Identificación de patrones de recurrencia en atenciones médicas de baja complejidad mediante técnicas de aprendizaje no supervisado: una propuesta para mejorar el acceso a la salud en la selva del Perú

Coello Vilcherrez Ariana
Maestría de Transformación Digital – UNI
Lima, Perú

Guzman Castilla Buddy
Maestría de Transformación Digital – UNI
Lima, Perú

Lazo Saavedra Himbher
Maestría de Transformación Digital – UNI
Lima, Perú

Rojas Granda Danny
Maestría de Transformación Digital – UNI
Lima, Perú

Sandoval Juárez Daniel
Maestría de Transformación Digital – UNI
Lima, Perú

Abstract

Este estudio aplica técnicas de aprendizaje automático no supervisado para identificar patrones de recurrencia en atenciones médicas, a partir del análisis de datos de encuestas anónimas en la región de selva de Perú. Se busca determinar si ciertas especialidades médicas —frecuentemente visitadas por pacientes con características similares— corresponden a casos de baja complejidad que podrían ser gestionados mediante plataformas digitales. El objetivo es proponer alternativas tecnológicas que descongestionen los servicios presenciales, permitiendo al sistema de salud focalizar sus recursos en atenciones de mayor complejidad. Se utilizaron variables como edad, sexo, peso, talla, frecuencia de atención, especialidad y motivo de consulta. Algoritmos como K-means y DBSCAN permitieron segmentar a los pacientes según sus patrones de atención.

INTRODUCCIÓN

El acceso equitativo a la salud es uno de los desafíos más persistentes en regiones geográficamente aisladas del Perú, como la región selvática. Estas zonas, ubicadas en la Amazonía peruana y limítrofes con Colombia, Brasil y Bolivia, presentan gran riqueza natural, biodiversidad y recursos, pero enfrentan limitaciones estructurales graves en la prestación de servicios básicos, particularmente en el sector salud. La dispersión poblacional, las largas distancias entre comunidades, las barreras naturales (ríos, selva densa) y la escasa presencia de infraestructura médica hacen que gran parte de la población tenga acceso limitado, irregular o tardío a servicios sanitarios.

Actualmente, la red de establecimientos de salud en estas regiones se concentra en las capitales departamentales como Puerto Maldonado y Iquitos, con una cobertura escasa en las zonas rurales e indígenas. Muchos pobladores deben recorrer largas horas por vía fluvial o terrestre para llegar al único centro hospitalario de referencia o postas médicas locales, que además suelen estar desabastecidas o contar con personal médico insuficiente. Esta situación se agrava con la falta de conectividad, medios de transporte, y disponibilidad de especialistas.

En este contexto de precariedad, se ha observado que muchas de las atenciones que logran realizarse en los establecimientos disponibles corresponden a **casos de baja complejidad**. Estos incluyen consultas por molestias leves, seguimiento de enfermedades crónicas ya diagnosticadas, consejería nutricional o psicológica, control de peso y renovación de recetas médicas. Este tipo de atenciones, que no requieren procedimientos especializados ni equipamiento avanzado, podría resolverse de manera remota mediante plataformas digitales, evitando así el colapso de los servicios presenciales y optimizando el uso de los recursos humanos y logísticos disponibles.

Se entiende por **casos de baja complejidad** aquellas situaciones clínicas que pueden ser abordadas sin intervención presencial obligatoria ni procedimientos invasivos. Son resolubles a través de seguimiento, orientación, prescripción o control, sin comprometer la seguridad del paciente. Algunos ejemplos incluyen dolores inespecíficos, malestares leves, dudas sobre tratamientos, seguimiento de pacientes crónicos estables, o consejerías preventivas.

El avance de la tecnología digital, particularmente la telemedicina, el uso de apps médicas y la integración de sistemas de información, ofrece una alternativa viable para ampliar la cobertura sanitaria en territorios como

la región selvática. Sin embargo, para que estas soluciones sean efectivas, es necesario comprender con claridad el comportamiento de los pacientes: ¿quiénes consultan con mayor frecuencia?, ¿por qué motivos?, ¿en qué especialidades se concentran las atenciones? y ¿cuáles podrían ser atendidas sin necesidad de presencialidad?

Frente a ello, este estudio propone la aplicación de técnicas de *machine learning* no supervisado, específicamente algoritmos de segmentación como **K-means** y **DBSCAN**, para identificar patrones de recurrencia en las atenciones médicas. A través del análisis de variables como edad, sexo, peso, talla, especialidad, frecuencia y motivo de consulta, se busca clasificar perfiles de pacientes cuyas atenciones podrían ser trasladadas a canales digitales. Esta propuesta no solo pretende descongestionar el sistema de salud presencial en la región selvática, sino también aportar evidencia útil para el diseño de políticas públicas orientadas a una atención más accesible, inclusiva y sostenible.

