UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE CIENCIAS QUÍMICAS, FÍSICAS Y MATEMÁTICAS

ESCUELA PROFESIONAL DE MATEMÁTICA CON MENCIÓN EN ESTADÍSTICA

TESIS

ANÁLISIS DE CLÚSTER PARA LA SEGMENTACIÓN DE PACIENTES CON DIABETES MELLITUS TIPO 2 MEDIANTE EL ALGORITMO FUZZY C-MEANS EN EL CUSCO, 2019-2022

PRESENTADO POR:

Br. JESUS RENAN HUACCANQUI CONDORI

PARA OPTAR AL TITULO PROFESIONAL DE LICENCIADO EN MATEMÁTICA MENCIÓN ESTADÍSTICA

ASESOR:

Mtro. ARTURO ZUÑIGA BLANCO

CUSCO – PERÚ

2025

INFORME DE ORIGINALIDAD

(Aprobado por Resolución Nro.CU-303-2020-UNSAAC)

El que suscribe, **Asesor** del trabajo de investigación/tesis titulada: ***ANÁLISIS DE CLÚSTER PARA LA SEGMENTACIÓN DE PACIENTES CON DIABETES MELLITUS TIPO 2 MEDIANTE EL ALGORITMO FUZZY C-MEANS EN EL CUSCO, 2019 - 2022*** .....

.....

Presentado por: ***JESUS RENAN HUACCANQUI CONDORI***..... DNI N° **74310813.** Para optar el título profesional/grado académico de ***LICENCIADO EN MATEMÁTICA MENCIÓN ESTADÍSTICA*** .....

.....

Informo que el trabajo de investigación ha sido sometido a revisión por ...2... veces, mediante el Software Antiplagio, conforme al Art. 6° del ***Reglamento para Uso de Sistema Antiplagio de la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de ...04.....**%.**

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No se considera plagio.	X
Del 11 al 30 %	Devolver al usuario para las correcciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, quien a su vez eleva el informe a la autoridad académica para que tome las acciones correspondientes. Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de asesor, firmo el presente informe en señal de conformidad y **adjunto** las primeras páginas del reporte del Sistema Antiplagio.

Cusco, 13 de Agosto de 20.25

.....

Firma

Post firma..Mtro. Arturo Zuñiga Blanco

Nro. de DNI....46452024.....

ORCID del Asesor.....0000-0002-8576-3415

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema Antiplagio: **oid:** 27259:482815694

JESUS RENAN HUACCANQUI CONDORI

Fuzzy C-Mans implementado a pacientes con diabetes mellitus tipo 2.pdf

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:482815694

Fecha de entrega
13 ago 2025, 8:45 a.m. GMT-5

Fecha de descarga
13 ago 2025, 8:49 a.m. GMT-5

Nombre de archivo
Fuzzy C-Mans implementado a pacientes con diabetes mellitus tipo 2.pdf

Tamaño de archivo
1.4 MB

91 Páginas

24.998 Palabras

129.882 Caracteres

4% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 15 palabras)

Fuentes principales

3%	Fuentes de Internet
1%	Publicaciones
3%	Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

PRESENTACIÓN

SEÑOR DECANO DE LA FACULTAD DE CIENCIAS QUÍMICAS, FÍSICAS Y MATEMÁTICAS,

SEÑOR DIRECTOR DE LA ESCUELA PROFESIONAL DE MATEMÁTICA,

SEÑORES DOCENTES MIEMBROS DEL JURADO:

En cumplimiento de lo establecido en el reglamento de grados y títulos de la Escuela Profesional de Matemática, me es grato presentar ante ustedes el trabajo de tesis titulado: "ANÁLISIS DE CLÚSTER PARA LA SEGMENTACIÓN DE PACIENTES CON DIABETES MELLITUS TIPO 2 MEDIANTE EL ALGORITMO FUZZY C-MEANS EN EL CUSCO, 2019-2022", elaborado con el objetivo de optar al título profesional de Licenciado en Matemática Mención en Estadística.

El presente estudio busca contribuir al análisis y caracterización de pacientes diagnosticados con diabetes mellitus tipo 2, mediante la aplicación del algoritmo de agrupamiento difuso Fuzzy C-Means, que permite identificar y representar de manera más realista los perfiles clínicos de los pacientes, evidenciando una heterogeneidad de las características clínicas presentes en la población diabética de la región Cusco, lo que puede aportar al diseño de estrategias focalizadas de atención a los pacientes diagnosticados con esta enfermedad crónica.

Agradezco de antemano a las autoridades, docentes y miembros del jurado por su tiempo, dedicación y valiosas observaciones que, sin duda enriquecerán esta investigación. Espero que el contenido expuesto sea de utilidad y sirva como base para futuras investigaciones tanto en el campo de la estadística aplicada como en el estudio de enfermedades crónicas. Atentamente,

Br. Jesus Renan Huaccanqui Condori

DEDICATORIA

A mi mamá, Julia, pilar fundamental en mi vida. La que me crio y educó con amor, valentía y sacrificio; quien asumió sola tantas responsabilidades y no se rindió, ni siquiera ante las más duras adversidades.

Tus palabras: “Tú estudia, yo veré qué hago para ayudarte” me acompañaron en cada paso, dándome la fuerza que necesitaba para seguir, incluso cuando sentía que no podía más.

Esta meta también es tuya, porque es el fruto de tu esfuerzo incansable, de tu entrega silenciosa y del amor inmenso que siempre me diste.

Te amo profundamente.

AGRADECIMIENTO

Agradezco profundamente a Dios por brindarme la oportunidad de culminar esta etapa tan importante de mi vida, llena de aprendizajes, desafíos y crecimiento personal.

A mi mamá, ejemplo de esfuerzo, valentía, superación y amor incondicional, gracias por estar siempre a mi lado y por tu apoyo constante. Tu presencia ha sido fundamental en cada etapa de mi vida y siéntete orgullosa, porque lo estoy logrando.

A mis hermanos Natalia, Renaldo, María y Renaldiño, gracias por estar presentes en cada paso de este camino, por sus palabras de aliento, por creer en mí incluso cuando yo mismo dudaba y por ser una fuente constante de motivación. Con ustedes crecí, reí, lloré y aprendí el verdadero valor de la familia. Gracias por acompañarme siempre, por escucharme con paciencia y por impulsarme a seguir adelante. Su apoyo incondicional ha sido una de las mayores fortalezas en mi vida.

A mis tíos, tías, amistades y a todas las personas que, en algún momento, me brindaron su apoyo y palabras de ánimo, incluso antes de mi formación profesional, les expreso mi más sincero agradecimiento.

De manera muy especial, expreso mi más sincero agradecimiento a:

Mtro. Arturo Zuñiga Blanco, mi asesor

Mtra. Carla Patricia Zuñiga Vilca

Dra. Natalie Veronika Rondinel Mendoza

PhD. Walter Quispe Vargas

Por sus valiosas enseñanzas, sus palabras de aliento, su confianza y por el apoyo brindado durante mi formación profesional. Cada uno de ustedes ha contribuido significativamente en este proceso y me siento honrado de haber contado con su guía.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El presente estudio consiste en segmentar clínicamente a pacientes con diabetes mellitus tipo 2, registrados en la DIRESA Cusco durante el periodo 2019-2022, mediante la aplicación del algoritmo de agrupamiento difuso Fuzzy C-Means. Para ello, se depuró una base de datos conformada por 2750 pacientes, seleccionando cinco variables de mayor relevancia clínica: edad, índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y nivel de glicemia. Se excluyeron variables sin utilidad clínica, con altos porcentajes de datos faltantes o con colinealidad elevada. El análisis se desarrolló en el lenguaje de programación R, donde se estandarizaron las variables y se aplicaron índices de validación interna para determinar el número óptimo de clústeres, estableciéndose cuatro segmentos clínicamente diferenciados. Los resultados evidencian la capacidad del algoritmo para capturar la heterogeneidad de la población estudiada, permitiendo identificar tanto perfiles específicos como pacientes con pertenencia difusa entre clústeres, los cuales podrían representan estados transicionales de la enfermedad. Se concluye que el enfoque basado en agrupamiento difuso representa una herramienta estadística eficaz para la segmentación clínica de pacientes, con potencial para orientar intervenciones más focalizadas y contribuir a una mejor planificación sanitaria basada en datos a nivel regional.

Palabras clave: *Diabetes mellitus tipo 2, Fuzzy C-Means, agrupamiento difuso, índices de validez interna.*

ABSTRACT

This study aims to clinically segment patients diagnosed with type 2 diabetes mellitus, registered in DIRESA Cusco between 2019 and 2022, through the application of the fuzzy clustering algorithm Fuzzy C-Means. A database of 2750 patients was cleaned, selecting five clinically relevant variables: age, body mass index (BMI), systolic blood pressure, diastolic blood pressure and blood glucosa level. Variables lacking clinical utility, with a high percentage of missing data or exhibiting strong collinerity were excluded. The analyses was conducted using the R programming language, where variables were standardized and internal validation índices were applied to determine the optimal number of clusters, resulting in four clinically distinct segments. The results demonstrate the ability of the algorithm to capture the heterogeneity of the studied population, allowing the identification of both specific profiles and patients with fuzzy membership between clústeres, which could represent transitional disease states. It is concluded that the fuzzy clustering approach represents an effective statistical tool for clinical segmentation of patients, with potential to guide more targeted interventions and contribute to better data-driven health planning at the regional level.

Keywords: *Type 2 diabetes mellitus, Fuzzy C-Means, fuzzy clustering, internal validity índices.*

INTRODUCCIÓN

La diabetes mellitus tipo 2 (DM2) representa uno de los principales desafíos para la salud pública a nivel mundial y nacional, no solo por su creciente prevalencia, sino también por las complicaciones agudas y crónicas que conlleva. En el Perú, esta enfermedad crónica afecta a una proporción considerable de la población, generando un impacto negativo tanto en la calidad de vida de los pacientes como en la demanda de recursos del sistema de salud.

En este contexto, el presente estudio tiene objetivo principal caracterizar a los pacientes diagnosticados con DM2 en la región Cusco durante el periodo 2019-2022, mediante la implementación del algoritmo de clúster difuso Fuzzy C-Means (FCM). Esta investigación se desarrolla para comprender la heterogeneidad clínica de esta enfermedad crónica, con el fin de identificar perfiles diferenciados de pacientes y contribuir al diseño de estrategias de atención más personalizadas, enfocadas a la prevención y tratamiento según el perfil de riesgo identificado en cada grupo.

El análisis se realizó sobre una base de datos de 2750 pacientes registrados por la Dirección Regional de Salud (DIRESA) Cusco, información obtenida a través del Centro Nacional de Epidemiologia, Prevención y Control de Enfermedades del Perú (CDC Perú). El algoritmo FCM fue seleccionado por su capacidad de generar agrupamientos difusos, permitiendo que cada paciente tenga distintos grados de pertenencia a múltiples clústeres. La implementación de FCM permitió representar con mayor precisión la complejidad clínica de la población estudiada, facilitando una segmentación más flexible y realista.

Para la identificación de perfiles clínicos, se emplearon variables relevantes del estado de salud de los pacientes y se aplicaron índices de validez como Xie-Beni, Fuzzy Simplified Silhouette, entre otros, para determinar el número óptimo de clústeres y evaluar la calidad del

agrupamiento. Uno de los hallazgos más importantes fue la identificación de pacientes con pertenencia difusa a más de un clúster, lo que revela la existencia de perfiles intermedios o transicionales. Estos casos evidencian condiciones clínicas mixtas que podrían requerir una atención más personalizada.

El proceso metodológico comprendido desde la depuración y el análisis exploratorio de los datos hasta la implementación del algoritmo Fuzzy C-Means mediante el lenguaje de programación R, permitió la identificación de los clústeres y una interpretación de las características distintivas de cada grupo de pacientes.

La tesis está estructurada en cinco apartados. El primero presenta el planteamiento del problema, incluyendo la situación problemática, la formulación de los problemas de investigación, los objetivos y la justificación. El segundo aborda el marco teórico y conceptual, explicando los fundamentos del análisis de clúster, el método Fuzzy C-Means y aspectos relevantes sobre la diabetes mellitus tipo 2, también los antecedentes. El tercero expone las hipótesis y variables del estudio. El cuarto describe la metodología, incluyendo el ámbito geográfico, el tipo de estudio y la población. El quinto apartado presenta los resultados obtenidos, desde la depuración de la base de datos y el análisis exploratorio, hasta la implementación del algoritmo FCM y la interpretación de los clústeres generados. Finalmente, se presentan las conclusiones y recomendaciones.

1. PLANTEAMIENTO DEL PROBLEMA

1.1 Situación problemática

La diabetes mellitus, comúnmente conocida como diabetes, es una enfermedad crónica de causas diversas, como factores genéticos, sociales, obesidad, sedentarismo y alimentación (Floreano Solano, Paccha Tamay, Gordillo Quizhpe, & Zambrano Villamar, 2017), se presenta cuando se elevan los niveles de glucosa en la sangre, causado por un deterioro absoluto o relativo de la secreción de Insulina (Naranjo Hernández, 2016). Los principales tipos de diabetes son tres: tipo 1, tipo 2 representando 90% de los casos y la diabetes gestacional (Organización Panamericana de la Salud [OPS], 2023).

A nivel mundial, en 2021 se estimó que 537 millones de adultos (10.6%) de 20 a 79 años padecían diabetes, lo que genero un gasto en salud de 966 mil millones de USD para este grupo de edad; además, se prevé que más de 6.7 millones de este grupo perderán la vida debido a causas relacionadas con la diabetes (International Diabetes Federation [IDF], 2021). El incremento en la prevalencia de la diabetes se debe a factores como el crecimiento y envejecimiento de la población, el incremento de sobrepeso y la obesidad, además, cuando no se atiende adecuadamente, la diabetes puede dar lugar a complicaciones graves como ceguera, insuficiencia renal, enfermedades cardiovasculares, amputación de miembros inferiores, entre otras (Organización Mundial de la Salud [OMS], 2016).

En el Perú, en 2023 se reportaron 37 919 casos de diabetes, de los cuales el 27.2% fueron casos nuevos, el 97.6% correspondían a diabetes de tipo 2, el 0.8% diabetes tipo 1 y el 1.6% a diabetes gestacional, con un 62.4% de los casos pertenecientes al sexo femenino, y los datos fueron recolectados de 1 169 establecimientos de salud (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades(CDC Perú), 2024).

Por otra parte, el 5.5% de la población peruana de 15 años o más ha sido diagnosticada con diabetes, con un porcentaje mayor en las mujeres (6.1%) que en varones (4.9%), y en cuanto a las regiones naturales, la costa presenta el porcentaje más alto (6.8%), mientras que la sierra (3.0%) y en la selva (4.1%) registran porcentajes más bajos; sin embargo, solo el 70.3% de la población diagnostica recibe tratamiento médico en los últimos 12 meses (Instituto Nacional de Estadística e Informática[INEI], 2024).

Por otro lado, en el año 2024, hasta el mes de octubre, se registraron 308 casos de diabetes en el Hospital Regional del Cusco, de los cuales el 22.1% son casos nuevos, y en cuanto a la evaluación nutricional, el 38.3% de los pacientes tienen sobrepeso, el 26.6% tienen obesidad, además, el 9.6% de estos casos resultaron en fallecimientos (Vigilancia Epidemiológica de Diabetes - HRC, 2024).

Además, la Federación Internacional de Diabetes (2021) estima que para 2030 habrá 643 millones de personas con diabetes y que el gasto sanitario total asociado a la enfermedad alcanzará los 1.03 billones de USD, proyecciones se hacen a nivel mundial. La diabetes y sus complicaciones generan grandes pérdidas económicas para los pacientes, sus familias, los sistemas de salud y las economías nacionales debido a gastos médicos y pérdida de ingresos, por lo que es crucial el diagnóstico temprano, el acceso a pruebas diagnósticas básicas en atención primaria y la existencia de sistemas de remisión y seguimiento para pacientes diabéticos (Organización Mundial de la Salud [OMS], 2016)

Ante el impacto de la diabetes, en 2021 la OMS estableció el Pacto Mundial contra la Diabetes (PMD) para mitigar el riesgo de la enfermedad y asegurar el acceso equitativo a tratamientos de calidad, con el objetivo de apoyar a los países en la implementación de programas efectivos de prevención, control y atención, con un enfoque en las poblaciones más vulnerables

(Organización Panamericana de la Salud [OPS], 2021). En este sentido, la diabetes representa un desafío global que exige enfoques innovadores, puesto que los tratamientos estándar son insuficientes (Sugandh, y otros, 2023). Esto implica la realización de investigaciones que permitan comprender mejor la magnitud del problema (Xie, Chan, & Ma, 2018). Además, es fundamental sensibilizar a la población sobre los riesgos de la diabetes, promover hábitos saludables, fortalecer la atención primaria, garantizar el acceso a medicamentos esenciales, fomentar la investigación, recopilar datos sobre la prevalencia y factores de riesgo, y promover la colaboración de todos los sectores para mitigar su impacto y mejorar la calidad de vida de los pacientes (Organización Mundial de la Salud [OMS], 2016). Estas acciones contribuirán no solo a la prevención y el control de la diabetes, sino también a la mejora de la calidad de vida de los pacientes y a la reducción de la carga económica asociada a la enfermedad.

1.2 Formulación del Problema

1.2.1 Problema General

¿Cómo segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 mediante el algoritmo de clúster Fuzzy C-Means en la región Cusco, durante el periodo 2019-2022?

1.2.2 Problemas Específicos

1. ¿Cuál es el número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2, según los indicadores de validación interna aplicados al algoritmo Fuzzy C-Means?
2. ¿Qué diferencias presentan los perfiles clínicos de los segmentos identificados mediante el algoritmo Fuzzy C-Means?
3. ¿Qué características clínicas presentan los pacientes con pertenencia difusa elevada en los segmentos identificados mediante el algoritmo Fuzzy C-Means?

1.3 Justificación de la investigación

La diabetes tipo 2 es una enfermedad compleja que se manifiesta de diversas formas en los pacientes, lo que dificulta la efectividad de las estrategias de tratamiento y prevención de complicaciones. El algoritmo Fuzzy C-Means se presenta como una alternativa para abordar esta variabilidad entre los pacientes, ya que permite una segmentación flexible en la que los pacientes pueden pertenecer simultáneamente a múltiples clústeres con distintos grados de pertenencia. Este enfoque, aun poco explorado en el contexto peruano, facilita la identificación de grupos con características específicas, proporcionando una visión más precisa de los factores de riesgo y patrones de comportamiento. La aplicación de este algoritmo en este estudio contribuirá a mejorar la comprensión teórica de la segmentación de pacientes diabéticos y a fundamentar estrategias de intervención más personalizados y focalizados.

Desde una perspectiva metodológica, esta investigación adopta un enfoque innovador al emplear Fuzzy C-Means para segmentar pacientes con diabetes tipo 2 en la región Cusco. Su capacidad para manejar incertidumbre y datos difusos lo convierte en una herramienta adecuada y eficiente para identificar patrones ocultos en la población diabética, permitiendo una caracterización más realista de los diferentes perfiles de pacientes. Además, el uso de indicadores de validación interna de clustering difuso garantizará la determinación del número óptimo de clústeres, optimizando la segmentación y reflejando de manera más precisa las características clínicas de los pacientes diabéticos.

Desde una perspectiva social, la diabetes tipo 2 no solo afecta significativamente la calidad de vida de los pacientes, sino que también representa una carga creciente para el sistema de salud peruano debido al alto costo de su tratamiento y a las complicaciones asociadas. Una segmentación adecuada de los pacientes permitirá diseñar intervenciones más estratégicas y

personalizadas, facilitando un mejor manejo de la enfermedad y optimizando los recursos sanitarios del sistema de salud. Al identificar clústeres con características y necesidades comunes, este estudio contribuiría a mejorar la atención médica y a fortalecer las estrategias de prevención de las complicaciones que conlleva, con un impacto positivo en la reducción de complicaciones y la mejora de la calidad de vida de los pacientes con diabetes tipo 2 en la región Cusco.

1.4 Objetivos de la investigación

1.4.1 Objetivo general

Caracterizar los segmentos clínicos identificados entre los pacientes con diabetes mellitus tipo 2 mediante el algoritmo de clúster Fuzzy C-Means en la región Cusco, durante el periodo 2019-2022.

1.4.2 Objetivos específicos

1. Determinar el número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2, mediante indicadores de validación interna aplicados al algoritmo Fuzzy C-Means.
2. Comparar los perfiles clínicos de los segmentos identificados mediante el algoritmo Fuzzy C-Means.
3. Analizar las características clínicas de los pacientes con pertenencia difusa elevada en los clústeres identificados mediante el algoritmo Fuzzy C-Means.

2. MARCO TEÓRICO CONCEPTUAL

2.1 Bases teóricas

2.1.1 Clustering

El clustering, o análisis de conglomerados, consiste en un conjunto de métodos estadísticos y computacionales pertenecientes al aprendizaje no supervisado, empleadas para agrupar un conjunto de datos en subgrupos denominados clústeres, donde las observaciones dentro de cada clúster presentan una mayor similitud entre sí que con las de otros clústeres; esta similitud se evalúa mediante métricas de distancia, medidas de correlaciones o de densidad, según el tipo de datos y el objetivo de análisis (James et al., 2013).

Dado un conjunto de datos con m observaciones, $X = \{x_1, x_2, \dots, x_m\}$, donde cada observación x_i está representada por un vector de dimensión n, es decir, $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, el objetivo del clustering es particionar X en k subconjuntos o clústeres $C_1, C_2, \dots, C_k$, de manera que los elementos dentro de cada clúster C_i presenten la máxima homogeneidad interna y la máxima heterogeneidad entre grupos; para ello, existen dos enfoques principales de particionamiento en clustering (Rokach, 2024):

1. Particionamiento duro: Se define como un conjunto de clústeres $P = \{C_1, C_2, \dots, C_k\}$, tal que:

$$P = \bigcup_{i=1}^{k} C_i, \quad C_i \cap C_i = \emptyset \ \ \forall i \neq j \tag{1}$$

   Esto implica que cada observación pertenece exclusivamente a un único clúster.

2. Particionamiento suave: Se representa mediante una matriz de pertenencia $U \in \mathbb{R}^{mxk}$, donde cada elemento u_{ij} indica el grado de pertenencia de la observación x_i al clúster C_j, con:

$$u_{ij} \in [0,1] \ y \ \sum_{j=1}^{c} u_{ij} = 1, \ \forall i, j \qquad (2)$$

En este caso, una observación puede pertenecer simultáneamente a varios clústeres con diferentes grados de pertenencia.

El clustering es una técnica de enfoque no paramétrico que agrupa observaciones similares sin requerir suposiciones sobre la distribución de los datos, permitiendo identificar patrones, sin necesidad de conocer previamente la existencia o el número de grupos, sin embargo, existen algunas técnicas que pueden ayudar a estimar el número ideal de clústeres (Zelterman, 2015).

Por otro lado, Pérez López (2004) distingue dos tipos principales de clustering:

- Jerárquico: Los clústeres se organizan de forma anidada en una estructura jerárquica similar a un árbol. Se clasifica en:
    - Aglomerativo: Comienza con cada observación como un clúster independiente y que fusiona los clústeres con menor disimilitud hasta formar un único grupo.
    - Divisivo: Inicia con un solo clúster que agrupa todos los datos y se divide iterativamente en subgrupos según la disimilitud entre ellos.
- Particional o no jerárquico: Los clústeres son independientes entre sí y no siguen una estructura jerárquica. Se clasifica en:
    - Disjunto: Cada observación pertenece exclusivamente a un único clúster.
    - Solapado: Una observación puede pertenecer a varios clústeres con diferentes grados de pertenencia.

