







Cluster analysis for the prevention of hospital readmission of diabetic patients

Análisis de clústeres para la prevención del reingreso hospitalario de pacientes diabéticos

César G. Villanueva-Rueda ¹, Oscar Chávez-Bosquez ², Betania Hernández-Ocaña ³ y José Hernández-Torruco ⁴

Fecha de Recepción: 28 de agosto de 2025

Fecha de Aceptación: 13 de noviembre de 2025

Cómo citar: Villanueva-Rueda, C.G; Chávez-Bosquez, O; Hernández-Ocaña, B; and Hernández Torruco, J. (2025). Análisis de clústeres para la prevención del reingreso hospitalario de pacientes diabéticos. *Tecnura*, 29(86), 1–16. <https://doi.org/10.14483/22487638.24061>

Abstract


Objective: Hospital readmission in diabetic patients represents a significant challenge for both healthcare systems and patients' quality of life. That is why the main objective of this work is to identify common patterns associated with a higher risk of readmission.


Methodology: This study presents a clustering analysis applied to a clinical dataset of diabetic patients who were readmitted to hospitals across 130 medical institutions in the United States. The analysis employs unsupervised clustering algorithms, such as k-means and PAM, to segment patients based on their clinical and demographic characteristics. The study also evaluates different feature selection methods, identifying Simulated Annealing (SA) as the most effective for selecting optimal subsets of variables.

Results: Recurring factors such as length of hospital stay, associated medical conditions, and types of treatment received, significantly influence the probability of readmission.

Conclusions: The clustering results provide valuable insights to guide the development of personalized intervention strategies aimed at reducing hospital readmission rates among diabetic patients.

Keywords: clustering analysis, hospital readmission, diabetes, machine learning, feature selection.

¹ Estudiante de la Maestría en Ciencias de la Computación, División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México.  Email: 232H21010@alumno.ujat.mx

² Profesor-Investigador, División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México. [Perfil UJAT](#).  Email: oscar.chavez@ujat.mx

³ Profesora-Investigadora, División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México. [Perfil UJAT](#).  Email: betania.hernandez@ujat.mx

⁴ Profesor-Investigador, División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México. [Perfil UJAT](#).  Email: jose.hernandezt@ujat.mx

Resumen

Objetivo: El reingreso hospitalario en pacientes con diabetes representa un reto significativo tanto para los sistemas de salud como para la calidad de vida del paciente. El objetivo principal de este trabajo es identificar patrones comunes asociados a un mayor riesgo de reingreso.

Métodología: Este estudio presenta un análisis de conglomerados (clusters) aplicado a un conjunto de datos clínicos de pacientes con diabetes que fueron reingresados en hospitales de 130 instituciones médicas en Estados Unidos. El análisis emplea algoritmos de conglomerados no supervisados, como k-means y PAM, con el fin de segmentar a los pacientes según sus características clínicas y demográficas. Este estudio también evalúa diferentes métodos de selección de características, identificando el Recocido Simulado (Simulated Annealing) como el más efectivo para seleccionar subconjuntos óptimos de variables.

Resultados: Factores recurrentes, tales como: duración de la estancia hospitalaria, afecciones médicas asociadas y los tipos de tratamiento recibido, influyen significativamente en la probabilidad de reingreso.

Conclusiones: Los resultados del conglomerado proporcionan información valiosa para guiar el desarrollo de intervenciones personalizadas dirigidas a reducir las tasas de reingreso hospitalario en pacientes con diabetes.

Palabras clave: análisis de agrupamiento, reingreso hospitalario, diabetes, aprendizaje automático, selección de características.

1. Introduction

Diabetes is one of the most prevalent chronic diseases worldwide. According to the International Diabetes Federation (IDF) (1), 537 million people between the ages of 20 and 79 live with this condition. An alarming increase in the prevalence of this disease is projected, estimating that by the year 2046, around 784 million people could be affected—representing a 46 % increase compared to 2019 (2). This growth presents a significant challenge for healthcare systems, not only in terms of diagnosis and treatment, but also in managing complications and recurrent hospital readmissions.

Indeed, hospital readmissions are a key indicator of the quality of medical services, as they reflect the effectiveness of clinical management, and the patient's ability to stabilize outside the hospital setting. In the context of diabetes, multiple factors, including the type of treatment received, length of hospital stay, and the presence of comorbidities, considerably influence the likelihood of readmission (3). Reducing these readmissions not only decreases costs for healthcare systems but also improves patients' quality of life.

