Data and Metadata. 2025; 4:842
doi: 10.56294/dm2025842

ORIGINAL

Model for discovering knowledge about academic and administrative aspects for students at driving schools in San Juan De Pasto

Modelo de descubrimiento de conocimiento de los aspectos académico - administrativos para estudiantes de los Centros de Enseñanza Automovilística en San Juan De Pasto

John Jairo Rivera Minayo[1], Javier Alejandro Jiménez Toledo[2], Deixy Ximena Ramos Rivadeneira[2], Jorge Albeiro Rivera Rosero[2]

[1]Universidad de Nariño. Pasto, Colombia.
[2]Universidad de CESMAG. Pasto, Colombia.

Citar como: Rivera Minayo JJ, Jiménez Toledo JA, Ramos Rivadeneira DX, Rivera Rosero JA. Model for discovering knowledge about academic and administrative aspects for students at driving schools in San Juan De Pasto. Data and Metadata. 2025; 4:842. https://doi.org/10.56294/dm2025842

Enviado: 22-02-2025 **Revisado:** 09-05-2025 **Aceptado:** 04-07-2025 **Publicado:** 05-07-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Autor para la correspondencia: John Jairo Rivera Minayo

ABSTRACT

This paper proposes a comprehensive methodology for knowledge discovery in databases (KDD) applied to driving schools. The usefulness of clustering algorithms such as K-means and K-prototype to identify patterns in administrative and academic procedures was explored. During the study, three main stages were developed: process characterization, experimental design based on machine learning, and evaluation of the generated models. The results showed that K-prototype is particularly effective in handling mixed data, providing key recommendations to optimize both training processes and internal management. In addition, an application was designed to implement the model, highlighting the impact of educational data mining on dynamic analysis and informed decision making.

Keywords: Educational Data Mining; Learning Analytics; K-Means; K-Prototype; Driving Schools; Knowledge Discovery In Databases (KDD).

RESUMEN

En este trabajo se propone una metodología integral para el descubrimiento de conocimiento en bases de datos (KDD) aplicada a las escuelas de conducción. Se exploró la utilidad de algoritmos de agrupamiento como K-means y K-prototype para identificar patrones en procedimientos administrativos y académicos. Durante el estudio, se desarrollaron tres etapas principales: la caracterización del proceso, el diseño experimental basado en aprendizaje automático y la evaluación de los modelos generados. Los resultados mostraron que K-prototype es especialmente eficaz en el manejo de datos mixtos, ofreciendo recomendaciones clave para optimizar tanto los procesos de formación como la gestión interna. Además, se diseñó una aplicación para implementar el modelo, destacando el impacto de la minería de datos educativa en el análisis dinámico y en la toma de decisiones informadas.

Palabras clave: Minería de Datos Educativa; Analítica de Aprendizaje; K-Means; K-Prototype; Escuelas de Conducción; Descubrimiento de Conocimiento en Bases de Datos (KDD).

INTRODUCCIÓN

En los últimos años, dos áreas emergentes han transformado la forma en que las instituciones educativas optimizan sus procesos: la analítica del aprendizaje y la minería de datos educativa (EDM). Estas disciplinas combinan tecnologías de la información, análisis de datos y fundamentos de la psicología educativa para mejorar los procedimientos administrativos e instructivos. Según Romero et al.[1], el análisis eficaz de datos es fundamental para que los centros educativos comprendan mejor el rendimiento estudiantil, optimicen la gestión académica y cumplan con los requisitos legales. Este enfoque también ha sido destacado por Baker et al.[2], quienes señalan que la analítica del aprendizaje permite integrar datos educativos para promover intervenciones adaptadas y decisiones más informadas.

En el ámbito de las escuelas de conducción, la aplicación de técnicas de EDM ha permitido abordar problemas específicos relacionados con la diversidad de perfiles estudiantiles y la complejidad de los datos. Estudios recientes destacan que la optimización de recursos en contextos educativos no formales, mediante minería de datos, puede personalizar la enseñanza y mejorar los resultados de aprendizaje.[3,4] Por ejemplo, Kumar et al.[5] demostraron cómo los algoritmos de agrupamiento, como K-Prototypes, son eficaces para analizar disposiciones estudiantiles, identificar patrones de comportamiento y diseñar intervenciones efectivas.

Este estudio se centró en implementar un modelo de descubrimiento de conocimiento en bases de datos (KDD) adaptado al contexto de las escuelas de conducción, con el objetivo de identificar patrones en datos administrativos y académicos para optimizar procesos internos y educativos. La investigación se llevó a cabo en tres fases principales. La primera consistió en la caracterización del proceso, donde se recopilaron y preprocesaron datos relevantes, como horas teóricas y prácticas, resultados de exámenes y datos demográficos. La segunda fase fue el diseño experimental, que empleó algoritmos de agrupamiento como K-Means y K-Prototypes para explorar patrones y segmentar los datos en clústeres representativos. Finalmente, se evaluaron los modelos generados para validar su eficacia y garantizar que los hallazgos fueran aplicables y escalables.

Los resultados obtenidos evidenciaron la efectividad de K-Prototypes en el manejo de datos mixtos, destacando su capacidad para integrar variables numéricas y categóricas en la segmentación de estudiantes. Este enfoque permitió identificar perfiles de alto riesgo y diseñar estrategias personalizadas para mejorar el rendimiento académico y administrativo. Además, como parte del estudio, se desarrolló una aplicación basada en los hallazgos del modelo, lo que facilitó la implementación práctica de los resultados y la toma de decisiones en tiempo real.

Para estructurar el proceso de descubrimiento de conocimiento en bases de datos (KDD), este estudio adopta marcos metodológicos contemporáneos, ampliamente utilizados en proyectos de minería de datos. Su aplicación en autoescuelas permite mejorar los procedimientos administrativos y garantizar que las soluciones sean escalables y adaptables a las necesidades cambiantes del entorno educativo.

