Identificación de perfiles clínicos en pacientes con obesidad mediante clustering difuso y análisis de consistencia estadística

Fernando Arturo Varilla Mendoza

Asesor

Julio Eduardo Mejia Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Julio Eduardo Mejia Manzano

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

Agradecimientos

Agradezco primeramente a Dios por darme la fortaleza y sabiduría para culminar este proyecto aplicado. A mi madre y padre, por su constante apoyo. Al docente julio Eduardo manzano que me orientó durante este proceso y cuyas correcciones contribuyeron significativamente a la mejora de este estudio. A Elis loana jimenez peña y Myladis rocio cogollo por su apoyo incondicional.

Resumen

El presente trabajo tiene como objetivo caracterizar perfiles clínicos en pacientes con obesidad mediante la aplicación de técnicas de agrupamiento difuso, específicamente el algoritmo Fuzzy C-Means (FCM) desde un enfoque estadístico. Para esto, se utilizó un conjunto de datos clínicos recolectados de pacientes con diagnóstico de obesidad en la ciudad de Bogota-Colombia durante el periodo 2024, la cual incluye variables bioquímicas y antropométricas relevantes como índice de masa corporal (IMC), edad, peso, talla y variables simuladas seleccionadas de acuerdo con la literatura tales como niveles de glucosa en ayunas, HbA1c, lípidos, entre otros. Como parte del desarrollo se realizó un análisis preliminar que incluyó estadísticas descriptivas y un análisis de Correspondencias Múltiples (ACM), con el fin de explorar la estructura subyacente de las variables numéricas y categóricas y así enriquecer la interpretación de los clústeres obtenidos. Posteriormente, se examinó la correlación entre variables con el propósito de identificar relaciones estructurales y seleccionar la medida de distancia más adecuada.

Se identificaron dos clusters con perfiles claramente diferenciados, donde el grupo 2 mostró un mayor riesgo asociado a obesidad con una edad promedio de 60 años y constituido por mujeres, evidenciado por un índice de masa corporal (IMC) significativamente más alto. Este grupo también presentó mayores niveles de glucosa en ayunas y HbA1c, indicadores relacionados con posibles condiciones de prediabetes. El grupo 1, por su parte, presentó niveles más bajos de IMC, edad promedio de 65 años, sugiriendo un perfil de riesgo moderado de enfermedades relacionadas a la obesidad. Las métricas de desempeño del modelo indicaron una partición aceptable y consistente con el enfoque fuzzy, que permite solapamiento entre grupos. Entre ellas encontramos el Índice de Silueta Difuso, que indica una calidad de partición

moderadamente buena, lo que sugiere una separación aceptable entre los clústeres. El Coeficiente de Partición de 0.69 y el Coeficiente de Partición Modificado de 0.38 respaldan una estructura de clústeres definida, aunque con cierto grado de traslapamiento, característico de este tipo de metodologías basadas en incertidumbre. Asimismo, la Entropía de Partición de 0.47 refleja un nivel medio de ambigüedad en la asignación de los datos a los clústeres.

Palabras claves**:** C-means , fuzzy clustering, incertidumbre,obesidad, logica difusa

Abstract

The present work aims to characterize clinical profiles in patients with obesity by applying fuzzy clustering techniques, specifically the Fuzzy C-Means (FCM) algorithm from a statistical perspective. To this end, a clinical dataset collected from patients diagnosed with obesity in the city of Bogotá, Colombia, during the period from 2024 to 2024 was used. This dataset includes relevant biochemical and anthropometric variables such as body mass index (BMI), age, weight, height, and simulated variables selected from the literature, such as fasting glucose levels, HbA1c, lipids, among others. As part of the development, a preliminary analysis was performed, including descriptive statistics and a Multiple Correspondence Analysis (MCA), to explore the underlying structure of the numerical and categorical variables and thus enrich the interpretation of the clusters obtained. Subsequently, the evaluation between variables is examined to identify structural relationships and select the most appropriate distance measure.

Two clusters with clearly differentiated profiles were identified. Group 2 showed a higher risk of obesity, with an average age of 60 years and comprised of women, as evidenced by a significantly higher body mass index (BMI). This group also had higher levels of fasting glucose and HbA1c, indicators associated with possible prediabetes. Group 1, on the other hand, had lower BMI levels and an average age of 65 years, suggesting a moderate risk profile for obesity-related diseases. The model's performance metrics indicated an acceptable partitioning consistent with the fuzzy approach, which allows for overlap between groups.shows these metrics, including the Fuzzy Silhouette Index, which indicates moderately good partitioning quality, suggesting acceptable separation between clusters. The Partition Coefficient of 0.69 and the Modified Partition Coefficient of 0.38 support a defined cluster structure, albeit with a certain degree of overlap, characteristic of this type of uncertainty-based methodologies. Furthermore,

the Partition Entropy of 0.47 reflects a medium level of ambiguity in the assignment of data to clusters.

Keywords: C-means , fuzzy clustering, fuzzy logic, obesity, uncertainty

Tabla de Contenido

Introducción 13

Despricion del Problema 15

Planteamiento del Problema 15

Sistematización del Problema 16

Preguntas Específicas 16

Justificación 17

Objetivos 18

Objetivo General 18

Objetivos Específicos 18

Marco de Referencia 19

Estado del Arte 19

Proceso de Búsqueda 20

Criterios de Inclusión y Exclusión 21

Resultados de la Revisión Bibliográfica 21

Marco Contextual 26

Marco Teorico 27

Medidas de Distancia 27

Clustering 29

Fuzzy Clustering 29

Algoritmo C-MEANS (FCM) 30

Marco Conceptual 32

Obesidad 32

Determinación y Evaluación de la Obesidad 33
Clustering Tradicional vs Difuso 34
Marco Normativo 35
Metodologia 36
Método 36
Descripción de los Datos Reales 37
Descripción de los Datos Simulados 38
Caracterizacion de Pacientes con Obesidad 40
Estrategia de Construcción del Algoritmo (FCM) 42
Tipo de Estudio 44
Recolección de los Datos 45
Resultados 46
Analisis Preliminar de los Datos 46
Aplicación del Método C-means Clustering (FCM) 53
Evaluación del Modelo 56
Robustez de los Clusters 57
Validación del Modelo de Agrupamiento Seleccionado 58
Caracterización de los Grupos 59
Análisis de Perfiles 63
Conclusiónes 65
Recomendaciones 67
Referencias Bibliográficas 68

Lista de Tablas

Tabla 1 *Descripción de los Artículos Seleccionados en la Revisión Bibliográfica* 22

Tabla 2 *Clasificación del IMC y sus Indicadores Clínicos* 38

Tabla 3 *Rangos de Valores Simulados* 39

Tabla 4 *Resumen Estadístico de las Variables Clínicas* 47

Tabla 5 *Valor P* 51

Tabla 6 *Análisis de Multicolinealidad (VIF)* 52

Tabla 7 *Encabezado variables estandarizadas* 53

Tabla 8 *Matriz de Pertenencia de la Partición en Dos Clústeres* 54

Tabla 9 *Métricas de Desempeño* 57

Tabla 10 *Consistencia de los Clusters* 57

Tabla 11 *Variabilidad de los Indices Difusos* 58

Tabla 12 . *Comparacion de los Modelos* 59

Tabla 13 *Caracterización de los Clusters* 59

Tabla 14 *Intervalos de Confianza* 62

Lista de Figuras

Figura 1 *Esquema de Clusterización Tradicional y Difusa* 34

Figura 2 *Cantidad de pacientes por genero* 48

Figura 3 *Análisis de Correspondencia Multiple* 49

Figura 4 *Correlación entre las Variables* 50

Figura 5 *Numero Optimo de Cluster* 53

Figura 6 *Visualización de los Clusters* 56

Lista de Ápendices

Apéndice A *Sofware Utilizado*..... 74

Apéndice B *Recursos Tecnicos*..... 74

Introducción

En las ciencias de la salud, es usual analizar grandes volúmenes de datos, por lo que se ha hecho indispensable contar con técnicas estadísticas automatizadas que permitan organizar, agrupar, inferir y extraer información relevante. Para manipular este tipo de situaciones surge la minería de datos, destacándose el método clustering que permite descubrir patrones ocultos y obtener información útil acerca del comportamiento de los datos (Rojas Diaz et al., 2009).