Given the complex and heterogeneous nature of clinical data, the use of machine learning techniques has become an essential tool for extracting meaningful patterns from large volumes of medical information. Although supervised models have been widely applied to predict readmissions, they depend on prelabeled data and may limit the exploration of hidden relationships between variables. In contrast, unsupervised methods, such as clustering algorithms, allow identifying subpopulations of

patients with similar characteristics without requiring predefined output information, facilitating hypothesis generation and the comprehension of underlying factors associated with readmission risk.

In this study, k -means and Partitioning Around Medoids (PAM) algorithms were applied to a clinical dataset of diabetic patients to detect homogeneous groups, according to their clinical and demographic characteristics. To improve cluster stability and quality, an optimization stage using Simulated annealing, a metaheuristic technique, was incorporated. This approach allowed for better centroid convergence and lower sensitivity to initial conditions. The results showed that the Simulated Annealing-based approach provided the best outcomes, in terms of cluster cohesion and separation, demonstrating its effectiveness in segmenting patients within complex clinical contexts.

Along these lines, this work aims at contributing to more interpretable and adaptive analytical models, allowing healthcare professionals to better understand the diversity of profiles among diabetic patients, and to design personalized strategies accordingly.

2. Related work

Several studies have addressed hospital readmission prediction in diabetic patients using machine learning techniques. One of the most relevant works is titled “Impact of HbA1c Measurement on Hospital Readmission Rates in Patients with Diabetes”, which analyzed more than 100,000 clinical records from 130 hospitals in the United States. This means it employed the same dataset than us. In this study, supervised models such as decision trees, logistic regression, and neural networks were applied to estimate the probability of hospital readmission, demonstrating the relevance of factors such as length of stay, glycemic control, and comorbidities (4).

However, supervised models present limitations as they rely on predefined labels and fail to capture latent structures in data. To overcome this issue, further recent research has incorporated unsupervised methods such as clustering algorithms: “Unsupervised Learning for Diabetes Subgroup Identification Based on Clinical Features and Glycemic Control”. This work applied the k -means algorithm to group type 2 diabetic patients according to clinical and demographic variables. The authors identified patterns associated with glycemic control levels and pharmacological treatment, enabling the definition of subgroups with different complication risks (5).

Similarly, the work “Enhanced Clinical Interpretation of Diabetes Clustering via PAM and PCA Integration” combined Partitioning Around Medoids (PAM) with Principal Component Analysis (PCA) to improve the interpretability of the resulting clusters. The results showed that PAM demonstrated greater robustness to outliers and revealed clinically coherent profiles consistent with medical practice (6).

Finally, “Optimization of Medical Data Clustering Using Simulated annealing” demonstrates that integrating optimization metaheuristics, such as Simulated annealing, significantly improves intra-cluster cohesion and inter-cluster separation, in contrast to k -means and PAM. This hybrid approach achieved more stable solutions with lower sensitivity to initial conditions (7). These studies highlight the effectiveness of unsupervised methods and metaheuristics in exploring complex clinical data, facilitating patient stratification, and contributing to personalized diabetes treatment.

3. Materials and Methods

3.1 Dataset

This study uses the “Diabetes 130-US Hospitals for Years 1999–2008” dataset (8) from the UC Irvine Machine Learning Repository. The dataset includes data from 130 U.S. hospitals on diabetic patients who experienced hospital or outpatient readmission. It consists of 50 variables and 101,766 instances, with 33 categorical and 17 numerical variables.

For this study, a sample of 5,000 instances was extracted to perform clustering analyses using different algorithms, whose results are discussed in later sections. The original dataset includes 50 variables and 101,766 instances of hospitalized diabetic patients between 1999 and 2008. From these, 17 variables are numerical and 33 categorical. Table 1 summarizes the main numerical variables.

Table 1. Statistical Description of Numeric Variables

Variable	Min	Max	Mean	Std. Dev
time_in_hospital	1	14	4.39	2.47
num_lab_procedures	1	132	43.09	19.67
num_procedures	0	6	1.33	1.70
num_medications	1	81	16.02	8.12
number_outpatient	0	42	0.37	1.29
number_emergency	0	76	0.19	0.86
number_inpatient	0	21	0.10	0.64
number_diagnoses	1	9	7.42	1.35

4. Evaluation and Validation metrics

The evaluation of the results obtained through clustering algorithms requires the use of specific metrics that allow for quantifying the segmentation quality performed. A metric can be defined as a mathematical function used to assess similarity, cohesion, or separation among the grouped elements,

depending on the analysis approach and the type of data (9). These metrics are fundamental for comparing the effectiveness of different clustering models.