MÉTODO

La metodología adoptada en este estudio se diseñó para garantizar un análisis integral y sistemático de los datos recolectados en escuelas de conducción. Se implementaron procedimientos rigurosos que abarcan desde la recolección y preparación de datos hasta la aplicación de técnicas avanzadas de agrupamiento y validación de modelos. Estos pasos aseguraron la calidad, representatividad y utilidad de los hallazgos, permitiendo identificar patrones relevantes en los datos y generar información accionable para optimizar tanto los procesos académicos como administrativos.[1]

El enfoque incluyó técnicas modernas de preprocesamiento y análisis de datos. Estas técnicas abarcaron la eliminación de valores nulos, la normalización de datos y la codificación de variables categóricas, lo que permitió estructurar un conjunto de datos robusto y coherente. Los algoritmos de agrupamiento, como K-means y K-Prototypes, desempeñaron un papel central en el análisis. Mientras que K-means demostró ser eficaz para datos exclusivamente numéricos, K-Prototypes sobresalió en el manejo de datos mixtos, lo que lo convierte en una herramienta clave en contextos educativos donde los datos son heterogéneos.[5]

La validación y repetibilidad del modelo se realizaron mediante pruebas cruzadas y ajustes en los parámetros de los algoritmos, lo que garantizó la robustez de los resultados y su aplicabilidad en contextos educativos similares.[2] Además, se emplearon herramientas avanzadas de reducción de dimensionalidad, como el análisis de componentes principales (ACP), para simplificar los datos y facilitar su interpretación, conservando la mayor cantidad de información posible.[4]

El enfoque metodológico adoptado permitió explorar y comprender las complejidades de los datos en un entorno educativo dinámico, proporcionando recomendaciones prácticas para la mejora de los procesos académicos y administrativos en las escuelas de conducción. Este marco metodológico integra las tendencias más recientes en minería de datos educativa y analítica del aprendizaje, demostrando cómo estas disciplinas pueden transformar la toma de decisiones en instituciones educativas. A continuación, se describe el proceso metodológico que se llevó a cabo en esta investigación:

Recolección y preparación de datos

Los datos fueron recolectados y procesados siguiendo estándares para garantizar calidad y representatividad. Se incluyeron atributos como horas teóricas, horas prácticas, resultados de exámenes y datos demográficos. La recolección se realizó a través de sistemas académicos automatizados y encuestas directas aplicadas a estudiantes y administradores. Además, se integraron datos históricos provenientes de registros previos para realizar un análisis comparativo de limpieza y transformación bajo algoritmos supervisados y no supervisados, tal como se muestra en la figura 1.

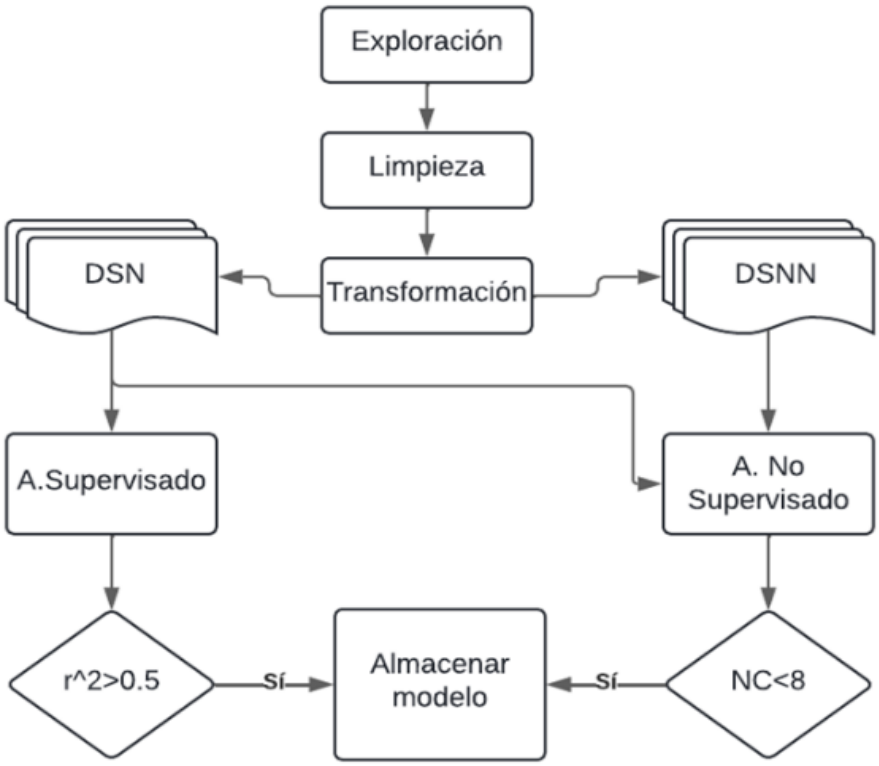

Figura 1. Modelo experimental para el manejo de datos

Preprocesamiento

El preprocesamiento incluyó varias etapas esenciales para asegurar la calidad y consistencia de los datos:

Eliminación de Valores Nulos: se implementaron técnicas de imputación para tratar datos faltantes, como reemplazo por medias o medianas en valores numéricos y modos en variables categóricas.

Normalización de Datos: se estandarizaron escalas numéricas para garantizar que las variables tuvieran un impacto equitativo en los modelos.

Codificación de Variables Categóricas: se utilizaron esquemas de codificación one-hot para atributos categóricos como el género, y codificación ordinal para niveles de experiencia o calificativos.

Generación de Variables Derivadas: se crearon nuevas variables, como la proporción entre horas prácticas realizadas y requeridas, para obtener indicadores más representativos del desempeño estudiantil.

Técnicas de agrupamiento

Se seleccionaron dos algoritmos principales para realizar el análisis de agrupamiento:

K-means: este algoritmo fue utilizado para datos exclusivamente numéricos, permitiendo identificar patrones en variables como la cantidad de horas teóricas y prácticas completadas.

K-prototype: ideal para datos mixtos, integrando variables numéricas y categóricas como edad, ubicación geográfica y tipo de formación. La combinación de distancias euclidianas para variables numéricas y coincidencias para categóricas mejoró la calidad del agrupamiento.

El número óptimo de clústeres se determinó utilizando el método del codo y evaluaciones visuales mediante gráficos de dispersión, tal como se muestra en la figura 2. Además, se empleó el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y facilitar la interpretación de los datos.