REVISIÓN DE LITERATURA

Pfeifer et al. (2024) desarrollaron un enfoque innovador de clustering no supervisado utilizando bosques aleatorios en un entorno federado, con el objetivo de facilitar la estratificación de pacientes a partir de datos multi-ómicos sin comprometer la privacidad. El estudio propone una metodología que permite a múltiples instituciones colaborar en el análisis sin necesidad de compartir datos sensibles, mediante la construcción local de modelos y el intercambio de matrices de afinidad.
La herramienta desarrollada no solo demostró un rendimiento competitivo frente a métodos centralizados, sino que también mejoró la interpretabilidad al identificar características relevantes por grupo de pacientes. Validado con conjuntos de datos públicos y clínicos como los del The Cancer Genome Atlas (TCGA), este enfoque representa una contribución significativa a la medicina personalizada, especialmente en contextos donde la confidencialidad de los datos es crítica.

Elhussein & Gursoy (2023) propusieron un marco de *Clustered Federated Learning* que incorpora criptografía para preservar la privacidad en la agrupación de pacientes distribuidos (Privacy-preserving Community-Based Federated Machine Learning – PCBFL). Este enfoque utiliza Secure Multiparty Computation (SMC) para calcular de forma segura las similitudes entre pacientes de diferentes hospitales. Con ello, forman clústeres clínicamente significativos de riesgo (bajo, medio, alto) y luego entrenan modelos federados personalizados en cada clúster. Al evaluar el enfoque en 20 sitios del conjunto de datos eICU, los autores reportan mejoras en la predicción de mortalidad: un aumento de 4.3 % en AUC y 7.8 % en AUPRC, en comparación con métodos federados tradicionales .

Momahhed et al. (2023) llevaron a cabo un estudio en Irán que aplicó el algoritmo K-means para segmentar a casi 200 000 asegurados según sus patrones de prescripción ambulatoria. Utilizando variables como edad, sexo, número de medicamentos, frecuencia de recetas y gastos asociados, se identificaron subgrupos dentro de tres clases de riesgo (bajo, medio y alto). Cada clase fue subdividida en tres clusters distintos, validados mediante coeficientes de silueta y Davies–Bouldin. Los resultados permitieron caracterizar perfiles diferenciados de consumo de medicamentos, lo cual aporta evidencia útil para ajustar políticas de aseguramiento y gestionar riesgos financieros en sistemas de salud. Aunque centrado en prescripciones, este estudio demuestra el valor de las técnicas no supervisadas para identificar patrones de comportamiento en grandes poblaciones de pacientes.

Ruiz-Ramos et al. (2025) aplicaron el algoritmo K-means para identificar perfiles clínicos entre pacientes que acudieron al servicio de urgencias por problemas relacionados con medicamentos (DRP). Utilizando datos de más de 1,600 pacientes en un hospital universitario español, se emplearon variables como el grupo de morbimortalidad ajustada (GMA), la edad y el número de medicamentos al ingreso para conformar seis clústeres clínicamente diferenciados. Los resultados revelaron diferencias significativas en las tasas de reconsulta a urgencias dentro de los 30 días, siendo el grupo con mayor carga farmacológica y GMA más alto el que presentó el mayor riesgo de reingreso (24.5 %). El estudio demuestra la utilidad del aprendizaje no supervisado para

identificar pacientes de alto riesgo y orientar intervenciones clínicas preventivas, constituyéndose en un referente aplicable para el diseño de estrategias similares en otros contextos hospitalarios o de atención ambulatoria.

ESTADO DEL ARTE

El aprendizaje automático no supervisado constituye una rama del machine learning orientada al descubrimiento de estructuras ocultas en datos sin etiquetas previas. A diferencia de los métodos supervisados, que requieren ejemplos con salidas conocidas para entrenar modelos predictivos, el aprendizaje no supervisado se basa en la exploración de regularidades y similitudes internas en los datos para agruparlos, reducir su dimensionalidad o detectar anomalías.

1. Clustering (agrupamiento)

El clustering es una de las técnicas más representativas del aprendizaje no supervisado. Su objetivo es agrupar elementos con características similares en clústeres, de modo que los objetos dentro de un mismo grupo sean más parecidos entre sí que respecto a los de otros grupos.

