en el desempeño escolar, se pudieron agregar más posibles indicadores que podrían ayudar a detectar a los estudiantes en riesgo de deserción escolar del ITESG. Para este análisis se dividieron las pruebas en dos fases, la primera usando 16 factores o características de los estudiantes y la segunda fase usando 8 características relevantes usando herramientas como Prueba de Chi-cuadrado.

A estos datos se les aplicó algoritmos de clasificación no supervisada como K-Means y Fuzzy C-Means, los cuales segmentaron a los estudiantes en 3 grupos de riesgo: bajo, medio y alto.

De acuerdo con las pruebas, se pudo evidenciar la superioridad del algoritmo Fuzzy C-Means en la detección precisa de estudiantes con distintos niveles de riesgo de deserción, en comparación con K-Means. Además, se corroboró que dividiendo en 3 clústeres se podían obtener grupos claramente segmentados tanto en el caso de 16 características como con 8 características relevantes, tal como lo denota el análisis del coeficiente de Silhouette.

Los hallazgos obtenidos coinciden con lo reportado en la literatura, especialmente en estudios donde las variables académicas son las de mayor peso en la predicción de deserción escolar. Tal como señalan [13], [15] y [17], factores como el número de asignaturas reprobadas, el avance académico y la acumulación de cursos pendientes son determinantes para identificar a los estudiantes en riesgo. Estos resultados se alinean con el presente estudio, donde las características con mayor impacto (asignaturas en repetición, asignaturas en especial, promedio actual y residencia actual) mostraron ser variables decisivas en la formación de clústeres. Asimismo, la relevancia de la asistencia y del nivel educativo de los padres coincide con las tendencias descritas en [3] y [16], reafirmando que tanto las condiciones personales como socioeducativas influyen de manera significativa en la permanencia escolar. En contraste, variables comúnmente consideradas en el estado del arte, como el género, la edad o el ingreso familiar, no mostraron impacto en el contexto del ITESG, lo que constituye una aportación relevante al evidenciar la necesidad de adaptar los modelos predictivos al entorno local. Este contraste con la literatura permite fortalecer la interpretación de los resultados y profundizar en la comprensión de los factores que influyen en la deserción estudiantil en el nivel superior.

Como trabajo a futuro, se plantea el ampliar el conjunto de datos de los estudiantes y la

incorporación de algoritmos de clasificación supervisada para poder integrar esto a un Sistema predictivo de deserción estudiantil y así poder implementar estrategias personalizadas que permitan a las instituciones educativas identificar y apoyar a los estudiantes en riesgo de manera proactiva, contribuyendo a mitigar la deserción estudiantil a nivel superior.

5. Referencias

[1] Subsecretaría de Educación Superior, "Criterios Generales para la Distribución de los Recursos Autorizados al Programa Presupuestario U079", Programa de Expansión de la Educación Media Superior y Superior, 2023.

[2] Secretaria de Educación Pública, "Programa Sectorial de Educación 2020-2024", 2020.

[3] A. Urbina Nájera, "Deserción escolar universitaria, Patrones para prevenirla aplicando minería de datos educativa", *e-Journal of Educational Research, Assessment and Evaluation,* 2020.

[4] H. Sadiq, "Educational data mining and analysis of students' academic performance using WEKA", *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 9, no. 2, pp. 447-459, 2018.

[5] F. Guiyun, "Analysis and prediction of students' academic performance based on educational data mining", *IEEE Access,* vol. 10, pp. 19558-19571, 2022.

[6] A. Ani, "Enhancing the clustering of student performance using the variation in confidence", *Intelligent Tutoring Systems: 14th International Conference,* pp. 274-279, 2018.

[7] R. González, "Massive LMS log data analysis for the early prediction of course-agnostic student performance", *Computers & Education,* vol. 163, 2021.

[8] F. I. Moreira, "Moodle Predicta: A Data Mining Tool for Student Follow Up", *CSEDU,* pp. 339-346, 2017.

[9] M. Saqr, "How learning analytics can early predict under-achieving students in a blended medical education course", *Medical Teacher,* pp. 757-767, 2017.

[10] D. Y. Hooshyar, "Clustering algorithms in an educational context: An automatic comparative approach", *IEEE Access,* vol. 8, pp. 146994-147014, 2018.

[11] Y. G. Li, "Educational data mining for students' performance based on fuzzy C-means clustering", *The Journal of Engineering,* pp. 8245-8250, 2019.

[12] M. Jessica, "DETECT: a hierarchical clustering

algorithm for behavioural trends in temporal educational data", *Artificial Intelligence in Education: 21st International Conference*, pp. 374-385, 2020.

[13] B.-A. Norka, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students", *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 11, 2020.

[14] B. Andreas, "Early Prediction of University Dropouts – A Random Forest Approach", *Jahrbücher für Nationalökonomie und Statistik*, vol. 240, no. 6, pp. 743-789, 2020.

[15] K. Lorenz, "Predicting student dropout: A machine learning approach", *European Journal of Higher Education*, 2020.

[16] V. Matti, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education", *Technology in Society*, vol. 76, 2024.

[17] S. Martín, "Perspectives to Predict Dropout in University Students with Machine Learning", *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018.

[18] H. Sadiq, "Classification, clustering and association rule mining in educational datasets using data mining tools: A case study", *Cybernetics and Algorithms in Intelligent Systems: Proceedings of 7th Computer Science On-line Conference 2018*, pp. 196-211, 2019.

[19] B. J.C., "Pattern Recognition with Fuzzy Objective Function", Springer, 1981.

[20] R. &. S. R. Suganya, "Fuzzy c-means algorithm-a review", *International Journal of Scientific and Research Publications*, vol. 2, no. 11, 2012.

[21] T. J., "Feature selection for classification: A review", *Data classification: Algorithms and applications*, 2014.