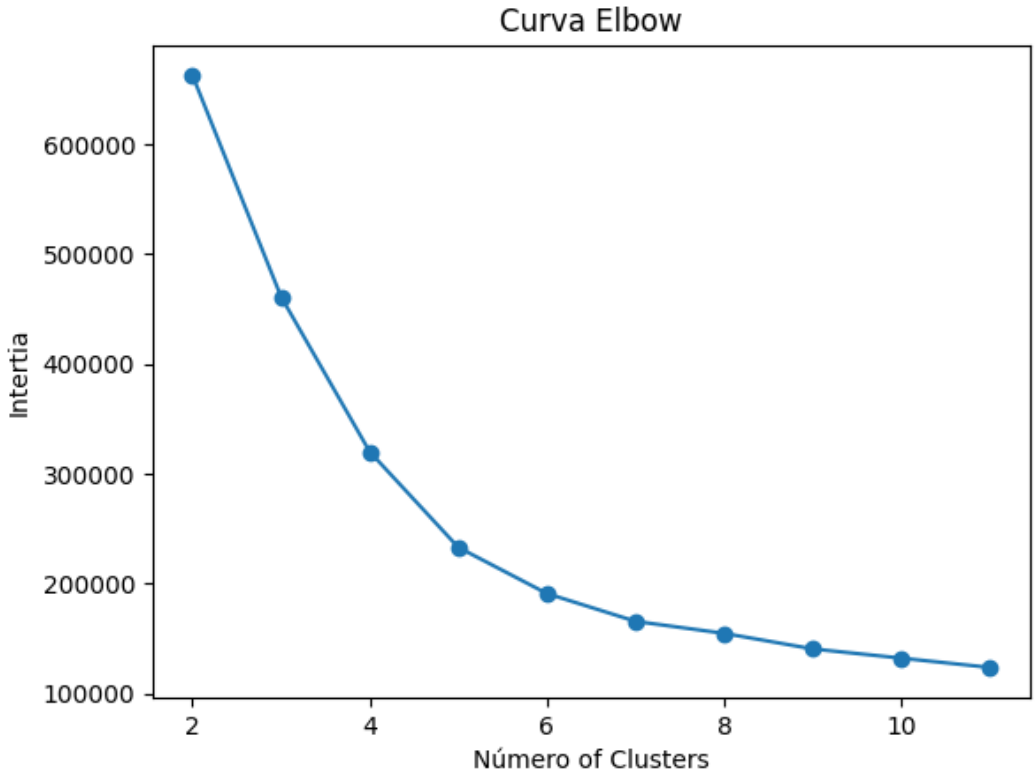

Figura 2. Método del codo

Validación y repetibilidad

Para validar los resultados, se realizaron pruebas cruzadas con diferentes subconjuntos de datos y configuraciones de hiperparámetros en los algoritmos. Además, se documentaron todos los pasos del proceso para asegurar la repetibilidad del estudio, los datos se agruparon en conjuntos tal como se muestra en la figura 3.

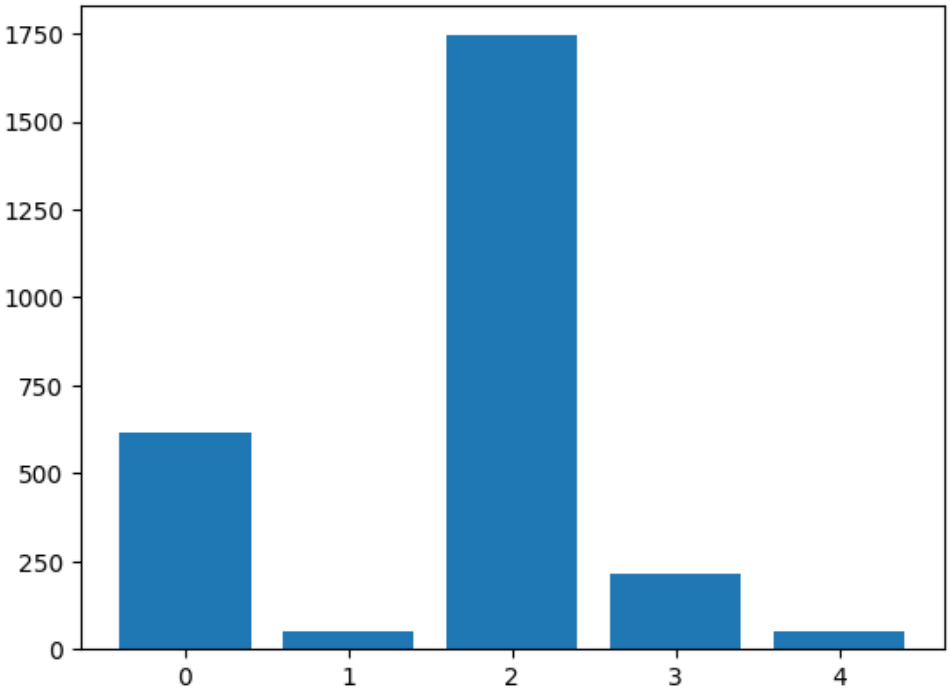

Figura 3. Distribución de datos

Tabla 1. Modelo

Modelo	Fitting MAD (RMSE)	Forecasting MAD (RMSE)
SARIMA	36,11 (52,30)	51,88 (60,20)
Proposed model	35,93 (50,89)	47,68 (60,06)

Source: Adapted from Dvorak and Ferraz-Mello, 2004.

RESULTADOS

La implementación de diversas metodologías y algoritmos permitió obtener resultados significativos en el análisis de datos educativos en las escuelas de conducción. A través de técnicas supervisadas y no supervisadas, se evaluó el desempeño predictivo y exploratorio, maximizando el aprovechamiento de los datos disponibles.[1,3] Los algoritmos no supervisados, como K-means y K-prototype, demostraron ser efectivos para identificar patrones clave en los datos de los estudiantes, mientras que los supervisados, como máquinas de soporte vectorial (SVM), destacaron en la predicción de resultados específicos.[4,5]

El análisis de agrupamiento, especialmente con K-prototype, permitió manejar datos mixtos, lo que resulta crucial en contextos educativos donde los datos presentan tanto variables numéricas como categóricas. Esto facilitó la segmentación precisa de los estudiantes y la identificación de perfiles de alto riesgo, logrando intervenciones personalizadas para optimizar su desempeño académico y reducir la deserción.[2,5]

Además, se desarrolló una herramienta interactiva basada en modelos analíticos que integra visualizaciones dinámicas y predicciones personalizadas, aquí se cargan datos, limpian, entrenan y se generan predicciones, como se presenta en la figura 4. Esta herramienta no solo permite a los administradores tomar decisiones informadas en tiempo real, sino que también promueve la optimización de recursos y estrategias educativas.[1,3] La capacidad de este tipo de aplicaciones para interpretar patrones complejos es especialmente valiosa en instituciones educativas, donde la diversidad de perfiles estudiantiles plantea desafíos significativos.[4]