Dos algoritmos relevantes son:

- K-means: Parte de un número predefinido de clústeres (K) y asigna cada instancia al centroide más cercano. Iterativamente ajusta los centroides para minimizar la varianza intra-cluster. Es eficiente y fácil de implementar, aunque sensible a la elección de K y a valores atípicos.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Forma clústeres basados en la densidad de puntos, permitiendo detectar agrupaciones de forma arbitraria y filtrar ruidos. Es útil cuando los datos no se distribuyen de manera uniforme y hay valores atípicos.

2. Estandarización de datos clínicos

Los modelos de clustering requieren una adecuada preparación de los datos. Variables como edad, sexo, peso, talla, frecuencia de atención o motivo de consulta deben ser normalizadas, codificadas y validadas para asegurar la coherencia analítica. Esta fase incluye la imputación de valores faltantes, la estandarización de unidades y la conversión de datos categóricos.

3. Medidas de validación interna

Evaluar la calidad de los clústeres es clave para garantizar la utilidad del modelo. Entre las métricas más comunes están:

- Índice de Silueta: Mide cuán similar es un punto a su propio clúster frente a otros.
- Índice de Davies–Bouldin: Evalúa la compacidad y separación de los clústeres.

Estas herramientas ayudan a determinar la cantidad óptima de clústeres y a interpretar la coherencia clínica de los grupos formados.

4. Aplicaciones en salud

Los algoritmos no supervisados han sido utilizados en la segmentación de pacientes por riesgo (Momahhed et al., 2023), en la identificación de problemas relacionados con medicamentos (Ruiz-Ramos et al., 2025) y en el análisis federado de datos clínicos sin vulnerar la privacidad (Pfeifer et al., 2024; Elhussein & Gursoy, 2023). Estas experiencias demuestran que es posible generar conocimiento clínico valioso incluso sin supervisión directa, lo cual resulta clave en contextos de baja disponibilidad de etiquetas o diagnósticos confirmados.

5. Atención digital en salud pública

El concepto de atención diferenciada a través de plataformas digitales —telemedicina, chatbots, apps móviles— surge como una solución para zonas rurales o con alta demanda y pocos recursos. La segmentación de pacientes con base en su recurrencia y motivos de consulta puede permitir asignar ciertos grupos a canales virtuales, reservando los recursos presenciales para casos de mayor complejidad.

PLANTEAMIENTO DEL PROBLEMA

Problema a resolver

¿Cómo segmentar a los pacientes que presentan recurrencia en especialidades médicas de baja complejidad, utilizando técnicas no supervisadas, para proponer estrategias de atención virtual en la región selva

Justificación

Las especialidades con alta recurrencia por temas menores representan una oportunidad para intervención digital, reduciendo presión sobre el sistema presencial y mejorando la experiencia del paciente, especialmente en regiones con acceso limitado como la selva peruana.

Objetivo General

Segmentar y caracterizar a los pacientes que asisten reiteradamente a especialidades de baja complejidad en la región selvática, utilizando técnicas de aprendizaje no supervisado, específicamente mediante el modelo de clustering **K-means**, con el fin de proponer soluciones digitales que mejoren la eficiencia del sistema de salud.

Objetivos Específicos

- Analizar la frecuencia de atención por especialidad en la región.
- Identificar agrupaciones de pacientes según sus características clínicas y demográficas.
- Determinar qué especialidades y motivos de atención son más aptos para gestión remota.
- Formular recomendaciones para el uso de plataformas digitales como mecanismo de atención complementario en zonas de difícil acceso.

METODOLOGÍA

Fase	Actividades
Comprensión del negocio	Identificar especialidades críticas y necesidades del sistema de salud en la región selvática.
Comprensión de los datos	Revisión de datos de pacientes: edad, sexo, peso, talla, especialidad, etc.
Preparación de los datos	Limpieza, codificación, estandarización de unidades, imputación de valores.
Modelado	Aplicación de **K-means** para segmentar perfiles de pacientes.
Evaluación	Validación de clusters por coherencia clínica y frecuencia de atención.
Despliegue	Propuesta de atención digital diferenciada por perfil detectado.

VARIABLES DEL ESTUDIO

Variable	Tipo	Descripción
Edad	Numérica	Años del paciente
Sexo	Categórica	Hombre / Mujer
Frecuencia	Numérica	Número de atenciones en un periodo definido
Especialidad	Categórica	Área médica atendida (p. ej. Medicina General)

INFORMACIÓN GENERAL DEL MODELO

- **Algoritmo**: SimpleKMeans
- **Número de Clusters (K)**: 7
- **Distancia**: Euclidiana (weka.core.EuclideanDistance)
- **Iteraciones realizadas**: 4
- **Reemplazo de valores nulos**: con media/moda
- **Número de registros**: 1009
- **Modo de prueba**: 70% entrenamiento, 30% prueba (split aleatorio)

MEDICIÓN Y RESULTADOS

Modelamiento

El modelado del estudio se enfocó en aplicar técnicas de Machine Learning, por lo que se optó el modelo Kmeans, para poder segmentar los pacientes que realizan una cita médica de una especialidad especifica.

