



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 Minayo¹ , Javier Alejandro Jiménez Toledo² , Deixy Ximena Ramos Rivadeneira² , Jorge Albeiro Rivera Rosero² 

¹Universidad de Nariño. Pasto, Colombia.

²Universidad de CESMAG. Pasto, Colombia.

Cite as: 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>

Submitted: 22-02-2025

Revised: 09-05-2025

Accepted: 04-07-2025

Published: 05-07-2025

Editor: Dr. Adrián Alejandro Vitón Castillo 

Corresponding Author: 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).

INTRODUCTION

In recent years, two emerging areas have transformed the way educational institutions optimize their processes: learning analytics and educational data mining (EDM). These disciplines combine information technologies, data analysis, and foundations of educational psychology to improve administrative and instructional procedures. According to Romero et al.⁽¹⁾, effective data analysis is critical for schools to better understand student performance, optimize academic management, and comply with legal requirements. This approach has also been highlighted by Baker et al.⁽²⁾, who note that learning analytics enables the integration of educational data to promote tailored interventions and more informed decisions.

In the field of driving schools, the application of EDM techniques has allowed addressing specific problems related to the diversity of student profiles and data complexity. Recent studies highlight that resource optimization in non-formal educational contexts, using data mining, can personalize instruction and improve learning outcomes.^(3,4) For example, Kumar et al.⁽⁵⁾ demonstrated how clustering algorithms, such as K-Prototypes, are effective in analyzing student dispositions, identifying behavioral patterns, and designing effective interventions.

This study focused on implementing a database knowledge discovery (KDD) model tailored to the context of driving schools, with the goal of identifying patterns in administrative and academic data to optimize internal and educational processes. The research was carried out in three main phases. The first consisted of process characterization, where relevant data, such as theoretical and practical hours, exam results and demographic data, were collected and preprocessed. The second phase was the experimental design, which employed clustering algorithms such as K-Means and K-Prototypes to explore patterns and segment the data into representative clusters. Finally, the generated models were evaluated to validate their effectiveness and ensure that the findings were applicable and scalable.

The results obtained evidenced the effectiveness of K-Prototypes in handling mixed data, highlighting its ability to integrate numerical and categorical variables in the segmentation of students. This approach made it possible to identify high-risk profiles and design personalized strategies to improve academic and administrative performance. In addition, as part of the study, an application was developed based on the model's findings, which facilitated the practical implementation of the results and real-time decision making.

To structure the knowledge discovery process in databases (KDD), this study adopts contemporary methodological frameworks, widely used in data mining projects. Its application in driving schools allows improving administrative procedures and ensuring that solutions are scalable and adaptable to the changing needs of the educational environment.

METHOD

The methodology adopted in this study was designed to ensure a comprehensive and systematic analysis of the data collected in driving schools. Rigorous procedures were implemented, ranging from data collection and preparation to the application of advanced clustering and model validation techniques. These steps ensured the quality, representativeness and usefulness of the findings, allowing the identification of relevant patterns in the data and the generation of actionable information to optimize both academic and administrative processes.⁽¹⁾

The approach included modern data preprocessing and analysis techniques. These techniques encompassed the elimination of null values, data normalization and coding of categorical variables, which allowed structuring a robust and coherent data set. Clustering algorithms, such as K-means and K-Prototypes, played a central role in the analysis. While K-means proved to be effective for exclusively numerical data, K-Prototypes excelled in handling mixed data, making it a key tool in educational contexts where data are heterogeneous.⁽⁵⁾

Model validation and repeatability were performed through cross-testing and adjustments to algorithm parameters, which ensured the robustness of the results and their applicability in similar educational contexts.

⁽²⁾ In addition, advanced dimensionality reduction tools, such as principal component analysis (PCA), were employed to simplify the data and facilitate their interpretation, while retaining as much information as possible.⁽⁴⁾

The methodological approach adopted allowed exploring and understanding the complexities of the data in a dynamic educational environment, providing practical recommendations for the improvement of academic and administrative processes in driving schools. This methodological framework integrates the most recent trends in educational data mining and learning analytics, demonstrating how these disciplines can transform decision making in educational institutions.