4.1 Elbow Method

One of the most commonly used techniques to determine the optimal number of clusters is the Elbow method. This technique is based on analyzing the total sum of squared errors (SSE) for different values of k , which represents the number of clusters. When plotting the SSE values against k , a progressive decrease is typically observed until a point where the improvement becomes marginal. This inflection point is interpreted as the optimal number of clusters, and it is known as the “elbow” of the graph. This metric is especially useful for centroid-based algorithms such as k -means, as it provides a visual way to balance between model accuracy and complexity (10).

As shown in Fig. 1, the Elbow method analysis suggests that the optimal value of k for this dataset lies within the range of 2 to 5. Starting from $k = 5$, the reduction in the Within-cluster sum of squares (WCSS) becomes marginal, indicating little improvement in cluster compactness. This suggests that increasing the number of clusters beyond this point would not provide substantial benefits to the model and could result in unnecessary oversegmentation.

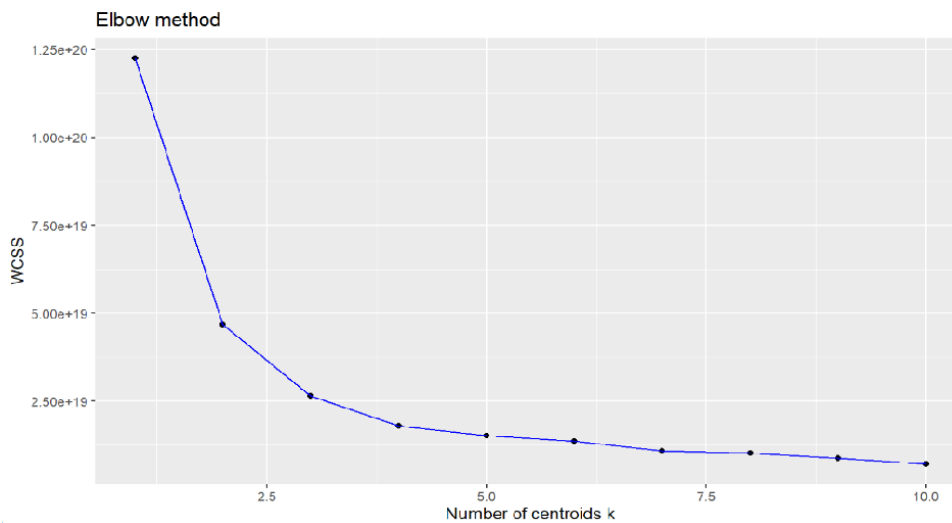


Figure 1. Elbow method for determining the Optimal number of centroids

WCSS (Within-Cluster Sum of Squares) is a metric that measures the sum of the squared distances from each data point to its corresponding centroid within a cluster. In other words, it is a measure of cluster compactness. The goal of the k -means algorithm is to minimize this value in order to achieve the most compact clusters as much as possible.

The formula to calculate WCSS is produced as follows:

$$WCSS = \sum_{k=1}^K \sum_{x \in C_k} \|x - \mu_k\|^2$$

Where:

- K : number of clusters.
- C_k : representa el clúster cluster k . ■ x : data point in cluster k .
- μ_k : centroid of cluster k .
- $\|x - \mu_k\|^2$: squared Euclidean distance between the point x and its centroid.

4.2 Silhouette

The silhouette coefficient is another widely used metric for evaluating clustering quality. This measure combines the concepts of cohesion (how close the points are within the same cluster) and separation (how far the points are from other clusters). The silhouette coefficient ranges from -1 to 1, where a value close to 1 indicates that the instances are well clustered, while values close to 0 or negative suggest incorrect or ambiguous assignments.

The silhouette coefficient provides an overall view of clustering effectiveness and can also be used to determine the appropriate number of clusters by evaluating the average silhouette score across different partitions (11).

5. Data Analysis

Exploratory Data Analysis (EDA) represents a fundamental stage in the development of predictive models. Through statistical techniques and graphical visualization, a deep understanding of the dataset's characteristics is achieved, allowing for the identification of patterns, detection of anomalies, and evaluation of relationships between variables. This prior understanding is essential for performing proper preprocessing, thereby optimizing the performance and interpretability of subsequent models.

5.1 Data Visualization

Data visualization is a powerful tool for EDA, facilitating the interpretation and detection of potential issues in the dataset. The most commonly used techniques include:

- **Histograms:** Allow for the analysis of the distribution of numerical variables, identifying skewness, multimodality, and possible deviations from a normal distribution.
- **Boxplots:** Useful for identifying outliers and evaluating data dispersion in relation to quartiles.
- **Bar Charts:** Used to represent the frequency of categorical variables, providing information about class or category distribution.
- **Scatter Plots:** Allow the evaluation of relationships between two variables, identifying linear or nonlinear patterns that could influence predictive modeling.