2.1.2 Medidas de similitud y disimilaridad

Las medidas de similitud y disimilaridad (también conocidas como medidas de distancia) permiten cuantificar el grado de semejanza o diferencia entre observaciones dentro de un conjunto de datos, donde la similitud indica cuán parecidas son dos observaciones, mientras que la disimilaridad mide que tan diferentes son entre sí (Tan et al., 2019).

Medidas de distancia. Sea un conjunto de datos con m observaciones $X = \{x_1, x_2, \dots, x_m\}$, donde cada observación x_i es un vector en un espacio de dimensión n, es decir:

$$x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$$

La distancia entre dos observaciones x_i y x_j se denota como $d(x_i, x_j)$ y es una función de los valores de sus respectivas variables. Para que $d(x_i, x_j)$ sea considerada una métrica de distancia, debe cumplir con las siguientes propiedades:

1. Positividad:

$$d(x_i, x_j) \geq 0, \qquad \forall x_i, x_j \in X \tag{3}$$

$$d(x_i, x_j) = 0 \iff x_i = x_j \tag{4}$$

2. Simetría:

$$d(x_i, x_j) = d(x_j, x_i), \qquad \forall x_i, x_j \in X \tag{5}$$

3. Desigualdad triangular:

$$d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k), \qquad \forall x_i, x_j, x_k \in X \tag{6}$$

Medidas de similitud. A diferencia de las métricas de distancia, las funciones de similitud no suelen cumplir la desigualdad triangular, aunque generalmente satisfacen la positividad y simetría. La similitud entre dos observaciones x_i y x_j se denota como $s(x_i, x_j)$ y debe cumplir con las propiedades:

$$s(x_i, x_j) = 1 \iff x_i = x_j, con\ 0 \leq s(x_i, x_j) \leq 1 \quad (7)$$

$$s(x_i, x_j) = s(x_j, x_i), \ \forall\ x_i, x_j \in X \quad (8)$$

Una medida de disimilaridad toma valores mayores conforme las observaciones son más diferentes entre sí, mientras que una medida de similitud toma valores más altos cuanto mayor sea la semejanza entre observaciones (Zelterman, 2015).

En este estudio, se emplean exclusivamente medidas de distancia para cuantificar la disimilaridad entre observaciones. Esta decisión se fundamenta en la naturaleza del algoritmo de clúster implementado, el Fuzzy C-Means (FCM), el cual requiere de una métrica de distancia para estimar el grado de pertenencia de cada observación a los distintos clusteres.

Matriz de distancias. También denominada matriz de disimilitud, es una matriz cuadrada que contiene las distancias entre cada par de observaciones dentro de un conjunto de datos, proporcionando una representación estructurada que permite cuantificar y analizar las diferencias entre observaciones (Abonyi & Feil, 2007). Se representa como:

$$M = \begin{bmatrix} d(x_1, x_1) & d(x_1, x_2) & \dots & d(x_1, x_m) \\ d(x_2, x_1) & d(x_2, x_2) & \dots & d(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_m, x_1) & d(x_m, x_2) & \dots & d(x_m, x_m) \end{bmatrix}$$

Donde cada elemento $d(x_i, x_j)$ indica la distancia entre las observaciones x_i y x_j, cumpliendo con las siguientes propiedades:

$$d(x_i, x_i) = 0, \qquad \forall\ x_i \in X \quad (9)$$

$$d(x_i, x_j) = d(x_j, x_i), \qquad \forall\ x_i, x_j \in X \quad (10)$$

Medidas de distancia para variables numéricos. Dado que el algoritmo Fuzzy C-Means calcula los grados de pertenencia de cada observación en función de su proximidad a los centroides difusos, es indispensable que las variables sean de tipo numérico, es decir, que se encuentren en

escala de intervalo o de razón. Esto se debe a que el cálculo de distancias requiere operaciones algebraicas y métricas bien definidas sobre los vectores de características. A continuación, se presentan las métricas de distancia más utilizadas para este tipo de variables (Aldás & Uriel, 2017):

1. Distancia Euclidiana. Es la distancia ordinaria entre dos puntos en el espacio $\mathbb{R}^n$. Es apropiada cuando las variables están en la misma escala y no presentan correlaciones importantes.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (11)$$

   Es la métrica más utilizada en algoritmos como FCM debido a su simplicidad y eficiencia computacional.

2. Distancia de Manhattan. Suma las diferencias absolutas entre los valores de las variables. Es menos sensible a valores extremos, adecuada cuando las variables son independientes entre sí y especialmente útil en espacios de alta dimensión.

$$d(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}| \quad (12)$$

3. Distancia de Chebyshev. Considera únicamente la mayor diferencia absoluta entre variables. Es útil cuando se desea priorizar la variable con mayor discrepancia.

$$d(x_i, x_j) = max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \dots, |x_{in} - x_{jn}|) \quad (13)$$

4. Distancia de Minkowski. Generaliza las métricas anteriores y permite ajustar la manera en que se agregan las diferencias entre variables mediante un parámetro $g \geq 1$, lo que ofrece flexibilidad para adaptarse a distintas configuraciones del espacio de datos. Se define como:

$$d(x_i, x_j) = \sqrt[g]{|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \cdots + |x_{in} - x_{jn}|^g} \quad (14)$$

Donde el parámetro g controla el grado de penalización frente a grandes discrepancias entre los valores de las distintas variables.

- Si $g = 1$, se obtiene la distancia de Manhattan.
- Si $g = 2$, se obtiene la distancia euclidiana.
- Si $g \to \infty$, se obtiene la distancia de Chebyshev.

5. Distancia de Mahalanobis. Esta métrica tiene en cuenta la correlación entre variables y la varianza entre cada variable, lo que hace especialmente adecuada en contextos donde las variables presentan diferentes escalas o están correlacionadas.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1}(x_i - x_j)} \tag{15}$$

donde S es la matriz de covarianza del conjunto de datos. Esta distancia es especialmente útil en análisis multivariado, ya que transforma el espacio de variables para que todas contribuyan equitativamente al cálculo de disimilaridad.

2.1.3 Fuzzy Clustering

El agrupamiento difuso o fuzzy clustering es un enfoque de análisis de clúster fundamentado en la teoría de conjuntos difusos, el cual permite que una observación pertenezca simultáneamente a múltiples clústeres con distintos grados de pertenencia, expresados mediante valores en el intervalo $[0,1]$, siendo especialmente útil cuando los límites entre clústeres no están claramente definidos o presentan solapamientos, proporcionando una representación más flexible y realista de la estructura de los datos (Tan et al., 2019).

Conjuntos difusos (Fuzzy Sets). Sea X un conjunto universo no vacío. Un conjunto difuso A, definido sobre X, es una generalización de los conjuntos clásicos propuesta por Zadeh (1965). Se caracteriza por una función de pertenencia $f_A: X \to [0,1]$ que asigna a cada elemento $x \in X$ un valor real en el intervalo $[0,1]$, el cual representa el grado de pertenencia del elemento x al conjunto

difuso A. A diferencia de los conjuntos clásicos, donde la pertenencia es dicotómica (0 o 1), en los conjuntos difusos se permite pertenencia parcial, de modo que:

- $f_A(x) = 1$ indica que x pertenece completamente a A.
- $f_A(x) = 0$ indica que x no pertenece a A.
- $0 < f_A(x) < 1$ indica que x pertenece parcialmente a A con un grado proporcional.

Partición difusa (Fuzzy partition). Una partición difusa de un conjunto de datos $X = \{x_1, x_2, \dots, x_m\}$, se representa mediante una matriz de pertenencia difusa $U = [u_{ij}] \in \mathbb{R}^{mxk}$, donde cada elemento u_{ij} indica el grado de pertenencia de la observación x_i al clúster C_j, para que la matriz U sea considerada una partición difusa valida, debe cumplir las siguientes condiciones (Abonyi & Feil, 2007):

1. Las observaciones tienen una pertenencia parcial acotada a los clústeres:

$$0 \leq u_{ij} \leq 1, \qquad i = 1, \dots, m, \quad j = 1, \dots, k \tag{16}$$

2. Cada observación pertenece completamente a alguna combinación de clústeres, distribuyendo su pertenencia total entre clústeres:

$$\sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \dots, m \tag{17}$$

3. Cada clúster debe tener al menos una observación con grado de pertenencia positivo y no puede contener a todas las observaciones con pertenencia total:

$$0 < \sum_{i=1}^{m} u_{ij} < m, \; j = 1, \dots, k \tag{18}$$

Espacio de partición difusa. Es el conjunto de todas las matrices de pertenencia U que cumplen con los requisitos de una partición difusa valida. Dado un conjunto de datos $X = \{x_1, x_2, \dots, x_m\}$ y un numero de clústeres k, donde $2 \leq k \leq m$, se define como:

$$M_s = \left\{ U = \left[u_{ij}\right] \in \mathbb{R}^{mxk} / u_{ij} \in [0,1]\ \forall i,j; \sum_{i=1}^{k} u_{ij} = 1, \forall i; 0 < \sum_{i=1}^{m} u_{ij} < m, \forall j \right\} \quad (19)$$

Este conjunto M_s garantiza que:

- Cada grado de pertenencia u_{ij} está acotado en el intervalo $[0,1]$.
- La suma de pertenencias para cada observación a los k clústeres sea igual a 1.
- Cada clúster tenga al menos una observación parcialmente asignada, pero no todas.

2.1.4 Fuzzy C-Means (FCM)

El algoritmo Fuzzy C-Means (FCM) es una técnica de agrupamiento difuso que extiende el algoritmo clásico k-means mediante la incorporación de pertenencias parciales, en la cual cada observación puede asociarse simultáneamente a múltiples clústeres con distintos grados de pertenencia representados por valores en el intervalo $[0,1]$ y cuyo fundamento se basa en la minimización de una función objetivo que pondera las distancias cuadráticas generalizadas entre las observaciones y los centroides de los clústeres, según sus respectivos grados de pertenencia elevados a una potencia de difusividad (Rokach, 2024).

Función objetivo del algoritmo FCM. El algoritmo FCM, busca minimizar la siguiente función objetivo (Bezdek, Ehrlich, & Full, 1984):

$$J_\alpha(U,V) = \sum_{i=1}^{m}\sum_{j=1}^{k} \left(\mu_{ij}\right)^{\alpha} \left\|x_i - v_j\right\|_A^2 \quad (20)$$

Donde:

- $U = \left[u_{ij}\right] \in \mathbb{R}^{mxk}$ es la matriz de partición difusa, cuyos elementos $u_{ij} \in [0,1]$ representan el grado de pertenencia de la observación x_i al clúster C_j.

- $V = [v_1, v_2 \dots, v_k]^T$ con $v_j \in \mathbb{R}^n$, representa los centroides (o prototipos) de los clústeres C_j.
- $\|x_i - v_j\|_A^2 = (x_i - v_j)^T A(x_i - v_j)$ es la distancia cuadrática generalizada entre la observación x_i y el centroide v_j, inducida por una matriz de pesos $A \in \mathbb{R}^{n \times n}$, simétrica y definida positiva. Si $A = I$, se reduce a la distancia euclidiana al cuadrado.
- $\alpha \in [1, \infty)$ es el parámetro de difusividad o fuzzificación, que regula el grado de solapamiento entre clústeres.

Optimización de la función objetivo de FCM. El objetivo del algoritmo Fuzzy C-Means (FCM) es encontrar los valores óptimos de la matriz de pertenencias difusas U y de la matriz de centroides V que minimicen la función objetivo $J_\alpha(U, V)$. Para ello, se derivan las expresiones parciales de J_α con respecto a U y V, luego se igualan a cero para encontrar los puntos críticos donde la función alcanza su mínimo, lo cual permite obtener las fórmulas de actualización tanto para los centroides como para los grados de pertenencia (Wierzchoń & Kłopotek, 2018).

Para estimar los centroides v_j difusos, se deriva la función objetivo respecto a cada v_j, manteniendo fija la matriz de pertenencias $U = [u_{ij}]$:

$$\frac{\partial}{\partial v_j} J_\alpha(U, V) = \frac{\partial}{\partial v_j} \sum_{i=1}^{m} \sum_{j=1}^{k} (\mu_{ij})^\alpha \|x_i - v_j\|_A^2 \tag{21}$$

Donde:

$$\|x_i - v_j\|_A^2 = (x_i - v_j)^T A(x_i - v_j) \tag{22}$$

La derivada parcial de esta distancia (22) con respecto a v_j es:

$$\frac{\partial}{\partial v_j}\left\|x_i - v_j\right\|_A^2 = -2A(x_i - v_j) \tag{23}$$

Sustituyendo (23) en la expresión (21), tenemos:

$$\frac{\partial}{\partial v_j}J_\alpha(U,V) = \sum_{i=1}^{m}\left(\mu_{ij}\right)^\alpha(-2A\left(x_i - v_j\right)) = -2A\sum_{i=1}^{m}\left(\mu_{ij}\right)^\alpha\left(x_i - v_j\right) \tag{24}$$

Igualando (24) a cero para obtener el mínimo:

$$\sum_{i=1}^{m}\left(\mu_{ij}\right)^\alpha\left(x_i - v_j\right) = 0 \tag{25}$$

Finalmente, despejando v_j:

$$v_j = \frac{\sum_{i=1}^{m}\left(\mu_{ij}\right)^\alpha x_i}{\sum_{i=1}^{m}\left(\mu_{ij}\right)^\alpha} \tag{26}$$

Este resultado muestra que el centroide de cada clúster v_j es una media ponderada de las observaciones, donde los pesos son los grados de pertenencia μ_{ij} elevados a la potencia α. Esto permite una representación más flexible de los clústeres, especialmente útil en situaciones donde las fronteras entre ellos son difusas.

Para determinar los grados de pertenencias u_{ij} que conforman la matriz de pertenencia difusa U, se utiliza el método de multiplicadores de Lagrange, bajo la restricción de que la suma de los grados de pertenencia de cada observación x_i a todos los clústeres debe ser igual a uno:

$$\sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \dots, m \tag{27}$$

Se define la función de Lagrange como:

$$L(J_\alpha, \lambda) = \sum_{i=1}^{m}\sum_{j=1}^{k}(\mu_{ij})^{\alpha}\|x_i - v_j\|_A^2 - \sum_{i=1}^{m}\lambda_i\left(\sum_{j=1}^{k}(\mu_{ij} - 1)\right) \tag{28}$$

Derivando (28) respecto a μ_{ij} e igualando a cero:

$$\frac{\partial}{\partial \mu_{ij}} L(J_\alpha, \lambda) = \alpha(\mu_{ij})^{\alpha-1}\|x_i - v_j\|_A^2 - \lambda_i = 0 \tag{29}$$

Despejamos λ_i a partir de (29):

$$\lambda_i = \alpha(\mu_{ij})^{\alpha-1}\|x_i - v_j\|_A^2 \tag{30}$$

Como esta relación es válida para cualquier j, sumando sobre j en (19) se obtiene:

$$\lambda_i = \sum_{l=1}^{k}\alpha(\mu_{ij})^{\alpha-1}\|x_i - v_l\|_A^2 \tag{31}$$

Sustituyendo (31) en (29):

$$\alpha(\mu_{ij})^{\alpha-1}\|x_i - v_j\|_A^2 = \sum_{l=1}^{k}\alpha(\mu_{il})^{\alpha-1}\|x_i - v_l\|_A^2 \tag{32}$$

Dividiendo ambos lados de (32) por $\alpha\|x_i - v_j\|_A^2$, se tiene:

$$(\mu_{ij})^{\alpha-1} = \frac{\|x_i - v_j\|_A^{-2}}{\sum_{l=1}^{k}\|x_i - v_l\|_A^{-2}} \tag{33}$$

Finalmente, elevando (33) a la potencia $\frac{-1}{\alpha-1}$, se obtiene:

$$\mu_{ij} = \frac{\|x_i - v_j\|_A^{\frac{-2}{(\alpha-1)}}}{\sum_{l=1}^{k}\|x_i - v_l\|_A^{\frac{-2}{(\alpha-1)}}} \tag{34}$$

Este resultado muestra que la pertenencia μ_{ij} de una observación x_i al clúster j depende inversamente de la distancia (en norma A) entre x_i y el centroide v_j, ponderada por el parámetro de difusividad α. A medida que α aumenta, se incrementa la dispersión de las pertenencias,

permitiendo un mayor solapamiento entre clústeres; en cambio, cuando $\alpha \to 1$, la asignación se vuelve más nítida, aproximándose al comportamiento del método k-means.

Parámetros del algoritmo FCM. Antes de iniciar el proceso de agrupamiento mediante el algoritmo Fuzzy C-Means (FCM), es necesario especificar varios parámetros fundamentales que afectan directamente el desempeño y la calidad de la partición obtenida:

Número de clústeres (k)**.** Es uno de los parámetros más relevantes en el algoritmo FCM; sin embargo, su valor óptimo no suele conocerse a priori y puede variar según la estructura interna del conjunto de datos. Por esta razón, se utilizan diversos métodos para determinarlo, siendo una estrategia común generar múltiples particiones para distintos valores de k y seleccionar aquella que optimice un determinado criterio de validez (Höppner et al., 1999).

Un método frecuentemente utilizado para determinar k es el método del codo (Rokach, 2024), que consiste en calcular el valor de la función objetivo del algoritmo FCM para distintos valores de k, graficar la relación entre k y el valor de la función objetivo, y seleccionar el valor de k en el que la disminución de la función objetivo se vuelve menos pronunciada, lo que sugiere que agregar más clústeres no proporciona mejoras significativas a la calidad del agrupamiento.

Exponente de ponderación (α)**.** También conocido como parámetro de difusividad o fuzzificación, este valor controla el grado de solapamiento entre clústeres. Según Kaushik & Hemanta (2013), cuando $\alpha \to 1$, la partición tiende a ser más rígida o crisp, con valores de pertenencia $\mu_{ij} \in \{0,1\}$, es decir, cada observación se asigna exclusivamente a un único clúster, mientras que cuando $\alpha \to \infty$, la partición se vuelve completamente difusa, donde $\mu_{ij} = \frac{1}{k}$, lo que implica una pertenencia equitativa de cada observación a todos los clústeres, siendo la elección del valor óptimo de α experimental y recomendándose un rango de $1.5 < \alpha < 3$.

Norma inducida por una matriz. La distancia entre una observación x_i y un centroide v_j se mide mediante una norma cuadrática generalizada definida como (Celebi, 2015):

$$\|x_i - v_j\|_A^2 = (x_i - v_j)^T A (x_i - v_j) \tag{35}$$

donde $A \in \mathbb{R}^{n \times n}$ es una matriz simétrica ($A = A^T$) y definida positiva ($x^T A x > 0, \forall x \neq 0$), la cual determina la métrica de distancia utilizada, por tanto, influye en la forma geométrica de los clústeres, permitiendo adaptarse a formas geométricas esféricas, elipsoidales o más complejas, lo cual mejora la flexibilidad del algoritmo de agrupamiento para capturar estructuras subyacentes de los datos (Abonyi & Feil, 2007).

Tolerancia de terminación (ϵ)**.** Este parámetro establece el criterio de convergencia del algoritmo, para finalizar el proceso iterativo cuando la norma de la diferencia entre las matrices de pertenencia obtenidas en dos iteraciones consecutivas es menor a un umbral preestablecido ϵ (Bezdek, Ehrlich, & Full, 1984).

$$\left\|U^{(b)} - U^{(b+1)}\right\| < \epsilon \tag{36}$$

Generalmente se utiliza $\epsilon = 0.001$, aunque un valor de $\epsilon = 0.01$ también resulta efectivo en muchos casos, reduciendo el tiempo computacional sin afectar significativamente la calidad del agrupamiento obtenido (Kaushik & Hemanta, 2013).

Procedimiento del algoritmo FCM. El algoritmo Fuzzy C-Means (FCM) optimiza de manera iterativa la función objetivo $J_\alpha(U, V)$ para obtener una partición difusa del conjunto de datos $X = \{x_1, x_2, \dots, x_m\}$, siguiendo los pasos descritos a continuación (Cannon, Dave, & Bezdek, 1986):

1. Inicialización: Se establece una matriz de pertenencias inicial $U^{(0)}$ de forma aleatoria, cumpliendo la condición:

$$\sum_{i=1}^{k} u_{ik} = 1, \quad \forall k \tag{37}$$

2. Iteración. El proceso iterativo se repite hasta que se cumpla el criterio de convergencia. En cada iteración $b = 0,1,2,...$, se realizan los siguientes pasos:

   - Actualización de los centroides. Se actualizan los centroides $v_i^{(b)}$ de los clústeres utilizando la matriz de pertenencia obtenida en la iteración previa:

$$v_i^{(b)} = \frac{\sum_{k=1}^{n} \left(u_{ik}^{(b)}\right)^m x_k}{\sum_{k=1}^{n} \left(u_{ik}^{(b)}\right)^m}, \forall i \tag{38}$$

     Este cálculo pondera cada observación x_k de acuerdo con su grado de partencia al clúster i.

   - Actualización de los grados de pertenencia. Se recalculan los valores de la matriz de pertenecía $u_{ik}^{(b+1)}$ mediante la siguiente expresión:

$$u_{ik}^{(b+1)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}, \forall i, k \tag{39}$$

     donde d_{ik} representa la distancia entre la observación x_k y el centroide v_i. Esta actualización garantiza que cada observación conserve una distribución de pertenencia total igual a 1 entre todos los clústeres.

3. Criterio de convergencia. Se evalúa si el cambio entre iteraciones consecutivas es inferior al umbral de tolerancia ϵ; para ello, se compara la matriz de pertenencia actual $U^{(b+1)}$ con la de la iteración anterior $U^{(b)}$ utilizando una norma matricial:

$$\left\| U^{(b)} - U^{(b+1)} \right\| < \epsilon \tag{40}$$

Si se cumple esta condición, el algoritmo se detiene y se considera que la partición ha alcanzado la convergencia. En caso contrario, se incrementa el número de iteración b y el proceso se repite hasta satisfacer el criterio.

2.1.5 Validación de Clustering

La validación de clustering es el proceso mediante el cual se evalúa si una partición generada representa de manera adecuada la estructura interna del conjunto de datos, con el objetivo de identificar aquella que refleje de forma más precisa los patrones o relaciones subyacentes entre las observaciones (Abonyi & Feil, 2007). Existen dos enfoques principales para validar los resultados del agrupamiento (Reddy & Aggarwal, 2014):

1. Validación externa del agrupamiento. Consiste en comparar los clústeres obtenidos con una estructura de referencia conocida, generalmente mediante etiquetas de clase previamente definidas, para medir en qué medida la partición generada coincide con una agrupación verdadera de las observaciones, asumiendo que se conoce de antemano el número y la composición real de los clústeres.

2. Validación interna del agrupamiento. Se basa únicamente en las características del propio conjunto de datos sin utilizar información externa y evalúa la calidad de la partición considerando criterios como la compacidad de clústeres, es decir, que tan agrupados están los elementos dentro de un mismo grupo, y la separación entre ellos, entendida como el grado de diferenciación entre los distintos grupos.

Dado que el número de clústeres k, generalmente es desconocido y puede variar según la naturaleza de los datos, se recurre a diferentes métodos, siendo una estrategia frecuentemente utilizada generar múltiples particiones para diferentes valores de k y luego seleccionar la que proporcione el mejor resultado según un criterio de calidad especifico (Höppner et al., 1999).