The following is a description of the methodological process that was carried out in this research:

Data collection and preparation

Data were collected and processed following standards to ensure quality and representativeness. Attributes such as theoretical hours, practical hours, exam results and demographic data were included. The collection was done through automated academic systems and direct surveys applied to students and administrators. In

addition, historical data from previous records were integrated to perform a comparative analysis of cleaning and transformation under supervised and unsupervised algorithms, as shown in figure 1.

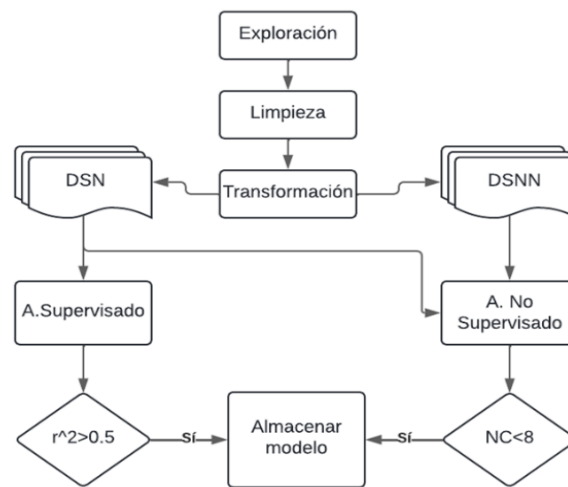


Figure 1. Experimental model for data management

Preprocessing

Preprocessing included several essential steps to ensure data quality and consistency:

Null Value Elimination: imputation techniques were implemented to deal with missing data, such as replacement by means or medians in numerical values and modes in categorical variables.

Data Normalization: numerical scales were standardized to ensure that variables had an equal impact on the models.

Coding of Categorical Variables: one-hot coding schemes were used for categorical attributes such as gender, and ordinal coding for experience levels or qualifiers.

Generation of Derived Variables: new variables were created, such as the ratio of practical hours completed to hours required, to obtain more representative indicators of student performance.

Grouping Techniques

Two main algorithms were selected to perform the clustering analysis:

K-means: this algorithm was used for exclusively numerical data, allowing to identify patterns in variables such as the number of theoretical and practical hours completed.

K-prototype: ideal for mixed data, integrating numerical and categorical variables such as age, geographic location and type of education. The combination of Euclidean distances for numerical variables and matching for categorical variables improved clustering quality.

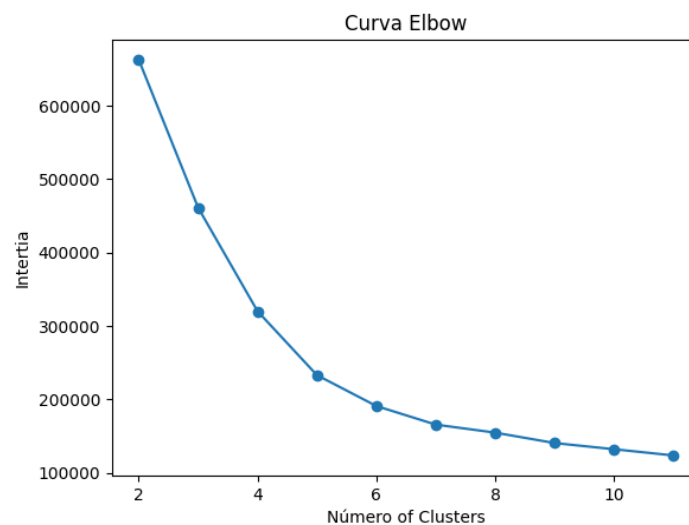


Figure 2. Elbow method

The optimal number of clusters was determined using the elbow method and visual assessments using scatter plots, as shown in figure 2. In addition, Principal Component Analysis (PCA) was used to reduce dimensionality and facilitate data interpretation.

Validation and repeatability

To validate the results, cross tests were performed with different subsets of data and hyperparameter settings in the algorithms. In addition, all the steps of the process were documented to ensure the repeatability of the study, the data were grouped into sets as shown in figure 3.