Below is a visual exploration of the dataset using various visualization techniques that help identify patterns, distributions, and potential anomalies in the variables. The following images illustrate these graphical representations, highlighting the structure and behavior of the data, thus facilitating interpretation and analysis.

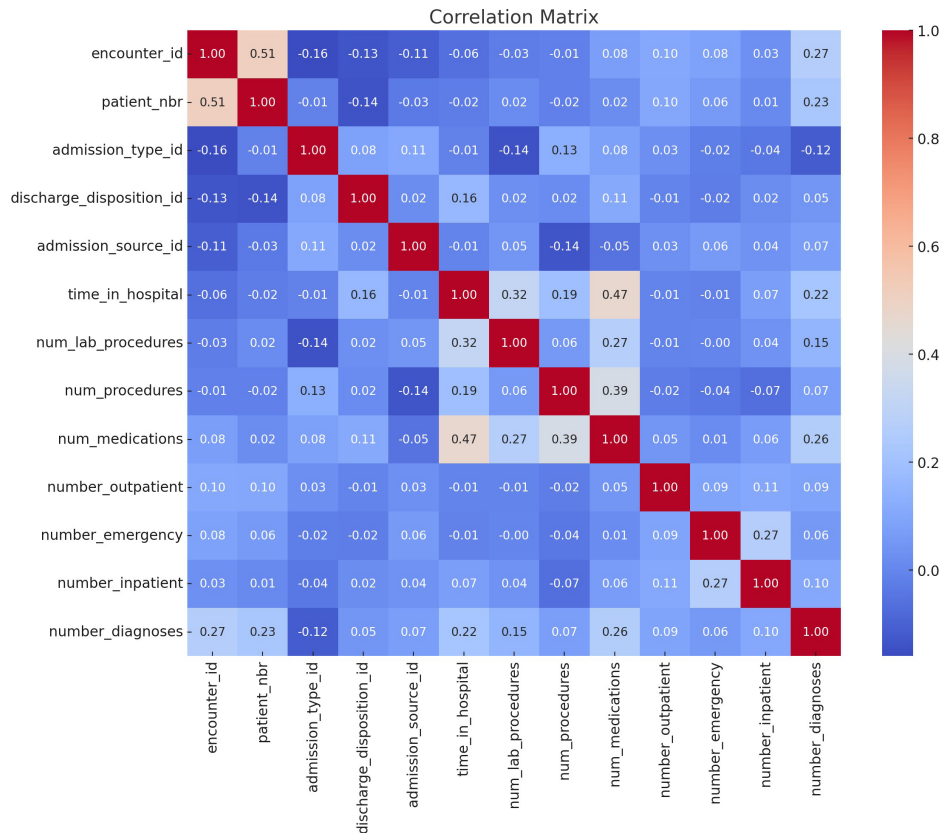
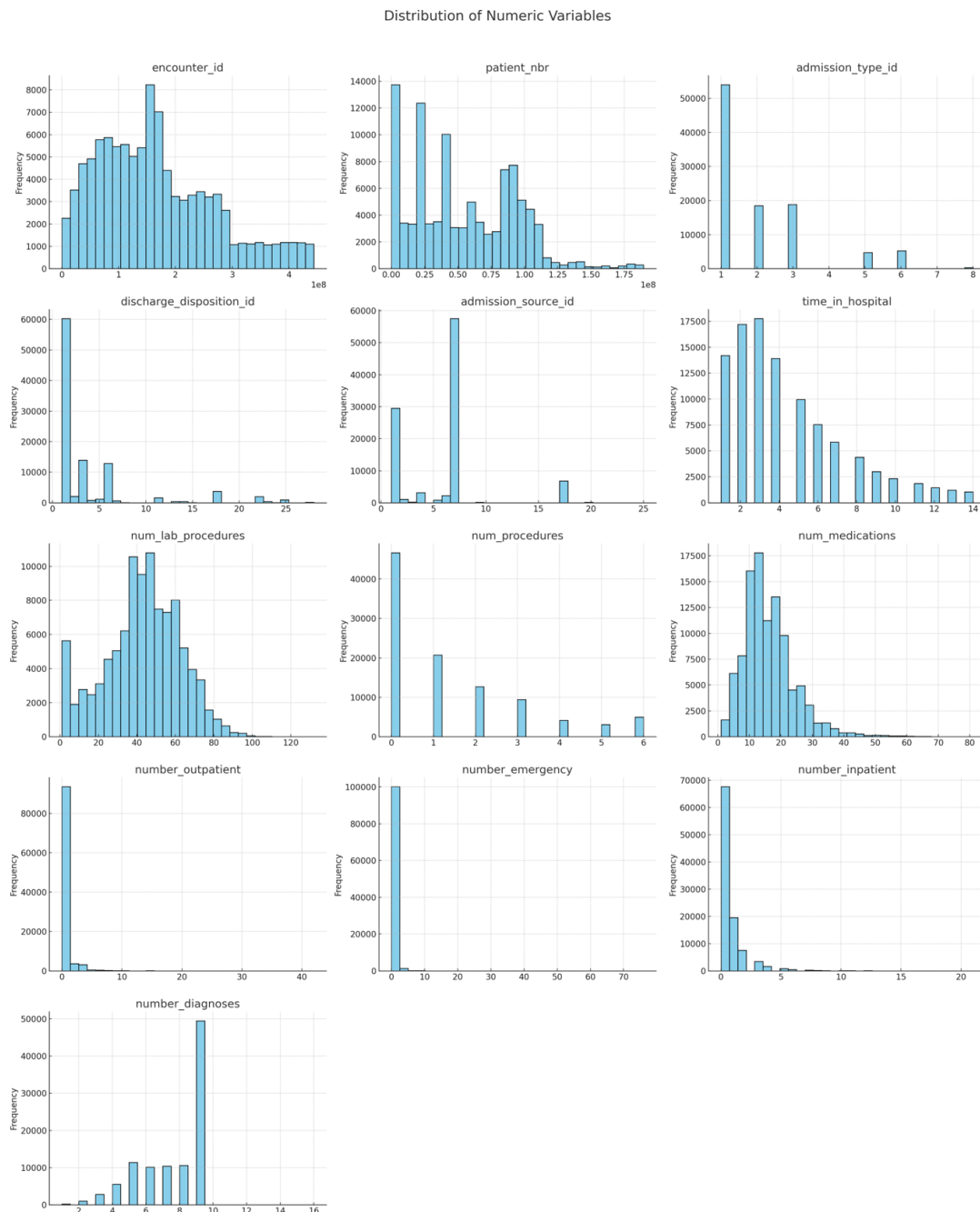