Uno de los principales objetivos del análisis de clustering es agrupar un conjunto de observaciones en subconjuntos exclusivos, es decir, que se puede identificar claramente si un objeto pertenece o no al clúster; sin embargo, tal partición difícil obtenerla en algunas situaciones reales. Por ello se ofrece un método difuso para construir grupos con limites inciertos, de modo que un individuo puede pertenecer a varios grupos con cierto grado de pertenencia (Sato-Ilic & Jain, 2006).

Los métodos de agrupación difusa o fuzzy clustering, tienen como propósito capturar resultados precisos y consistentes, reduciendo el ruido o error de los datos; ya que utilizan las propiedades esenciales para resolver la situación de incertidumbre de los datos. En la literatura especializada se han reportado varios estudios de aplicación de los métodos de agrupación difusa a datos clínicos, con resultados más favorables que los obtenidos con los métodos convencionales clustering (William et al., 2019).

En este trabajo se Desarrolla una metodología para obtener perfiles clínicos a partir de pacientes con enfermedades relacionadas con la obesidad, en la ciudad de Bogotá, Colombia, atendidos en centros de salud. Para ello, se emplearán técnicas de fuzzy clustering.

Los resultados obtenidos en esta investigación representan un primer paso para demostrar, en el contexto local, la relevancia del uso de técnicas de agrupamiento difuso (fuzzy clustering) en el apoyo a la toma de decisiones clínicas.

El documento está organizado de la siguiente manera: en la seccion 1 se presenta una revisión de la literatura sobre las técnicas de agrupamiento difuso más utilizadas. Esto se realiza con el objetivo de fundamentar la selección de métodos, dado que, al trabajar con datos caracterizados por incertidumbre, no se puede determinar a priori cuál técnica es la más adecuada, por lo que se tomará como referencia la literatura existente. En la sección 2, correspondiente a Materiales y Métodos, se describen las fuentes de datos utilizadas, se expone la propuesta de modelado aplicada al contexto colombiano, y se define la metodología y las técnicas de agrupamiento difuso seleccionadas para el estudio. Finalmente, en el sección 3 se presentan los resultados obtenidos y se formulan las conclusiones derivadas de la investigación.

Despricion del Problema

Planteamiento del Problema

En algunas investigaciones científicas es de interés identificar patrones de agrupamiento en los individuos de acuerdo a la similaridad de las características de los datos. En la mayoría de los casos una de las técnicas más utilizadas es el agrupamiento convencional, en la que se pueden clasificar a los individuos en grupos excluyentes entre sí. Sin embargo, existen situaciones en que se necesita agrupar datos que están asociados a incertidumbre, es decir, a la hora de asignar un individuo a un grupo no se tiene certeza a cuál pueda pertenecer. En este tipo de situaciones la incertidumbre se considera mediante la fusión de la lógica difusa con las técnicas convencionales, la cual se denomina fuzzy Clustering (Contreras Contreras et al., 2022).

El análisis de patrones de agrupamiento difuso cobra especial relevancia en el ámbito de las enfermedades crónicas, como la obesidad, que representan un desafío importante para la salud pública. Este enfoque permite abordar la complejidad inherente a los datos clínicos, caracterizados por su alta dimensionalidad y variabilidad. Identificar perfiles clínicos a partir de técnicas de fuzzy clustering no solo facilita una comprensión más precisa de las relaciones entre las características individuales y los resultados clínicos, sino que también ofrece oportunidades para desarrollar intervenciones personalizadas y dirigidas. Además, el análisis propuesto contribuirá a generar conocimientos útiles para la investigación médica, mejorando la identificación de grupos de riesgo y potencialmente apoyando estrategias de prevención y manejo más eficaces en el tratamiento de la obesidad.

La obesidad, como problema de salud pública en Colombia, tiene una alta prevalencia, con cifras recientes que indican que más del 50% de la población presenta sobrepeso u obesidad, según los informes del (Ministerio de Salud y Protección Social, 2023). Esta condición está

estrechamente vinculada con enfermedades crónicas como la diabetes tipo 2, la hipertensión arterial y las afecciones cardiovasculares, lo que pone de manifiesto su complejidad. En este contexto, la técnica de fuzzy clustering resulta especialmente adecuada, ya que permite trabajar con la incertidumbre y variabilidad características de los datos clínicos. A diferencia de otros métodos de agrupamiento, esta técnica ofrece una representación más flexible al permitir que los datos pertenezcan a varios grupos con distintos niveles de confianza, lo que facilita la identificación de patrones complejos y relaciones subyacentes.

Sistematización del Problema

Teniendo en cuenta la problemática indicada y con el ánimo de mejorar el dominio, se propone la siguiente pregunta de investigación. ¿Cómo puede el uso de fuzzy clustering contribuir a identificar perfiles clínicos asociados con la obesidad en Bogota, Colombia, y de qué manera esta técnica puede mejorar las estrategias de prevención y manejo de esta enfermedad?.

Preguntas Específicas

¿Cuáles son las variables clínicas y sociodemográficas más relevantes en la caracterización de pacientes con obesidad ?

¿Qué nivel de efectividad presentan las técnicas de agrupamiento difuso (fuzzy clustering) en la segmentación de estos pacientes según sus perfiles clínicos?

¿Qué tipos de perfiles clínicos pueden identificarse a partir del análisis de datos utilizando el enfoque de agrupamiento difuso?

Justificación

En el contexto clínico, es común que las técnicas convencionales de agrupamiento no sean efectivas para segmentar a los pacientes según sus características personales, clínicas y sociales, debido a los traslapes entre grupos. Esto hace necesario incluir la incertidumbre de pertenencia a los diferentes grupos en los algoritmos de agrupamiento (Ahmadi et al., 2018).

Adicionalmente, el volumen considerable de información generada por los pacientes con enfermedades crónicas, como la obesidad, presenta un desafío en términos de análisis integral. Este análisis resulta complejo al considerar las múltiples dimensiones de los datos. Por lo tanto, surge la necesidad de implementar herramientas estadísticas y algoritmos avanzados que permitan identificar patrones y similitudes en los perfiles clínicos (Sato-Ilic & Jain, 2006).

La aplicación de algoritmos de agrupamiento difuso (fuzzy clustering) ofrece una solución potencial para identificar perfiles clínicos de pacientes con obesidad. Este enfoque no solo permite manejar de manera adecuada la incertidumbre en los datos, sino que también facilita la segmentación en grupos con características similares. Los resultados obtenidos pueden ser una base sólida para estudios futuros relacionados con la prevención y tratamiento de la obesidad, ofreciendo una visión integral del problema a nivel poblacional.

Objetivos

Objetivo General

Identificar perfiles clínicos de pacientes con obesidad mediante técnicas de clustering difuso, con el fin de facilitar la comprensión de los patrones asociados y su consistencia estadística.