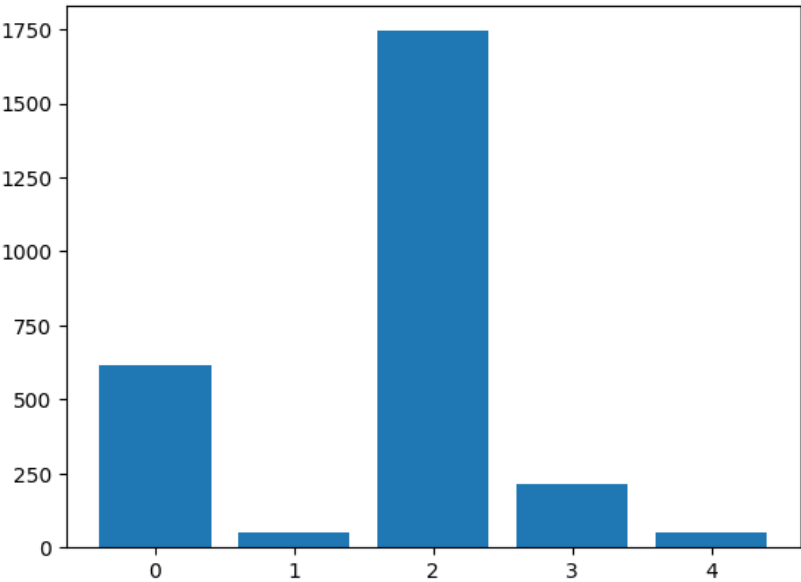


Figure 3. Data distribution

Table 1. Model		
Model	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.		

RESULTS

The implementation of several methodologies and algorithms allowed obtaining significant results in the analysis of educational data in driving schools. Through supervised and unsupervised techniques, predictive and exploratory performance was evaluated, maximizing the use of the available data.^(1,3) Unsupervised algorithms, such as K-means and K-prototype, proved effective in identifying key patterns in student data, while supervised ones, such as support vector machines (SVMs), excelled in predicting specific outcomes.^(4,5)

The clustering analysis, especially with K-prototype, allowed handling mixed data, which is crucial in educational contexts where data present both numerical and categorical variables. This facilitated the accurate segmentation of students and the identification of high-risk profiles, achieving personalized interventions to optimize their academic performance and reduce dropout.^(2,5)

In addition, an interactive tool was developed based on analytical models that integrates dynamic visualizations and personalized predictions, here data are loaded, cleaned, trained and predictions are generated, as presented in figure 4. This tool not only allows administrators to make informed decisions in real time, but also promotes the optimization of resources and educational strategies.^(1,3) The ability of such applications to interpret complex patterns is especially valuable in educational institutions, where the diversity of student profiles poses significant challenges.⁽⁴⁾

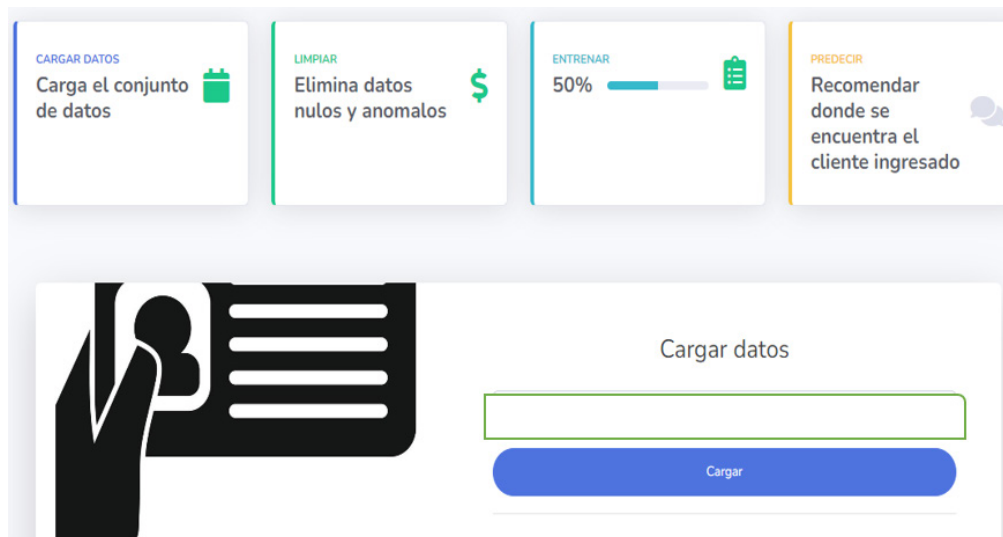


Figure 4. Developed tool

Algorithm testing

Supervised and unsupervised algorithms were tested in order to evaluate different analysis methodologies. The supervised algorithms included logistic regression and support vector machines (SVM), while the unsupervised algorithms included K-means and K-prototype. This mixed approach allowed comparing predictive and exploratory performance, maximizing the potential of the data.⁽⁶⁾

```
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)
```

Figure 5. Code in K Means

As we can see in figure 5, the following information analysis was performed for the research.