En el contexto del clústering difuso, la evaluación de la calidad de las particiones generadas se fundamenta en diversas métricas que consideran aspectos clave como la separación entre clústeres, definida como la distancia cuadrada mínima entre los centros de los grupos y la compacidad intraclúster, entendida como la distancia cuadrada media entre cada observación y el centro del clúster al que pertenece (Reddy & Aggarwal, 2014). Además, se pueden emplear medidas complementarias de calidad de partición, clasificadas en dos grandes categorías: medidas de difusividad y medidas de compacidad /separación (Giordani , Ferraro, & Martella, 2020).

Medidas de difusividad. Evalúan el grado de precisión con que se asignan las observaciones a los clústeres, resumiendo la información contenida en la matriz de pertenencia en un solo valor. Entre las más utilizadas destacan:

Coeficiente de Partición (PC). Propuesto por Bezdek (1973), se define como:

$$PC = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(u_{ij}\right)^2 \tag{41}$$

Sus valores se encuentran en el intervalo $\left[\frac{1}{k}, 1\right]$ y se maximiza para determinar el número óptimo de clústeres.

Entropía de Partición (PE). También propuesto por Bezdek (1973), se expresa como:

$$PE = -\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\mu_{ij}\log\left(\mu_{ij}\right)}{n} \tag{42}$$

Con valores en el intervalo $[0, \log k]$. Se minimiza para encontrar la mejor partición.

Coeficiente Modificado de Partición (MPC). Introducido por Dave (1996), corrige la dependencia del PC respecto al numero de clusteres k:

$$MPC = 1 - \frac{k}{k-1}(1 - PC) \tag{43}$$

Sus valores están en $[0,1]$ y se maximiza para identificar el valor óptimo de clústeres.

Medidas de compacidad y separación. Consideran simultáneamente la cohesión interna de los clústeres y su diferenciación respecto a los demás. Entre ellas destacan los siguientes:

Índice Xie-Beni (XB). Evalúa la relación entre la dispersión intraclúster y la distancia mínima interclúster (Xie & Beni, 1991):

$$V_{XB} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(\mu_{ij})^{\alpha}\|x_i - v_j\|_A^2}{n \,.\, min_{l\neq s}\|v_l - v_s\|_A^2} \tag{44}$$

El numerador indica la dispersión dentro de los grupos y es equivalente a la función objetivo $J_\alpha(U,V)$ normalizada por el número total de puntos. El denominador mide la separación entre los prototipos de los clústeres. Un valor bajo de V_{XB} indica una mejor partición, ya que sugiere alta compactación y una buena separación entre los grupos.

Fuzzy Simplified Silhouette (FSS). El índice FSS es una extensión difusa del coeficiente de silueta, propuesto por Campello y Hruschka (2006). Se define como:

$$V_{FSS} = \frac{\sum_{i=1}^{m}(\mu_{pi} - \mu_{qi})^{\alpha} s_i}{\sum_{i=1}^{m}(\mu_{pi} - \mu_{qi})^{\alpha}} \tag{45}$$

donde μ_{pi} y μ_{qi} son las dos mayores pertenencias de x_i y s_i es el coeficiente de silueta definido por:

$$s_i = \frac{b_{pi} - a_{pi}}{max\{a_{pi}, b_{pi}\}} \tag{46}$$

siendo a_{pi} es la distancia entre x_i y su centroide más cercano v_p, y b_{pi} la distancia entre x_i y el segundo centroide más cercano. Valores más altos de V_{FSS}, indican mayor calidad del agrupamiento.

Índice de Kwon (K). Corrige el comportamiento del índice XB ante un número creciente de clústeres (Kwon, 1998):

$$V_K = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(\mu_{ij})^{\alpha}\|x_i - v_j\|_A^2 + \frac{1}{k}\sum_{j=1}^{k}\|v_j - \bar{x}\|_A^2}{min_{l\neq s}\|v_l - v_s\|_A^2} \tag{47}$$

donde $\bar{x}$ es el centroide global del conjunto de datos. El primer término del numerador evalúa la similitud intra-clúster, mientras que el segundo introduce un término de penalización que compensa la tendencia del índice a favorecer particiones con un número excesivo de grupos. Se minimiza para obtener particiones compactas y bien separadas.

Índice de Tang-Sun-Sun (TSS). Ajusta el índice XB incorporando un término adicional de penalización que mejora su estabilidad frente al aumento de k (Tang, Sun, & Sun, 2005):

$$V_{TSS} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(\mu_{ij})^{\alpha}\|x_i - v_j\|_A^2 + \frac{1}{k(k-1)}\sum_{j=1}^{k}\sum_{l=1,l\neq j}^{k}\|v_l - v_s\|_A^2}{min_{l\neq s}\|v_l - v_s\|_A^2 + \frac{1}{k}} \tag{48}$$

El primer término en el numerador mide la similitud dentro de los grupos, mientras que el segundo termino es un factor de penalización y el denominador mide la separación de los grupos. Este índice también se minimiza, siendo preferible valores bajos que reflejen una estructura de clústeres bien definida.

2.2 Marco conceptual

2.2.1 Diabetes Mellitus (DM)

La diabetes mellitus (DM) es una enfermedad crónica caracterizada por hiperglucemia, resultado de una producción insuficiente de insulina o de la incapacidad del organismo para utilizarla eficazmente, que si no se trata de forma adecuada puede causar daños en diversos órganos y complicaciones graves, aunque con un tratamiento oportuno es posible controlar la enfermedad y prevenir sus consecuencias (Organización Panamericana de la Salud [OPS], 2023).

2.2.2 Tipos de diabetes

Según la Federación Internacional de Diabetes (IDF, 2021), existen tres tipos principales:

Diabetes mellitus tipo 1 (DM1). Es una enfermedad autoinmune en la que el sistema inmunológico destruye las células beta del páncreas, responsables de la producción de insulina, por lo que las personas que la padecen requieren inyecciones diarias de esta hormona para mantener niveles adecuados de glucosa en sangre; generalmente se manifiesta en la infancia o adolescencia, su causa exacta aún no se conoce, y actualmente no existe una forma comprobada para su prevención.

Diabetes mellitus tipo 2 (DM2). Es el tipo más común de diabetes y corresponde a un trastorno metabólico en el que el organismo presenta resistencia a la insulina o no la utiliza de manera eficiente, lo que genera niveles elevados de glucosa en sangre; está fuertemente relacionada con factores genéticos y de estilo de vida, como el sedentarismo y obesidad, suele ser asintomática en sus etapas iniciales, dificultando el diagnóstico oportuno y favoreciendo la aparición de complicaciones crónicas.

Diabetes mellitus gestacional (DMG). Hiperglucemia que se presenta durante la gestación y que no cumple los criterios diagnósticos para diabetes tipo 2, suele diagnosticarse entre

las semanas 24 y 28 de embarazo y está asociada con un mayor riesgo de complicaciones obstétricas y de desarrollo posterior de DM2 tanto en la madre como en la descendencia.

2.2.3 Factores de riesgo de la diabetes

De acuerdo con la Organización Mundial de la Salud (OMS, 2016), los factores de riesgo varían según el tipo de diabetes:

Diabetes mellitus tipo 1 (DM1). Las causas exactas de este tipo de diabetes aún no se conocen con certeza, se presume que está asociada a una iteración compleja entre factores hereditarios y ambientales, que podrían desencadenar una respuesta autoinmune dirigida contra las células beta del páncreas.

Diabetes mellitus tipo 2 (DM2). Su desarrollo esta influenciado por una combinación de factores genéticos, metabólicos y de estilo de vida. Este riesgo se incrementa significativamente en personas con antecedentes familiares de diabetes, edad superior a 45 años, sobrepeso u obesidad, alimentación inadecuada, sedentarismo y tabaquismo.

Diabetes mellitus gestacional (DMG). Los principales factores de riesgo incluyen la edad materna avanzada, el sobrepeso previo al embarazo, el aumento excesivo de peso durante la gestación, antecedentes familiares de diabetes, historial de diabetes gestacional en embarazos anteriores.

2.2.4 Criterios de diagnóstico de diabetes

Los criterios clínicos y bioquímicos establecidos para el diagnóstico de las distintas formas de diabetes constituyen una herramienta fundamental para una intervención médica oportuna y eficaz, por lo que se recomienda su aplicación rigurosa y repetición las pruebas en caso de ausencia de síntomas evidentes para confirmar el diagnóstico (Organización Mundial de la Salud [OMS], 2016):

Diabetes mellitus tipo 2. Una persona puede ser diagnosticada con diabetes mellitus tipo 2 si cumple al menos uno de los siguientes criterios:

- Glucosa plasmática en ayunas ≥ 126 mg/dL (7,0 mmol/L), después de al menos 8 horas de ayuno.
- Glucosa plasmática ≥ 200 mg/dL (11,1 mmol/L) dos horas después de una carga de 75 g de glucosa durante una prueba de tolerancia a la glucosa oral (PTGO).
- Hemoglobina glucosilada (HbA1c) ≥ 6,5%, esta prueba de debe realizarse en laboratorios certificados y estandarizados.

Diabetes mellitus gestacional. El diagnóstico de diabetes gestacional suele realizarse entre las semanas 24 y 28 de gestación, o antes si existen factores de riesgo evidentes; se considera diagnostico positivo si se cumple al menos uno de los siguientes criterios:

- Glucosa plasmática entre 92 y 125 mg/dL (5,1 a 6,9 mmol/L), tras al menos 8 de ayuno.
- Glucosa plasmática entre 153 y 199 mg/dL (8.5 a 11,0 mmol/L) dos horas después de una carga de 75 g de glucosa durante una prueba de tolerancia a la glucosa oral (PTGO).

Diabetes mellitus tipo 1. En presencia de síntomas característicos como aumento excesivo de la micción (poliuria), sed excesiva (polidipsia) y pérdida de peso inexplicada, el diagnóstico de la diabetes mellitus tipo 1 puede realizarse sin necesidad de una PTGO, si se cumple al menos uno de los siguientes criterios (International Diabetes Federation [IDF], 2021):

- Glucosa plasmática aleatoria ≥ 200 mg/dL (11.1 mmol/L).
- Glucosa plasmática en ayunas ≥ 126 mg/dL (7,0 mmol/L).

- Hemoglobina glucosilada (HbA1c) ≥ 6,5%.

2.2.5 Complicaciones de la diabetes

La diabetes mellitus, cuando no se trata adecuadamente, puede generar complicaciones agudas y crónicas que afectan de manera significativa la salud y calidad de vida de quienes la padecen (Organización Panamericana de la Salud [OPS], 2021):

Complicaciones agudas. Estas complicaciones aparecen de forma repentina y si no se tratan de inmediato, pueden ser potencialmente mortales. A continuación, se describen las más frecuentes (Organización Mundial de la Salud [OMS], 2016):

Cetoacidosis diabética (CAD). Es más frecuente en personas con diabetes tipo 1 y se produce cuando una deficiencia absoluta o relativa de insulina impide que la glucosa ingrese a las células para ser utilizada como fuente de energía, lo que obliga al organismo a recurrir a la lipólisis como vía alternativa, generando cuerpos cetónicos que se acumulan en la sangre y provocan una acidosis metabólica (Fuks & Vaisberg, 2022):

Causas más comunes: incluyen la omisión o suspensión de dosis de insulina, infecciones agudas (como urinarias o respiratorias), estrés físico o emocional y errores en el manejo del tratamiento.

Síntomas principales: náuseas, vómitos, dolor abdominal, respiración rápida y profunda, confusión o pérdida de conciencia en casos severos.

Tratamiento: requiere atención hospitalaria urgente con reposición de líquidos intravenosos, administración de insulina y corrección de los desequilibrios electrolíticos.

Estado hiperglucémico hiperosmolar (EHH). Es más frecuente en personas con diabetes tipo 2, sobre todo en adultos mayores y se caracteriza por una hiperglucemia extrema (glucosa > 600 mg/dL), acompañada de hiperosmolaridad y con ausencia de cetosis (Honório et al., 2024):

Causas más comunes: incluyen infecciones graves (como neumonía o infección urinaria), deshidratación prolongada, omisión o uso inadecuado de medicamentos hipoglucemiantes y enfermedades agudas descompensadas como infarto agudo de miocardio o accidente cerebrovascular.

Síntomas principales: sed intensa, micción frecuente al inicio seguida por oliguria, debilidad marcada, signos de deshidratación, alteración del estado mental como confusión, convoluciones y en casos graves, coma.

Tratamiento: requiere intervención médica urgente, con reposición intensiva de líquidos intravenosos, administración de insulina y corrección de los desequilibrios hidroelectrolíticos.

Hipoglucemia severa. Es más frecuente en personas con diabetes tipo 1, se caracteriza por una disminución anormal de los niveles de glucosa en sangre, generalmente por debajo de 70 mg/dL (Catón, Barrón, & De Lobera Martínez, 2024):

Causas más comunes: administración excesiva de insulina o antidiabéticos orales, omisión de comidas, ejercicio físico intenso o no habitual sin ajuste de medicación, consumo de alcohol.

Síntomas principales: sudoración, temblores, taquicardia, visión borrosa, confusión, perdida de conciencia o convoluciones en casos graves.

Tratamiento: en casos leves, se recomienda ingesta inmediata de carbohidratos de absorción rápida (como jugo azucarado, azúcar o tabletas de glucosa), pero si el paciente ya está inconsciente, se debe administrar glucagón por vía intramuscular o glucosa intravenosa en un establecimiento de salud.

Complicaciones crónicas. Se desarrollan de manera lenta y progresiva como consecuencia de niveles elevados y sostenidos de glucosa en sangre, lo que produce daño en los vasos sanguíneos y los nervios, afectando a distintos órganos y sistemas del cuerpo, siendo inicialmente silenciosas,

pero pueden tener consecuencias graves sino se detectan y controlan a tiempo (Organización Panamericana de la Salud [OPS], 2023):

Enfermedades cardiovasculares. Las personas adultas con diabetes presentan un riesgo dos a tres veces mayor de desarrollar enfermedades cardiovasculares (ECV) en comparación con la población general (Organización Mundial de la Salud [OMS], 2016). Las ECV más frecuentes son cardiopatía coronaria, la enfermedad cerebrovascular, la arteriopatía periférica y la insuficiencia cardiaca congestiva, las cuales pueden conducir a eventos clínicos graves como infarto agudo de miocardio, accidente cerebrovascular isquémico o hemorrágico, hospitalizaciones, intervenciones terapéuticas o muerte súbita, constituyendo una de las principales causas de morbilidad y mortalidad en las personas con diabetes (International Diabetes Federation [IDF], 2021).

Nefropatía diabética. Es una de las complicaciones más frecuentes y graves de la diabetes, caracterizada por el deterioro progresivo de la función renal debido al daño en los glomérulos, encargados de filtran los desechos de la sangre; iniciando con microalbuminuria (presencia de pequeñas cantidades de albúmina en la orina) y avanzando silenciosamente hacia una insuficiencia renal crónica que puede requerir diálisis o trasplante renal, siendo su aparición favorecida por el mal control de la glucemia, hipertensión arterial, la dislipidemia, la obesidad y el tabaquismo, con síntomas tempranos mínimos o ausentes, por lo que se recomienda realizar controles periódicos de función renal (Bravo-Zúñiga & Solari-Yokota, 2024).

Retinopatía diabética. La retinopatía diabética es una de las causas principales de ceguera en adultos, su desarrollo esta influenciado por diversos factores de riesgo, entre los que destacan la duración prolongada de la enfermedad, un deficiente control glucémico, la presencia de daño real, hipertensión arterial y trastornos de perfil lipídico, los cuales contribuyen al deterioro

progresivo de los vasos sanguíneos de la retina (Organización Panamericana de la Salud [OPS], 2023).

Neuropatía diabética. Se manifiesta principalmente en las extremidades inferiores, con síntomas como ardor, dolor, hormigueo y pérdida progresiva de la sensibilidad, lo que favorece la aparición de úlceras, infecciones, deformidades óseas como el pie de Charcot y en casos graves, amputaciones; su desarrollo se relaciona con la duración de la enfermedad, el mal control glucémico y factores como hipertensión arterial, dislipidemia, tabaquismo y edad avanzada (Cobos-Palacios, Sampalo, & Carmona, 2020).

Por otro lado, aunque suele resolverse después del parto, la diabetes gestacional se asocia a múltiples riesgos, tanto para la madre como para el recién nacido (International Diabetes Federation [IDF], 2021):

- En las madres, aumenta la probabilidad de desarrollar diabetes tipo 2 en los años posteriores al embarazo y eleva el riesgo de complicaciones como la preeclampsia, el parto prematuro y necesidad de cesaría.
- En cuanto al bebé, la exposición a niveles elevados de glucosa en el útero puede provocar un mayor peso al nacer (macrosomía), hipoglucemia neonatal, necesidad de cuidados intensivos y lesiones relacionadas con el parto, como la distocia de hombros. A largo plazo, estos niños tienen más probabilidad de desarrollar sobrepeso, obesidad y resistencia a la insulina, lo que los predispone a padecer diabetes tipo 2 durante la niñez o la adolescencia.

2.3 Antecedentes

2.3.1 Antecedentes internacionales

A nivel internacional, se han desarrollado investigaciones que implementan algoritmos de clustering para la segmentación de pacientes con diabetes mellitus tipo 2, con el objetivo de comprender mejor los perfiles clínicos y demográficos de los pacientes.

En este contexto, el estudio de Carrillo-Larco et al. (2021) titulado “Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in Latin America and the caribbean”, tuvo como objetivo identificar grupos de personas con diabetes mellitus tipo 2 (DM2) y evaluar si su frecuencia era consistente entre paises de America Latina y el Caribe (ACL). Para ello, se analizaron 13 encuestas nacionales (n=8361), correspondientes a distintos países y años: Argentina (2018), Barbados (2007), Chile (2003, 2010 y 2017), Costa Rica (2005), El Salvador (2016), Mexico (2016 y 2019), Perú (2005), Uruguay (2006 y 2014) e Islas Vírgenes Británicas (2009), considerando los siguientes predictores como la edad, sexo, el indice de masa corporal (IMC), el perímetro de cintura (CC), la presion arterial sistólica (PAS), la presion arterial diastólica (PAD) y los antecedentes familiares de diabetes. Para la identificacion de grupos, se implemento el algoritmo de k-means y el numero optimo de clusteres se determino mediante los metodos gráficos del codo y de la silueta, estableciendo cuatro conglomerados. El primero clúster agrupó principalmente a personas con valores promedio elevados de PAS, PAD y una mayor proporcion de genero masculino. El segundo clúster incluyo a personas con valores promedio altos de IMC, CC y mayor frecuencia de antecedentes familiares de diabetes. El tercer clúster se caracterizó por presentar los valores promedio más bajos de IMC, CC, PAS y menor proporcion de genero masculino. Finalmente, el cuarto clúster agrupó a personas con la media de edad más alta. Al aplicar esta agrupacion a cada conjunto de datos por pais y año,

se observaron diferencias en la distribucion de los clústeres, lo que sugiere la existencia de patrones epidemiológicos diferenciados según el contexto geografico y temporal.

Lomo et al. (2023) en su estudio titulado “Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method”, cuyo objetivo fue clasificar a pacientes con diabetes mellitus tipo 2 (DM2) mediante técnicas de aprendizaje automático no supervisado, con el fin de identificar perfiles clínicos asociados a la supervivencia y características terapéuticas. Para ello, se analizaron 447 registros médicos de pacientes hospitalizados con diagnóstico de DM2 en el Hospital PKU Muhammadiyah Gamping (Indonesia) durante el periodo 2015-2019. Se considero como predictores las siguientes variables: sexo, edad, presencia de comorbilidades, tipo de complicaciones, tipo de tratamiento antidiabético (insulina, antidiabéticos orales o ambos), niveles de glucosa en la sangre y estado de supervivencia (vivo o fallecido). Para la identificación de grupos, se aplicó una estrategia combinada de reducción de dimensiones mediante Análisis de Componentes Principales (PCA) y agrupamiento mediante el algoritmo Fuzzy C-Means, evaluando la calidad de clústeres a través del índice de Davies-Bouldin (DBI). El número óptimo de clusteres se estableció en tres, siendo la mejor solución la que obtuvo un DBI de 2,2645. El primer clúster agrupo exclusivamente a pacientes sobrevivientes, principalmente hombres ≥ 45 años, sin comorbilidades, con niveles de glucosa moderados o altos y tratados mayoritariamente con combinación de insulina y antidiabéticos orales. El segundo clúster incluyo a un pequeño grupo de hombres ≥ 45 años, todos con comorbilidades y complicaciones circulatorias y con tratamiento combinado. El tercer clúster el más numeroso, se caracterizó por una mayor proporción de mujeres ≥ 45 años, con alta prevalencia de comorbilidades, glucosa elevada y mayor mortalidad (46 fallecidos).

Marhamah et al. (2023) en su investigacion titulada “ The Risk Cluster in Type 2 Diabetes Mellitus Based on Risk Parameters Using Fuzzy C-Menas Algorithm”, en que tiene como objetivo identificar los grupos de riesgo de la diabetes mellitus tipo 2 (DM2) en funcion de los parametros de riesgo utilizando el algoritmo Fuzzy C-Means. El estudio se baso en los datos de 905 personas, de los cuales 562 era hombres y 343 mujeres. Para la identificacion de los grupos de riesgo, se aplico el algoritmo FCM, determinando el numero optimo de clusteres y realizando un analisis de correlacion de Pearson para evaluar la relacion entre los parametros de riesgo y los grupos identificados. Como resultado se obtuvieron dos clusteres de riesgo: el cluster 1, de alto riesgo y el cluster 2, de bajo riesgo. El clúster 1 incluyó principalemnete a personas mayores de 60 años, con antecedentes familiares de DM2, con hipertencion, que ingerían medicamentos regularmente, con actividad fisica insuficiente (< 30 minutos diarios), niveles elevados de presion arterial, IMC elevado y tiempo de sueño insuficiente (<7 horas). El cluster 2 se caracterizo por personas menores de 40 años, sin antecedentes familiares de DM2, sin hipertencion, que no tomaban medicamentos regularmente, que realizaban mas de 30 minutos de actividad fisica y mantenian niveles normales de presion arterial. Por otro lado, el anális de correlacion de Person mostro que la edad, el uso regular demedicamentos, la hipertencion y el nivel de presion arterial tenian correlaciones significativas con la pertenencia al cluster 1.

2.3.2 Antecedentes nacionales

Bernabe-Ortiz et al. (2022) en su estudio titulado “Multimorbidity Patterns among People with Type 2 Diabetes Mellitus: Findings from Lima, Perú”, tuvo como objetivo identificar patrones de morbilidades crónicas en personas con diagnóstico de diabetes mellitus tipo 2 (DM2), utilizando datos provenientes de historias clínicas electrónicas del Hospital de Emergencias Villa El Salvador, un centro de segundo nivel ubicado en Lima, Perú. Para ello, se analizaron 9582 registros

correspondientes al periodo 2016-2021, considerando variables clínicas y demográficas, así como la presencia de comorbilidades identificadas mediante códigos CEI-10 y medidas clínicas como el índice de masa corporal (IMC). En el estudio se incluyeron enfermedades crónicas con una prevalencia igual o mayor al 1%, entre ellas obesidad, hipertensión, dislipidemia, hipotiroidismo, enfermedad renal crónica, entre otras. Para la identificación de grupos, se implementó el algoritmo de K-means, empleando como métrica de similitud el coeficiente de Jaccard, dado el carácter binario de las variables. El número óptimo de conglomerados se definió a partir del criterio de Calinski/Harabasz, estableciéndose cuatro clústeres. El primer clúster agrupo a personas con DM2 sin presencia de otras enfermedades crónicas. El segundo clúster estuvo conformado por personas con DM2 y obesidad como única comorbilidad. El tercer clúster incluyo a personas con DM2 y enfermedades cardiovasculares, en especial hipertensión, pero sin obesidad. Por último, el cuarto clúster agrupo a personas con DM2 y otras condiciones crónicas diversas, como hipotiroidismo, dislipidemia o enfermedad renal crónica.