Objetivos Específicos

Seleccionar las técnicas de agrupamiento difuso propuesta en la literatura para segmentar pacientes clínicos.

Establecer un algoritmo clustering difuso que permita agrupar individuos con características clínicas, considerando el proceso de selección de las variables reales e incorporando un proceso de simulación de variables clínicas basado en patrones reales.

Caracterizar los perfiles clínicos obtenidos mediante la metodología propuesta, utilizando análisis internos y métricas que aseguren su consistencia y relevancia**.**

Marco de Referencia

Estado del Arte

La obesidad es una enfermedad crónica caracterizada por un exceso de tejido adiposo en el cuerpo, considerada actualmente uno de los principales problemas de salud pública a nivel mundial. Esta condición se asocia a un aumento significativo en el riesgo de desarrollar diversas enfermedades, como diabetes tipo 2, hipertensión arterial, enfermedades cardiovasculares, entre otras (World Health Organization, 2000). En Colombia, la obesidad ha mostrado una tendencia creciente en los últimos años, según el Departamento Administrativo Nacional de Estadística (DANE) la prevalencia de exceso de peso en adultos fue del 56,4%, mientras que en niños y adolescentes fue del 24,4%. Evidenciando una tendencia creciente en los últimos años, afectando tanto a la población adulta como infantil.

El análisis de perfiles clínicos en pacientes con enfermedades asociadas a la obesidad es clave para apoyar mejores decisiones médicas y desarrollar tratamientos más específicos. No obstante, debido a la variedad y la incertidumbre presentes en los datos clínicos, los métodos tradicionales de análisis presentan limitaciones. Por ello, las técnicas de agrupamiento difuso (fuzzy clustering) han cobrado relevancia, ya que ofrecen una forma más flexible de interpretar datos inciertos y diversos (Ahmadi et al., 2018) . A partir de la revisión de trabajos previos, se seleccionarán las técnicas de agrupamiento difuso más relevantes que puedan adaptarse a las necesidades de este estudio.

Por otra parte, en la literatura especializada se encuentra que el modelado de clustering se ha abordado desde varios enfoques: Por ejemplo, (Zheng et al., 2024) realiza una revisión sistemática de los enfoques utilizados para la segmentación de datos médicos inciertos en el periodo 2014-2022, encontrando un total de 513 estudios relevantes. De estos, un 28 % emplea

modelos de agrupamiento difuso, principalmente C-Medias Difusas (Fuzzy C-Means, FCM), combinados con técnicas tradicionales de análisis clínico, mientras que el 72 % integra enfoques híbridos con machine learning y análisis estadístico avanzado. Además, se observa un crecimiento acelerado en la investigación, pasando de 28 estudios en 2018 a 172 en 2021, con 83 artículos publicados hasta junio de 2022.Este aumento refleja la creciente necesidad de abordar la incertidumbre inherente en los datos clínicos, ya que los registros médicos suelen contener información ambigua, imprecisa o incompleta. (Aslan & Hızıroğlu, 2024) , muestran que en el año 2024 hubo un incremento significativo en la cantidad de estudios reportados para segmentar pacientes clinicos. Dentro de las técnicas de lógica difusa más empleadas se destacan los sistemas basados en reglas difusas, el algoritmo de agrupación en clústeres de C-medias difusas y los sistemas de inferencia difusa. Por otra parte, el autor encontró que las tasas de precisión reportadas en estos estudios varían entre el 85 % y el 98 % en tareas de predicción y diagnóstico.

En este capítulo se expone la metodología seguida para identificar las principales técnicas de agrupamiento difuso aplicadas en el área clínica, específicamente en el análisis de perfiles de pacientes con enfermedades relacionadas con la obesidad. El enfoque metodológico es de tipo documental, basado en la recopilación, actualización, organización y clasificación de material bibliográfico proveniente de bases de datos académicas. Los resultados de esta revisión permiten actualizar el conocimiento sobre el uso de métodos de fuzzy clustering en el ámbito de la salud y servirán como base para definir la metodología empleada en este estudio, centrado en el contexto colombiano.

Proceso de Búsqueda

El proceso de búsqueda se llevó a cabo combinando estrategias manuales y la técnica de bola de nieve, enfocándose en artículos publicados en journals indexados entre el año 2019 y

2025. Para ello, se consultaron bases de datos académicas como ScienceDirect, MDPI, IEEE y PubMed. Las palabras clave empleadas fueron: "obesity", "clinical profiles", "fuzzy clustering" y "health data analysis".

Para asegurar una búsqueda sistemática, se formuló la siguiente ecuación de búsqueda utilizando operadores booleanos:

("obesity" OR "overweight") AND ("clinical profiles" OR "health characterization") AND ("fuzzy clustering" OR "soft clustering") AND ("health data analysis" OR "healthcare analytics")

Criterios de Inclusión y Exclusión

En este estudio fueron considerados todos los artículos publicados entre 2019 y 2025 que abordaran el uso de técnicas de agrupamiento o análisis de datos clínicos relacionados con enfermedades metabólicas, especialmente obesidad. Además, se incluyeron únicamente estudios aplicados al área de la salud, disponibles en texto completo y publicados en inglés o español.

A sí mismo, se excluyeron aquellos artículos que utilizaran exclusivamente métodos tradicionales de agrupamiento (como k-means clásico) sin aplicación de técnicas difusas, investigaciones que no estuvieran relacionadas directamente con perfiles clínicos o enfermedades de interés, publicaciones anteriores al año 2019 y documentos incompletos.

Resultados de la Revisión Bibliográfica

La Tabla 1 presenta los estudios elegidos tras aplicar los criterios de inclusión y exclusión previamente descritos. En ella, la primera columna indica el identificador asignado a cada artículo dentro de este trabajo, mientras que la segunda columna detalla la referencia bibliográfica correspondiente.

A continuación, se incluyen columnas que especifican el tipo de estudio, si es una revisión sistematica (R) o un articulo (A), las características de los datos utilizados, las variables consideradas, y la técnica de agrupamiento difuso empleada. Estas categorías permiten comparar de manera estructurada los enfoques metodológicos y los resultados obtenidos en cada investigación.

Tabla 1

Descripción de los Artículos Seleccionados en la Revisión Bibliográfica

Estudio	Referencias	Tipo		Datos	Variables	Tecnica utilizada
		A	R			
E1	(Nedyalkova et al., 2020)	x		Diabetes mellitus tipo 2	Edad, sexo, peso, talla, IMC, cintura, cadera, relación peso/altura creatinina, ácido úrico, proteínas totales, albúmina, datos de colesterol (HDL, VHDL, LDL, colesterol total), electrolitos (sodio, potasio), triglicéridos, niveles de glucosa	Fuzzy C-Means, FCM
E2	(Takeshita et al., 2024)	x		Obesidad, Diabetes	IMC, circunferencia de la cintura, hemoglobina A1c	Fuzzy C-Means, FCM