In the analysis with K-Prototypes, we worked with a data set that included both numerical and categorical variables. The numerical variables, such as theoretical and practical hours completed, as well as the age of the students, were processed to ensure their fit to the model. Unlike K-Means, K-Prototypes does not require prior scaling of the numerical variables, as it uses a combination of Euclidean distances for numerical data and matching for categorical data. This eliminates the need to apply techniques such as StandardScaler, which

are normally used in K-Means to ensure that numerical variables have a mean of zero and a unit variance. The ability of K-Prototypes to handle heterogeneous data without complex transformations is especially valuable in educational contexts, where datasets often include mixed features that need to be analyzed together in a consistent manner.

```
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)
```

Figure 6. Code in K Prototype

As shown in figure 6, the code execution process was performed as follows. Mixed Data:

A DataFrame is created that combines numerical (theoretical and practical hours) and categorical (gender) data.

Data Conversion

`to_numpy()`: Converts the DataFrame to a Numpy array, a format required by K-prototype.

Initialization and Tuning:

`KPrototypes(n_clusters=2)`: Indicates that you want to form 2 clusters.

`categorical=[2]`: Specifies that column 2 (0-based index) contains categorical data.

Centroids:

`kproto.cluster_centroids_`: Displays the cluster centroids, which include average values for numeric variables and most frequent values for categorical variables.

Labels:

`kproto_labels`: Labels assigned to each point, indicating to which cluster each record belongs. `categorical=[2]`: Specifies that column 2 (0-based index) contains categorical data.

Centroids:

`kproto.cluster_centroids_`: Displays the cluster centroids, which include average values for numeric variables and most frequent values for categorical variables.

Labels:

`kproto_labels_`: Assigned labels.

Configuration and evaluation

Evaluation was performed using metrics such as intracluster cohesion, intercluster separation, and cross-validation. The elbow method was key to determine the optimal number of clusters, complemented with indices such as the Silhouette coefficient. For the supervised models, metrics such as precision, recall, and F1-score were used. The results showed that K-prototype was more effective when handling mixed variables,

while SVM obtained a high performance in specific predictions, as shown in figure 7.

```
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}")
```

Figure 7. Silhouette coefficient

Practical examples

A notable example was the clustering of students according to their performance and course completion time. The clusters identified high-risk profiles, such as students with low grades in theoretical modules who required additional support. It was also found that students with previous road safety training were able to complete their courses in less time, which allowed for more efficient tailoring of educational resources.

Integration of tools

The findings were integrated into an application based on Flask and Dash that allows for dynamic data visualization. This tool provides real-time simulations, facilitating administrators to make informed decisions on resource allocation and course planning.

Main findings

Table 2. Comparative aspects		
Aspect	K Means	K Prototype
Supported Data	Numeric only.	Mixed (numeric and categorical).
Preprocessing	Mandatory scaling (StandardScaler).	Not required for categorical data.
Initialization	KMeans(n_clusters=N).	KPrototypes(n_clusters=N, categorical=[]).
Centroids	Numerical coordinates.	Most frequent numerical average and categorical values.
Libraries	sklearn	kmodes

The analysis conducted allowed us to identify key patterns and generate valuable information for both administrative and academic process optimization in automotive education centers. Some of the most salient findings include.

As shown in table 2, the comparison between the K-Means and K-Prototypes algorithms highlights key differences that are critical when choosing a clustering method in educational contexts with heterogeneous data. While K-Means proves to be effective when working exclusively with numerical data, its application is limited in scenarios where categorical variables play an important role. On the other hand, K-Prototypes is presented as a robust solution for handling mixed data, since it combines Euclidean distances for numerical variables and matches for categorical variables, allowing a more complete and accurate representation of the patterns in the data.