3. HIPÓTESIS Y VARIABLES

3.1 Hipótesis

3.1.1 Hipótesis general

El algoritmo de clúster Fuzzy C-Means permite segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 en la región Cusco, durante el periodo 2019-2022.

3.1.2 Hipótesis especificas

1. El número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 en la región Cusco se encuentra entre dos y cinco, según los índices de validación interna aplicados al algoritmo Fuzzy C-Means.
2. Cada segmento identificado mediante el algoritmo Fuzzy C-Means agrupa a pacientes con perfiles clínicos diferenciados.
3. Los pacientes con pertenencia difusa elevada presentan combinaciones de características clínicas propias de más de un segmento identificado mediante el algoritmo Fuzzy C-Means.

3.2 Identificación de variables e indicadores

Dado que el algoritmo Fuzzy C-Means pertenece al enfoque de aprendizaje no supervisado, este estudio no considera variables dependientes ni independientes, sin embargo, se utilizan variables intervinientes que representan las características clínicas relevantes de los pacientes con diabetes mellitus tipo 2: edad, peso, talla, índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y glicemia. La selección de estas variables se fundamenta en su relevancia clínica para caracterizar el estado de salud de los pacientes y facilitar la identificación de segmentos con perfiles clínicos diferenciados.

3.3 Operacionalización de variables

Variable	Definición conceptual	Definición operacional	Escala	Unidad de medida	Tipo de variable
Edad	Tiempo transcurrido desde el nacimiento hasta el momento de evaluación clínica.	Se registra en años complidos según la fecha de nacimiento consignada en la historia clínica del paciente.	De razón	Años	Cuantitativa discreta
Peso	Medida de la masa corporal del paciente.	Se mide en kilogramos mediante una balanza calibrada durante la consulta médica.	De razón	Kilogramos (kg)	Cuantitativa continua
Talla	Estatura del paciente desde el talón hasta la parte superior de la cabeza.	Se mide en metros con un tallímetro, en posición erguida durante la evaluación clínica.	De razón	Metros (m)	Cuantitativa continua
Índice de masa corporal (IMC)	Indicador del estado nutricional del paciente que relaciona el peso con la talla.	Se calcula dividiendo el peso entre la talla al cuadrado.	De razón	kg/m^2	Cuantitativa continua
Presión arterial sistólica	Presión ejercida por la sangre sobre las paredes arteriales durante la contracción del corazón (sístole).	Se mide en reposo con un esfigmomanómetro, registrando el valor máximo.	De razón	Milímetro de mercurio (mmHg)	Cuantitativa continua
Presión arterial diastólica	Presión ejercida por la sangre sobre las paredes arteriales durante la relajación del corazón (diástole).	Se mide en reposo con un esfigmomanómetro, registrando el valor mínimo.	De razón	Milímetro de mercurio (mmHg)	Cuantitativa continua
Glicemia	Concentración de glucosa en sangre, indicador metabólico clave en el diagnostico de diabetes.	Se mide en ayunas mediante prueba de laboratorio o glucómetro calibrado.	De razón	Miligramos por decilitro (mg/dL)	Cuantitativa continua
Sexo	Condición biológica que distingue a los seres humanos como masculinos y femeninos.	Se registra según lo indicado en el historial clínica del paciente (1= masculino, 2= femenino).	Nominal	No aplica	Cualitativa nominal

4. METODOLOGIA

4.1 Ámbito de estudio: localización política y geográfica

El estudio se desarrolló con pacientes diagnosticados con diabetes mellitus tipo 2, registrados en establecimientos de salud pertenecientes a la Dirección Regional de Salud (DIRESA) Cusco, Perú.

4.2 Tipo, enfoque, nivel y diseño de investigación

La presente investigación es de tipo aplicada, dado que busca aportar conocimientos prácticos y metodológicos al campo de la salud pública desde una perspectiva matemática y estadística. Mediante la implementación del algoritmo Fuzzy C-Means, un método innovador en análisis de datos clínicos, se realiza la segmentación de pacientes con diabetes mellitus tipo 2 para identificar perfiles clínicos diferenciados.

El enfoque es cuantitativo, dado que se fundamenta en la medición numérica de variables clínicas y en el análisis estadístico de los datos a través del algoritmo Fuzz C-Means, lo que facilita la identificación de patrones en la segmentación de pacientes.

En cuanto al nivel, es descriptiva, porque tiene como propósito caracterizar los segmentos identificados y analizar patrones comunes entre los pacientes, incluyendo aquellos con pertenencia difusa alta a más de un clúster.

El diseño es no experimental, porque no se manipulan las variables de estudio, sino que se trabaja con datos secundarios previamente registrados. Además, es de corte transversal, porque los datos analizados corresponden a un periodo especifico (2019-2022), sin seguimiento longitudinal de los pacientes.

4.3 Unidad de análisis

Cada paciente con diagnóstico de diabetes mellitus tipo 2 registrado por la DIRESA Cusco durante el periodo de estudio constituye una unidad de análisis.

4.4 Población de estudio

La población está conformada por 2750 pacientes diagnosticados con diabetes mellitus tipo 2, registrados por la DIRESA Cusco entre los años 2019 y 2022. En el Anexo C, se presentan los primeros registros del año 2019, así como la totalidad de variables disponibles en la base de datos.

4.5 Tamaño de muestra

No se aplicó muestreo, ya que se trabajó con el total de registros disponibles en la base de datos, correspondientes a la población objetivo.

4.6 Técnicas de recolección de información

Los datos fueron obtenidos de fuentes secundarias, específicamente registros clínicos anonimizados proporcionados por el Centro Nacional de Epidemiología, Prevención y Control de Enfermedades del Perú (CDC Perú). La base de datos incluye variables clínicas, sociodemográficas y tipo de tratamiento recibido por los pacientes.

4.7 Técnicas de análisis e interpretación de la información

Para el análisis de la base de datos proporcionada por el CDC Perú, se realizó un proceso de preprocesamiento que incluyó: exclusión de variables no clínicas (información geográfica, sociodemográfica y administrativa), eliminación de variables con más del 20% de valores faltantes, imputación de datos faltantes en variables con menos del 20% de datos faltantes mediante el algoritmo k-Nearest Neighbors (kNN) y tratamiento de datos atípicos. Asimismo, se eliminaron variables con alta correlación.

Posteriormente, se estandarizaron las variables y se implementó el algoritmo de clúster Fuzzy C-Means utilizando el paquete fclust del lenguaje de programación R, con los siguientes parámetros: número de clústeres igual = 4, parámetro de difusividad =1.4 y métrica de distancia euclidiana.

La determinación del número óptimo de clústeres se realizó mediante medidas de validación interna:

- Medidas de difusividad: Coeficiente de Partición (PC), Coeficiente de Entropía (PE), Coeficiente Modificado de Partición (MPC).
- Medidas de separación/compactación: Índice de partición de Xie-Beni (XB), Fuzzy Simplified Silhouette (FSS), Índice de Kwon (K) e Índice de Tang-Sun-Sun (TSS).

Finalmente, los segmentos definidos fueron caracterizados clínicamente con el fin de evaluar la existencia de perfiles diferenciados entre los pacientes.

4.8 Técnicas para demostrarla verdad o falsedad de las hipótesis planteadas

La validación de las hipótesis se realizó a partir del análisis de los resultados obtenidos con el algoritmo Fuzzy C-Means, en conjunto con los índices de validación aplicados. Además, se analizaron los casos con pertenencia difusa elevada, a fin de identificar patrones clínicos compartidos entre distintos segmentos y evaluar su relevancia para un enfoque de atención más personalizado.

5. RESULTADOS Y DISCUSIÓN

5.1 Procesamiento, análisis, interpretación y discusión de resultados

Este estudio se realizó a partir de una base de datos compuesta por 2750 registros de pacientes diagnosticados con diabetes mellitus tipo 2 y 38 variables que incluían información clínica, sociodemográfica y relacionada con el tipo de tratamiento. El procesamiento, análisis e interpretación de los datos se realizaron mediante el lenguaje de programación estadística R.

5.1.1 Preprocesamiento de la data

Con el objetivo de realizar una segmentación clínica adecuada, se llevó a cabo un proceso de preprocesamiento de la data, el cual se desarrolló en varias etapas: exclusión de variables no clínicas, eliminación de variables clínicas con elevada proporción de datos faltantes, imputación de valores faltantes y atípicos mediante el algoritmo kNN (k-Nearest Neighbors), evaluación de colinealidad entre las variables clínicas. Estas etapas fueron fundamentales para garantizar la calidad y consistencia de los datos utilizados en el análisis de segmentos mediante el algoritmo Fuzzy C-Means.

En una primera etapa, se descartaron aquellas variables que no aportaban información relevante desde el punto de vista clínico, tales como variables de tipo administrativo, geográfico, institucional o sociodemográfico, así como aquellas redundantes o sin registros disponibles. La Tabla 1 presenta un resumen de las variables eliminadas junto con su respectiva justificación.

En la segunda etapa del preprocesamiento, se eliminaron del conjunto de variables clínicas aquellas con un porcentaje de valores faltantes superior al 20%, dado que una proporción elevada de datos faltantes puede introducir sesgos y afectar negativamente a la calidad de la segmentación. La Tabla 2 presenta un resumen de los datos faltantes por variable.

Tabla 1

Variables descartadas de la base de datos

Variable	Justificación
diresas, establecimiento, ubigeo_nac, ubigeo_res	Información geográfica o institucional, sin utilidad clínica.
país_nac	Sin registros disponibles.
fecha_nac, fecha_cap, semana	Información contenida en variables como “edad” y “año”.
tdiabetes	Todos los registros corresponden a DM2, variable redundante.
asegurado, tseguro	Información administrativa sin influencia directa en el perfil clínico.
tcaso	Indica si es caso nuevo o prevalente, no aporta valor clínico.
tratamiento	Representa el estado terapéutico, influido por factores externos no clínicos.
metformina, sulfolinurea, inhibidores, insul_humana, insul_analoga, glitazonas, glifozinas, agonistas	Información terapéutica posterior al diagnóstico, susceptible a sesgo institucional.
cumplimiento	Influenciado por factores externos, no clínicos.
instrucción	Variable sociodemográfica que no incide directamente en el perfil clínico.

Nota. Se conservaron las variables “sexo” y “edad” que, aunque son clasificadas como sociodemográficas, se consideran esenciales en el análisis clínico por su relación con la evolución, riesgo y características de la DM2.

Tabla 2

Datos faltantes por variable

Variable	Datos faltantes (N)	Datos faltantes (%)
sexo	0	0.00
edad	0	0.00
peso	207	7.53
talla	326	11.85
imc	352	12.80
pcint	894	32.51
sistólica	6	0.22
diastólica	6	0.22
glicemia	104	3.78
hemoglic	2084	75.78
ldl	2386	86.76
col_total	2594	94.33
triglicéridos	2526	91.85
hdl	2595	94.36

Nota. Esta tabla muestra la cantidad absoluta y el porcentaje de datos faltantes por variable.

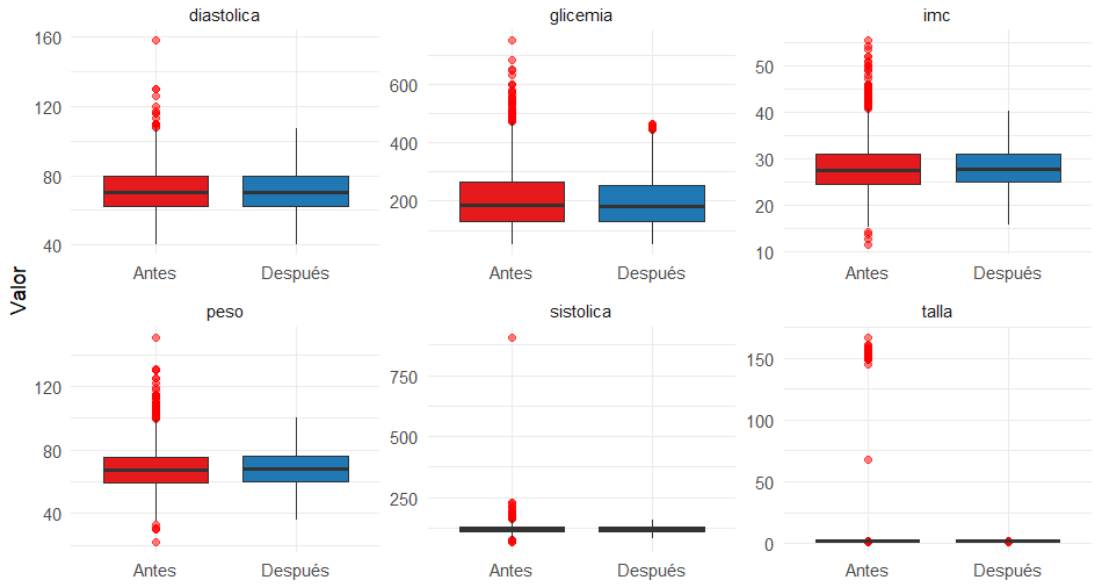
A partir de los valores mostrados en la Tabla 2, se excluyeron del análisis las siguientes variables por superar el umbral del 20% de datos faltantes: pcint (32.51%), hemoglic (75.78%), ldl (86.76%), col_total (94.33%), triglicéridos (91.85%) y hdl (94.36%).

Luego, se procedió con la imputación de valores faltantes en las variables que presentaban menos del 20% de datos faltantes: “peso”, “talla”, “imc”, “sistólica”, “diastólica” y “glicemia”. Para ello, se utilizó el algoritmo kNN (k-Nearest Neighbors), el cual estima los valores faltantes considerando la similitud entre observaciones. A diferencia de métodos univariados como la mediana, el kNN emplea las observaciones más cercanas, según la distancia euclidiana y realiza la imputación ponderada, lo que permite preservar la estructura multivariada de los datos. Esta característica resulta especialmente útil en contextos clínicos, donde suele haber correlación entre variables y es más adecuada para análisis de clusterring, ya que contribuye a mantener las relaciones naturales entre las variables.

Posteriormente, se identificaron valores atípicos (outliers) en las variables “peso”, “talla”, “imc”, “sistólica”, “diastólica” y “glicemia”, los cuales se visualizaron mediante diagramas de cajas (boxplots), como se observa en la Figura 1. Con el fin de reducir su impacto en el análisis, estos valores extremos fueron tratados como datos faltantes y posteriormente imputados utilizando nuevamente el algoritmo kNN, Este procedimieto permitió mitigar su influencia sin distorsionar la estructura subyacente de los datos, lo cual es esencial en estudios basados en agrupamiento difuso como el Fuzzy C-Means. Tras la imputación con el algoritmo kNN, la mayoría de las variables clínicas ya no presentan valores outliers, con excepción de glicemia, donde aún se identifican algunos valores extremos. Se decidió conservarlos, dado que podrían representar condiciones clínicas reales de hiperglicemia, propias del comportamiento esperado en pacientes con diabetes mellitus tipo 2. La Figura 2 muestra las distribuciones de las variables clínicas antes y después de la imputación de valores faltantes y valores atípicos.

Figura 1

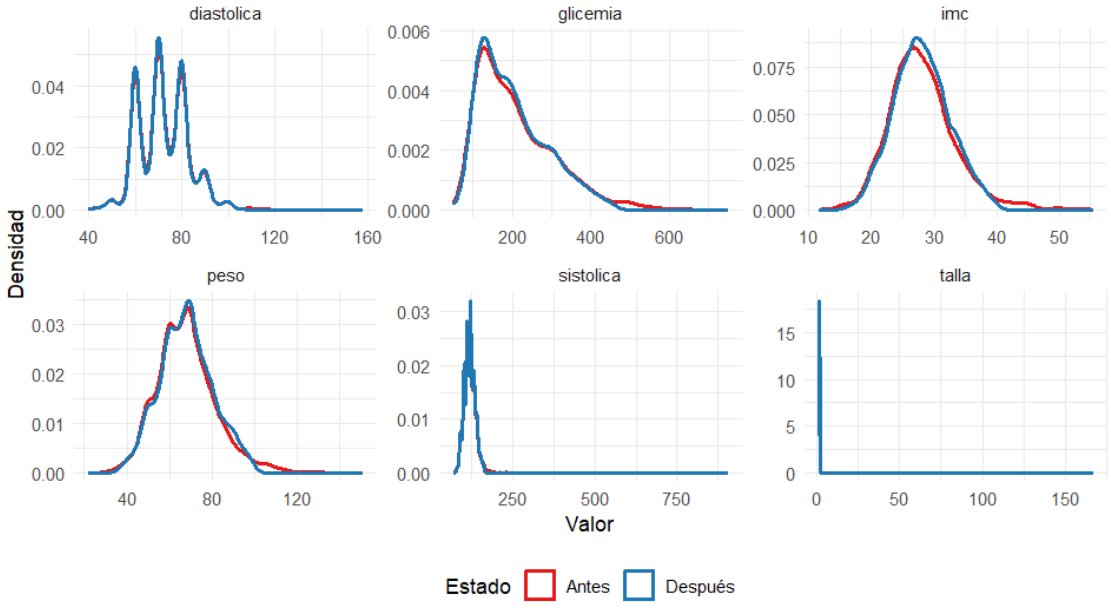
Comparación de distribución y outliers antes y después de la imputación con kNN



Nota. Se identificaron 74 outliers en peso, 350 en talla, 84 en IMC, 51 en presión arterial sistólica, 17 en presión arterial diastólica y 63 en glicemia.

Figura 2

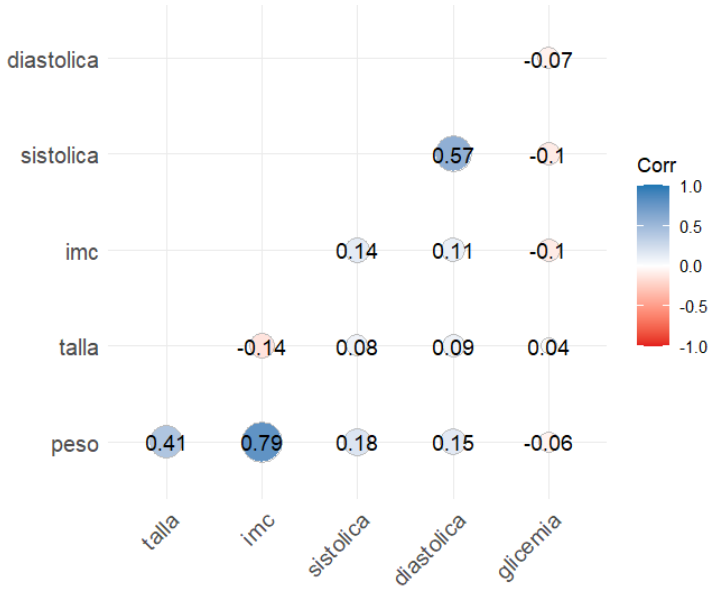
Comparación de densidad antes y después de la imputación con kNN



Nota. Las curvas de densidad muestran una alta similitud antes y después de la imputación, lo que sugiere que la estructura subyacente de las variables clínicas de ha conservado.

Figura 3

Matriz de correlación entre variables clínicas



Nota. Se observan correlaciones débiles entre la mayoría de las variables, lo que indica una baja colinealidad general.

Finalmente, se evaluó la colinealidad entre las variables clínicas mediante la matriz de correlación de Spearman, dado que es más robusto frente a asimetrías y valores extremos. El objetivo de este análisis fue identificar y eliminar variables con una correlación alta, con el fin de reducir la redundancia informativa y mejorando la interpretabilidad de los clústeres. Como se muestra en la Figura 3, se identificó una correlación positiva fuerte entre “peso” e “IMC” (r=0.79), una correlación positiva moderada entre la presión “sistólica” y “diastólica” (r=0.57) y una correlación positiva moderada entre “peso” y “talla” (r=0.41). Dado que el IMC se calcula a partir del peso y la talla, y resume de manera eficiente la relación entre ambas, se decidió conservar únicamente el IMC y eliminar “peso” y “talla” con el fin de reducir la redundancia informativa en

el conjunto de datos. Esta decisión es especialmente pertinente en el contexto del algoritmo Fuzzy C-Means, puesto que, al basarse en distancias euclidianas, la inclusión de variables altamente correlacionadas puede duplicar la información y sesgar la formación de clústeres. Sin embrago, aunque la presión sistólica y diastólica presentan una correlación moderada, ambas variables reflejan aspectos fisiológicos distintos del funcionamiento cardiovascular (presión durante la contracción y relajación del corazón, respectivamente), por lo que su inclusión conjunta se considera clínicamente justificada y relevante para el análisis.

5.1.2 Análisis exploratorio de datos (EDA)

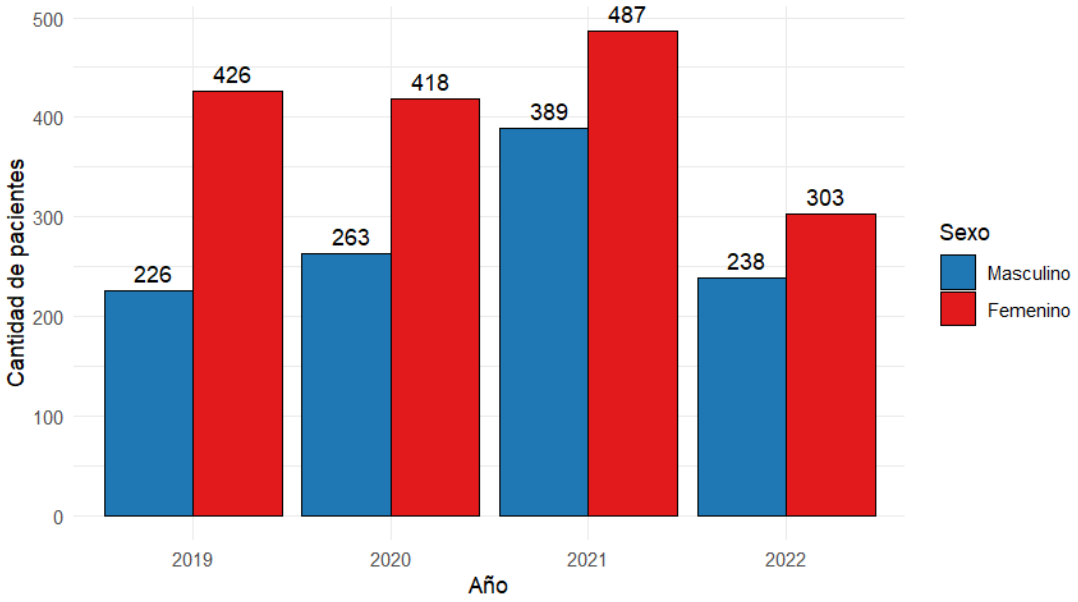
Una vez concluido el preprocesamiento de la base de datos, se realizó un análisis exploratorio de datos con el objetivo de identificar patrones generales y describir características iniciales de la población de estudio. En primer lugar, se examinó la distribución de pacientes diagnosticados con diabetes mellitus tipo 2 según el año y el sexo, con el objetivo de observar la evolución temporal de los registros en el periodo de 2019-2020.