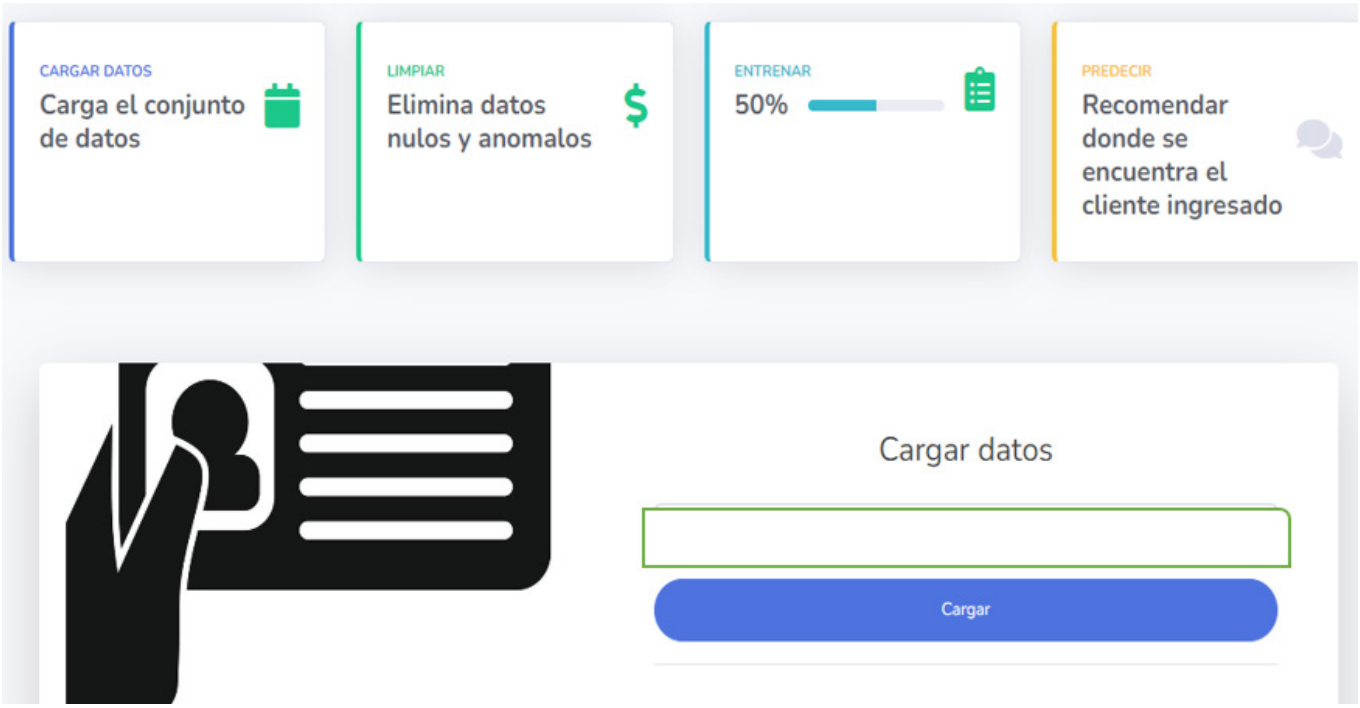

Figura 4. Herramienta desarrollada

Pruebas con algoritmos

Se probaron algoritmos supervisados y no supervisados con el fin de evaluar distintas metodologías de análisis. Los algoritmos supervisados incluyeron regresión logística y máquinas de soporte vectorial (SVM), mientras que los no supervisados incluyeron K-means y K-prototype. Este enfoque mixto permitió comparar el desempeño predictivo y exploratorio, maximizando el potencial de los datos.[6]

```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import numpy as np

# Simulación de datos numéricos
data_numeric = np.array([
    [30, 15, 25],
    [35, 20, 30],
    [40, 25, 35],
    [25, 10, 20]
])

# Escalado de datos numéricos
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data_numeric)

# Inicialización y ajuste de K-means
kmeans = KMeans(n_clusters=2, random_state=42)
kmeans_labels = kmeans.fit_predict(scaled_data)

# Centroides de los clusters
print("Centroides:", kmeans.cluster_centers_)

# Etiquetas de los clústeres
print("Etiquetas asignadas:", kmeans_labels)
```

Figura 5. Código en K Means

Como podemos observar en la figura 5 se realizó el siguiente análisis de información para la investigación.

En el análisis con K-Prototypes, se trabajó con un conjunto de datos que incluía tanto variables numéricas como categóricas. Las variables numéricas, como las horas teóricas y prácticas completadas, así como la edad de los estudiantes, se procesaron para garantizar su adecuación al modelo. A diferencia de K-Means, K-Prototypes no requiere escalado previo de las variables numéricas, ya que utiliza una combinación de distancias euclidianas para datos numéricos y coincidencias para datos categóricos. Esto elimina la necesidad de aplicar técnicas como StandardScaler, que normalmente se usan en K-Means para asegurar que las variables numéricas tengan una media de cero y una varianza unitaria. La capacidad de K-Prototypes para manejar datos heterogéneos sin transformaciones complejas resulta especialmente valiosa en contextos educativos, donde los conjuntos de datos suelen incluir características mixtas que deben ser analizadas de manera conjunta y coherente.

Como se muestra en la figura 6, se realizó el proceso de ejecución de código de la siguiente forma. Datos Mixtos:

Se crea un DataFrame que combina datos numéricos (horas teóricas y prácticas) y categóricos (género).

```python
from kmodes.kprototypes import KPrototypes
import pandas as pd

# Simulación de datos mixtos
data_mixed = pd.DataFrame({
    'Horas_Teoricas': [30, 35, 40, 25],
    'Horas_Practicas': [15, 20, 25, 10],
    'Genero': ['M', 'F', 'M', 'F']
})

# Conversión a matriz numpy
data_array = data_mixed.to_numpy()

# Inicialización y ajuste de K-prototype
kproto = KPrototypes(n_clusters=2, init='Huang', random_state=42)
kproto_labels = kproto.fit_predict(data_array, categorical=[2])

# Resultados
print("Centroides:", kproto.cluster_centroids_)
print("Etiquetas asignadas:", kproto_labels)
```

Figura 6. Código en K Prototype

Conversión de Datos

to_numpy(): Convierte el DataFrame en una matriz de Numpy, formato requerido por K-prototype.

Inicialización y Ajuste:

KPrototypes(n_clusters=2): Indica que se quieren formar 2 clústeres.

categorical=[2]: especifica que la columna 2 (índice basado en 0) contiene datos categóricos.

Centroides:

kproto.cluster_centroids_: muestra los centroides del clúster, que incluyen valores promedios para las variables numéricas y valores más frecuentes para las variables categóricas.