				(HbA1c) y tasa de filtración glomerular estimada (TFGe), tabaquismo	
E3	(Nagase et al., 2023)	x	Hígado graso, obesidad	Ácidos grasicos, IMC, edad y género.	Fuzzy C-Means, FCM
E4	(Sümbül-Şekerci et al., 2024)	x	diabetes mellitus tipo 2	Género, edad, IMC, hemoglobina, creatina, antecedentes, TSH, colesterol HDL y LDL	Fuzzy C-Means, FCM y arboles de decisión.
E5	(Hantoli & others, 2021)	x	Obesidad infantil	IMC, Edad, Relación cintura-estatura (RCE), Relación cintura-cadera (RCC), Porcentaje de masa grasa, Peso, Estatura, Circunferencia de la cadera, Circunferencia de la cintura y Ciudad.	Fuzzy C-Means, FCM, arboles de decisión y sistemas vectoriales
E6	(Velmurugan & Emayavaramban, 2025)	x	Diabetes	Glucosa en sangre, Glucosa en sangre en ayunas (antes del desayuno), Glucosa en sangre después	Fuzzy C-Means, FCM

				del desayuno, Hemoglobina glicosilada (HbA1c), Edad, Índice de Masa Corporal (IMC), Triglicéridos, Presión arterial.	
E7	(Zheng et al., 2024)	x	Datos clinicos	No reporta	Fuzzy C-Means, FCM, regression lineal, maching learning
E8	(Khamis et al., 2024)	x	Enfermedades cardiovasculares y obesidad	Edad, genero, presión arterial sistólica, hábitos de fumar, peso, talla, IMC, circunferencia de cintura, HbA1c, colesterol (LDL, HDL), relación hdl,ldl, Triglicéridos.	Fuzzy C-Means, FCM y análisis de componentes principales
E9	(Sulla-Torres et al., s. f.)	x	Obesidad en Niños y Adolescentes	Edad, peso, estatura, IMC, obesidad, HbA1c, colesterol (LDL, HDL), relación hdl,ldl, Triglicéridos.	Fuzzy C-Means, FCM y sistemas neuro difusos

E10	(Bhatti et al., 2023)	x	Diabetes	Glucosa Presión arterial, Índice de Masa Corporal (IMC) antecedentes hereditarios de diabetes, edad, genero. HbA1c, colesterol (LDL, HDL), relación hdl,ldl, Triglicéridos.	Fuzzy C-Means, FCM y análisis de componentes principales y aprendizaje conjunto de bosque aleatorio

Nota. Todos los referentes fueron seleccionados por su relevancia, actualidad (2019–2025) y contribución directa al análisis de datos clínicos y agrupamiento difuso en salud.

A partir de la Tabla 1 se encuentra que el 100% de los estudios analizados emplean la técnica de agrupamiento difuso FCM (Fuzzy C-Means) como método principal para la identificación de patrones o agrupamientos. Además, en un 5% de los estudios, esta técnica fue combinada con otros enfoques metodológicos, lo que evidencia su robustez y versatilidad en el análisis de datos relacionados con obesidad y enfermedades asociadas como la diabetes.

En cuanto a las variables utilizadas, se observa que el 90% de los estudios incluyen de forma recurrente variables antropométricas clave, tales como: edad, sexo, peso, estatura, índice de masa corporal (IMC). Esto confirma su relevancia como factores fundamentales en el análisis del estudio.

Por otro lado, también se destaca el uso frecuente de exámenes médicos, el 80% de los artículos se encuentran que utilizaron variables como: niveles de glucosa en sangre (ayuno), hemoglobina glicosilada (HbA1c), triglicéridos, colesterol, HDL, LDL.

Los resultados de esta revisión bibliográfica sirven como base para la selección del método que se aplicará en este estudio, así como para identificar las variables más relevantes a utilizar.

Marco Contextual

Según la Organización Mundial de la salud (OMS), la obesidad se considera un problema de salud pública a nivel global debido a su alta prevalencia y su impacto en la calidad de vida de las personas. En Colombia, la prevalencia de la obesidad ha aumentado significativamente, siendo impulsada por factores como cambios en los patrones de alimentación, disminución de la actividad física, y predisposición genética. Este fenómeno no solo afecta a los países desarrollados, sino que también ha alcanzado proporciones alarmantes en países de ingresos medios y bajos (MedlinePlus, 2021)

Bogotá, al ser la capital de Colombia y una de las ciudades con mayor densidad poblacional, presenta una alta concentración de personas afectadas por enfermedades crónicas, como la obesidad. Esta situación se presenta por diversos factores, entre los que destacan el estilo de vida sedentario, los patrones alimenticios, las barreras en el acceso equitativo a los servicios de salud y las desigualdades socioeconómicas (Sánchez et al., 2020).

En este contexto, durante el año 2024, se examinara los datos clínicos de pacientes con diagnóstico de obesidad en Bogotá, para caracterizar sus perfiles clínicos utilizando un enfoque de agrupamiento difuso (fuzzy clustering), que permite abordar la complejidad y la incertidumbre inherente a los datos de salud, identificando patrones que podrían no ser evidentes mediante métodos tradicionales.

Marco Teorico

Medidas de Distancia

Una medida de distancia es aquella que mide la disimilitud o diferencia entre las observaciones de los individuos, siendo está considerada como una métrica que permite interpretar geométricamente muchas técnicas de análisis multivariante, y a su vez es la base de muchos algoritmos de la ciencia de datos. A continuación, se presentan algunas de las medidas de distancias más frecuentes (Giordani et al., 2020).

Distancia Euclidiana: Dada una matriz X de tamaño n × p, siendo p el número de variables observadas, n el número de clusters y x_i=(x_i1,…,x_ij ,…,x_ip) los elementos de la matriz, entonces se define la distancia euclidiana entre dos vectores de objetos observados x_i, xi′ para las unidades i, i′ = 1, ..., n como:

$$d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2} \quad (1)$$

Una de las ventajas de usar la distancia euclidiana es la sencillez en su cálculo y rapidez al realizar cualquier tipo de algoritmo que requiera de ella, pero posee inconvenientes cuando se tienen variables con diferentes unidades de medida, de modo que se requiere una estandarización.

Distancia de Manhattan: Es una métrica utilizada para calcular la distancia entre dos puntos en un espacio multidimensional, sumando las diferencias absolutas de sus coordenadas.

La distancia de Manhattan entre dos vectores de objetos observados d(xi,x') se expresa en (*2*):

$$d_M(x_i, x_i') = \sum_{j=1}^{p} \left| x_{ij} - x_{ij}' \right| \quad (2)$$

Esta medida se ve menos afectada por valores outliers debido a que las diferencia entre pares de observaciones no están elevadas al cuadrado, además funciona de manera correcta para datos de alta dimensión, pero es menos intuitiva que la distancia euclidiana.

Distancia de Minkowski: Es una métrica generalizada utilizada para medir la distancia entre dos puntos en un espacio multidimensional. Es una extensión de las distancias euclidiana y Manhattan, La distancia de Minkowski se define como:

$$d_M^q(x_i, x_i') = \left(\sum_{j=1}^{p} \left| x_{ij} - x_{ij}' \right|^q \right)^{\frac{1}{q}} \quad (3)$$

cuando q = 2 y q = 1 se obtienen la distancia euclidiana y de Manhattan respectivamente, por tanto, la distancia de Minkowski es un caso particular de las dos distancias mencionadas anteriormente.

Distancia de Mahalanobis: La distancia de Mahalanobis es una medida que evalúa la distancia entre un punto y una distribución en un espacio multidimensional. A diferencia de la distancia euclidiana, tiene en cuenta las correlaciones entre las variables y las escalas de las mismas, lo que la hace ideal para datos donde las variables no son independientes o tienen varianzas diferentes.

sea Σ−1 una matriz de covarianzas y (xi,x') dos objetos observados, entonces se define la distancia de Mahalanobis como:

$$d(x_i, x_i') = \sqrt{\left((x_i - x_i')^T \Sigma^{-1} (x_i - x_i') \right)} \quad (4)$$

Esta distancia presenta la ventajosa propiedad de que considera la correlación entre las variables, lo que soluciona el inconveniente de la aplicación de los casos particulares de la distancia de Minkowski; debido a que es invariante ante los cambios de escala y no depende, por tanto, de las unidades de medición.