Figure 2. Correlation matrix of numeric variables

In Figure 2, the correlation matrix of the variables selected for analysis is shown. This representation allows the identification of linear relationships between numerical variables, facilitating the detection of collinearities that could affect the performance of predictive models. High positive values (close to 1) indicate a strong direct correlation, while negative values reflect an inverse relationship.

The matrix highlights some significant correlations, such as those between `encounter_id` and `patient_nbr`, which may suggest a dependency between these variables in the hospital records. On the other hand, the variable `num_medications` shows a moderate correlation with `num_procedures`, indicating a possible pattern in the clinical management of patients. Additionally, variables like `time_in_hospital` and `num_lab_procedures` present a positive correlation, which could reflect an increase in laboratory tests as the hospital stay lengthens.

Figure 3. Histograms of numeric variables



In Figure 3, the histograms of the selected numerical variables for analysis are presented. This visualization allows for the observation of data distributions and the detection of potential skewness, outliers, and asymmetries in each variable.

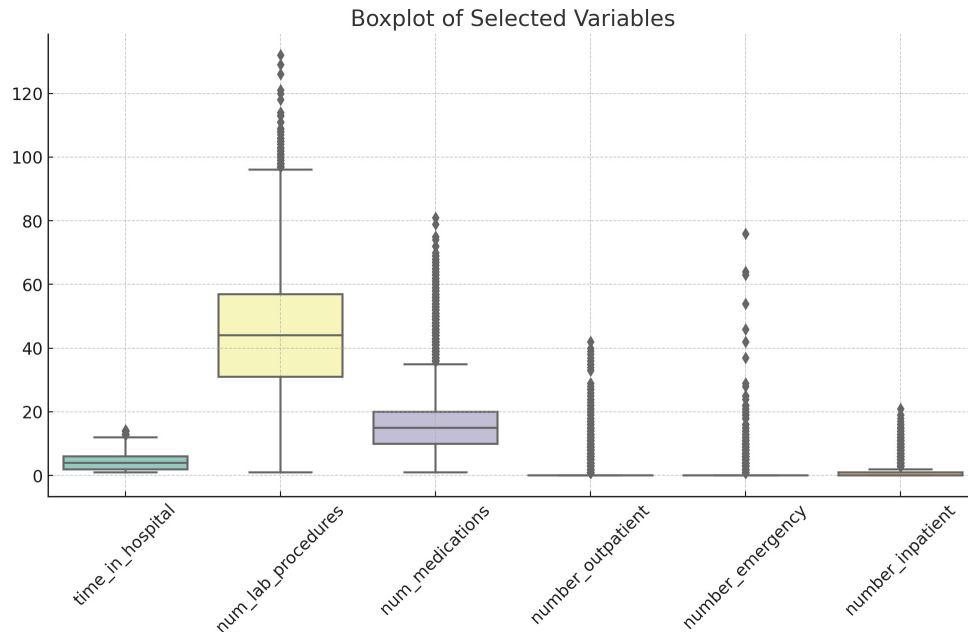


Figure 4. Boxplots for detecting outliers

Comparatively, the histograms reveal distinctive behaviors across variables. For example, variables such as time_in_hospital and num_lab_procedures exhibit slightly skewed distributions, suggesting a higher concentration within specific intervals. Additionally, the variable number_emergency shows a left-skewed distribution, indicating that most patients had few recorded emergencies, while a smaller number presented exceptionally high values.

In Figure 4, the boxplots of the selected variables for analysis are shown. This graphical representation allows for identifying the dispersion of the data, as well as the presence of outliers in each variable. These boxplots reveal notable differences in data variability. For example, the variable num_lab_procedures exhibits a wider interquartile range compared to the others, indicating greater dispersion in the number of laboratory procedures performed. In contrast, the variable time_in_hospital shows a tighter concentration around its central values, even when several outliers stand out above the 75th percentile.

6. Data Preprocessing

After visualizing the data, we observed the presence of missing values represented by question marks. Therefore, when loading our dataset, we specified that values containing question marks should

be replaced with NA, as this is recognized by the R programming language. Similarly, we noted that these missing values were numerical.

Table 2. Table of variables with missing data

Variable Name	Number of Records	Missing Values
weight	101766	98569
payer.code	101766	40256
medical.specialty	101766	49949