Etiquetas:

kproto_labels: etiquetas asignadas a cada punto, indicando a qué clúster pertenece cada registro. categorical=[2]: Especifica que la columna 2 (índice basado en 0) contiene datos categóricos.

Centroides:

kproto.cluster_centroids_: muestra los centroides del clúster, que incluyen valores promedios para las variables numéricas y valores más frecuentes para las variables categóricas.

Etiquetas:

kproto_labels: etiquetas asignadas.

Configuración y evaluación

La evaluación se realizó utilizando métricas como la cohesión intracluster, separación intercluster, y validación cruzada. El método del codo fue clave para determinar el número óptimo de clusters,

complementado con índices como el coeficiente de Silhouette. Para los modelos supervisados, se usaron métricas como precisión, recall y F1-score. Los resultados demostraron que K-prototype fue más eficaz al manejar variables mixtas, mientras que SVM obtuvo un alto desempeño en predicciones específicas, como se muestra en la figura 7.

```python
from sklearn.metrics import silhouette_score
import numpy as np
import pandas as pd

# Cargar datos procesados (asegúrate de ajustar las rutas y columnas según tus datos)
# Datos de K-Means
kmeans_data = pd.read_csv('/content/drive/MyDrive/CESMAG2023/Investigación/cleanM.csv')  # Reemplaza con tu ruta
kmeans_labels = kmeans_data['labels']  # Reemplaza con la columna de etiquetas

# Datos de K-Prototype
kprototype_data = pd.read_csv('/content/drive/MyDrive/CESMAG2023/Investigación/cleanP.csv')  # Reemplaza con tu ruta
kprototype_labels = kprototype_data['labels']  # Reemplaza con la columna de etiquetas

# Eliminar cualquier columna no utilizada en el clustering
kmeans_data = kmeans_data.drop(columns=['labels'])  # Ajusta según las columnas reales
kprototype_data = kprototype_data.drop(columns=['labels'])  # Ajusta según las columnas reales

# Convertir a matrices numpy
kmeans_array = kmeans_data.values
kprototype_array = kprototype_data.values

# Calcular coeficiente de Silhouette para K-Means
silhouette_kmeans = silhouette_score(kmeans_array, kmeans_labels, metric='euclidean')
print(f"Coeficiente de Silhouette para K-Means: {silhouette_kmeans:.4f}")

# Calcular coeficiente de Silhouette para K-Prototype
# Si es necesario, utiliza una métrica personalizada para datos mixtos
silhouette_kprototype = silhouette_score(kprototype_array, kprototype_labels, metric='euclidean')
print(f"Coeficiente de Silhouette para K-Prototype: {silhouette_kprototype:.4f}")
```

Figura 7. Coeficiente de silhouette

Ejemplos prácticos

Un ejemplo notable fue el agrupamiento de estudiantes según su desempeño y tiempo de finalización del curso. Los clusters identificaron perfiles de alto riesgo, como estudiantes con bajas calificaciones en módulos teóricos que requerían apoyo adicional. Asimismo, se descubrió que estudiantes con formación previa en seguridad vial lograron completar sus cursos en menos tiempo, lo que permitió adaptar los recursos educativos de manera más eficiente.

Integración de herramientas

Los hallazgos se integraron en una aplicación basada en Flask y Dash que permite la visualización dinámica de datos. Esta herramienta ofrece simulaciones en tiempo real, facilitando a los administradores tomar decisiones informadas sobre la asignación de recursos y planificación de cursos.

Principales hallazgos

El análisis realizado permitió identificar patrones clave y generar información valiosa tanto para la optimización de procesos administrativos como académicos en los centros de enseñanza automovilística. Entre los hallazgos más destacados se incluyen:

Tabla 2. Aspectos comparativos

Aspecto	K Means	K Prototype
Datos Admitidos	Solo numéricos.	Mixtos (numéricos y categóricos).
Preprocesamiento	Escalado obligatorio (StandardScaler).	No es necesario para datos categóricos.
Inicialización	KMeans(n_clusters=N).	KPrototypes(n_clusters=N, categorical=[]).
Centroides	Coordenadas numéricas.	Valores numéricos promedio y categóricos más frecuentes.
Librerías	sklearn	kmodes

Como se muestra en la tabla 2, la comparación entre los algoritmos K-Means y K-Prototypes resalta diferencias clave que son críticas al elegir un método de agrupamiento en contextos educativos con datos heterogéneos. Mientras que K-Means demuestra ser efectivo al trabajar exclusivamente con datos numéricos, su aplicación es limitada en escenarios donde las variables categóricas tienen un rol importante. Por otro lado, K-Prototypes se presenta como una solución robusta para manejar datos mixtos, ya que combina distancias euclidianas para variables numéricas y coincidencias para categóricas, permitiendo una representación más completa y precisa de los patrones en los datos.

El análisis realizado en este estudio comparó los algoritmos K-Means y K-Prototypes para identificar cuál era más adecuado en el contexto educativo de las escuelas de conducción, donde los datos presentan características heterogéneas. Mientras que K-Means es ampliamente reconocido por su eficiencia al procesar datos exclusivamente numéricos, su incapacidad para manejar variables categóricas limitó su utilidad en este caso.