Distancia basada en correlaciones**:** Mide la disimilitud entre dos vectores basándose en el grado de correlación que existe entre ellos. Se define como :

$$d_c(x, y) = 1 - p(x, y) \tag{5}$$

Esta distancia es útil cuando las variables presentan correlaciones altas.

Clustering

El análisis clustering hace parte de los métodos de aprendizaje no supervisado más usados en el análisis de datos; el cual consiste en ordenar un conjunto de individuos en grupos o clusters, de tal manera que los miembros de un mismo grupo tengan las características más homogéneas entre sí, pero lo más heterogéneas a la de los miembros de otros grupos (Bijuraj, 2013)

Fuzzy Clustering

Fuzzy clustering es un método de agrupación de datos borrosos, o también conocidos difusos que considera no solo el estado de pertenencia a los clusters, sino también considera en qué grado los objetos pertenecen a estos, es decir, cada objeto tiene un grado de pertenencia a clusteres, en lugar de pertenecer completamente a un solo grupo (Sato-Ilic & Jain, 2006).

Suponga que $X = \{x_1, x_2, \dots, x_n\}$ es un conjunto dado de n objetos, y $K = 1, \dots, n; k \in N$ es el número de conglomerados. Entonces un grupo difuso, que es un subconjunto difuso en X, se muestra en (*5*):

$$u_k: X \rightarrow [0,1], k = 1, \ldots, K \quad (5)$$

de modo que el grado de pertenencia de un objeto *i* a un grupo *k*, está dado por la ecuación (*6*):

$$u_{ik} = u_k(x_i),\ i = 1, \ldots, n, k = 1, \ldots, K \quad (6)$$

En general u_{ik} satisface las siguientes condiciones:

$$u_{ik} \in [0,1], \forall i, k$$

$$\sum_{k=1}^{K} u_{ik} = 1, \forall i$$

El estado de la agrupación difusa se representa mediante una matriz de partición $U = (u_{ik})$ y un conjunto de las matrices que se define como:

$$M_{fnK} = \{U \in Rnk \mid u_{ik} \text{ satisface } \forall i, k\} \quad (7)$$

En particular si para cualquier *i* y *k*, $u_{ik} \in \{0, 1\}$, que corresponde al supuesto que se emplea en el clustering clásico, entonces el conjunto de las matrices de partición *U* es:

$$M_{nK} = \{U \in M_{fnK} \mid u_{ik} \in \{0,1\}\ \forall i, k\}. \quad (8)$$

Lo cual indica que $M_{nK} = \subset M_{fnK}$ es decir, el agrupamiento clásico se puede considerar como un caso particular del agrupamiento difuso.

En la literatura se han propuesto diversos métodos para realizar el agrupamiento difuso; sin embargo, se optó para el estudio el método de agrupamiento difuso Fuzzy C-Means (FCM), el cual se describe a continuación.

Algoritmo C-MEANS (FCM)

Se emplea el método de agrupación de datos borrosos, o también conocidos difusos que considera no solo el estado de pertenencia a los clusters, sino también considera en qué grado los

objetos pertenecen a estos, es decir, cada objeto tiene un grado de pertenencia a clusteres, en lugar de pertenecer completamente a un solo grupo (Sato-Ilic & Jain, 2006).

El algoritmo minimiza la función objetivo descrita a continuación, la cual se puede entender como una ponderación del error cuadrático que se comete al establecer los elementos c_k como centrodides de los k clusters como se expresa en (*9).*

$$J(X, U, C) = \sum_{i=1}^{n} \sum_{k=1}^{K} (u_{ik})^m |x_i - c_k|_B^2 \tag{9}$$

donde X es el número de objetos, U la matriz de pertenencia, cuyos elementos u_{ik} están elevados a un factor de borrosidad $\mathrm{m}(\mathrm{m} > 1)$, y C la matriz de centroides de los clusters. El termino $| xi - ck |^2$_es la distancia entre los puntos de los datos x_i y los centroides c_k. Cuando B toma el valor de la matriz identidad I, se obtiene la distancia euclidiana elevada al cuadrado.

Al minimizar la función objetivo, igualando a cero las respectivas derivadas parciales, se obtiene como resultado las siguientes dos ecuaciones las cuales se utilizan para obtener los valores de los centroides y de las pertenencias como se expresa en (*10*) y (*11*):

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{|x_i - c_k|}{|x_i - c_j|} \right)^{\frac{2}{m-1}}} \tag{10}$$

$$c_k = \frac{\sum_{i=1}^{n} (u_{ik})^m x_i}{\sum_{i=1}^{n} (u_{ik})^m}, \quad 1 \leq i \leq c \tag{11}$$

Partiendo de las expresiones el algoritmo FCM puede describirse de la siguiente manera (Edla et al., 2020). Fije un valor de umbral $\epsilon > 0$, el cual debe ser pequeño, normalmente 0.001 o menor

1. Inicializar la matriz de pertenencias U = $[u_{ik}]$, con valores aleatorios
2. Calcular los centros de los clusters usando c_k

3. Actualizar u_{ik}
4. Si el cambio de U entre dos iteraciones $\mathrm{U(i)} - \mathrm{U(i-1)} < \epsilon$, entonces el algoritmo se detiene. En caso contrario vuelve al paso 2

Marco Conceptual

Obesidad

La obesidad es una enfermedad multifactorial y crónica que se define por un exceso de acumulación de grasa corporal que puede ser perjudicial para la salud. La obesidad no es solo un problema estético, sino que está vinculada a una variedad de trastornos metabólicos, cardiovasculares, y psicológicos. De hecho, se considera un factor de riesgo primario para enfermedades como diabetes tipo 2, hipertensión arterial, apnea del sueño, y ciertos tipos de cáncer (Alarcón et al., 2021)

Enfermedades Cardiovasculares: La obesidad incrementa significativamente el riesgo de enfermedades cardiovasculares, incluyendo hipertensión arterial, enfermedad coronaria y accidentes cerebrovasculares. El exceso de tejido adiposo contribuye a la disfunción endotelial y a la aterosclerosis, factores clave en estas patologías.

Las enfermedades cardiovasculares son un concepto amplio que se utiliza para definir todas las enfermedades que corresponden a los trastornos del sistema circulatorio, que incluye el corazón, los vasos sanguíneos y la sangre. Entre las principales enfermedades cardiovasculares encontramos:

Miocardiopatía: También conocidas como cardiomiopatía, son todas aquellas enfermedades que afectan el musculo cardiaco, lo cual provoca que el corazón le cueste bombear sangre al cuerpo, causando paros cardiacos, insuficiencia cardiaca, infartos, entre otras complicaciones.

Diabetes mellitus**:** La diabetes mellitus es un grupo de enfermedades en que el cuerpo no puede controlar la cantidad de azúcar en la sangre. Esto se debe a que la célula del páncreas no produce la cantidad suficiente de insulina que el cuerpo necesita para combatir el azúcar, provocando así enfermedades cardiovasculares. La diabetes mellitus se puede clasificar en dos grupos entre los cuales encontramos(Clínica Universidad de los Andes, 2021).

Diabetes mellitus tipo 1: También conocidos como diabetes mellitus insulinodependiente, en ella se clasifican las personas que no producen la cantidad adecuada de insulina que el cuerpo necesita, por lo que el azúcar en la sangre es bastante alto. Generalmente la mayoría de los casos se presentan en jóvenes y niños.