As shown in Table 2, the variable weight contains 98,569 missing values out of a total of 101,766 instances, which accounts for more than 80 % missing data. Likewise, the other variables also exceed 20 % of missing data (12). Due to the high proportion of missing values, we chose to remove these variables, resulting in a total of 47 variables remaining for our model.

We found that 33 of the variables were categorical, so it was necessary to transform them from categorical to numerical, as it is essential for modeling that the data be numeric.

Table 3 presents the variables that were excluded from the analysis, resulting in a total of 42 variables selected for experimental testing. This reduction was carried out with the purpose of optimizing model performance and mitigating possible effects of multicollinearity or redundancy in the dataset.

Table 3. Removed variables

Removed Variables
Encounter_id Pa- tient_nbr Weight Admission_type_id Discharge_disposition_id Payer_code Medi- cal_specialty Num_medications

7. Modeling

7.1 K-means

The k -means algorithm is a clustering technique that partitions a dataset into K distinct groups by minimizing intra-cluster variation and maximizing inter-cluster variation. This method is especially useful in applications such as customer segmentation, image compression, and pattern detection in unlabeled data (13).

7.2 PAM

The Partitioning Around Medoids (PAM) algorithm is a clustering technique that, unlike *k*-means, uses actual objects from the dataset as the centers of clusters, called medoids. This makes it more robust to outliers and allows it to work with arbitrary dissimilarity metrics (14).

7.3 FSelector

Feature selection is a crucial process for reducing data dimensionality and improving the efficiency of machine learning models (15). The FSelector library in R provides several techniques for this purpose, including:

- **Chi-square:** Evaluates the independence between each feature and the target variable, making it useful for categorical data.
- **Information Gain:** Measures the reduction in entropy when the value of a feature is known, favoring those that provide the most information about the target variable.
- **Relief:** Estimates the relevance of features by considering the difference between nearby instances of different classes.
- **Random Forest Importance:** Uses Random forest models to assess each feature's contribution to prediction, identifying the most influential ones.
- **CFS (Correlation-based Feature Selection):** Selects subsets of features that are highly correlated with the target variable and have low intercorrelation, using a heuristic search to maximize relevance and minimize redundancy.
- **OneR:** Creates simple classification rules based on a single feature, selecting the one with the lowest error rate.

7.4 Simulated annealing

Simulated annealing is a metaheuristic technique inspired by the process of metal cooling, used to find optimal solutions in complex search spaces. In the context of feature selection, this method enables the exploration of different combinations of variables to identify subsets that optimize the performance of the predictive model.