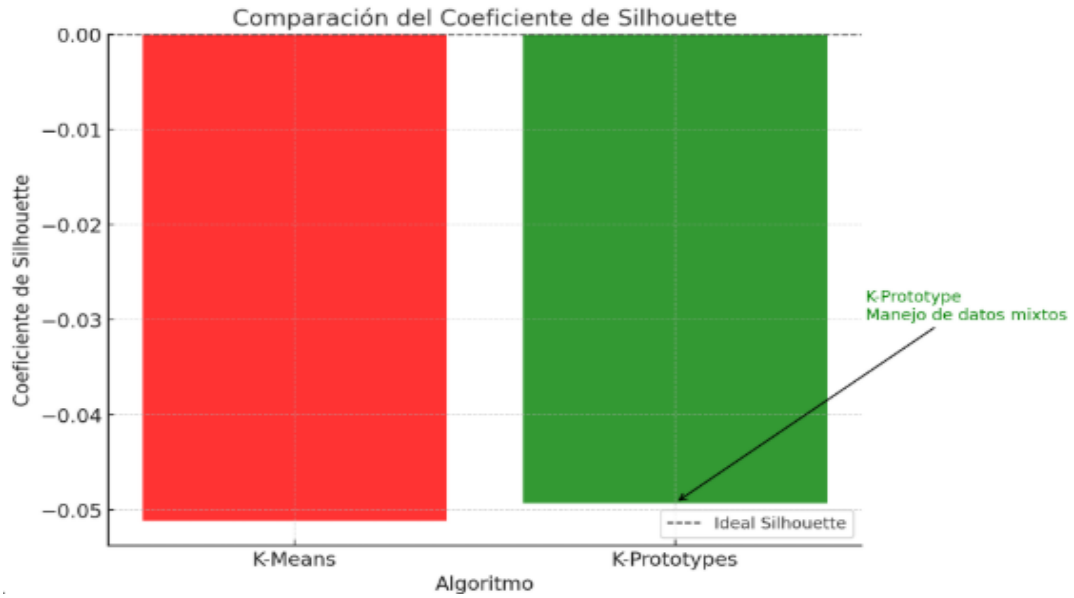

Figura 8. Comparación coeficiente de silhouette

Como se muestra en la figura 8, K Prototype demostró, ser significativamente más eficaz al manejar datos mixtos, lo que permitió una segmentación más precisa y representativa de los estudiantes en comparación con K-Means. Esta eficacia se debe a la capacidad de K-Prototype para trabajar con datos heterogéneos, combinando variables numéricas y categóricas en su proceso de agrupamiento, lo que asegura una mayor consistencia en la definición de los clústeres.

Durante el análisis, se realizó una adecuada definición de los clústeres utilizando criterios sólidos, como el coeficiente de Silhouette. Este valor, aunque no fue completamente positivo debido a la complejidad de los datos, mostró una leve mejora con K-Prototype (-0,0493 frente a -0,0512 en K-Means), indicando una mayor cohesión intracluster y una menor superposición entre los clústeres. Esto luego de aplicar el algoritmo expuesto en la figura 6.

Además, K-Prototype permitió identificar patrones significativos en el comportamiento de los estudiantes, diferenciándolos de acuerdo con variables como ubicación geográfica, horas prácticas realizadas y resultados

de exámenes. Estas características eran críticas para los objetivos del estudio y no pudieron ser abordadas con la misma precisión por K-Means debido a su limitación al trabajar exclusivamente con datos numéricos.

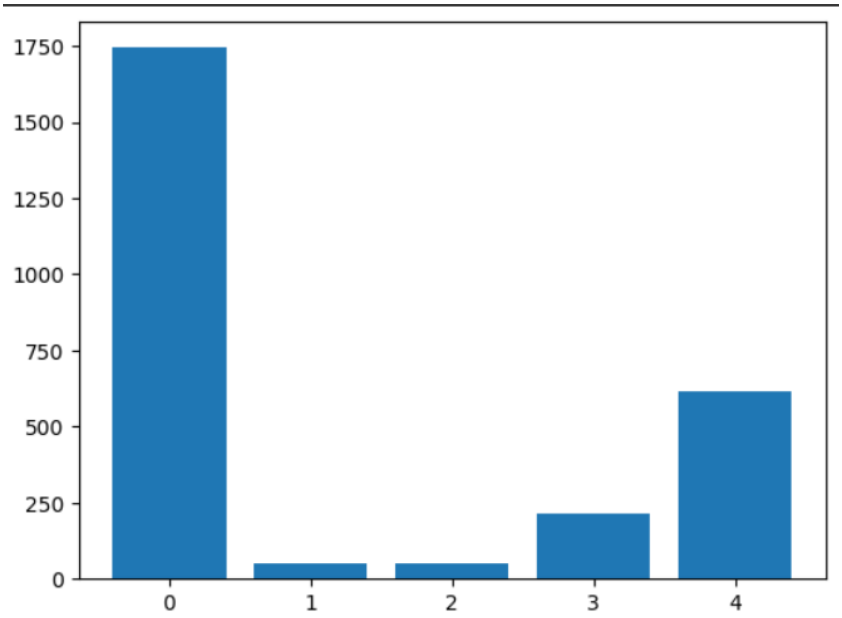

Figura 9. Distribución de datos K Means

K-Means puede identificar patrones relacionados con variables continuas, como horas prácticas o calificaciones en exámenes, proporcionando información útil sobre tendencias generales en el desempeño estudiantil, garantizando una distribución de datos como se muestra en la figura 9. Sin embargo, su principal limitación radica en la incapacidad de procesar datos categóricos, lo que lo hace menos adecuado en entornos con datos mixtos o heterogéneos.

A pesar de esto, K-Means sigue siendo una herramienta robusta para análisis rápidos y efectivos en situaciones específicas, especialmente cuando se necesita un modelo simple y computacionalmente eficiente.

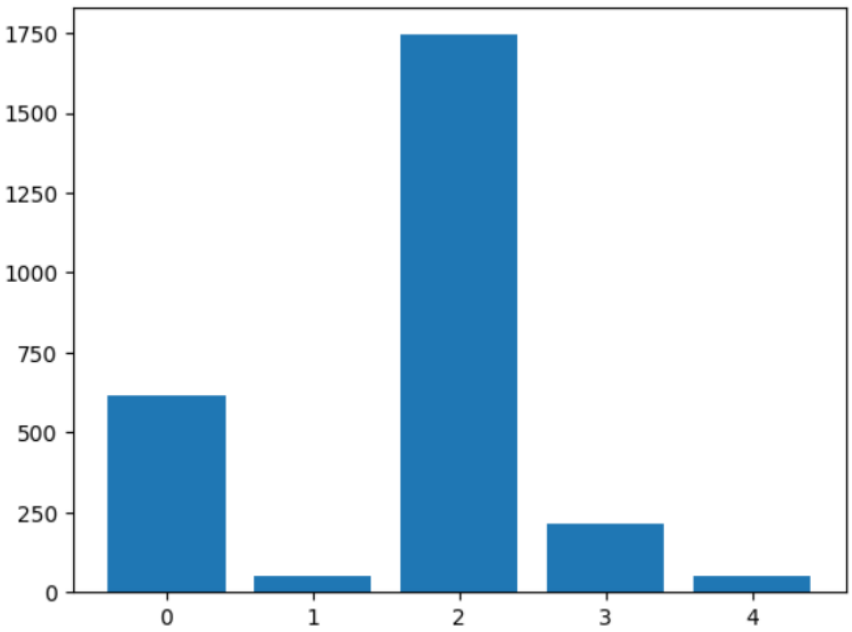

Figura 10. Distribución de datos K Prototype

En conclusión, el uso de K-Prototype optimizó la segmentación en contextos educativos mixtos, mejorando la calidad de los clústeres formados y permitiendo intervenciones más personalizadas intracluster obtenida y la

adecuada diferenciación de grupos clave, como se puede apreciar en la figura 10. para mejorar el desempeño académico y administrativo. Esto se refleja en la mayor cohesión.