The analysis performed in this study compared the K-Means and K-Prototypes algorithms to identify which one was more suitable in the educational context of driving schools, where the data present heterogeneous characteristics. While K-Means is widely recognized for its efficiency in processing exclusively numerical data, its inability to handle categorical variables limited its usefulness in this case.

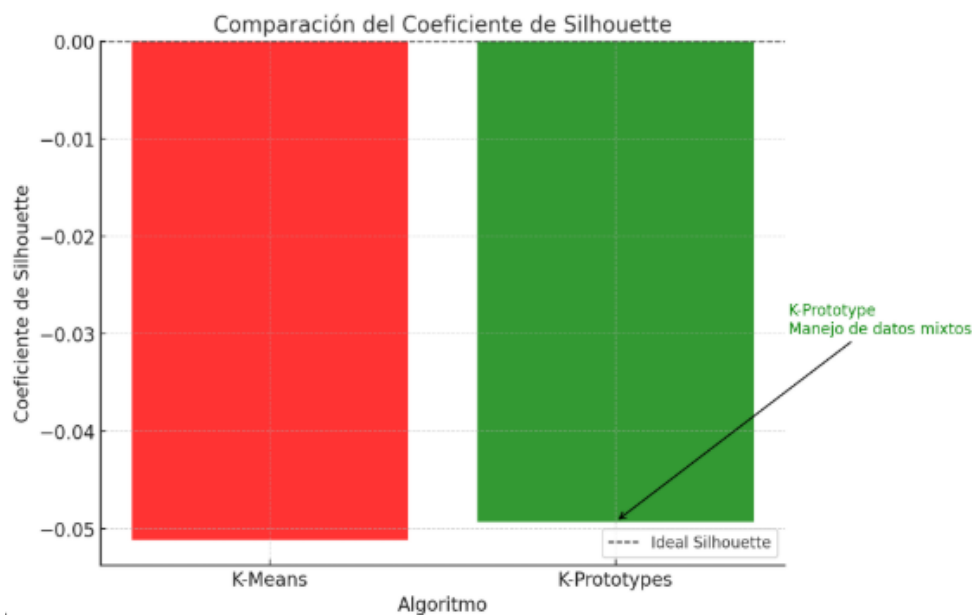


Figure 8. silhouette coefficient comparison

As shown in figure 8, K Prototype proved to be significantly more effective in handling mixed data, which allowed a more accurate and representative segmentation of students compared to K-Means. This efficiency is due to K-Prototype’s ability to work with heterogeneous data, combining numerical and categorical variables in its clustering process, which ensures greater consistency in the definition of clusters.

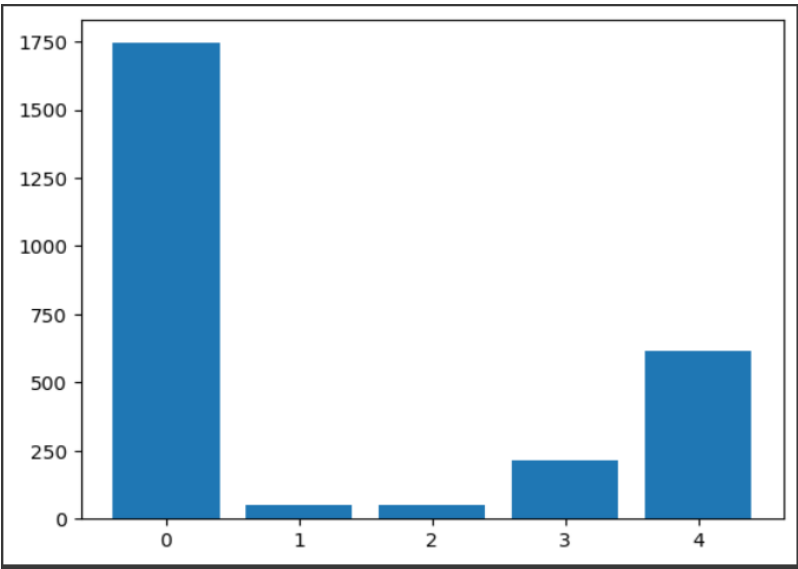


Figure 9. K Means data distribution

During the analysis, an adequate definition of the clusters was performed using solid criteria, such as the Silhouette coefficient. This value, although not completely positive due to the complexity of the data, showed a slight improvement with K-Prototype (-0,0493 versus -0,0512 in K-Means), indicating greater intracluster cohesion and less overlap between clusters. This after applying the algorithm shown in figure 6.