The implementation of Simulated annealing in the caret package in R, through the safs function, efficiently performs the search for the optimal combination of features. The algorithm evaluates the model's performance at each iteration and accepts suboptimal solutions with a decreasing probability, which helps avoid local optima and improves the quality of the final solution.

8. Results

For the clustering analysis, the k -means and PAM algorithms were implemented, selecting a consistent configuration to ensure reproducibility and stability of the results. Both techniques were executed using a fixed seed value of 3, allowing for deterministic initialization of centroids and medoids, respectively. Moreover, 1000 iterations were configured to ensure convergence in each run, and 10 restarts ($nstart = 10$) were applied to evaluate multiple initializations and select the best partition based on the lowest intra-cluster distance.

Table 4 shows the silhouette analysis results obtained through the k -means and PAM algorithms, using the 42 selected variables from the dataset.

Table 4. Silhouette scores for different values of k using k -means and PAM

Algorithm	K Measures			
	K = 2	K = 3	K = 4	K = 5
k -means	0.53	0.47	0.46	0.40
PAM	0.50	0.48	0.43	0.38

Table 5 presents the results of the feature selection process using the CFS (Correlation-based Feature Selection) method from the FSelector library. This technique evaluates the relevance of variables, based on their correlation with the target variable and low redundancy with other selected variables. As a result, five optimal variables were identified and used to calculate the silhouette scores under different values of k . The evaluation was conducted incrementally, starting with the first two selected variables and progressively adding one variable at a time, evaluating the model's performance at each step. This approach allows us to observe how the quality of the clusters varies as new dimensions are introduced into the analysis.

Regarding table 6, it presents the results of the variable selectors generated by FSelector algorithms. These algorithms were executed incrementally, starting with the first two selected variables and progressively adding one more at each iteration until all features identified by each method were considered. However, only the best results for each algorithm are shown. Additionally, the variables that significantly contributed to achieving those values are specified, helping to identify the most representative subsets for each evaluated model.

Within table 7, there is a list of variables selected during the iterations that achieved the best silhouette results. As shown, three of the selectors initially identified the same two variables; however, as the iterations progressed, the selected variables diverged, revealing different selection behavior in subsequent stages.

Table 8 shows the results obtained using the Simulated annealing algorithm. At the end of the process, the model identified three optimal variables: Nateglinide, Tolbutamide, and A1Cresult. These variables represented the best performance when evaluated with the silhouette coefficient.

Table 5. Cluster analysis results for different subsets and *k* values

Subset	K	Silhouette Mean
gender, metformin	2	0.7189350
gender, metformin	3	0.8133443
gender, metformin	4	0.8938578
gender, metformin	5	0.9846291
gender, metformin, glyburide	2	0.6175232
gender, metformin, glyburide	3	0.7064362
gender, metformin, glyburide	4	0.6856388
gender, metformin, glyburide	5	0.7778758
gender, metformin, glyburide, change	2	0.4085565
gender, metformin, glyburide, change	3	0.4832919
gender, metformin, glyburide, change	4	0.5104543
gender, metformin, glyburide, change	5	0.5557880
gender, metformin, glyburide, change, cluster	2	0.3766780
gender, metformin, glyburide, change, cluster	3	0.3390272
gender, metformin, glyburide, change, cluster	4	0.3728791
gender, metformin, glyburide, change, cluster	5	0.3716568

9. Discussion

The results obtained confirm that the use of clustering algorithms effectively segments diabetic patients into clinically relevant groups. Compared to previous studies, the silhouette values obtained (up to 0.99 with Simulated annealing) exceed those reported by Zeng et al. (2021), who achieved maximum values of 0.78 using only *k*-means (5). This improvement occurs due to the use of Simulated annealing, which reduces dependence on initial conditions and optimizes the global search for centroids.