Como se muestra en la Figura 4, el mayor número de diagnosticados se registró en el año 2021 (876 pacientes), seguido por 2020 (681), 2019 (652) y 2022 (541). El incremento en 2021 podría estar relacionado con la reactivación progresiva de los servicios de salud tras las restricciones implementadas durante la pandemia de COVID-19 en 2020, lo que habría generado una acumulación de casos no detectados oportunamente durante ese año. Además, se observa una mayor proporción de pacientes femeninas diagnosticadas con diabetes mellitus tipo 2 en todos los años del estudio. Esta predominancia podría estar asociada a diversos factores, entre ellos, los antecedentes de diabetes gestacional, condición exclusiva de las mujeres que incrementa el riesgo de desarrollar diabetes tipo 2 en etapas posteriores de la vida. Asimismo, las mujeres tienden a

realizarse controles médicos de forma más regular que los varones, lo que favorece a una detección oportuna y temprana de la diabetes.

Figura 4

Distribución de pacientes según año y sexo



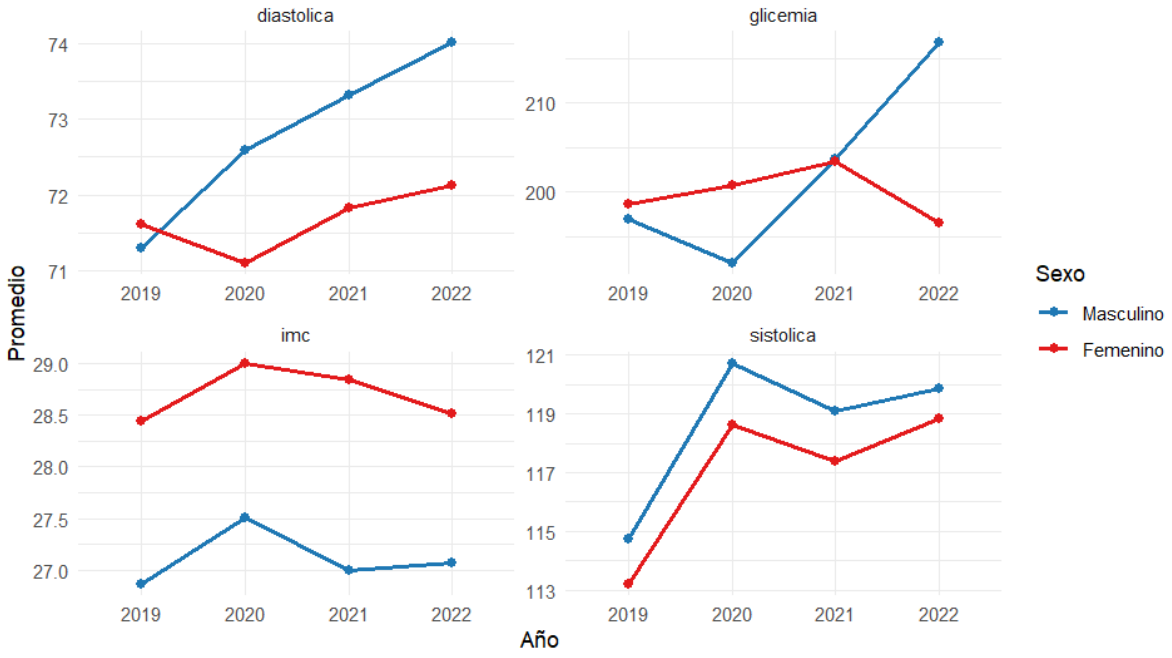
Nota. La proporción de mujeres fue superior en todos los años del estudio: 65.3% en 2019, 61.4% en 2020, 55.6% en 2021 y 56.0% en 2022.

Por otro lado, se estimaron los valores promedio de las variables clínicas: índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y glicemia, desagregados por sexo y año, cuyas tendencias interanuales se presentan en la Figura 5. Entre 2019 y 2022, se observa un incremento progresivo en la presión arterial diastólica para ambos sexos, siendo más pronunciado en varones. La arterial presión sistólica muestra una elevación importante entre 2019 y 2020, con una posterior estabilización en los siguientes años en ambos sexos. Los niveles de glicemia presentan una tendencia creciente hasta 2021 en mujeres; sin embargo, en 2022 se observa una disminución en mujeres, en cambio en varones se observa una disminución en 2020 y un

aumento marcado en 2021 y 2022. Por otro lado, el IMC permanece relativamente estable en ambos sexos, mostrando variaciones con un pico en 2020.

Figura 5

Promedio de variables clínicas según el sexo

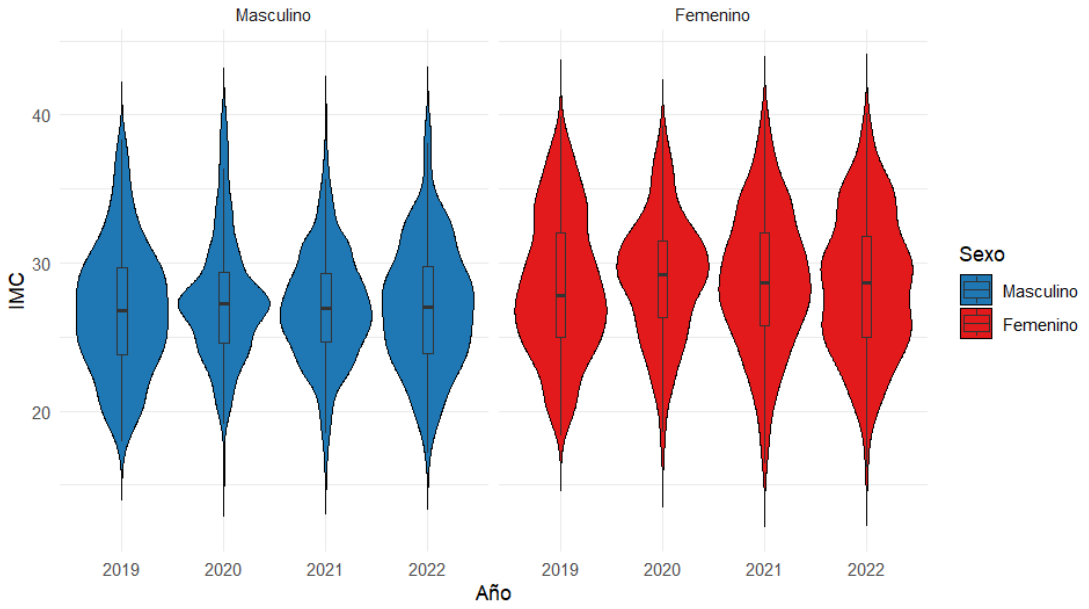


Nota. Se observa una tendencia creciente en los valores de presión arterial, especialmente en varones, mientras que el IMC se mantienen relativamente estable y glicemia con una tendencia creciente en varones y una disminución en el último año en mujeres.

Como parte del análisis exploratorio de datos, se examinaron las distribuciones del índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y glicemia, desagregados por sexo y año. Para ello, se utilizaron gráficos de violín, que permiten visualizar simultáneamente la forma de la distribución, su densidad, simetría, mediana y dispersión. Este análisis preliminar tuvo como finalidad identificar indicios de posibles diferencias clínicas entre varones y mujeres que podrían anticipar la existencia de perfiles diferenciados en el proceso de segmentación mediante el algoritmo Fuzzy C-Means.

Figura 6

Distribución del índice de masa corporal (IMC) por sexo y año



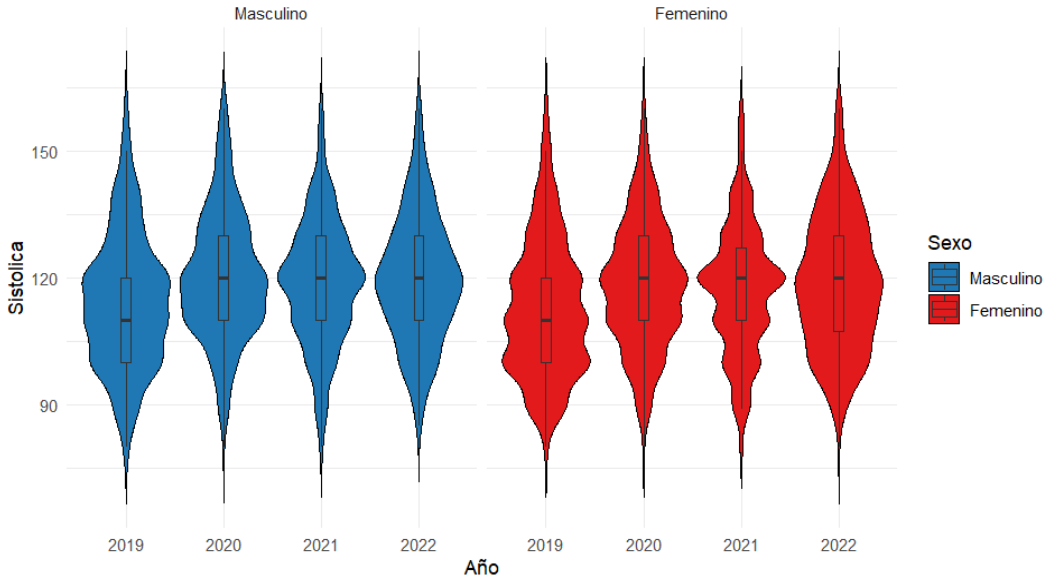
Nota. Las colas alargadas hacia valores elevados en ambos sexos, indican presencia de casos con obesidad severa.

Las distribuciones del IMC son similares en ambos sexos a lo largo del periodo 2019-2022, con la mediana ligeramente más alta en mujeres. En ambos grupos, la mayor concentración de valores está alrededor de 25-30 kg/m^2, lo que sugiere una alta prevalencia de sobrepeso en la población de estudio. Además, en ambos grupos también se identifican colas alargadas hacia valores más altos de IMC, lo que sugiere la presencia de subgrupos de obesidad.

La presión arterial sistólica presenta distribuciones más extendidas hacia valores altos y dispersas en varones. Esto indica que la presión sistólica tiende a estar más elevada y con mayor variabilidad entre los varones, lo que podría corresponder a un patrón de mayor riesgo cardiovascular en ese grupo. En cambio, las mujeres muestran distribuciones más concentradas, con menor presencia de valores extremos.

Figura 7

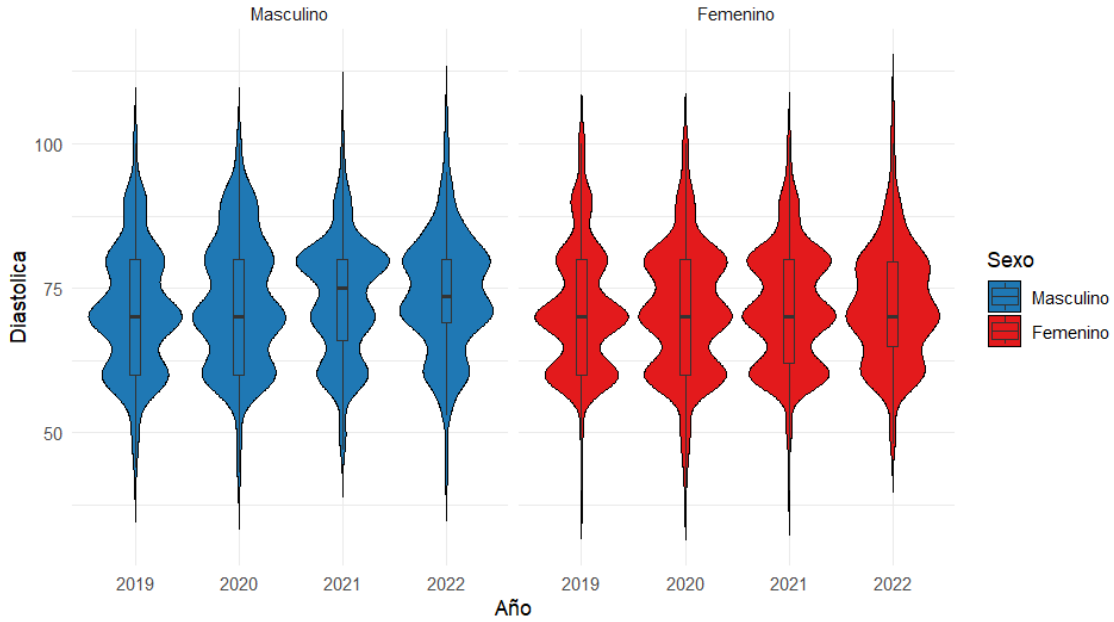
Distribución de presión arterial sistólica por sexo y año



Nota. Hay una mayor dispersión hacia niveles hipertensivos (≥ 140 mmHg) en 2020.

Figura 8

Distribución de la presión arterial diastólica por sexo y año



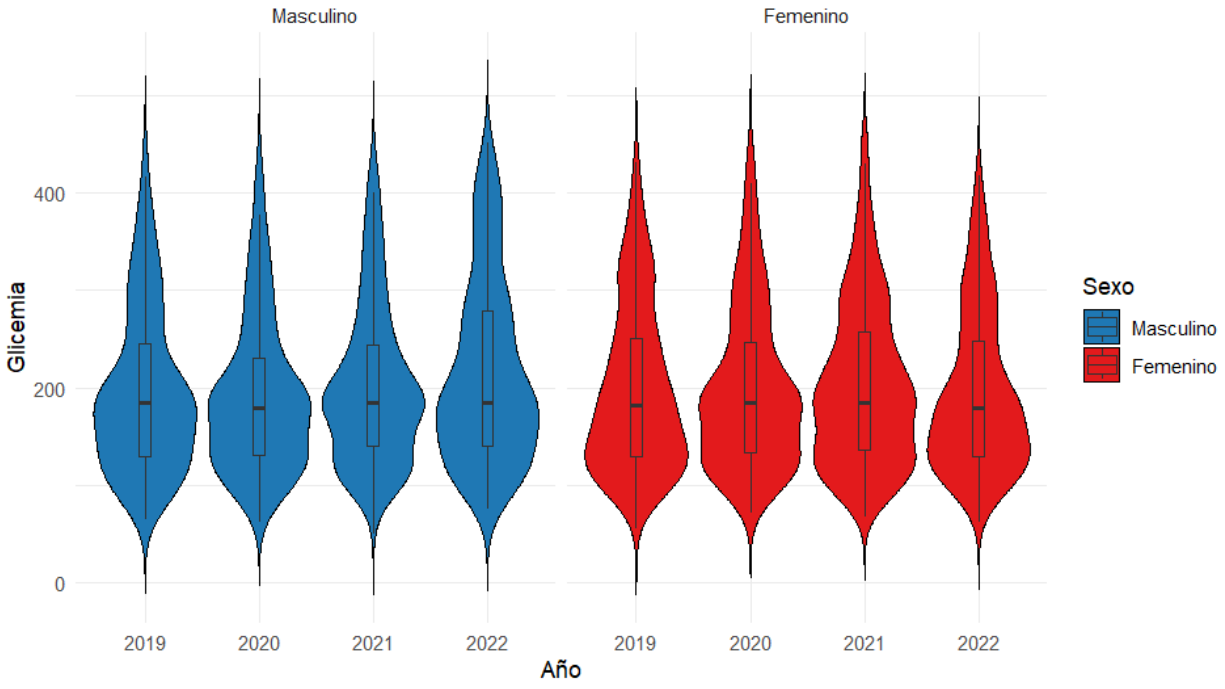
Nota. Las medianas son ligeramente superiores en varones en todos los años.

La presión arterial diastólica refleja un comportamiento similar al observado en la presión arterial sistólica. Las distribuciones en varones son más anchas y con mayor densidad por encima de 80 mmHg, lo que sugiere una mayor prevalencia de hipertensión diastólica en ese grupo. En contraste, las mujeres presentan distribuciones son más centradas, lo que podría reflejar un mayor control clínico o detección temprana.

La glicemia muestra una distribución más asimétrica y dispersa en los varones, con colas alargadas hacia valores muy elevados (hiperglucemias severas). Aunque las medianas de glicemia entre varones y mujeres son similares, la mayor presencia de valores extremos en los varones podría asociarse a un control glucémico más deficiente o irregular. En contraste, las mujeres presentan distribuciones más centradas y menos dispersas.

Figura 9

Distribución de la glicemia por sexo y año



Nota. Las colas alargadas en varones reflejan mayor frecuencia de hiperglucemias severas, particularmente en el año 2022.

Las diferencias observadas, entre sexos en la forma y dispersión de las distribuciones de las variables clínicas sugieren que la variabilidad interna dentro de cada grupo es distinta. En particular, los varones presentan una mayor heterogeneidad clínica, evidenciada en distribuciones más amplias y con colas más pronunciadas, especialmente en las variables glicemia, presión arterial sistólica y diastólica. Estas diferencias pueden constituir indicios de perfiles clínicos diferenciados por sexo, lo cual resulta relevante en el contexto del análisis de segmentación de pacientes. Aunque la variable sexo no será incluida en la implementación del algoritmo Fuzzy C-Means, debido a que se busca una segmentación basada exclusivamente en variables clínicas, su consideración en el análisis interpretativo permitirá una mejor caracterización de los clústeres resultantes.

5.1.3 Implementación del algoritmo Fuzzy C-Means

Dado que el objetivo del análisis es identificar perfiles clínicos entre los pacientes, se decidió excluir la variable categórica sexo en la fase de agrupamiento, ya que no representa una característica clínica cuantificable como las demás variables consideradas: edad, índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y nivel de glicemia. Sin embargo, la distribución de la variable sexo será analizada posteriormente dentro de cada segmento, con fines interpretativos.

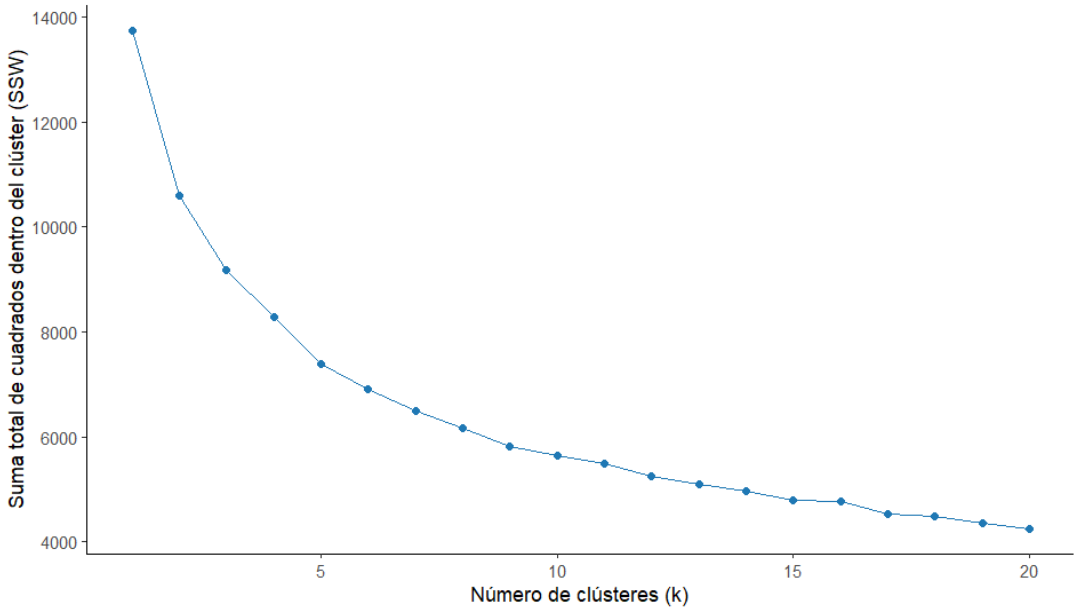
La implementación del algoritmo Fuzzy C-Means se realizó utilizando el paquete fclust del lenguaje de programación R, el cual se basa en el cálculo de distancias euclidianas. Todas las variables fueron previamente estandarizadas con el propósito de eliminar las diferencias de escala entre ellas y asegurar una contribución equitativa de cada variable en el proceso de agrupamiento.

Para obtener una referencia preliminar sobre el número óptimo de clústeres, se utilizó el método del codo. Como se muestra en la Figura 10, la curva de la suma de los cuadrados dentro

de los clústeres (SSW) presenta una disminución menos pronunciada entre $k = 3$ y $k = 4$, lo que sugiere que el número óptimo de clústeres podría estar en esos valores.

Figura 10

Gráfico del método del codo



Nota. El grafico muestra como varía la suma de los cuadrados dentro de los clústeres (SSW) a medida que aumenta el número de clústeres (k).

Para determinar con mayor precisión el número óptimo de clústeres, se implementó el algoritmo Fuzzy C-Means con diferentes cantidades de clústeres: $K = 2, 3, 4, 5$ y un grado de difusividad $m = 1.5$, por ser un valor ampliamente recomendado en la literatura para bases de datos con ruido moderado y sin estructura claramente definida, permitiendo así una representación más flexible y realista. La calidad del agrupamiento se evaluó mediante múltiples índices de validez específicos para clústeres difusos, los cuales consideran tanto la compactación como la separación entre clústeres, así como el grado de pertenencia de cada observación a los distintos clústeres. Los resultados obtenidos se presentan en la Tabla 3.

Tabla 3

Índices de validez para diferentes valores del número de clúster

k	PC	PE	MPC	XB	FS	Kwon	TSS
2	0.7080	0.4521	0.4159	0.7424	-0.7800	2041.77	1792.72
3	0.5812	0.7242	0.3719	0.7957	-0.9061	2188.63	1917.43
4	0.5243	0.8928	0.3658	0.5230	-0.9525	1438.86	1322.34
5	0.4831	1.0264	0.3539	0.4661	-0.9698	1282.53	1192.72

Nota. Estos índices permiten comparar la calidad del agrupamiento y la selección del número óptimo de clústeres.

La selección del número óptimo de clústeres se realizó mediante la evaluación simultanea de siete índices de validez interna: Coeficiente de Partición (PC), Entropía de la Partición (PE), Coeficiente de Partición Modificado (MPC), Xie-Beni (XB), Fuzzy Silhouette (FS), Kwon (K) y Tan-Sun-Sun (TSS). Los tres primeros índices PC, PE y MPC, que favorecen asignaciones más definidas, indicaron un mejor desempeño para k=2. Por otro lado, los índices XB, FS, Kwon y TSS, que evalúan la estructura del agrupamiento considerando tanto la compacidad intra-clúster como la separación inter-clúster sugieren un mejor desempeño con k=5. Sin embargo, la diferencia entre los valores obtenidos para k=4 y k=5 en dichos índices fue mínima, especialmente si se compara con la mejora más marcada observada entre k=3 y k=4. Por ejemplo, el índice XB disminuyo de 0.7957 a 0.5230 al pasar de k=3 a k=4, mientras que solo se redujo a 0.4661 en k=5. Esto indica que k=4 constituye un punto de equilibrio, en el cual se logra una mejora relevante en la calidad del agrupamiento, sin incrementar la complejidad interpretativa que dificulte el análisis. En este contexto, se seleccionó k=4 como numero óptimo de clústeres, por representar un

equilibrio entre la calidad estadística del agrupamiento y utilidad interpretativa en el contexto del análisis clínico.