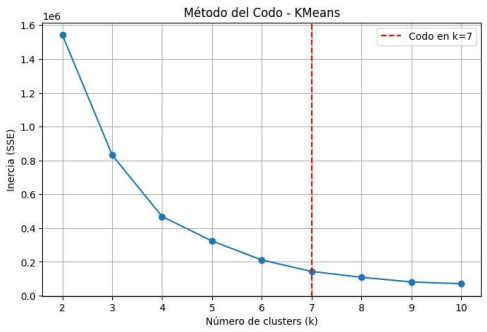
K-means

Es un algoritmo no supervisado de agrupación que divide un conjunto de datos en k grupos o clusters. Los datos se agrupan de tal manera que los puntos en el mismo cluster sean lo más similares.

Selección del Número de Clusters: Aplicando el método de codo

El número de clústeres (k) define la cantidad de grupos en los que se particionan los datos durante el proceso de segmentación. La selección del valor óptimo de k se realiza mediante el método del codo (elbow method), el cual consiste en ejecutar el algoritmo K-means para distintos valores de k y calcular, en cada iteración, la suma de las distancias cuadradas intra-clúster (inertia). A medida que k incrementa, la inertia tiende a disminuir; no obstante, se identifica un punto a partir del cual la disminución marginal es poco significativa. Este punto de inflexión, conocido como "el codo", representa el valor óptimo de k.

En el presente estudio, se aplicó el método del codo considerando un rango de valores de k entre 2 y 10. Los resultados evidenciaron que el punto de inflexión se encontraba en k = 7, lo que indica que la partición de los jefes de hogar en cuatro clústeres constituye una solución balanceada entre complejidad y representatividad. Por tanto, este valor fue adoptado para los análisis subsiguientes.



Evaluación y Despliegue

Una vez determinado el número optimo de clusters se procedió a ejecutar el algoritmo K-means con K=7, cada dataset fue asignado a cada uno de los clusters, se identificó siete:

- Pacientes 1: Son pacientes del sexo femenino, de una edad aproximada de 37 años, cita en la especialidad de ginecología y van con una frecuencia de 10 veces.
- Pacientes 2: Son de sexo masculino, de una edad de 26 años, medicina general y una frecuencia de 5 veces.
- Pacientes 3: Son de sexo masculino, de una edad 21 años, van a enfermería con una frecuencia de una vez.
- Pacientes 4: Son de sexo masculino, de una edad de 20 años, van a la especialidad de enfermería y con una frecuencia de una vez.
- Pacientes 5: Son de sexo femenino, de una edad promedio de 34 años, especialidad de odontología y con una frecuencia de 12 veces.
- Pacientes 6: Son de sexo femenino, de una edad promedio de 29 años, en la especialidad de ginecología y con una frecuencia de 85 veces.
- Pacientes 7: Son de sexo femenino, con una edad promedio de 24 años, en la especialidad de odontología y con una frecuencia de 5 veces.

CONCLUSIONES

El estudio aplicó técnicas de aprendizaje automático no supervisado, específicamente el algoritmo K-means, para analizar datos en la región selvática, con el objetivo de identificar patrones de recurrencia en atenciones médicas de baja complejidad. Los resultados permitieron segmentar a los pacientes en siete grupos diferenciados según variables clínicas y demográficas, lo que facilitó la identificación de grupos con características similares en cuanto a edad, sexo, peso, talla, frecuencia de atención, especialidad y motivo de consulta.

La edad y sexo son comunes en las evaluaciones de las investigaciones de para describir patrones demográficos. La variable frecuencia es para

determinar las visitas realizadas por los pacientes.

El análisis demostró que muchas de las atenciones que actualmente se realizan de manera presencial corresponden a casos de baja complejidad, tales como consultas por molestias leves, seguimiento de enfermedades crónicas estables, consejería nutricional o psicológica, y renovación de recetas médicas. Estas atenciones podrían ser gestionadas eficientemente mediante plataformas digitales, lo que permitiría descongestionar los servicios presenciales y optimizar el uso de los recursos humanos y logísticos disponibles en la región.