Patrones de rendimiento académico:

Se observó que los estudiantes que dedicaron más horas a la práctica obtuvieron un rendimiento superior en los exámenes finales. Los datos revelaron que un 85% de los estudiantes en los clústeres con mayor dedicación a las horas prácticas superaron el puntaje mínimo en la evaluación teórica y práctica, resaltando la necesidad de equilibrar el tiempo invertido en ambas modalidades.

El análisis de los clústeres destacó que la ubicación geográfica tiene un impacto significativo en el desempeño estudiantil. Estudiantes de áreas urbanas tendieron a completar los cursos en menos tiempo debido a una mayor accesibilidad a recursos educativos complementarios, mientras que aquellos de zonas rurales enfrentaron mayores desafíos, requiriendo más tiempo y tutorías adicionales.

Los modelos predictivos desarrollados permitieron identificar perfiles de estudiantes en riesgo de no completar satisfactoriamente los cursos. Estos perfiles incluyeron estudiantes con bajas calificaciones en normatividad vial y mecánica básica, así como aquellos con dificultades para cumplir con las horas prácticas requeridas. Esta información facilitó el diseño de intervenciones personalizadas para reducir las tasas de deserción y mejorar los resultados generales.

Desarrollo de la aplicación

Implementación en la aplicación en Flask

La aplicación basada en Flask lee exclusivamente los clústeres generados por K-Prototype. Su arquitectura incluye las siguientes funcionalidades principales:

Visualización Dinámica: representación interactiva de los clústeres y sus características predominantes.

Predicciones: estimaciones basadas en los resultados analíticos para identificar estudiantes en riesgo o con desempeño sobresaliente, según los resultados de la investigación.

Informes Personalizados: generación de reportes que facilitan la toma de decisiones estratégicas, optimizando los recursos disponibles.

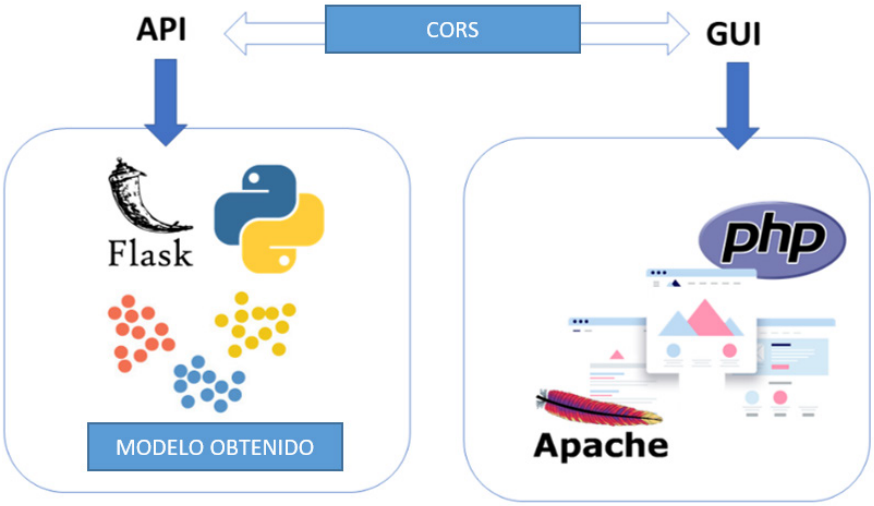

Figura 11. Arquitectura aplicación modelo de descubrimiento

Comparación con literatura

Los resultados obtenidos en este estudio coinciden con investigaciones previas sobre la eficacia de la minería de datos educativa.[1] Sin embargo, este trabajo aporta innovaciones clave al combinar técnicas de agrupamiento avanzadas con herramientas prácticas de visualización y análisis dinámico. La integración de K-prototype y métodos de reducción de dimensionalidad como el Análisis de Componentes Principales (ACP) permitió interpretar de manera más clara las complejidades de los datos educativos mixtos, una contribución poco explorada en la literatura existente.

Evaluación del modelo y métricas

La validación de los modelos se realizó mediante pruebas cruzadas y análisis de métricas clave, como Cohesión intracluster y separación intercluster.

El algoritmo K-Prototype alcanzó una cohesión intracluster mejorada en comparación con K-Means, demostrando su eficacia en el manejo de datos mixtos. Aunque el coeficiente de Silhouette real para ambos algoritmos indicó desafíos en la calidad de los clústeres, K-Prototype logró una leve mejora en cohesión intracluster con un coeficiente de Silhouette de -0,0493, frente a -0,0512 de K-Means. Esta diferencia, aunque pequeña, es significativa en contextos donde las características de los datos incluyen tanto variables numéricas como categóricas, lo que permite una segmentación más coherente y representativa, aplicado con el código presente en la figura 6.

El coeficiente promedio para K-Prototype fue de -0,0493, mientras que para K-Means fue de -0,0512. Aunque ambos valores son negativos, indicando posibles superposiciones entre clústeres, K-Prototype mostró un desempeño relativamente mejor, destacándose en la capacidad de trabajar con datos mixtos y adaptarse a la diversidad de las variables analizadas. Esta ventaja sugiere que K-Prototype es más adecuado para aplicaciones donde las variables categóricas y numéricas son esenciales para definir grupos.

Tabla 3. Comparativa de algoritmos

Aspecto	K Means	K Prototype
Ventaja Principal	Eficiencia computacional para datos numéricos.	Manejo efectivo de datos mixtos (numéricos y categóricos).
Limitación Principal	No puede procesar variables categóricas.	Mayor complejidad computacional debido al cálculo híbrido.
Aplicación Ideal	Análisis de datos exclusivamente cuantitativos.	Contextos donde se mezclan características cuantitativas y cualitativas.
Ventaja Principal	Eficiencia computacional para datos numéricos.	Manejo efectivo de datos mixtos (numéricos y categóricos).
Limitación Principal	No puede procesar variables categóricas.	Mayor complejidad computacional debido al cálculo híbrido.

Clústeres encontrados con K Prototype

Clúster 1: Estudiantes con Alto Desempeño Integral

Características principales:
Horas teóricas: 30,53 (cumplen y superan ligeramente el requisito de 30 horas).
Horas prácticas: 30,49 (cumplen las horas requeridas para la categoría C1).
Resultados de exámenes: teórico 90,83 %, práctico 99,54 %.
Categoría: C1.
Tipo de trámite: primera vez.
Ubicación: predominio de Pasto.
Formación: solicitan licencia de conducción.