Diabetes mellitus tipo 2: También conocidos como diabetes mellitus no insulinodependiente, en ella se clasifican las personas que no utilizan de manera eficaz o el cuerpo no responde a la insulina. La mayor parte de los casos se presentan en adultos.

Determinación y Evaluación de la Obesidad

El Índice de Masa Corporal (IMC) es una herramienta comúnmente empleada para evaluar si el peso de una persona se encuentra dentro de los rangos considerados saludables. Se calcula dividiendo el peso en kilogramos por el cuadrado de la altura en metros:

$$IMC = \frac{peso(kg)}{estatura(m)}$$

Según la Sociedad Española, el IMC se clasifica de la siguiente manera (Obesidad, 2023).

5. Bajo peso: IMC menor a 18.5.
6. Peso saludable: IMC entre 18.5 y 24.9.
7. Sobrepeso leve: IMC entre 25 y 26.9.
8. Sobrepeso moderado: IMC entre 27 y 29.9.
9. Obesidad leve: IMC entre 30 y 34.9.

10. Obesidad moderada: IMC entre 35 y 39.9.

11. 44Obesidad grave: IMC entre 40 y 49.9.

12. Obesidad mórbida: IMC de 50 o más.

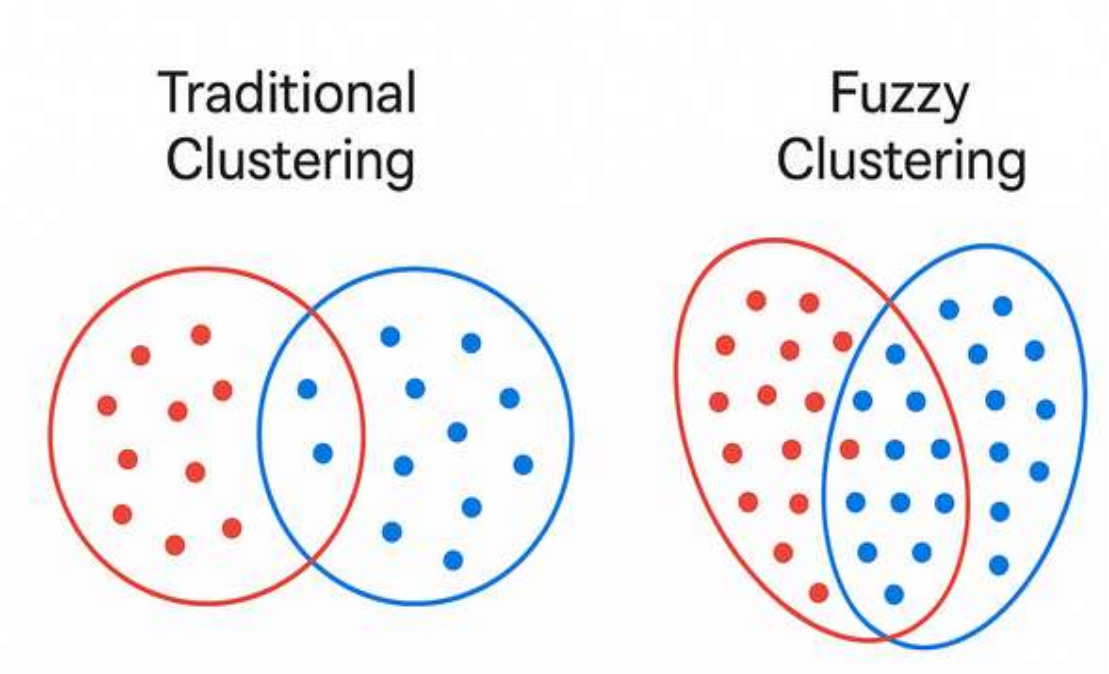
Con esta clasificación, se puede determinar el grado de obesidad y el riesgo de padecer enfermedades vinculadas a esta, como la diabetes tipo 2, la hipertensión y problemas cardiovasculares.

Clustering Tradicional vs Difuso

La construcción de un modelo de clustering difuso se basa en la agrupación de datos que presentan incertidumbre o ambigüedad en su pertenencia a un grupo específico. A diferencia del clustering tradicional (como K-means), donde cada observación pertenece estrictamente a un único grupo, en el clustering difuso se asignan grados de pertenencia a múltiples grupos de forma simultánea tal y como se muestra en la Figura 1.

Figura 1

Esquema de Clusterización Tradicional y Difusa



Marco Normativo

En Colombia, la obesidad como problema de salud pública se encuentra respaldado por un conjunto de normas y políticas que orientan las acciones de prevención, atención y control de las enfermedades crónicas no transmisibles.

Una de las principales disposiciones es la Ley 1355 de 2009, que reconoce oficialmente la obesidad como una enfermedad crónica y un asunto de interés prioritario en salud pública. Esta ley establece la necesidad de adoptar medidas para su control, incluyendo la divulgación de hábitos de vida saludables, la regulación de la publicidad de alimentos, y la generación de entornos más saludables.

Adicionalmente, el Plan Decenal de Salud Pública 2022–2031, impulsado por el Ministerio de Salud, plantea estrategias orientadas a mejorar la calidad de vida de la población colombiana mediante la reducción de los factores de riesgo asociados a enfermedades como la obesidad.

También se encuentra en vigencia la Política Nacional de Seguridad Alimentaria y Nutricional, la cual propone la disponibilidad, acceso y consumo de alimentos nutritivos como herramienta clave en la prevención del sobrepeso y la obesidad.

Metodologia

Método

En esta sección se presentan las particularidades metodológicas propuestas para segmentar pacientes clínicos mediante técnicas de Fuzzy Clustering en la ciudad de Bogota. Inicialmente, se describe la naturaleza de los datos utilizados y la construcción de la base de datos, los cuales provienen de registros clínicos recopilados de diversas fuentes de datos abiertos de Colombia, específicamente de repositorios como el Sistema de Información de Salud Pública (SISPRO) y la plataforma de datos abiertos del Ministerio de Salud y Protección Social.

Dado que algunas variables clínicas sensibles, como los resultados de exámenes médicos (glucosa en ayunas, HbA1c, triglicéridos, HDL y LDL), no estaban disponibles por restricciones de acceso y confidencialidad, se optó por simular estos valores. La simulación se realizó teniendo en cuenta los rangos clínicos establecidos por la Organización Mundial de la Salud (OMS) y otras referencias médicas, así como la distribución del IMC y el grado de obesidad de los pacientes. Esta decisión se tomó en cumplimiento de (Constitución Política de Colombia, 1991) artículo 15, que garantiza el derecho a la intimidad y a la protección de datos personales; la Ley 1581 de (Congreso de Colombia, 2012), que regula el tratamiento de datos personales sensibles; la resolución de (Ministerio de Salud de Colombia, 1999), que establece normas para la historia clínica y la confidencialidad de la información; y la Resolución 199 de 2021 del Ministerio de Salud, que reglamenta el manejo de datos en investigaciones en salud. De esta manera, el enfoque adoptado permitió enriquecer la base de datos y garantizar un análisis robusto sin comprometer la privacidad de la información.

A continuación, se describe la base de datos utilizadas y las variables y se expone el esquema de modelado del método aplicado en este estudio: la segmentación basada en técnicas

de clustering difuso (Fuzzy C-Means). El modelo fue implementado y evaluado utilizando el software RStudio (2021), con el fin de proporcionar una clasificación más precisa y adaptable de los pacientes clínicos en función de sus características médicas y patrones de comportamiento.