La segmentación de pacientes mediante técnicas de clustering no solo aportó evidencia sobre los perfiles de mayor recurrencia, sino que también permitió identificar especialidades y motivos de consulta susceptibles de ser atendidos de manera remota. Esto representa una oportunidad estratégica para el sistema de salud de la región, ya que la implementación de soluciones digitales, como la telemedicina y aplicaciones móviles, podría mejorar el acceso equitativo y sostenible a la atención en salud, especialmente en zonas geográficamente aisladas y con limitaciones estructurales.

En conclusión, el uso de técnicas de aprendizaje no supervisado constituye una herramienta valiosa para la toma de decisiones basada en datos en el sector salud. La identificación de patrones de recurrencia y la caracterización de grupos de pacientes permiten diseñar políticas públicas y estrategias de atención más eficientes, inclusivas y adaptadas a las necesidades reales de la población, contribuyendo así a la mejora del acceso y la calidad de los servicios de salud en la Amazonía peruana.

Para mejora la segmentación se puede incorporar información sobre nivel educativo, ocupación, acceso a Internet y situación socioeconómica, contar con mayor rango de años para las historias clínicas aportaría también en una mayor profundidad e identificación de patrones, también debemos considerar que al ser información sensible se deben considerar en su implementación cumplir con las normativas y regulaciones, asimismo es importante que los resultados sean validados por expertos médicos y del sector salud para potenciar los resultados

REFERENCIAS BIBLIOGRÁFICAS

Elhussein, A., & Gursoy, G. (2023). *Privacy-preserving patient clustering for personalized federated learning*. arXiv. https://arxiv.org/abs/2307.08847

Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Huang, L., & Liu, D. (2019). *Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed Electronic Medical Records.* https://doi.org/10.1016/j.jbi.2019.103291

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: An introduction to cluster analysis. Wiley.

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

Momahhed, S. S., Emamgholipour S., Sefiddashti, S. E., Minaei, B., & Shahali, Z. (2023). *K-means clustering of outpatient prescription claims for health insureds in Iran*. BMC Public Health, 23(1), 788. https://doi.org/10.1186/s12889-023-15753-1

Pfeifer, B., Sirocchi, C., Bloice, M. D., Kreuzthaler, M., & Urschler, M. (2024). *Federated unsupervised random forest for privacy-preserving patient stratification*. arXiv. https://doi.org/10.48550/arXiv.2401.16094

Ruiz-Ramos, J., Plaza-Díaz, A., Puig-Campmany, M., Sampol-Mayol, C., Blázquez-Andión, M., et al. (2025). *K-means clustering to identify high risk of early revisits in patients with drug-related problems attending the emergency department*. European Journal of Hospital Pharmacy. Advance online publication. https://doi.org/10.1136/ejhpharm-2024-004414

Yang, W.-C., Lai, J.-P., Liu, Y.-H., Lin, Y.-L., Hou, H.-P., & Pai, P.-F. (2024). *Using Medical Data and Clustering Techniques for a Smart Healthcare System. Electronics, 13(1), 140.* https://doi.org/10.3390/electronics13010140

Zhao, Y., et al. (2024). *A comparative analysis of clustering algorithms for EHR data.*

J. Innov. Data Sci., 2024. https://hal.science/hal-05037350v1/file/peerj-cs-2286.pdf

Miller AC, Arakkal AT, Koeneman SH, Cavanaugh JE, Polgreen PM. A clinically-guided unsupervised clustering approach to recommend symptoms of disease associated with diagnostic opportunities. Diagnosis (Berl). 2022 Sep 21;10(1):43-53. doi: 10.1515/dx-2022-0044. PMID: 36127310; PMCID: PMC9934811. https://pubmed.ncbi.nlm.nih.gov/36127310/

Bose E, Radhakrishnan K. Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients With Heart Failure Using Telehealth. Comput Inform Nurs. 2018 May;36(5):242-248. doi: 10.1097/CIN.0000000000000423. PMID: 29494361. https://pubmed.ncbi.nlm.nih.gov/29494361/

Miller AC, Arakkal AT, Koeneman SH, Cavanaugh JE, Polgreen PM. A clinically-guided unsupervised clustering approach to recommend symptoms of disease associated with diagnostic opportunities. Diagnosis (Berl). 2022 Sep 21;10(1):43-53. doi: 10.1515/dx-2022-0044. PMID: 36127310; PMCID: PMC9934811. https://pmc.ncbi.nlm.nih.gov/articles/PMC9934811/?utm_source=chatgpt.com

*Hu, Y., Yan, H., Liu, M. et al. Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records. BMC Med Res Methodol **24**, 309 (2024).* https://doi.org/10.1186/s12874-024-02422-z