Tabla 4

Índices de validez para diferentes valores del parámetro de difusividad

Índice	m=1.3	m=1.5	m=1.8	m=2
Coeficiente de Partición (PC)	0.7144	0.5243	0.3387	0.2820
Entropía de la Partición (PE)	0.5351	0.8928	1.2200	1.3227
Coeficiente de Partición Modificado (MPC)	0.6192	0.3658	0.1183	0.0427
Xie-Beni (XB)	0.4969	0.5230	1.1196	7.1962
Fuzzy Silhouette (FS)	-0.9915	-0.9528	-0.7913	-0.6054
Kwon (K)	1367.07	1438.86	3079.59	19792.23
Tan-Sun-Sun (TSS)	1284.35	1322.34	2518.76	8065.41

Nota. Valores bajos de m generan particiones más nítidas, mientras que valores altos generan particiones más difusas.

Por otro lado, la determinación del valor adecuado del parámetro de difusividad (m) es fundamental en el algoritmo Fuzzy C-Means, ya que influye directamente en el grado de superposición entre clústeres. Para este análisis, se evaluaron los valores $m = 1.3, 1.5, 1.8$ y 2.0 también mediante siete índices de validez interna. En la Tabla 4 se presentan los resultados obtenidos.

Aunque $m = 1.3$ mostró la mayor nitidez en la partición, como se evidencia en los valores más altos del PC=0.7144, MPC=0.6192 y el valor más bajo de PE=0.5351, generando una agrupación rígida, limitando la representación de perfiles con características mixtas. Por otro lado,

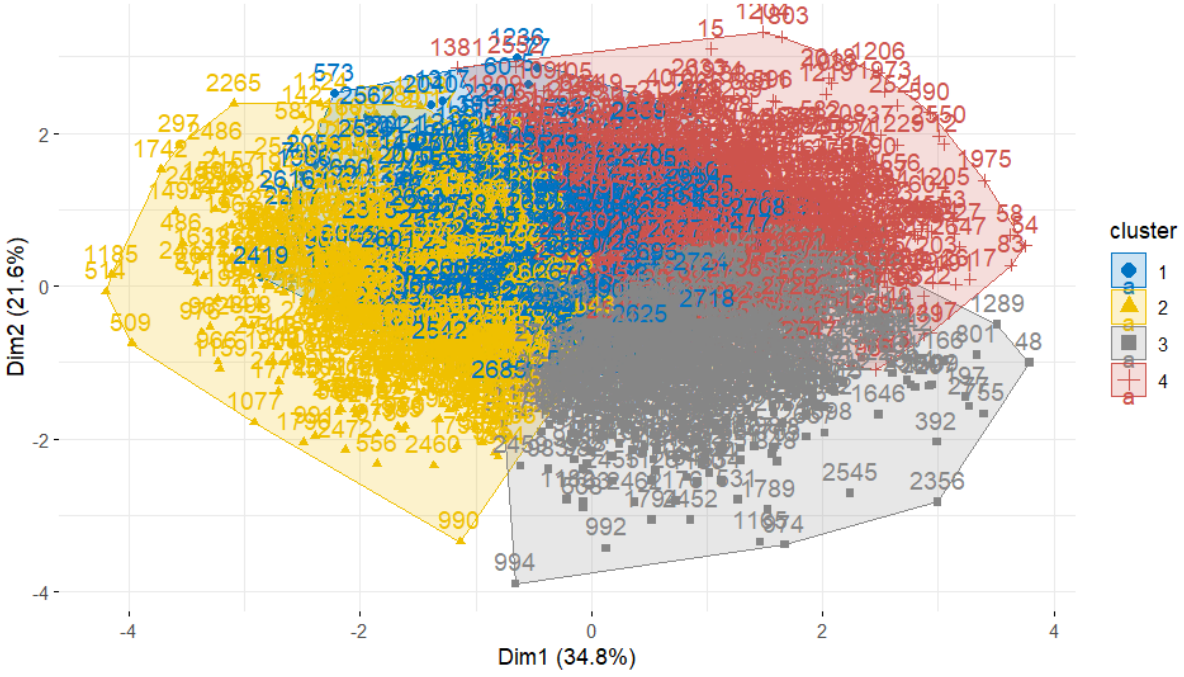
$m = 1.8$ y 2.0 produjeron particiones excesivamente difusas. Esto se evidencia en los bajos valores de PC y MPC, altos valores de PE y un fuerte deterioro en los índices estructurales: XB=1.1196 y 7.1962, Kwon=3079.59 y 19792, TSS=2518.76 y 8065.41, lo cual indica una pérdida de compacidad y separación entre clústeres. En este contexto, $m = 1.5$ resultó ser el valor más adecuado, ya que representa un equilibrio entre nitidez y difusividad entre los clústeres. Este valor presentó valores intermedios en los principales índices PC=0.5243, MPC=0.3658, PE=0.8928 y XB=0.5230, lo que indica una agrupación flexible sin comprometer la calidad del agrupamiento. Asimismo, los valores de Kwon=1438.86 y TSS=1322.34 reflejan una adecuada compacidad y separación entre clústeres. Por tanto, $m = 1.5$ fue seleccionado como valor óptimo, al favorecer una segmentación realista con el enfoque difuso y clínicamente interpretable.

5.1.4 Análisis e interpretación de los clústeres

Tras la aplicación del algoritmo Fuzzy C-Means, con $k = 4$ clústeres y un parámetro de difusividad $m = 1.5$, se procedió a caracterizar los perfiles clínicos de los pacientes agrupados en cada segmento. Para ello, se analizaron los valores promedios y las distribuciones de las variables incluidas en el análisis: edad, índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y glicemia. La Figura 11 presenta una visualización bidimensional de los clústeres, obtenido mediante una reducción de dimensión a través del Análisis de Componentes Principales (PCA, por sus siglas en ingles). El primer componente (Dim1) explicó el 34.8% de la variabilidad total, mientras que el segundo componente (Dim2) explicó el 21.6%, alcanzando juntos un 56.4% de la información contenida en las variables clínicas consideradas.

Figura 11

Visualización bidimensional de los clústeres obtenidos mediante PCA



Nota. Los colores asignados a los puntos indican el clúster con mayor grado de pertenencia para cada paciente.

Cada punto en la Figura 11 representa a un paciente diagnosticado con diabetes mellitus tipo 2, cuya ubicación en el plano se determina a partir de combinaciones lineales de sus características clínicas. A continuación, se describen las regiones asociadas a cada segmento:

- Clúster 1 (azul): se ubica principalmente en la zona central del plano y presenta una superposición con los demás clústeres, lo que refleja la naturaleza difusa del algoritmo FCM y sugiere un perfil clínico intermedio.
- Clúster 2 (amarillo): se sitúa al lado izquierdo del plano y se superpone parcialmente con los clústeres 1 y 4, lo cual indica similitud parcial en algunas variables clínicas.

- Clúster 3 (gris): se localiza en la parte inferior del plano, con limites más definidos respecto a los otros clústeres, lo que sugiere un perfil clínico más específico.
- Clúster 4 (rojo): se encuentra en la parte superior derecha del gráfico, con superposición con los clústeres 1 y 3, lo que sugiere la presencia de patrones clínicos mixtos.

A continuación, en la Figura 12 se presentan los diagramas de caja para cada variable clínica, lo que permite observar la dispersión, tendencia central y presencia de valores atípicos dentro de cada clúster. Asimismo, la Tabla 6 resume los valores promedio de cada variable clínica por grupo. A partir del análisis conjunto de estas representaciones gráficas y estadísticas, se identificaron los siguientes perfiles clínicos diferenciados:

Clúster 1 (Perfil metabólico con obesidad predominante): Este segmento incluye pacientes con edad promedio intermedia (54.2 años) y el valor más elevado de IMC (33.0 kg/m^2), lo que indica obesidad. Los niveles promedio de presión arterial sistólica y diastólica fueron ligeramente elevados (119/74 mmHg) y la glicemia media alcanzó los 166mg/dL. Los diagramas de caja muestran una amplia dispersión en las variables IMC y glicemia, lo que sugiere heterogeneidad clínica dentro del clúster. Este perfil puede estar asociado a un riesgo metabólico elevado relacionado con el exceso de peso.

Clúster 2 (Perfil hipertensivo en adultos mayores): Este clúster agrupa a los pacientes de mayor edad (66.9 años), quienes presentan los valores más altos de presión arterial sistólica (134 mmHg) y diastólica (81.6 mmHg), indicadores clínicos de hipertensión arterial. El IMC promedio es de 27.3 kg/m^2, clasificado como sobrepeso y la glicemia media asciende a 181 mg/dL. Este segmento representa un perfil clínico de riesgo

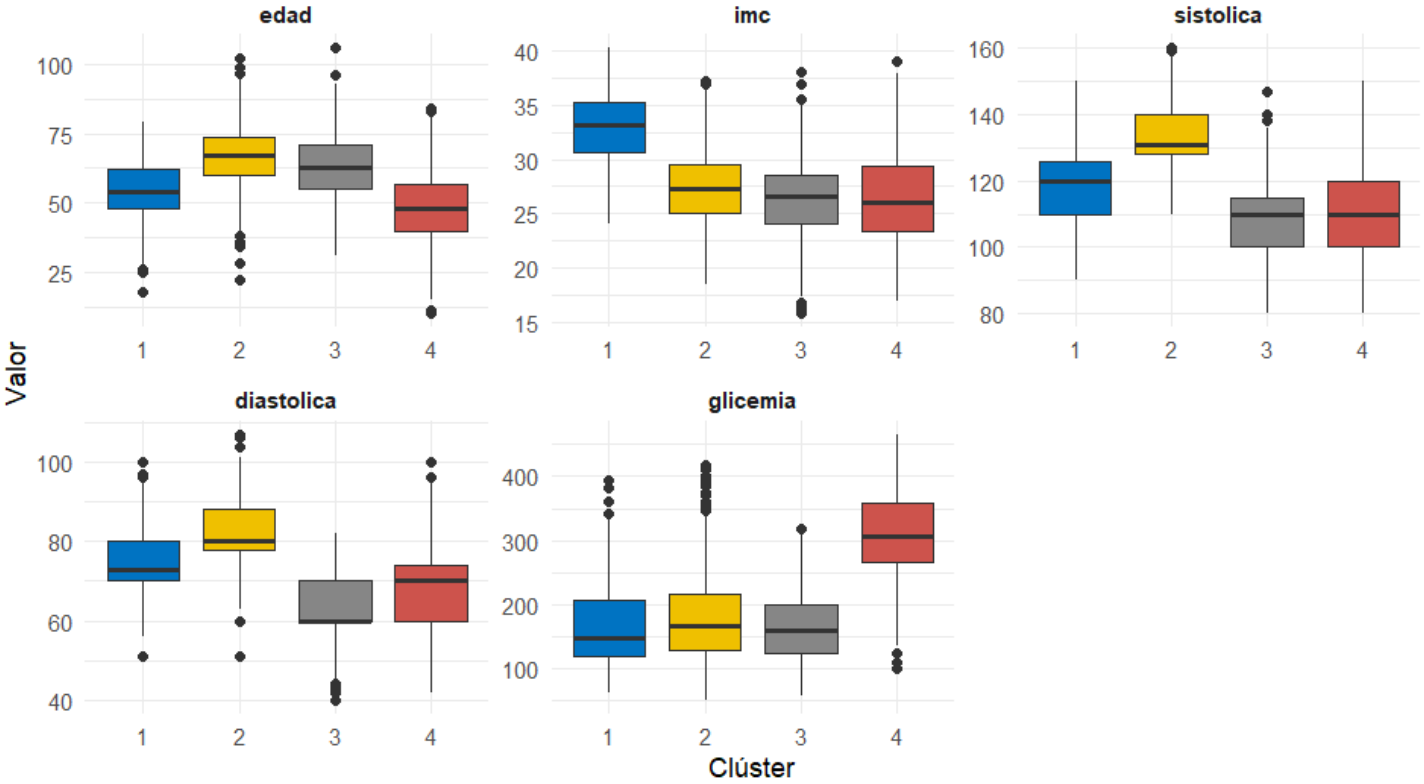
cardiovascular asociado a la edad, caracterizado por hipertensión arterial y desregulación glucémica moderada.

Clúster 3 (Perfil clínico con presión arterial baja y parámetros metabólicos más estables): Los pacientes de este segmento presentan niveles promedio de presión arterial sistólica y diastólica más bajos (107/63 mmHg), un IMC casi normal (26.2 kg/m^2) y una glicemia moderada (164 mg/dL), con una edad promedio de 63 años. La menor variabilidad observada en los diagramas de caja sugiere mayor homogeneidad clínica. Este perfil podría corresponder a pacientes con un estado metabólico más controlado.

Clúster 4 (Perfil hiperglucémico severo en pacientes jóvenes): Este segmento agrupa a pacientes más jóvenes (48.8 años en promedio) con niveles de glicemia considerablemente más altos (309 mg/dL), lo que representa un descontrol glucémico severo. A pesar de ello, se observan valores normales de presión arterial sistólica y diastólica (109/69 mmHg) y un IMC moderado (26.3 kg/m^2). Los diagramas de caja evidencian una alta dispersión de la glicemia, con presencia de múltiples valores atípicos. Este grupo puede reflejar un perfil clínico de alta severidad metabólica en personas jóvenes, posiblemente asociado a diagnostico reciente, falta de adherencia terapéutica o progresión acelerada de la enfermedad.

Figura 12

Boxplots de variables clínicas por clúster



Nota. Los boxplots muestran la distribución, variabilidad y posibles valores atípicos de cada variable en cada clúster.

Tabla 5

Promedio de variables clínicas por clúster

Variable	Clúster 1	Clúster 2	Clúster 3	Clúster 4
Edad (años)	54.2	66.9	63.0	48.8
IMC (kg/m^2)	33.0	27.3	26.2	26.3
Sistólica (mmHg)	119.0	134.0	107.0	109.0
Diastólica (mmHg)	74.2	81.6	63.8	69.0
Glicemia (mg/dL)	166.0	181.0	164.0	309.0

Nota. Los valores promedio se encuentran expresadas en sus unidades originales.

El análisis de clústeres permitió identificar cuatro perfiles diferenciados entre los pacientes con diabetes mellitus tipo 2, lo cual pone de manifiesto la heterogeneidad en la manifestación de la enfermedad. Este enfoque demostró ser adecuado para captar la complejidad y heterogeneidad de la enfermedad, proporcionando información relevante para orientar estrategias de intervención en salud a nivel regional. La segmentación obtenida permite proponer intervenciones más específicas y eficaces, de acuerdo con las características clínicas predominantes en cada segmento identificado:

Clúster 1. Se recomienda enfocar las intervenciones en el control del exceso de peso y la prevención de complicaciones metabólicas. La obesidad observada en este segmento constituye un factor de riesgo importante para el desarrollo de enfermedades cardiovasculares, dislipidemias y nefropatía diabética. Es fundamental implementar programas de educación nutricional, promoción de actividad física y monitoreo del índice de masa corporal, con el fin de reducir la carga metabólica y prevenir complicaciones crónicas de progresión silenciosa.

Clúster 2. Las intervenciones deben orientarse al manejo integral de la hipertensión arterial, especialmente en adultos mayores, a fin de prevenir complicaciones agudas como infarto agudo de miocardio, accidente cerebrovascular e insuficiencia cardiaca. Es fundamental garantizar el control regular de la presión arterial, promover la adherencia terapéutica y evaluar periódicamente el riesgo cardiovascular. Asimismo, se recomienda el seguimiento del estado renal y visual, dad la asociación entre hipertensión y el desarrollo de nefropatía y retinopatía diabética.

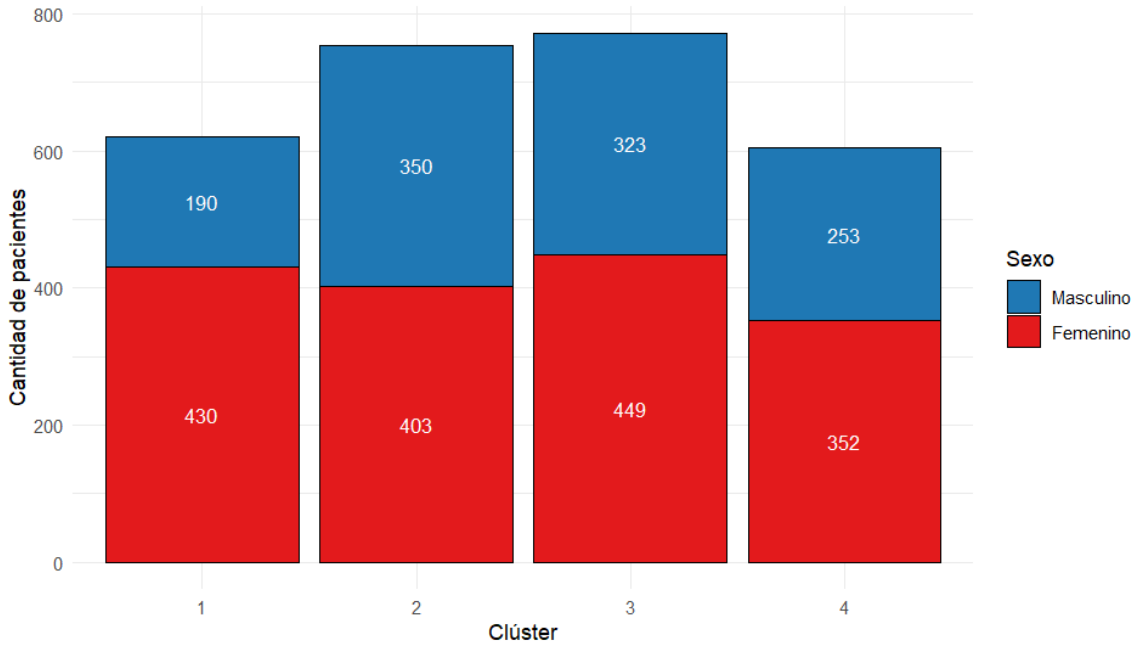
Clúster 3. Aunque presenta parámetros clínicos más estables, es importante mantener el control metabólico y reforzar el seguimiento preventivo. Se sugiere continuar con controles periódicos de glicemia, presión arterial y función renal, junto con evaluaciones oftalmológicas. El

objetivo es prevenir la progresión de la enfermedad hacia complicaciones crónicas mediante estrategias de educación, autocuidado y adherencia sostenida al tratamiento.

Clúster 4. Las intervenciones deben centrarse en el control intensivo de la hiperglucemia y en la detección oportuna de signos de descompensación. Este grupo, conformado mayoritariamente por personas jóvenes con glicemia muy elevada, presenta alto riesgo de desarrollar complicaciones agudas como el estado hiperglucémico hiperosmolar y la cetoacidosis diabética. Se recomienda fortalecer la educación terapéutica, promover el cumplimiento del tratamiento farmacológico e implementar estrategias de seguimiento continuo, especialmente en pacientes recién diagnosticados o con baja adherencia.

Figura 13

Distribución de sexo por clúster



Nota. El grafico de barras apiladas evidencia una mayor proporción de mujeres en todos los clústeres: en el Clúster 1 el 69.4%, en el Clúster 2 el 53.5%, en el Clúster 3 y 4 el 58.2%.

Por otro lado, se analizó la distribución por sexo en los clústeres identificados mediante el algoritmo Fuzzy C-Means. Como se muestra en la Figura 13, en todos los clústeres se observa una mayor proporción de pacientes de sexo femenino. Este patrón es especialmente marcado en el Clúster 1, donde las mujeres representan aproximadamente el 69% del total de pacientes del grupo (430 de 620). El Clúster 3, que agrupa el mayor número de pacientes (772), también presenta una mayoría femenina (449 mujeres frente a 323 varones). En cambio, el Clúster 2 muestra una distribución más equilibrada, aunque aún con predominancia femenina (403 mujeres y 350 varones). Estos resultados sugieren que la segmentación clínica realizada tiende a agrupar más a mujeres en todos los segmentos identificados, lo cual podría reflejar una mayor prevalencia de diagnóstico o atención en mujeres dentro de la población estudiada.

Dado que el algoritmo Fuzzy C-Means permite un agrupamiento difuso, se definieron como casos de pertenencia difusa aquellos pacientes cuya pertenencia a dos clústeres fue superior al umbral de 0.4, lo cual indica una ambigüedad en su asignación a un único segmento. Del total de pacientes analizadas, se identificaron 78 casos con pertenencia difusa. En la Tabla 7 se presentan los primeros diez pacientes con esta característica. A continuación, se detallan algunos casos relevantes:

Paciente 33: Persona de 35 años, con IMC normal (24.38 kg/m^2), presión arterial baja (90/60 mmHg) y glicemia elevada (155 mg/dL). Su pertenencia al Clúster 3 (0.420) perfil de presión baja y parámetros estables como al Clúster 4 (0.404) perfil hiperglucémico severo en jóvenes, sugiere una condición mixta. Aunque sus niveles de glicemia no alcanzan los valores críticos del Clúster 4, comparte con dicho segmento el componente juvenil y la alteración glucémica incipiente. La presión baja y el IMC, sin embrago, son más compatibles con el perfil

estable del Clúster 3. Esto sugiere un estado clínico intermedio entre estabilidad metabólica y riesgo emergente, que puede requerir monitoreo frecuente.

Paciente 79: Persona de 44 años con obesidad (IMC=32 kg/m^2), presión arterial elevada (120/79 mmHg) y glicemia severamente elevada (300 mg/dL). Registra alta pertenencia tanto al Clúster 1 (0.421) perfil de obesidad con riesgo metabólico como al Clúster 4 (0.404) perfil hiperglucémico severo en jóvenes. Esta asignación difusa revela que el paciente combina exceso de peso característico del Clúster 1 con una glicemia que se aproxima al patrón critico observado en el Clúster 4. Clínicamente, representa un caso de riesgo metabólico severo por la combinación de obesidad y descontrol glucémico avanzado en una persona joven, con potencial necesidad de intervención intensiva.

Paciente 138: Persona de 50 años, con obesidad marcada (IMC=33.6 kg/m^2), presión arterial baja (100/60 mmHg) y glicemia moderada (140 mg/dL). Sus grados de pertenencia al Clúster 1 (0.418) y al Clúster 3 (0.401) muestran una combinación entre riesgo metabólico por obesidad y un perfil más estable. Este tipo de ambigüedad refleja que, si bien el paciente presenta un componente de riesgo vinculado al peso, sus niveles de glicemia y presión arterial aún se mantienen dentro de márgenes aceptables. Puede tratarse de un caso de riesgo metabólico incipiente, en una fase que permite aún acciones preventivas eficaces.

Paciente 253: Persona de 58 años con sobrepeso (IMC=27.8 kg/m^2), presión arterial normal (120/80 mmHg) y glicemia dentro de rangos aceptables (126 mg/dL). Su pertenencia elevada al Clúster 1 (0.429) y al Clúster 2 (0.449) sugiere una superposición entre un perfil metabólico con sobrepeso y un perfil hipertensivo en adultos mayores. Aunque su presión arterial no esta elevada, la edad y el IMC lo vinculan parcialmente con el perfil del Clúster 2. Este paciente

se ubica en una zona clínica de riesgo cardiovascular moderado, compatible con las primeras etapas de transición hacia un perfil más hipertensivo o metabólicamente alterado.

Tabla 6

Pacientes con pertenencia difusa en clústeres

Paciente ID	Edad	IMC	Sistólica	Diastólica	Glicemia	Pertenencia Cluster1	Pertenencia Cluster2	Pertenencia Cluster3	Pertenencia Cluster4
33	35	24.38	90	60	155	0.127	0.049	0.420	0.404
76	43	19.47	100	70	172	0.093	0.077	0.402	0.428
79	44	32.01	120	79	300	0.421	0.107	0.069	0.404
135	50	23.8	100	70	198	0.053	0.027	0.470	0.450
138	50	33.59	100	60	140	0.418	0.044	0.401	0.137
239	57	20.96	80	60	263	0.068	0.047	0.416	0.469
253	58	27.83	120	80	126	0.429	0.449	0.095	0.027
271	59	29.47	110	70	112	0.405	0.043	0.525	0.028
273	59	23.05	120	60	247	0.052	0.056	0.488	0.404
308	61	29.78	110	70	139	0.401	0.033	0.543	0.024

Nota. La inclusión de pacientes con pertenencia difusa permite explorar situaciones complejas y transicionales que no podrían ser captadas por métodos clásicos.