Descripción de los Datos Reales

Se consideran los datos reportados por (Datos.gov.co 2023), disponibles en la plataforma de datos abiertos del Gobierno de Colombia. Para este estudio, se tomó como referencia la información correspondiente al año 2024. Para la extracción de los datos, se descargó la base de datos original la cual contenia información de pacientes con enfermedades cronocias en la ciudad de Bogota y se procedió a una limpieza preliminar, en la cual se filtró únicamente la información de pacientes con diagnóstico de obesidad y se eliminarón valores duplicados. Asimismo, se eliminaron aquellas variables consideradas no relevantes para los fines del estudio, tales como el tipo de afiliación al sistema de salud (subsidiado o contributivo), el plan de beneficos al que el paciente pertenece, el estado mental y neurologico, el estado de epoc y de disnea.La base de datos no se encontró valores faltntes por lo que no fue necesario realizar imputación de datos, por tanto el resultado fue la consolidación de una base de datos única, estructurada con las variables esenciales para la segmentación de pacientes clínicos. Estas variables fueron seleccionadas con base en la revisión de la literatura especializada, priorizando aquellas que han demostrado mayor relevancia en estudios previos sobre análisis de obesidad y segmentación clínio.

La base de datos final contiene 158 observaciones y variables como la edad, peso, talla, género, IMC y los siguientes indicadores asociados a la obesidad, los cuales se derivaron de la clasificación oficial del Índice de Masa Corporal (IMC) publicada por la Organización Mundial de la Salud (World Health Organization, 2000). Adicionalmente, se incorporaron variables

clínicas relacionadas con exámenes médicos como glucosa en ayunas, HbA1c, triglicéridos, HDL y LDL, cuyos valores fueron simulados con base en los rangos clínicos establecidos y de acuerdo con el grado de obesidad de los datos reales. La descripción detallada del proceso de simulación se presenta en la siguiente sección.

Tabla 2

Clasificación del IMC y sus Indicadores Clínicos

Rango de IMC (kg/m^2)	Indicador Clínico	Descripción
25.0 – 29.9	Sobrepeso	Riesgo moderado de enfermedades cardiovasculares, hipertensión y diabetes tipo 2.
30.0 – 34.9	Obesidad tipo I	Riesgo alto de complicaciones metabólicas y enfermedades crónicas.
35.0 – 39.9	Obesidad tipo II	Riesgo muy alto de comorbilidades graves como apnea del sueño o diabetes severa.
40.0 o más	Obesidad tipo III (mórbida)	Riesgo extremo de complicaciones médicas severas y mortalidad prematura.

Nota. La tabla presenta el rango de obesidad y su indicador clínico para el riesgo que pueda presentar cada paciente.Tomado de Minsalud, (2024)

Descripción de los Datos Simulados

Dado que las bases de datos públicas disponibles en Colombia no incluyen información detallada sobre exámenes médicos individuales debido a restricciones de acceso derivadas de políticas de seguridad y protección de datos sensibles, y en cumplimiento con la legislación vigente sobre la protección de datos personales Ley 1581 de 2012 (Congreso de Colombia, 2012), se decidió simular los valores de ciertos indicadores clínicos de los pacientes. Entre estos

se encuentran los niveles de glucosa en ayunas, HbA1c (%), triglicéridos (mg/dL), HDL (mg/dL) y LDL (mg/dL). Para ello, se definieron rangos de valores normales y alterados con base en guías clínicas oficiales, tales como las publicadas por la (Ministerio de Salud de Colombia, 1999), (American Diabetes Association, 2024) y el informe NCEP ATP III. A partir del grado de obesidad (sobrepeso, obesidad tipo I, II o III), se establecieron intervalos realistas de variación para cada parámetro, permitiendo una simulación coherente con los perfiles clínicos esperados.

La técnica utilizada fue la generación de datos sintéticos, implementada en Microsoft Excel a través de funciones condicionales y aleatorias. Por ejemplo, se emplearon fórmulas como =SI(I2="SOBREPESO";ALEATORIO.ENTRE(45;55);SI(I2="OBESIDAD I";ALEATORIO.ENTRE(50;60);...)) para asignar valores simulados a cada paciente según su categoría de obesidad. De esta forma, los valores de cada biomarcador fueron generados dentro de intervalos clínicamente en función del grado de obesidad (sobrepeso, obesidad tipo I, II o III).

Como resultado del proceso de simulación, se obtuvo una base de datos ampliada, donde los pacientes clasificados con obesidad tipo II y III presentaron niveles promedio simulados de glucosa, triglicéridos y HbA1c más elevados en comparación con los de sobrepeso u obesidad tipo I, lo cual era de esperarse. Además, en promedio, los valores simulados para cada variable clínica se mantuvieron dentro de los rangos establecidos según las guías médicas de referencia, lo que garantiza la coherencia estadística y realismo clínico de los datos generados.

Tabla 3

Rangos de Valores Simulados

Tipo Obesidad	Glucosa Ayunas	HbA1c (%)	Triglicéridos (mg/dL)	HDL (mg/dL)	LDL (mg/dL)
Sobrepeso	85 – 110	5.2 – 5.8	120 – 150	45 – 55	100 – 130
Obesidad I	100 – 125	5.7 – 6.4	150 – 200	40 – 50	120 – 160

Obesidad II	110 – 140	6.0 – 7.0	180 – 250	35 – 45	140 – 180
Obesidad III	130 – 160	6.5 – 8.5	220 – 350	30 – 40	160 – 200

*Nota***.** Rangos de valores simulados para Parámetros Clínicos según el Grado Obesidad.Tomado de ADA, NCEP ATP III, MinSalud.

La Tabla 3 presenta los rangos de valores simulados para los parámetros clínicos considerados, establecidos según el grado de obesidad (sobrepeso, obesidad tipo I, II y III). Es importante aclarar que estos datos fueron simulados a partir de los patrones observados en los datos reales, con el objetivo de mantener la coherencia clínica y la concordancia entre las variables. Tal como lo señalan estudios como los de (Amaro-López et al., 2019) & (Gómez Martínez et al., 2021) la simulación de datos clínicos es una estrategia válida y recomendable en contextos donde el acceso a información médica sensible está restringido. Esta estrategia permite generar conjuntos de datos que conservan la estructura y comportamiento esperados, siempre que se basen en referencias clínicas sólidas y bien documentadas garantizando la integridad del análisis.

Caracterizacion de Pacientes con Obesidad

La caracterización de pacientes con obesidad atendidos en la ciudad de Bogota, durante el periodo 2024 será realizada mediante el enfoque de agrupamiento difuso (Fuzzy Clustering), con el fin de identificar perfiles clínicos predominantes que permitan una mejor comprensión del comportamiento de esta condición en la población. Este enfoque permitirá clasificar a los pacientes considerando la naturaleza incierta y gradual de los datos clínicos, lo cual facilita una interpretación más realista de los perfiles observados.

Previo al análisis de clustering difuso, se llevará a cabo un análisis preliminar que permita comprender la estructura y comportamiento de los datos. Este análisis consiste en: (i) realizar un análisis descriptivo del conjunto de datos, calculando medidas de tendencia central (media, mediana) y de dispersión (varianza, desviación estándar), así como una exploración de la distribución de variables categóricas. Una vez finalizado el análisis descriptivo, se procederá a la normalización o estandarización de las variables numéricas, con el fin de que todas contribuyan equitativamente al proceso de agrupamiento, dado que la técnica de clustering está basada en distancias.