In addition, K-Prototype allowed us to identify significant patterns in the behavior of students, differentiating them according to variables such as geographical location, practical hours performed and exam results. These characteristics were critical to the objectives of the study and could not be addressed with the same precision by K-Means due to its limitation of working exclusively with numerical data.

K-Means can identify patterns related to continuous variables, such as practical hours or test scores, providing useful information on general trends in student performance, guaranteeing a data distribution as shown in figure 9. However, its main limitation lies in its inability to process categorical data, which makes it less suitable in environments with mixed or heterogeneous data.

Despite this, K-Means remains a robust tool for fast and effective analysis in specific situations, especially when a simple and computationally efficient model is needed.

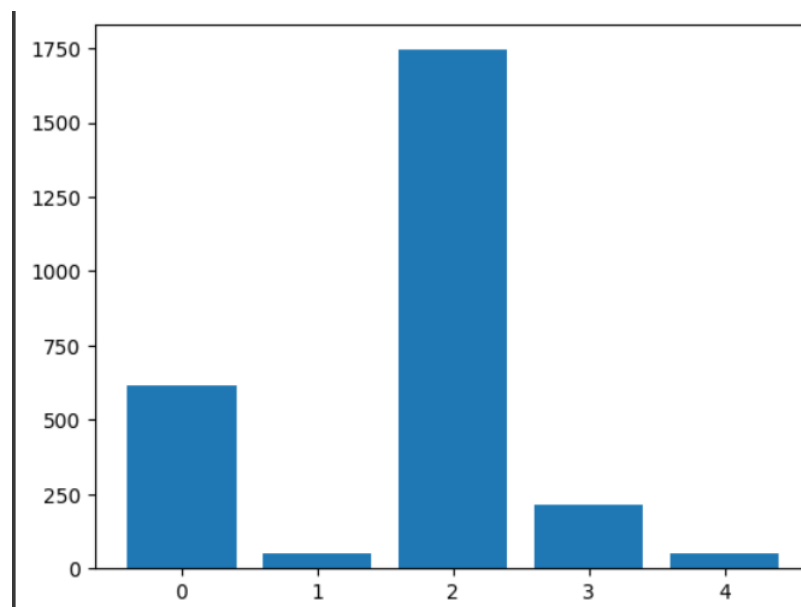


Figure 10. K Prototype data distribution

In conclusion, the use of K-Prototype optimized segmentation in mixed educational contexts, improving the quality of the clusters formed and allowing for more personalized intracluster interventions obtained and adequate differentiation of key groups, as can be seen in Figure 10. to improve academic and administrative performance. This is reflected in the increased cohesion

Patterns of academic performance:

It was observed that students who dedicated more hours to practice obtained higher performance in the final exams. The data revealed that 85 % of the students in the clusters with greater dedication to practical hours exceeded the minimum score in the theoretical and practical evaluation, highlighting the need to balance the time invested in both modalities.

The cluster analysis highlighted that geographic location has a significant impact on student performance. Students from urban areas tended to complete courses in less time due to greater accessibility to complementary educational resources, while those from rural areas faced greater challenges, requiring more time and additional tutoring.

The predictive models developed allowed us to identify profiles of students at risk of not successfully completing the courses. These profiles included students with low grades in road regulations and basic mechanics, as well as those with difficulties in meeting the required practical hours. This information facilitated the design of customized interventions to reduce dropout rates and improve overall outcomes.

Application Development

Implementation in the Flask application

The Flask-based application exclusively reads clusters generated by K-Prototype. Its architecture includes the following main functionalities:

Dynamic Visualization: interactive representation of the clusters and their predominant characteristics.

Predictions: estimates based on analytical results to identify at-risk or high-performing students based on research results.

Customized Reports: generation of reports that facilitate strategic decision making, optimizing available resources.