Table 6. Silhouette scores with different *k* values using FSelector algorithms

FSelector	K Measures			
	K = 2	K = 3	K = 4	K = 5
Chi-square	0.74	0.72	0.73	0.74
Information Gain	0.82	0.69	0.68	0.68
OneR	0.82	0.69	0.68	0.68
Random Forest	0.82	0.69	0.68	0.68
Relief	0.60	0.65	0.69	0.70

Table 7. Algorithms and selected variables

Algorithm	Selected variables
Chi-square	number_inpatient, number_emergency
Information Gain	number_inpatient, discharge_disposition_id
OneR	number_inpatient, discharge_disposition_id
Random Forest	number_inpatient, discharge_disposition_id
Relief	glimepiride, glipizide

Table 8. Silhouette scores for different k values using Simulated annealing

Algorithm	K Measures			
	K = 2	K = 3	K = 4	K = 5
Simulated annealing	0.72	0.82	0.98	0.99

Likewise, the results suggest that variables such as `AlCresult`, `Nateglinide`, and `Tolbutamide` are key determinants for characterizing risk profiles, consistent with the findings of Li and Zhang (2023) regarding the importance of glycemic control and the use of medications in predicting readmissions (7). These patterns support the hypothesis that combining unsupervised methods with metaheuristics constitutes an effective alternative for exploratory analysis of complex clinical data.

10. Conclusion

The results obtained demonstrate that Simulated annealing is the most effective approach for identifying the global optimum in the analyzed dataset, compared to the other evaluated algorithms such as `FSelector`, `k-means`, and `PAM`. Through its ability to explore the solution space in a random-controlled manner, this algorithm avoids becoming trapped in local optima, allowing it to reach a solution closer to the optimal value. Performance evaluation showed that, compared to other methods, Simulated Annealing exhibited greater consistency in selecting the most representative variables and optimizing the established criteria.

The analysis of results reveals that different feature selection algorithms identified certain variables as highly relevant. For example:

- Chi-square selected: `number_inpatient` and `number_emergency`.
- Information Gain, OneR, and Random Forest all selected: `number_inpatient` and `discharge_disposition_id`.

- Relief identified: glimepiride and glipizide.
- Simulated annealing selected: Nateglinide, Tolbutamide, and A1Cresult.

This overlap between algorithms in selecting `number_inpatient`, `ischarge_disposition_id` highlights their importance as highly relevant variables that influence hospital readmission. Furthermore, the number of variables selected by each method ranged from 2 to 5, depending on the selection criteria and mechanisms of each algorithm.

Finally, these findings reinforce the value of unsupervised models in clinical data exploration. From a social and economic perspective, they can be used for more efficient and patient-centered decision-making.

References

- [1] International Diabetes Federation (IDF). "International diabetes federation website." idf.org. (accessed Nov. 13, 2025)
- [2] Statista. "Global diabetes percentage." [es.statista.com](https://es.statista.com/grafico/6698/la-expansion-de-la-diabetes/). (accessed Nov. 13, 2025)
- [3] Sociedad Española de Diabetes (SED). "Artificial intelligence for improving diabetes treatment." [revistadiabetes.org](https://www.revistadiabetes.org/tecnologia/el-papel-de-la-inteligencia-artificial-en-el-tratamiento-de-la-diabetes-promesas-y-realidad/). (accessed Nov. 13, 2025)
- [4] B. Strack, J. P. Deshazo, C. Gennings, C. A. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates in patients with diabetes," *Biomedical Engineering Online*, vol. 13, no. 1, pp. 1–22, 2014. 3
- [5] W. Zeng, J. Li, X. He, and T. Xu, "Unsupervised learning for diabetes subgroup identification based on clinical features and glycemic control," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 210, 2021. 3, 13
- [6] M. Huang, L. Zhao, and C. Wang, "Enhanced clinical interpretation of diabetes clustering via pam and pca integration," *Computers in Biology and Medicine*, vol. 145, pp. 105–142, 2022. 3
- [7] P. Li and Y. Zhang, "Optimization of medical data clustering using simulated annealing," *Expert Systems with Applications*, vol. 212, p. 118716, 2023. 4, 14
- [8] C. K. D. J. Clore, John and B. Strack, "Diabetes 130-US Hospitals for Years 1999-2008," 2014, DOI: <https://doi.org/10.24432/C5230J>
- [9] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005. 4
- [10] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strategic management journal*, vol. 17, no. 6, pp. 441–458, 1996. 5

- [11] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. 6
- [12] R. G. Downey and C. V. King, "Missing data in likert ratings: A comparison of replacement methods," *The Journal of general psychology*, vol. 125, no. 2, pp. 175–191, 1998. 10
- [13] M. Suyal and S. Sharma, "A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning," *Journal of Artificial Intelligence and Systems*, vol. 6, pp. 8–95, 2024. 10
- [14] L. Kaufman, "Partitioning around medoids (program pam)," *Finding groups in data*, vol. 344, pp. 68–125, 1990. 11
- [15] P. Romanski, L. Kotthoff, and M. L. Kotthoff, "Package 'fselector'," URL <http://cran.r-project.org/web/packages/FSelector/index.html>, 2013. 11