Estos casos refuerzan la utilidad del enfoque difuso en contextos clínicos, permitiendo identificar pacientes con características intermedias entre segmentos. Esta información puede apoyar estrategias de atención más individualizada, particularmente en aquellos casos que no encajan de forma clara en un único perfil de riesgo.

En definitiva, el algoritmo Fuzzy C-Means, se presenta como una herramienta estadística útil y eficaz para la segmentación de pacientes con diabetes mellitus tipo 2. Su aplicación permite identificar perfiles clínicos específicos y contribuir a un manejo más focalizado de la enfermedad en el contexto regional, abriendo nuevas posibilidades para la planificación sanitaria y la toma de decisiones basadas en datos.

5.1.5 Discusiones

Los resultados obtenidos en este estudio, mediante la segmentación de pacientes con diabetes mellitus tipo 2 (DM2) registrados en la DIRESA Cusco entre 2019 y 2022, a través del algoritmo Fuzzy C-Means (FCM), revelan la existencia perfiles clínicos diferenciados, lo cual indica la heterogeneidad de la población diabética en la región Cusco. Se identificaron cuatro segmentos con patrones clínicos específicos, permitiendo también reconocer pacientes con pertenencia difusa entre clústeres, lo que sugiere la presencia de casos clínicamente intermedios o transicionales. Este enfoque y los resultados obtenidos, guardan relación con algunos estudios realizados a nivel internacional y nacional, los cuales también evidencian heterogeneidad entre los pacientes con DM2.

A nivel internacional, Carrillo-Larco et al. (2021) utilizo el algoritmo k-means en multiples paises de America Latina y el Caribe para identificar cuatro perfiles distintos de pacientes con DM2, considerando variables como edad, sexo, el indice de masa corporal (IMC), el perímetro de cintura (CC), la presion arterial sistólica (PAS), la presion arterial diastólica (PAD) y los antecedentes familiares de diabetes. Sin embargo, la presente investigacion aporta una novedad al aplicar un enfoque difuso, permitiendo captar la pertenencia parcial de los pacientes a más de un clúster, revelando perfiles clinicos intermedios o mixtos. Estos perfiles mixtos no fueron detectados es el estudio de Carrillo-Larco et al. (2021) debido a su enfoque de agrupamiento duro, donde cada paciente fue asignado exclusivamente a un unico clúster. Asi mismo, a diferencia de Carrillo-Larco et al. (2021), quienes trabajaron con encuestas poblacionales y datos autodeclarados, el presente estudio se basa en registros clinicos reales de pacientes atendidos en establecimientos de salud, lo cual otroga mayor solidez y validez a los resultados obtenidos.

Por otro lado, el estudio de Lomo et al. (2023) tiene una mayor similitud metodologica con el presente estudio, dado que tambien implementa el algoritmo Fuzzy C-Means, asi como tecnicas de validaciíon (indice de Davies-Bouldin). Sin embargo, su analisis fue limitado a 447 pacientes en un solo hospital de Indonesia, mientras que nuestro estudio considera una poblacion mayor de 2750 de la DIRESA Cusco, compuesto por diferentes establecimientos de salud, abarcando un peridodo de 4 años. En cuanto a los clusteres identificados, Lomo et al. (2023) identifico perfiles asociados al pronostico de supervivencia, mientras que nuestra investigacion se centro en variables de riesgo y caracteristicas clínicas. A pesar de estas diferencias, ambos estudios coinciden en identificar grupos con mayor carga de riesgo y otros con condiciones mas estables, auque nuestro enfoque no abordo directamente variables como mortalidad y tratamiento famacológico.

Asimismo, Marhamah et al. (2023) en su trabjo de investigacion tambien utilizó el algoritmo Fuzzy C-Means para agrupar pacientes con DM2 según los niveles de riesgo, identificando dos clusteres de riesgo alto y bajo. En cambio, en nuestro estudio se identificaron perfiles intermedios y tambien de riesgo alto, mostrando que existe un continuo de riesgo en el cual algunos pacientes comparten caracteristicas de más de un clúster. Esto podria reflejar de manera mas realista la situacion clinica de muchos pacientes con DM2.

A nivel nacional, el estudio realizado por Bernabe-Ortiz (2022) en Lima, tambien utilizó el algoritmo k-means para identificar patrones de multimorbilidad cronicas en pacientes con DM2. Al igual que en el presente estudio, se logro agrupar a los pacientes en segmentos diferenciados, como aquellos con diabetes sin otras comorbilidades, con obesidad, con enfermedades cardiovasculares y otro grupo con condiciones cronicas diversas. Aunque su enfoque se centro en la coexistencia de multiples patologias y no exclusivamente en la caracterizacion clinica de la DM2, su identificacion de grupos clinicamente diferenciables guarda relacion con los perfiles

obtenidos en este estudio. Sin embargo, cabe señalar que el uso de un algoritmo no difuso y la inclusion de multiples enfermedades limita la comparacion directa con nuestro estudio.

En conjunto, los resultados del presente estudio amplian la literatura existente al aplicar el algoritmo FCM utilizando variables clinicas. La posibilidad de identificar perfiles intermedios o difusos, representa un avance metodologico y practico. Este enfoque permite una compresion mas realista y detallada de la poblacion diabetica, especialmente util en contextos como el Perú, donde los recursos para atencion personalizada son limitados y se requiere priorizar acciones según caracteristicas clinicas especificas.

Por lo tanto, se reafirma que el uso de tecnicas de clustering difuso no solo ofrece ventajas analiticas frente a los metodos tradicionales de clustering, sino puede aportar un fundamento para la toma de desiciones clinicas. En particular, este enfoque permitiria adaptar las estrategias de tratamiento y seguimiento según la complejidad del perfil de cada paciente, especialmente en aquellos con pertenencia difusa entre clusteres, quienes podrian beneficiarse de una atencion mas personalizada según sus caracteristicas clinicas.

CONCLUSIONES

Los resultados del presente estudio permitieron caracterizar clínicamente a los pacientes con diabetes mellitus tipo 2 registrados en la DIRESA Cusco entre 2019 y 2022, mediante la segmentación realizada con el algoritmo Fuzzy C-Means. Este enfoque permitió identificar perfiles diferenciados, abordando de manera eficaz la heterogeneidad de esta enfermedad crónica en el contexto regional.

El número óptimo de clústeres fue determinado a partir de índices de validez interna para agrupamiento difuso, lo que permitió establecer una estructura de cuatro segmentos clínicamente diferenciados, validando la capacidad del enfoque difuso para representar la complejidad subyacente de los datos y evitando una agrupación rígida que podría ocultar la transición entre estados clínicos.

Los perfiles clínicos identificados presentaron diferencias en variables como glicemia, presión arterial, índice de masa corporal y edad. Se identificaron cuatro patrones clínicos: un perfil con obesidad predominante (clúster 1), un perfil hipertensivo en adultos mayores (clúster 2), un segmento con parámetros metabólicos más estables (clúster 3) y un perfil hiperglucémico severo en personas jóvenes (clúster 4). Esta segmentación puede orientar intervenciones más específicas, alineadas a las necesidades de cada segmento.

Finalmente, se identificaron pacientes con pertenencia difusa elevada a más de un clúster, reflejando la existencia de perfiles transicionales. Estos casos ofrecen información relevante para el seguimiento clínico, al representar etapas intermedias de la enfermedad que requieren una atención más personalizada. En este sentido, el enfoque difuso ofrece una ventaja relevante frente a métodos tradicionales de agrupamiento, al representar de forma más realista la complejidad clínica de los pacientes y facilitar intervenciones medicas más personalizadas.

RECOMENDACIONES

El presente estudio se centró principalmente en variables numéricas de tipo clínico, lo cual limito parcialmente la caracterización integral de los pacientes. Se sugiere incorporar en futuras investigaciones variables categóricas relevantes como el nivel de instrucción, tipo de seguro, adherencia al tratamiento, hábitos alimenticios y actividad física. Esto permitiría enriquecer la segmentación obtenida y comprender con mayor profundidad los factores asociados a cada perfil.

Asimismo, dado que el análisis se limitó a datos de la DIRESA Cusco, se sugiere replicar este tipo de estudios en otras regiones del país o a nivel nacional. Esto permitiría identificar patrones geográficos y socioeconómicos diferenciados, aportando evidencia útil para políticas públicas regionales y estrategias de intervención más focalizadas.

Por otro lado, se identificó un alto porcentaje de datos faltantes en varias variables de interés en la base de datos. En ese sentido, se recomienda fortalecer los mecanismos de recolección y registro de información en los establecimientos de salud, con el fin de mejorar la calidad de datos disponibles para futuras investigaciones.

Finalmente, considerando el potencial de métodos estadísticos como Fuzzy C-Means para apoyar en la toma de decisiones en salud, se sugiere promover la capacitación continua del personal médico y estadístico en análisis multivariado y herramientas como el lenguaje de programación R. El fortalecimiento de estas capacidades contribuirá al uso eficaz de los datos disponibles y a una mejor gestión de los servicios de salud.

BIBLIOGRAFÍA

Abonyi, J., & Feil, B. (2007). *Cluster Analysis for Data Mining and System Identification.* Birkhäuser.

Aldás, J., & Uriel, E. (2017). *Análisis multivariante aplicado con R* (2ª ed. ed.). Paraninfo.

Bernabe-Ortiz, A., Borjas-Cavero, D., Páucar-Alfaro, J., & Carrillo-Larco, R. (2022). Multimorbidity Patterns among People with Type 2 Diabetes Mellitus: Findings from Lima, Peru. *International Journal of Environmental Research and Public Health, 19*(15). doi:https://doi.org/10.3390/ijerph19159333

Bezdek, J. (1973). Cluster Validity with Fuzzy Sets. *Journal of Cybernetics, 3*(3), 58-73. doi:https://doi.org/10.1080/01969727308546047

Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2-3), 191-203. doi:https://doi.org/10.1016/0098-3004(84)90020-7

Bravo-Zúñiga, J., & Solari-Yokota, J. (2024). Nefropatía diabética. Historia natural, diagnóstico precoz y tratamiento. *Rev Soc Peru Med Interna, 37*(2), pp. 102-113. doi:https://doi.org/10.36393/spmi.v37i2.851

Cannon, R., Dave, J., & Bezdek, J. (1986). Efficient Implementation of the Fuzzy c-Means Clustering Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*(2), 248-255. doi:10.1109/TPAMI.1986.4767778

Carrillo-Larco, R., Castillo-Cara, M., Anza-Ramírez , C., & Bernabé-Ortiz, A. (2021). Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in Latin America and the Caribbean. *BMJ open diabetes research & care, 9*(1). doi:https://doi.org/10.1136/bmjdrc-2020-001889

Catón, V., Barrón, R., & De Lobera Martínez, N. (2024). Hipoglucemia. *FMC-Formación Médica Continuada en Atención Primaria, 31*(5), pp. 252-256. doi:https://doi.org/10.1016/j.fmc.2023.09.009

Celebi, M. E. (Ed.). (2015). *Partitional Clustering Algorithms.* Springer. doi:https://link.springer.com/book/10.1007/978-3-319-09259-1

Centro Nacional de Epidemiología, Prevención y Control de Enfermedades(CDC Perú). (2024). *BOLETÍN EPIDEMIOLÓGICO DEL PERÚ SE 12.* Obtenido de https://www.dge.gob.pe/epipublic/uploads/boletin/boletin_202412_29_153641.pdf

Cobos-Palacios, L., Sampalo, A., & Carmona, M. (2020). Neuropatía diabética. *Medicine-Programa de Formación Médica Continuada Acreditado, 13*(16), pp. 911-923. doi:https://doi.org/10.1016/j.med.2020.09.013

Dave, R. (1996). Validating fuzzy partitions obtained through c-shells clustering. *Pattern recognition letters, 17*(6), pp. 613-623. doi:https://doi.org/10.1016/0167-8655(96)00026-8

Floreano Solano, L. M., Paccha Tamay, C. L., Gordillo Quizhpe, I., & Zambrano Villamar, V. R. (2017). Factores de riesgo asociados a diabetes e hipertensión. *Conference Proceedings., 1*(1). Obtenido de http://investigacion.utmachala.edu.ec/proceedings/index.php/utmach

Fuks, A. G., & Vaisberg, M. (2022). Cetoacidose Diabética. *Anais da Academia Nacional de Medicina, 193*(1), pp. 74-83. Obtenido de https://www.anm.org.br/wp-content/uploads/2022/08/AANM2022v193n1p74-83.pdf

Giordani , P., Ferraro, M., & Martella, F. (2020). *An Introduction to Clustering with R.* Springer. doi:https://doi.org/10.1007/978-981-13-0553-5

Gower, J. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics, 27*(4), pp. 857-871.

Honório, D., Pereira, R., Aguiar, B., Rodrigues, B., & Dourado, J. (2024). Estado hiperglicêmico hiperosmolar: desafios diagnósticos e estratégias terapêuticas avançadas. *Brazilian Journal of Health Review, 7*(4), pp. 1-12. doi:https://doi.org/10.34119/bjhrv7n4-407

Höppner , F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition.* John Wiley & Sons.

Instituto Nacional de Estadística e Informática[INEI]. (2024). *PERÚ: ENFERMEDADES NO TRANSMISIBLES Y TRANSMISIBLES, 2023.* Obtenido de https://proyectos.inei.gob.pe/files/WEB_ENDES/SALUD/2023/ENFERMEDADES_ENDES_2023.pdf

International Diabetes Federation [IDF]. (2021). IDF Diabetes Atlas10th. Obtenido de https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

Kaushik, K., & Hemanta, K. (2013). Extension of the fuzzy c means clustering algorithm to fit with the composite graph model for web document representation. *International Journal of Cognitive Research in science, engineering and education, 1*(2), pp. 173-179.

Kwon, S. (1998). Cluster validity index for fuzzy clustering. *Electronics letters, 34*(22), pp. 2176-2177. doi:https://doi.org/10.1049/el:19981523

Lomo, S., Darmawan, E., & Sugiyarto. (2023). Cluster analysis of type IIDiabetes Mellitus Patients with the Fuzzy C-means method. *Annals of Mathematical Modeling, 3*(1), pp. 24-31. doi:https://doi.org/10.33292/amm.v3i1.28

Marhamah, Surono, S., & Darmawan, E. (2023). The Risk Cluster in Type 2 Diabetes Mellitus Based on Risk Parameters Using FuzzyC-Means Algorithm. *Science and Technology Indonesia, 8*(1), pp. 17-24. doi:https://doi.org/10.26554/sti.2023.8.1.17-24

Naranjo Hernández, Y. (2016). La diabetes mellitus: un reto para la Salud Pública. *6*(1), aprox. 1 p. Obtenido de http://revfinlay.sld.cu/index.php/finlay/article/view/399

Organización Mundial de la Salud [OMS]. (2016). Informe mundial sobre la diabetes. Obtenido de https://iris.who.int/handle/10665/254649

Organización Panamericana de la Salud [OPS]. (2021). *Pacto Mundial contra la Diabetes Implementación en la Región de las Américas.* Obtenido de https://iris.paho.org/handle/10665.2/54682

Organización Panamericana de la Salud [OPS]. (2023). *Panorama de la diabetes en la Región de las Américas.* Obtenido de https://doi.org/10.37774/9789275326336

Pérez López, C. (2004). *Técnicas de Análisis Multivariante de Datos.* PEARSON EDUCACIÓN, S.A.

Reddy, C., & Aggarwal, C. (Edits.). (2014). *Data Clustering: Algorithms and Applications.* CRC Press.

Rojas Díaz, J., Chavarro Porras, J. C., & Moreno Laverde, R. (2008). Tecnicas de logica difusa aplicadas a la mineria de datos. *Scientia Et Technica, XIV*(40), 1-6.

Rokach, L. (2024). *Cluster Analysis: A Primer Using R.* World Scientific. Obtenido de https://www.worldscientific.com/worldscibooks/10.1142/13968#t=suppl

Sugandh, F., Chandio, M., Raveena , F., Kumar , L., Karishma, F., Khuwaja, S., . . . Kumar, S. (2023). Advances in the Management of Diabetes Mellitus: A Focus on Personalized Medicine. *Coreus, 15*(8). doi:10.7759/cureus.43697

Tan , P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2ª ed. ed.). Pearson.

Tang, Y., Sun, F., & Sun, Z. (2005). Improved validation index for fuzzy clustering. *In Proceedings of the 2005, American Control Conference, 2005.* Portland, OR, USA: IEEE . doi:10.1109/ACC.2005.1470111

Vigilancia Epidemiológica de Diabetes - HRC. (2024). Hospital Regional del Cusco reporte de vigilancia epidemiologica de diabtes 2016-2024 (SE N° 1-44).

Wierzchoń, S., & Kłopotek, M. (2018). *Modern Algorithms of Cluster Analysis.* Springer Internacional Publishing. doi:https://doi.org/10.1007/978-3-319-69308-8

Xie, F., Chan, J., & Ma, R. (2018). Precision medicine in diabetes prevention, classification and management. *J Diabetes Investig, 9*(5), 998-1015. doi:10.1111/jdi.12830

Xie, X., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(8), pp. 841-847. doi:10.1109/34.85677

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*(3). doi:https://doi.org/10.1016/S0019-9958(65)90241-X

Zelterman, D. (2015). *Applied Multivariate Statistics with R.* Springer.

ANEXOS

A. Matriz de consistencia

PROBLEMA GENERAL	OBJETIVO GENERAL	HIPÓTESIS GENERAL	VARIABLES	METODOLOGIA
¿Cómo segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 mediante el algoritmo de clúster Fuzzy C-Means en la región Cusco, durante el periodo 2019-2022?	Caracterizar los segmentos clínicos identificados entre los pacientes con diabetes mellitus tipo 2 mediante el algoritmo de clúster Fuzzy C-Means en la región Cusco, durante el periodo 2019-2022.	El algoritmo de clúster Fuzzy C-Means permite segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 en la región Cusco, durante el periodo 2019-2022.	**Variable de estudio:** Perfil clínico de los pacientes con diabetes mellitus tipo 2. **Variables intervinientes:** Edad, peso, talla, índice de masa corporal (IMC), presión arterial sistólica, presión arterial diastólica y glicemia.	**ENFOQUE:** Cuantitativo **TIPO:** Aplicada **NIVEL:** Descriptiva **DISEÑO:** No experimental, de corte transversal **POBLACION:** 2750 pacientes diagnosticados con diabetes mellitus tipo 2 registrados por la Dirección Regionales de Salud (DIRESA) Cusco durante el periodo 2019-2022. **MUESTRA:** No se estableció una muestra especifica, dado que se trabajó con la totalidad de la población disponible en la base de datos. **TECNICAS E INSTRUMENTOS PARA LA RECOLECCION DE DATOS:** Fuente secundaria, registros clínicos anonimizados proporcionados por el Centro Nacional de Epidemiologia, Prevención y Control de enfermedades (CDC Perú).
PROBLEMAS ESPECÍFICOS	**OBJETIVOS ESPECIFICOS**	**HIPÓTESIS ESPECÍFICOS**		
1. ¿Cuál es el número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2, según los indicadores de validación interna aplicados al algoritmo Fuzzy C-Means? 2. ¿Qué diferencias presentan los perfiles clínicos de los segmentos identificados mediante el algoritmo Fuzzy C-Means? 3. ¿Qué características clínicas presentan los pacientes con pertenencia difusa elevada en los segmentos identificados mediante el algoritmo Fuzzy C-Means?	1. Determinar el número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2, mediante indicadores de validación interna aplicados al algoritmo Fuzzy C-Means. 2. Comparar los perfiles clínicos de los segmentos identificados mediante el algoritmo Fuzzy C-Means. 3. Analizar las características clínicas de los pacientes con pertenencia difusa elevada en los clústeres identificados mediante el algoritmo Fuzzy C-Means.	1. El número óptimo de clústeres para segmentar clínicamente a los pacientes con diabetes mellitus tipo 2 en la región Cusco se encuentra entre dos y cinco, según los índices de validación interna aplicados al algoritmo Fuzzy C-Means. 2. Cada segmento identificado mediante el algoritmo Fuzzy C-Means agrupa a pacientes con perfiles clínicos diferenciados. 3. Los pacientes con pertenencia difusa elevada presentan combinaciones de características clínicas propias de más de un segmento identificado mediante el algoritmo Fuzzy C-Means.		

B. Respuesta de Solicitud de Acceso a la Información Pública 24-012895

PERÚ Ministerio de Salud | DESPACHO VICEMINISTERIAL DE SALUD PÚBLICA | CENTRO NACIONAL DE EPIDEMIOLOGIA PREVENCION Y CONTROL DE

PERÚ MINSA — Firmado digitalmente por MUNAYCO ESCATE Cesar Vladimir FAU 20131373237 hard. Cargo: Director General. Motivo: Soy el autor del documento. Fecha: 16.01.2025 15:41:29 -05:00

"Decenio de la Igualdad de Oportunidades para Mujeres y H
"Año de la recuperación y consolidación de la economía peruana"

Jesus Maria, 16 de Enero del 2025

MEMORANDUM N° D000137-2025-CDC-MINSA

Para	:	**JEAN EDWIN CASTAÑEDA RIVERA** DIRECTOR EJECUTIVO OFICINA DE TRANSPARENCIA Y ANTICORRUPCION
De	:	**CESAR VLADIMIR MUNAYCO ESCATE** DIRECTOR GENERAL CENTRO NACIONAL DE EPIDEMIOLOGIA PREVENCION Y CONTROL DE ENFERMEDADES
Asunto	:	TRASLADA SOLICITUD DE INFORMACIÓN EN VIRTUD DE LA LEY N° 27806 (SAIP WEB N° 24-012895).
Referencia	:	NOTA INFORMATIVA N° D011120-2024-SG-OTRANS-MINSA N° Exp : SG-OTRANS20240013600
Fecha	:	Jesus Maria, 16 de enero de 2025

Es grato dirigirme a usted para expresar mi cordial saludo, y en atención al documento de la referencia a través del cual su despacho puso en conocimiento de este Centro Nacional la solicitud de acceso a la información pública (SAIP N° 24-012895), ingresado por el ciudadano **JESÚS RENÁN HUACCANQUI CONDORI**, referida a lo siguiente:

> "BASE DE DATOS SOBRE "DIABETES MELLITUS TIPO 2", CORRESPONDIENTE AL PERIODO 2019, 2020, 2021, 2022, 2023; LAS VARIABLES QUE REQUIERO SON: EDAD, INSTRUCCION, PESO, TALLA, IMC, SISTOLICA, DIASTOLICA, GLICEMIA, COL_TOTAL, TRIGLICERIDOS, ...ETC. ESTO CON EL OBJETIVO DE CLASIFICAR A LOS PACIENTES EN RIESGO ALTO Y BAJO, PARA ASI TOMAR PRECAUCIONES Y/O MEDIDAS. ADJUNTAR EL DICCIONARIO DE LOS DATOS, E INFORMCION DE LA POBLACION OBTENIDA DE LOS DATOS."