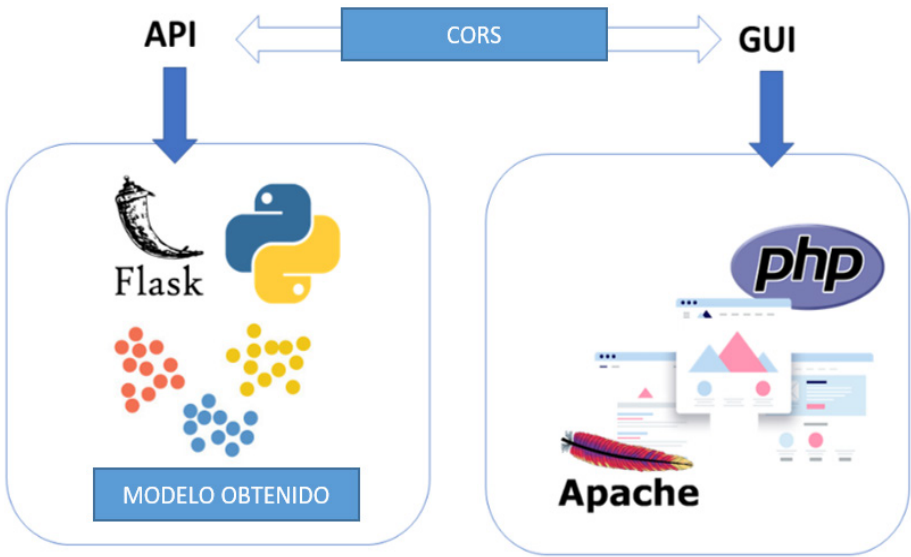


Figure 11. Discovery model application architecture

Comparison with literature

The results obtained in this study are consistent with previous research on the effectiveness of educational data mining.⁽¹⁾ However, this work brings key innovations by combining advanced clustering techniques with practical visualization and dynamic analysis tools. The integration of K-prototype and dimensionality reduction methods such as Principal Component Analysis (PCA) allowed for a clearer interpretation of the complexities of mixed educational data, a contribution little explored in the existing literature.

Model evaluation and metrics

Model validation was performed by cross-testing and analysis of key metrics, such as Intraclass Cohesion and Intercluster Separation.

The K-Prototype algorithm achieved improved intraclass cohesion compared to K-Means, demonstrating its effectiveness in handling mixed data. Although the actual Silhouette coefficient for both algorithms indicated challenges in cluster quality, K-Prototype achieved a slight improvement in intraclass cohesion with a Silhouette coefficient of -0,0493, versus -0,0512 for K-Means. This difference, although small, is significant in contexts where the data characteristics include both numerical and categorical variables, allowing a more coherent and representative segmentation, applied with the code present in figure 6.

Table 3. Comparison of algorithms		
Aspect	K Means	K Prototype
Main Advantage	Computational efficiency for numerical data.	Effective handling of mixed data (numeric and categorical).
Main Limitation	Cannot process categorical variables.	Higher computational complexity due to hybrid computation.
Ideal Application	Analysis of exclusively quantitative data.	Contexts where quantitative and qualitative characteristics are mixed.
Main Advantage	Computational efficiency for numerical data.	Effective handling of mixed data (numerical and categorical).
Main Limitation	Cannot process categorical variables.	Higher computational complexity due to hybrid computation.

The average coefficient for K-Prototype was -0,0493, while for K-Means it was -0,0512. Although both values are negative, indicating possible overlaps between clusters, K-Prototype showed a relatively better performance, standing out in the ability to work with mixed data and to adapt to the diversity of the variables analyzed. This advantage suggests that K-Prototype is more suitable for applications where categorical and numerical variables are essential to define clusters.

Clusters found with K Prototype

Cluster 1: Students with High Integral Performance

Main characteristics:

Theoretical hours: 30,53 (meet and slightly exceed the 30-hour requirement).

Practical hours: 30,49 (meets the hours required for the C1 category).

Exam results: Theoretical 90,83 %, practical 99,54 %.

Category: C1.

Type of procedure: First time.

Location: predominance of Pasto.

Training: applying for driver's license.

Interpretation: this cluster represents students with excellent performance, who meet or exceed all training requirements for category C1. They are ideal candidates for licensing and do not require additional intervention.

Cluster 2: Students with Insufficient Training.

Key characteristics:

Theory hours: 9,72 (well below the 30-hour requirement).