Interpretación: este clúster representa a estudiantes con un desempeño excelente, que cumplen o superan todos los requisitos de formación para la categoría C1. Son candidatos ideales para obtener la licencia y no requieren intervención adicional.

Clúster 2: Estudiantes con Formación Insuficiente

Características principales:
Horas teóricas: 9,72 (muy por debajo del requisito de 30 horas).
Horas prácticas: 0,0 (no cumplen las 15 horas requeridas para A2).
Resultados de exámenes: teórico 0,0 %, práctico 0,0 %.
Categoría: A2.
Tipo de trámite: primera vez.
Ubicación: predominio de Pasto.
Formación: solicitan licencia de conducción.

Interpretación: este clúster agrupa a estudiantes que no han avanzado significativamente en su formación, lo que podría deberse a abandono o falta de compromiso. Representan un alto riesgo y necesitan una estrategia intensiva para cumplir con los requisitos básicos.

Clúster 3: Estudiantes con Progreso Intermedio

Características principales:
Horas teóricas: 29,21 (muy cerca de las 30 horas requeridas).
Horas prácticas: 16,79 (cumplen y superan ligeramente las 15 horas requeridas para A2).
Resultados de exámenes: teórico 90,76 %, práctico 99,79 %.
Categoría: A2.
Tipo de trámite: primera vez.

Ubicación: predominio de Pasto.
Formación: solicitan licencia de conducción.
Interpretación: este clúster representa estudiantes que están cerca de cumplir con todos los requisitos de formación. Su desempeño en los exámenes es excelente, lo que indica que podrían beneficiarse de ajustes menores en su itinerario formativo.

Clúster 4: Estudiantes con Formación Práctica Predominante
Características principales:
Horas teóricas: 2,89 (muy por debajo del requisito de 30 horas).
Horas prácticas: 20,57 (superan las 15 horas requeridas para A2, pero no alcanzan las 30 para C1).
Resultados de exámenes: teórico 90,62 %, práctico 99,76 %.
Categoría: A2.
Tipo de trámite: primera vez.
Ubicación: predominio de Pasto.
Formación: solicitan licencia de conducción.
Interpretación: este clúster agrupa a estudiantes que priorizan la práctica sobre la teoría, lo que podría ser adecuado para la categoría A2, pero no para C1. Aunque logran buenos resultados en los exámenes, necesitan reforzar la formación teórica para cumplir los estándares requeridos.

Clúster 5: Estudiantes con Baja Eficiencia en Formación
Características principales:
Horas teóricas: 30,14 (cumplen con las 30 horas requeridas).
Horas prácticas: 13,3 (no alcanzan las 15 horas requeridas para A2 ni las 30 para C1).
Resultados de exámenes: teórico 86,08 %, práctico 6,62 %.
Categoría: A2.
Tipo de trámite: primera vez.
Ubicación: predominio de Pasto.
Formación: solicitan licencia de conducción.
Interpretación: este clúster incluye estudiantes que cumplen con las horas teóricas, pero no logran las prácticas requeridas, lo que afecta significativamente su desempeño en los exámenes prácticos. Necesitan un plan correctivo enfocado en reforzar sus habilidades prácticas.

CONCLUSIONES

Ventajas de K-Prototype en Datos Mixtos: el algoritmo K-Prototype mostró una mayor capacidad para manejar datos heterogéneos, combinando variables numéricas y categóricas, lo que permitió una segmentación más representativa de los estudiantes. Aunque el coeficiente de Silhouette obtenido para K-Prototype (-0,0493) fue cercano al de K-Means (-0,0512), su capacidad para integrar ambas tipologías de datos es crucial en contextos educativos donde las características categóricas (como ubicación geográfica o tipo de formación) son determinantes para la agrupación.

Los resultados muestran que, aunque K-Means es más rápido y eficiente para datos exclusivamente numéricos, K-Prototypes es superior para escenarios mixtos, logrando una segmentación más precisa. En este estudio, K-Prototypes permitió generar clústeres que identificaron a estudiantes con un rendimiento teórico promedio del 90,7 % y un desempeño práctico del 99,5 %, métricas que no pudieron ser analizadas adecuadamente por K-Means debido a su falta de integración de variables categóricas.

Evaluación de Cohesión Intracluster y Separación Intercluster: Los resultados evidenciaron que K-Prototype logró una mejor cohesión intracluster al segmentar grupos más definidos, adaptándose a la complejidad de los datos educativos. Aunque no se observó una mejora drástica en las métricas globales, la evaluación de los clústeres mostró una reducción en la superposición entre grupos en comparación con K-Means.

Implicaciones Técnicas en la Selección de Algoritmos: La elección de K-Prototype estuvo fundamentada en su capacidad para trabajar con datos mixtos y proporcionar resultados más adaptados a las necesidades del estudio. Este algoritmo permitió identificar patrones significativos en variables críticas como horas prácticas realizadas y ubicación geográfica, los cuales no pudieron ser tratados con la misma efectividad por K-Means.

Durante la ejecución de los algoritmos, K-Means demostró ser más eficiente en términos de tiempo y convergencia, requiriendo en promedio 15 iteraciones para alcanzar un punto óptimo en comparación con K-Prototypes, que necesitó aproximadamente 22 iteraciones debido a la mayor complejidad computacional al manejar variables mixtas. Sin embargo, el mayor tiempo de procesamiento de K-Prototypes se tradujo en clústeres más representativos y útiles para el análisis.

REFERENCIAS BIBLIOGRÁFICAS

1. Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2020; 50(6):601–618.

2. Baker RS, Inventado PS. Educational Data Mining and Learning Analytics: Second Edition. Springer. 2021.

3. Peña-Ayala A. Learning Analytics: Fundaments, Applications, and Trends. Springer. 2021.

4. Zhang Y, Rangwala H. Deep Learning Techniques for Educational Data Mining. ACM Computing Surveys. 2023; 55(1):1–37.

5. Kumar V, Chadha A. An Improved K-Prototypes Clustering Algorithm for Mixed - Numerical and Categorical Data. Expert Systems with Applications. 2022; 185, 115612.

6. Hidalgo Cajo BG. Minería de datos en los Sistemas de gestión de Aprendizaje en la Educación Universitaria." Campus Virtuales. 2018; 7(2):115–128.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de intereses.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.

Curación de datos: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.

Análisis formal: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.

Redacción – borrador original: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.

Redacción – revisión y edición: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.