Al respecto, el área técnica poseedora de la información, en base a las variables solicitadas, ha emitido respuesta a fin de entregar la información al solicitante, la cual se adjunta.

Sin otro particular, sea propicia la oportunidad para hacerle llegar las muestras de mi especial consideración y estima personal.

Atentamente,

Documento firmado digitalmente

CESAR VLADIMIR MUNAYCO ESCATE
DIRECTOR GENERAL
CENTRO NACIONAL DE EPIDEMIOLOGIA PREVENCION Y CONTROL DE ENFERMEDADES

(CME)

cc.:

C. Ficha de registro de datos

diresas	ablecimier	sexo	pais_nac	fecha_nac	edad	nstruccion	ubigeo_na	ubigeo_res	asegurado	tseguro	año	semana	fecha_cap	tcaso	tdiabetes	peso	talla	imc	pcint	sistolica	diastolica	glicemia	hemoglic	ldl	col_total	rigliceridos	hdl	ratamiento	netformina	ulfolinurea	nhibidores	sul_humar	sul_analog	glitazonas	glifozinas	agonistas	mplimient
CUSCO	HOSPITAL	2		1963-07-04	56	4	080901	080901	1	1	2019	38	2019-09-20	2	2	62	1.55	25.81	95	112	72	140		110				1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1953-01-28	66	4	080901	080901	1	1	2019	25	2019-06-18	2	2	53	1.53	22.64	91	138	62	132	6.9	177				1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1997-07-12	22	4		080601	1	1	2019	45	2019-11-05	1	2	58	1.52	25.1	78	135	100	112						1	0	0	0	0	1	0	0	0	3
CUSCO	HOSPITAL	1		1997-08-24	22	5	080105	080105	1	1	2019	43	2019-10-24	2	2	76	1.75	24.82	92	130	70							1	0	0	0	0	1	0	0	0	1
CUSCO	HOSPITAL	2		1993-09-17	25	4	080604	080604	1	1	2019	4	2019-01-22	1	2	98.5	1.51	43.2	123	130	80	213						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1994-01-01	25	4	150115	080901	1	1	2019	42	2019-10-16	2	2	65	1.61	25.08	98	106	59	269		160				0	0	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1992-03-17	26	3	080101	080101	1	1	2019	6	2019-02-07	2	2	50	1.5	22.22	80	110	60	320						1	1	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	2		1993-08-24	26	4	080105	080105	1	1	2019	45	2019-11-06	2	2	63	1.49	28.38	90	100	70	81						1	1	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	1		1993-09-08	26	4	080601	080601	1	1	2019	42	2019-10-15	2	2	57.5	1.61	22.18	90	110	70							1	0	0	0	0	1	0	0	0	1
CUSCO	HOSPITAL	2		1991-10-16	27	4	081105	081105	1	1	2019	11	2019-03-12	1	2	57.5	1.67	20.62	72	110	90	324						1	1	0	0	1	0	0	0	0	3
CUSCO	HOSPITAL	2		1991-11-28	27	4	080601	080601	1	1	2019	5	2019-01-28	2	2	63	1.51	27.63	92	110	70	344	7.8					1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1991-06-26	27	1	080901	080901	1	1	2019	12	2019-03-20	2	2	79	1.56	32.46	93	84	57	299		199				1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1991-10-19	28	4	080101	080101	1	1	2019	42	2019-10-17	1	2	80	1.63	30.11	97	130	80							1	1	0	0	0	1	0	0	0	1
CUSCO	HOSPITAL	1		1990-11-15	28	2	080101	080101	1	1	2019	9	2019-02-26	2	2	49	1.5	21.78	80	116	74	225						1	1	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	1		1990-03-09	29	8	080601	080601	1	1	2019	23	2019-06-06	1	2	96	1.59	37.97	111	110	70	358						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1989-07-23	29	4	080108	080106	1	1	2019	14	2019-04-04	2	2	64	1.44	30.86	94	120	70	233						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1988-05-11	30	4	080901	150101	1	1	2019	7	2019-02-10	2	2	60	1.6	23.44	85	110	70	302						1	0	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	2		1988-10-26	31	2		081102	1	1	2019	43	2019-10-22	1	2	52	1.43	25.43	82	120	60	168						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1987-10-01	31	2	081211	081105	1	1	2019	23	2019-06-04	2	2	57	1.5	25.33	88	130	80	324						1	1	0	0	0	1	0	0	0	3
CUSCO	HOSPITAL	2		1988-05-08	31	3	080105	080105	1	1	2019	46	2019-11-13	2	2	42	1.46	19.7	73	100	70	298						1	1	1	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1987-06-24	32	6	080101	080101	2	0	2019	12	2019-03-23	1	2	76	1.71	25.99	92	120	80	323		115		232		1	1	0	0	1	0	0	0	0	3
CUSCO	HOSPITAL	2		1986-05-02	32	3	080901	080901	1	1	2019	6	2019-02-07	1	2	75	1.5	33.33		118	76	180						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1987-10-13	32	4	080303	080303	1	1	2019	45	2019-11-08	2	2	46	1.52	19.91	72	100	80	355	11.5	23		83		1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	1		1985-12-15	33	4	080105	080105	1	1	2019	41	2019-10-10	1	2	108	1.72	36.51	120	120	70	127						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1985-05-07	33	4	080101	080101	1	6	2019	8	2019-02-22	2	2	53	1.53	22.64	72	110	70	102						1	1	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	1		1985-06-01	33	4	080903	080903	1	1	2019	21	2019-05-21	2	2	68	1.66	24.68	81	90	60	115						1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1985-04-11	33	3	080105	080105	1	1	2019	10	2019-03-05	2	2	101	1.68	35.79	118	110	70	286						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1986-03-06	33	1	081212	080107	1	1	2019	17	2019-04-26	2	2	56	1.51	24.56	95	125	70	212						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1985-05-27	34	3		080101	1	1	2019	50	2019-12-14	1	2	71	1.65	26.08	95	100	60	90						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1985-04-04	34	1	081210	081210	1	1	2019	20	2019-05-14	1	2	51	1.45	24.26	81	135	70	305						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1985-03-19	34	2	210101	080101	1	1	2019	31	2019-07-30	2	2	57	1.48	26.02	79	140	90	160						1	1	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	2		1984-11-11	35	4	080108		1	1	2019	46	2019-11-13	1	2	41	1.48	18.72	70	100	70	328						1	0	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	1		1984-07-23	35	4	080101	080101	1	1	2019	42	2019-10-13	2	2	68	1.67	24.38	88	90	60	155						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1984-09-11	35	4	080105	080105	1	1	2019	37	2019-09-13	2	2	66.5	1.61	25.65	94	120	70	147						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1981-11-28	37	4	080101	080101	1	1	2019	45	2019-11-05	1	2	91	1.71	31.12	110	110	60	298	13.9	165				1	1	0	0	1	0	0	0	0	3
CUSCO	HOSPITAL	2		1981-04-23	37	2	081302	080101	1	1	2019	8	2019-02-19	1	2	75	1.51	32.89	109	100	60	168	12					1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1981-06-21	37	4	080205	080205	1	1	2019	2	2019-01-11	1	2	65	1.51	28.51	88	100	60	408						1	1	1	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1983-03-18	37	2		080105	1	1	2019	43	2019-10-26	2	2	61	1.5	27.11	91	120	70	433						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1982-01-19	37	4	080101	080101	1	1	2019	5	2019-02-01	2	2	62	1.55	25.81	82	110	70	225						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1980-07-26	38	8	080101	080108	1	1	2019	12	2019-03-20	2	2	131	1.95	34.45	122	120	70	357						1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1980-11-05	38	6	080101	080101	1	1	2019	36	2019-09-04	2	2	73	1.55	30.39	97	110	70	92						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1980-12-02	38	2	080401	080401	1	1	2019	42	2019-10-19	2	2	70	1.47	32.39	130	110	60	107						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1980-03-14	38	2	080608	080902	1	1	2019	7	2019-02-11	2	2	85	1.46	39.88	114	110	90	135						1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1980-04-29	39	1	081007	081007	1	1	2019	44	2019-10-30	1	2	62	1.48	28.31	95	110	60	93						1	1	0	0	1	0	0	0	0	3
CUSCO	HOSPITAL	2		1980-01-01	39	1	080805	080805	1	1	2019	14	2019-04-02	1	2	53	1.51	23.24		90	60	307						0	0	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1979-12-26	39	4	081009	081009	1	1	2019	42	2019-10-18	1	2	70.5	1.45	33.53	101	115	75	244						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1979-10-07	39	3	081205	081205	1	1	2019	23	2019-06-07	2	2	58	1.5	25.78	80	120	70	200						1	0	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	1		1980-05-09	39	4	080607	080601	1	1	2019	22	2019-06-01	2	2	46	1.55	19.15	80	85	44	156						1	0	0	0	0	1	0	0	0	1
CUSCO	HOSPITAL	1		1979-06-09	40	4	030503	080101		0	2019	44	2019-10-29	1	2	71	1.58	28.44	94	100	80	159	12.3	85		104		1	1	1	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1978-03-13	40	1	081005	080106	1	1	2019	1	2019-01-04	1	2	75	1.53	32.04	104	110	60	289						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1978-08-05	40	2	080408	080408	1	1	2019	8	2019-02-18	1	2	79	1.49	35.58	95	100	60	133						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1979-07-11	40	4	080901	080901	1	1	2019	42	2019-10-16	1	2	79	1.65	29.02	99	112	68	140	5.6					0	0	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1979-02-11	40	4	080407	080407	1	1	2019	18	2019-04-29	2	2	48	1.53	20.5	67	90	60	353						1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1978-08-28	40	4	080101	080101	1	1	2019	6	2019-02-04	2	2	40	1.42	19.84	66	90	50	361						0	0	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1979-02-28	40	3		080401	1	1	2019	51	2019-12-19	2	2	55	1.47	25.45	89	120	70	386						1	1	1	0	0	1	0	0	0	2
CUSCO	HOSPITAL	1		1978-12-18	40	7	080705	080705	2	0	2019	15	2019-04-11	2	2	74	1.71	25.31	85	115	75	202						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1979-02-04	40	2	081202	081202	1	1	2019	21	2019-05-20	2	2	88	1.61	33.95	94	120	60	368						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1978-07-22	40	2	081212	081212	1	1	2019	12	2019-03-18	2	2	48.5	1.6	18.95	80	80	60	351						1	0	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1979-01-13	40	4	080702	080702	1	1	2019	10	2019-03-04	2	2	57	1.56	23.42	86	90	60	214						1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1977-11-09	41	3	080901	080901	1	1	2019	22	2019-05-27	1	2	93	1.6	36.33	102	138	77	328	9.3					1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1977-08-14	41	3	080104	080104	1	1	2019	10	2019-03-06	2	2	56	1.58	22.43	78	118	70	225						1	1	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	1		1978-10-08	41	4	080101	080101	1	1	2019	48	2019-11-27	2	2	71	1.68	25.16	92	110	60	311	9.2	150		142		1	1	1	0	0	0	0	0	0	
CUSCO	HOSPITAL	2		1977-10-19	41	4	080101	080101	1	1	2019	36	2019-09-06	2	2	65	1.57	26.37	97	90	60	311						1	0	0	0	0	1	0	0	0	2
CUSCO	HOSPITAL	1		1977-02-11	42			080306	1	1	2019	41	2019-10-10	1	2	67	1.65	24.61	86	110	70	314	14.9			90		1	1	1	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1976-05-31	42	3	080909	080106	1	1	2019	15	2019-04-12	1	2	57	1.64	21.19	93	110	70	133						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1977-06-17	42	1	081005	081005	1	1	2019	50	2019-12-09	1	2	47	1.54	19.82	78	90	60	217		128			43	1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1978-08-19	42	4		170101	1	1	2019	41	2019-10-12	2	2	75	1.7	25.95	92	140	90	65						1	0	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	2		1976-03-03	43	1	080601	080601	1	1	2019	11	2019-03-12	1	2	72	1.5	32	105	100	70	316						1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1976-08-02	43	8	080901	080108	1	1	2019	40	2019-10-01	1	2	80				120	70	141	8.8			135	212	0	0	0	0	0	0	0	0	0	
CUSCO	HOSPITAL	1		1976-07-19	43	4		080301	1	1	2019	48	2019-11-30	2	2	80	1.6	31.25	100	100	55	128						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1976-06-14	43	3	080103	080103	1	1	2019	44	2019-10-28	2	2	47	1.41	23.64	78	110	60	214						1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1976-06-30	43	2		080405	1	1	2019	51	2019-12-16	2	2	38	1.38	19.95	88	100	70	418						1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1975-10-18	43	4	080105	080105	1	1	2019	28	2019-07-09	2	2	59	1.46	27.68	81	100	60	264						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1976-05-24	43	3		080101	1	1	2019	52	2019-12-24	2	2	75	1.48	34.24	106	100	70	232						1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1976-10-13	43	3	080607	080607	1	1	2019	49	2019-12-04	2	2	51	1.47	23.6	86	110	70	355						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1975-07-19	43	2	080101	081201	1	1	2019	18	2019-04-30	2	2	48	1.57	19.47	70	100	70	484						1	0	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1978-06-13	44	4		170302	1	1	2019	26	2019-06-26	1	2	89	1.54	37.53	106	110	90	342	14.3	157		157		1	0	0	0	1	0	0	0	0	1
CUSCO	HOSPITAL	2		1975-10-16	44	3	080101	080101	1	1	2019	48	2019-11-30	1	2	68	1.51	29.82	98	100	80	141		125		110		1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1974-09-29	44	3	080901	080901	1	1	2019	28	2019-07-09	1	2	84	1.62	32.01		120	79	300		32				1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1974-06-15	44	4	080901	080904	1	1	2019	13	2019-03-30	1	2	77	1.47	35.63	98	100	70	213		144				1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1974-11-12	44	2	080401	080401	1	1	2019	38	2019-09-20	2	2	47	1.49	21.17	84	126	70	133						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1974-09-07	44	3	081101	081101	1	1	2019	23	2019-06-03	2	2	49	1.55	20.4	88	110	55	132						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1975-07-29	44	4	170101	170101	1	1	2019	40	2019-10-02	2	2	52.5	1.62	20	63	90	50							1	0	0	0	0	1	0	0	0	1
CUSCO	HOSPITAL	2		1975-11-13	44	5	080901		1	1	2019	46	2019-11-16	2	2	72	1.55	29.97	100	97	74	203			135			1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1975-10-28	44	2	030104	030104	1	1	2019	45	2019-11-05	2	2	52	1.6	20.31	70	100	70	98						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1976-03-30	44	4		080106	1	1	2019	45	2019-11-09	2	2	76	1.6	29.69	96	90	60	163						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1973-12-08	45	3	080301	080301	1	1	2019	23	2019-06-04	1	2	65	1.52	28.13	95	120	80	184	13.4	125		244		1	1	1	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1973-04-13	45	1	080704	170104	1	1	2019	2	2019-01-12	2	2	66	1.52	28.57	96	110	90	236						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	1		1974-10-07	45	3		250204	1	1	2019	48	2019-11-30	2	2	47	1.5	20.89	76	100	60	294						1	1	1	0	0	1	0	0	0	2
CUSCO	HOSPITAL	1		1975-07-19	45	4		080103	1	1	2019	43	2019-10-20	2	2	72	1.6	28.13	88	90	60	278						1	1	0	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1974-01-14	45	4	080104	080104	1	1	2019	17	2019-04-22	2	2	56	1.44	27.01	88	90	70	97						1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1972-11-10	46	1	080901	080901	1	1	2019	10	2019-03-09	1	2	65	1.44	31.35	99	110	60	258		164				1	1	0	0	0	0	0	0	0	1
CUSCO	HOSPITAL	2		1972-10-14	46	8	080101	080104	1	1	2019	18	2019-05-04	1	2	86	1.62	32.77	111	100	80	190				180		1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1972-06-21	46	6	080101	080108	1	1	2019	17	2019-04-26	1	2	85	1.64	31.6	104	110	80	246	11.7	198				1	0	0	1	0	0	0	0	0	3
CUSCO	HOSPITAL	1		1973-04-16	46	4	080101	080104	1	1	2019	15	2019-04-12	1	2	70	1.56	28.76	90	90	60	144		89				1	1	0	0	0	0	0	0	0	3
CUSCO	HOSPITAL	2		1972-10-29	46	3	081106	170101	1	1	2019	5	2019-02-01	2	2	49	1.51	21.49	100	80	82	136						1	1	0	0	1	0	0	0	0	2
CUSCO	HOSPITAL	2		1973-01-06	46	6	080101	080105	2	0	2019	8	2019-02-22	2	2	98	1.64	36.44	123	110	60	216	11.7	131				1	0	0	1	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1973-08-20	46	7		080603	1	1	2019	50	2019-12-14	2	2	73	1.63	27.48	90	120	70	84						1	1	1	0	0	0	0	0	0	1
CUSCO	HOSPITAL	1		1973-07-28	46	4		081212	1	1	2019	51	2019-12-19	2	2	78	1.62	29.72	102	120	80	313						1	1	1	0	0	0	0	0	0	2
CUSCO	HOSPITAL	2		1971-03-25	47	4	080601	080601	2	10	2019	10	2019-03-09	1	2	51	1.5	22.67	98	100	60	162						1	1	0	0	0	0	0	0	0	3

D. Pacientes con pertenencia difusa en clústeres

Paciente ID	edad	imc	sistolica	diastolica	glicemia	Pertenencia Cluster1	Pertenencia Cluster2	Pertenencia Cluster3	Pertenencia Cluster4
33	35	24.38	90	60	155	0.127	0.049	0.420	0.404
76	43	19.47	100	70	172	0.093	0.077	0.402	0.428
79	44	32.01	120	79	300	0.421	0.107	0.069	0.404
135	50	23.8	100	70	198	0.053	0.027	0.470	0.450
138	50	33.59	100	60	140	0.418	0.044	0.401	0.137
239	57	20.96	80	60	263	0.068	0.047	0.416	0.469
253	58	27.83	120	80	126	0.429	0.449	0.095	0.027
271	59	29.47	110	70	112	0.405	0.043	0.525	0.028
273	59	23.05	120	60	247	0.052	0.056	0.488	0.404
308	61	29.78	110	70	139	0.401	0.033	0.543	0.024
317	61	28.33	120	70	164	0.413	0.111	0.443	0.033
432	69	30.86	120	80	174	0.439	0.489	0.053	0.018
466	72	33.91	130	80	167	0.456	0.474	0.048	0.022
542	56	26.02	100	70	225	0.041	0.016	0.484	0.459
545	47	24.46	100	60	213	0.043	0.018	0.431	0.508
561	57	30.22	100	60	249	0.106	0.022	0.464	0.408
650	48	32.89	131	99	158	0.421	0.433	0.069	0.078
654	49	29.14	110	68	125	0.451	0.038	0.438	0.074
675	52	27.55	110	70	208	0.145	0.025	0.414	0.416
677	52	28.13	100	60	235	0.055	0.013	0.429	0.503
726	56	29.43	110	70	100	0.467	0.048	0.449	0.035
762	59	34.65	100	60	93	0.428	0.061	0.424	0.087
801	63	24.39	88	43	311	0.093	0.058	0.427	0.421
815	64	31.11	130	80	110	0.406	0.548	0.035	0.012
829	65	29.57	120	80	117	0.445	0.464	0.074	0.016
859	67	31.78	130	80	82	0.424	0.498	0.060	0.018
1057	68	32.05	130	80	106	0.410	0.529	0.046	0.015
1094	58	32.95	110	60	173	0.446	0.038	0.425	0.091
1118	71	36.39	140	80	158	0.470	0.429	0.063	0.037
1130	64	31.64	130	80	96	0.484	0.457	0.045	0.015
1186	64	35.55	140	86	118	0.459	0.456	0.052	0.032
1196	63	31.11	130	80	112	0.440	0.516	0.033	0.011
1266	40	25.56	89	59	178	0.105	0.036	0.444	0.415
1518	58	31.18	130	80	233	0.520	0.416	0.028	0.037
1580	62	21.36	87	59	268	0.056	0.042	0.469	0.433
1583	62	33.02	141	78	196	0.414	0.527	0.034	0.026
1594	63	23.12	96	66	264	0.037	0.029	0.460	0.474
1682	69	28.13	110	60	305	0.089	0.055	0.409	0.447
1720	72	18.26	80	60	305	0.087	0.087	0.425	0.401
1829	73	33.62	126	79	217	0.467	0.431	0.064	0.038
1838	75	31.86	110	70	148	0.406	0.131	0.416	0.047

1849	71	33.2	135	72	170	0.484	0.413	0.074	0.029
1928	58	30.76	130	80	228	0.463	0.479	0.026	0.032
1950	57	31.56	132	84	115	0.498	0.453	0.032	0.017
1962	60	37.22	149	90	112	0.429	0.447	0.071	0.054
2071	47	32.82	131	100	129	0.420	0.425	0.076	0.078
2073	47	26.57	100	53	218	0.073	0.024	0.476	0.426
2077	48	28.76	120	64	127	0.447	0.062	0.401	0.090
2081	48	24.03	99	55	218	0.052	0.023	0.483	0.441
2084	48	24.34	100	70	196	0.062	0.027	0.417	0.494
2085	48	32.84	146	91	90	0.404	0.474	0.067	0.055
2106	49	28.4	110	70	117	0.430	0.049	0.445	0.076
2116	50	31.11	109	62	130	0.420	0.034	0.461	0.085
2122	51	29.72	130	85	119	0.469	0.458	0.044	0.029
2164	54	30.85	124	93	117	0.429	0.467	0.060	0.044
2191	56	30.85	128	84	163	0.541	0.422	0.021	0.016
2325	65	22.37	100	60	288	0.041	0.036	0.436	0.487
2328	65	32.77	129	83	114	0.507	0.440	0.038	0.016
2330	66	25.71	106	68	272	0.044	0.036	0.452	0.469
2335	66	23.44	110	60	280	0.041	0.041	0.493	0.425
2339	66	23.52	98	66	286	0.041	0.036	0.401	0.522
2366	69	30.3	119	68	200	0.407	0.137	0.400	0.056
2369	69	30.86	120	80	141	0.468	0.453	0.063	0.016
2392	71	34.1	130	80	117	0.510	0.409	0.058	0.022
2419	75	38.76	134	90	126	0.445	0.411	0.089	0.055
2494	44	28.04	100	60	194	0.110	0.023	0.450	0.417
2497	50	33.3	100	60	130	0.406	0.044	0.422	0.128
2500	52	20.36	90	60	227	0.056	0.039	0.457	0.448
2503	56	33.31	140	85	150	0.462	0.476	0.035	0.027
2507	60	32.47	135	80	230	0.469	0.468	0.030	0.033
2616	53	34.89	130	100	101	0.448	0.405	0.080	0.067
2641	46	23.05	90	60	213	0.060	0.029	0.422	0.490
2677	71	20.93	90	60	301	0.066	0.063	0.455	0.416
2694	53	20.89	90	60	240	0.051	0.035	0.432	0.482
2705	36	30.04	120	70	222	0.418	0.062	0.110	0.410
2714	47	21	90	60	199	0.065	0.039	0.475	0.421
2725	59	26.06	100	60	261	0.030	0.014	0.463	0.494
2743	52	26.67	100	70	207	0.067	0.018	0.484	0.431

Nota. Se presentan los 78 pacientes que mostraron una pertenencia difusa mayor a 0.40 en al menos dos clústeres, según el algoritmo Fuzzy C-Means. Esta tabla amplia la información resumida en la Tabla 6 del capítulo de Resultados.