Practical hours: 0,0 (not meeting the 15 hours required for A2).

Exam results: theoretical 0,0 %, practical 0,0 %.

Category: A2.

Type of procedure: First time.

Location: predominantly Pasto.

Training: applying for driver's license.

Interpretation: this cluster groups students who have not advanced significantly in their training, which could be due to abandonment or lack of commitment. They represent a high risk and need an intensive strategy to meet basic requirements.

Cluster 3: Students with Intermediate Progress

Key characteristics:

Theoretical hours: 29,21 (very close to the required 30 hours).

Practical hours: 16,79 (meets and slightly exceeds the 15 hours required for A2).

Exam results: theoretical 90,76 %, practical 99,79 %.

Category: A2.

Type of procedure: first time.

Location: predominantly Pasto.

Training: applying for driver's license.

Interpretation: this cluster represents students who are close to meeting all training requirements. Their performance in the exams is excellent, indicating that they could benefit from minor adjustments in their training itinerary.

Cluster 4: Students with Predominant Practical Training

Key characteristics:

Theoretical hours: 2,89 (well below the 30-hour requirement).

Practical hours: 20,57 (exceeding the 15 hours required for A2, but falling short of the 30 for C1).

Exam results: theoretical 90,62 %, practical 99,76 %.

Category: A2.

Type of procedure: first time.

Location: predominantly Pasto.

Training: applying for driver's license.

Interpretation: this cluster groups students who prioritize practice over theory, which might be appropriate for the A2 category, but not for C1. Although they achieve good exam results, they need to reinforce theoretical training to meet the required standards.

Cluster 5: Students with Low Training Efficiency

Main characteristics:

Theoretical hours: 30,14 (they meet the required 30 hours).

Practical hours: 13,3 (do not reach the 15 hours required for A2 nor the 30 for C1).

Exam results: theoretical 86,08 %, practical 6,62 %.

Category: A2.

Type of procedure: first time.

Location: predominantly Pasto.

Training: applying for driver's license.

Interpretation: this cluster includes students who comply with the theoretical hours, but do not achieve the required practices, which significantly affects their performance in the practical exams. They need a corrective plan focused on reinforcing their practical skills.

CONCLUSIONS

Advantages of K-Prototype on Mixed Data: The K-Prototype algorithm showed a greater ability to handle heterogeneous data, combining numerical and categorical variables, which allowed a more representative segmentation of the students. Although the Silhouette coefficient obtained for K-Prototype (-0,0493) was close to that of K-Means (-0,0512), its ability to integrate both data typologies is crucial in educational contexts where categorical characteristics (such as geographic location or type of education) are determinant for clustering.

The results show that, although K-Means is faster and more efficient for exclusively numerical data, K-Prototypes is superior for mixed scenarios, achieving a more accurate segmentation. In this study, K-Prototypes allowed generating clusters that identified students with an average theoretical performance of 90,7 % and a practical performance of 99,5 %, metrics that could not be adequately analyzed by K-Means due to its lack of integration of categorical variables.

Evaluation of Intracluster Cohesion and Intercluster Separation: The results showed that K-Prototype achieved better intracluster cohesion by segmenting more defined groups, adapting to the complexity of the educational data. Although no drastic improvement in global metrics was observed, cluster evaluation showed a reduction in intercluster overlap compared to K-Means.

Technical Implications for Algorithm Selection: The choice of K-Prototype was based on its ability to work with mixed data and provide results more tailored to the needs of the study. This algorithm made it possible to identify significant patterns in critical variables such as practical hours performed and geographic location, which could not be treated with the same effectiveness by K-Means.

During the execution of the algorithms, K-Means proved to be more efficient in terms of time and convergence, requiring on average 15 iterations to reach an optimal point compared to K-Prototypes, which required approximately 22 iterations due to the higher computational complexity of handling mixed variables. However, the longer processing time of K-Prototypes resulted in more representative and useful clusters for analysis.

BIBLIOGRAPHIC REFERENCES

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: Fundamentals, 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.

FINANCING

None.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

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

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

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

Drafting - original draft: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.

Writing - proofreading and editing: John Jairo Rivera Minayo, Javier Alejandro Jiménez Toledo, Deixy Ximena Ramos Rivadeneira, Jorge Albeiro Rivera Rosero.