Posteriormente, se realizará una evaluación exploratoria de las relaciones entre variables, incluyendo un análisis de Correspondencias Múltiples (ACM) y un análisis de correlación mediante la función cor() de R-studio, este último con el fin de determinar la medida de distancia más adecuada, tal y como lo recomienda (Hastie et al., 2009). Si las variables no están fuertemente correlacionadas y han sido debidamente estandarizadas, el autor recomienda el uso de la distancia euclidiana. No obstante, cuando existe una correlación significativa entre las variables, es preferible utilizar la distancia de Mahalanobis o una distancia basada en la correlación, ya que estas métricas consideran la relación entre las variables. En particular, la distancia por correlación mide la similitud en la forma de los vectores, sin importar sus magnitudes, lo que la hace ideal para datos altamente correlacionados.

Una vez finalizado el análisis preliminar de los datos se procederá a llevar a cabo el algoritmo de agrupación difusa seleccionado de acuerdo a la literatura (véase capítulo 1), cuyo proceso se presenta a continuación

Estrategia de Construcción del Algoritmo (FCM)

Paso 1.Estandarización: Se realizo el proceso de estandarización de las variables para eliminar el efecto de diferencias de magnitud entre ellas. Esto se realizó utilizando la función scale() en R, la cual transforma los datos para que cada variable tenga media cero y desviación estándar uno.

Paso 2.Selección del número de clústeres (k): Para determinar un valor apropiado de grupos (k), se recurrió al método del índice de validez de particiones. se utilizó el paquete {NbClust} que sugiere el número óptimo de clústeres según el criterio del índice de Silhouette.

Paso 3.Selección de la distancia a utilizar: Se evalúa la posible correlación entre las variables, ya que esta puede influir en la elección de una métrica de distancia adecuada de acuerdo con lo establecido por (Hastie et al., 2009).

Paso 4.Aplicación del algoritmo**:** Una vez determinado k, se procedió a aplicar el algoritmo Fuzzy C-Means (FCM) mediante la función fcm() del paquete {ppclust}. Esta función implementa el algoritmo propuesto por (Dunn, 1973) y mejorado por (Bezdek, 1981) , el cual se basa en la minimización de una función objetivo difusa. Cabe Resaltar que dicha función inicializa automáticamente la matriz de pertenencia (U) con valores aleatorios, garantizando que las sumas de pertenencia de cada observación sean iguales a 1.

Paso 5. Conversión del modelo a estructura compatible con visualización: Como el resultado de fcm() no es directamente compatible con funciones de visualización como fviz_cluster(), se utiliza la función ppclust2() para convertir el modelo a un objeto tipo K-means.

Paso 6. Visualización de los clústeres: Los resultados se visualizaran mediante un gráfico de dispersión con elipses convexas que representan los grupos, usando la función fviz_cluster()

Paso 7. Validación de los resultados**:** La validación del modelo se realizó mediante índices de validez difusos, como:

- SIL.F (Índice Silhouette Fuzzy): Evalúa la cohesión e individualidad de los clusters, adaptado al contexto difuso. Se expresa como:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i)} \tag{12}$$

Donde b es la distancia del grupo más cercano y a la distancia media del grupo para cada muestra x_i. Valores cercanos a 1 indican particiones bien definidas, valores cercanos a -1 indican conglomerados equivocados.

- PE (Partition Entropy): Calcula la incertidumbre general de la partición. Cuanto más bajo sea el valor, más nítidos (menos ambiguos) son los clusters tal y como se muestra en *(13).*

$$V_{PE} = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} u_{j,i}\log_a(u_{j,i}) \tag{13}$$

Donde n es el número total de datos, u_{ji} el grado de pertenencia del dato i-ésimo al cluster j-ésimo.

- PC (Partition Coefficient): Mide el grado de nitidez de los cluster, no penaliza directamente la superposición de clusters. Se expresa como:

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} u_{j,i}^2 \tag{14}$$

- MPC (Modified Partition Coefficient): Corrige una debilidad del índice PC, haciendo que sus valores sean más comparables entre diferentes números de clusters. Se define como:

$$MPC = \frac{PC - \frac{1}{k}}{1 - \frac{1}{k}} \quad (15)$$

Donde k es el número de cluster.

Paso 8 . Analisis de robuztes de los cluster: Se realizo un análisis de consistencia de los cluster mediante repeticiones Bootstrap de 50 y se analiza su desviación estándar para detectar incosistencias.

Paso 9. Comparacion con otras variantes**:** El modelo seleccionado será comparado con otras variantes. Específicamente, el modelo Fuzzy C-Means (FCM) con distancia de correlación fue contrastado con el algoritmo Gustafson-Kessel (GK), el cual permite formas elipsoidales en los clústeres, y con una aproximación basada en la distancia de Mahalanobis. Esta comparación se realizó a partir de métricas descritas anteriormente para así evaluar la calidad, separación y ambigüedad de las particiones generadas por cada modelo.

Paso 10. Caracterización: Después de validar el modelo y verificar que las particiones se agrupan de manera adecuada,se procede a realizar la caracterización para describir y entender mejor los grupos o clusters obtenidos. Para ellos se utilizaron medidas como el mínimo, máximo, media y sesgo y se realizara un intervalo de confianza para la media con un nivel de significancia del 5%.

Paso 11. Interpretación de los cluster**:** La matriz de pertenencia fue analizada para

observar el grado de pertenencia de cada observación a los diferentes clústeres desde un enfoque netamente estadístico utilizando las métricas descritas anteriormente.

Tipo de Estudio

El tipo de estudio que se realizará es de carácter cuantitativo, dado que, según (Hernández et al., 2014) este enfoque busca generar conocimiento objetivo mediante un proceso

deductivo. A través de la recolección y el análisis estadístico de datos numéricos, se pueden obtener resultados confiables. En este sentido, se utilizarán variables clínicas (reales y simuladas) junto con técnicas de agrupamiento difuso para identificar perfiles en pacientes con obesidad.

Recolección de los Datos

El instrumento utilizado para la recolección de los datos fue de carácter secundario, ya que se trabajó con una base de datos pública que contenía información de pacientes con obesidad. Debido a las restricciones legales de acceso a datos clínicos sensibles, se complementó la base original mediante la simulación de variables médicas relevantes, como glucosa en ayunas, HbA1c, triglicéridos, HDL y LDL, conforme a parámetros definidos por organismos internacionales como la ADA y la OMS.

Resultados

Esta sección contiene los resultados obtenidos después de aplicar la metodología propuesta para segmentar perfiles clínicos de pacientes con obesidad mediante el algoritmo fuzzy clustering (FCM). Inicialmente se muestren los resultados obtenidos del análisis preliminar. Posteriormente se presentan los resultados obtenidos del algoritmo de segmentación empleado y finalmente se examina el desempeño del modelo con sus respectivos análisis estadísticos.

Analisis Preliminar de los Datos

La Tabla 4 muestra el resumen estadístico de cada variable numérica del dataset, conformado por 157 observaciones. Como se puede evidenciar, el peso promedio de los pacientes es de 80.2 kg, con un rango que va desde los 55 kg hasta los 117 kg, lo cual sugiere una notable dispersión en la masa corporal. En cuanto a la talla, la media se sitúa en 155.7 cm, siendo bastante homogénea, aunque se presentan algunos valores atípicos hacia los extremos (mínimo de 94 cm y máximo de 182 cm). La edad de los pacientes se concentra en adultos mayores, con un promedio de 63.4 años. El IMC medio es de 33.3, ubicando al grupo dentro del rango de obesidad clase I, con casos extremos que alcanzan hasta un IMC de 96.2, lo cual evidencia un grado de obesidad alto.

Con respecto a las variables metabólicas, la glucosa en ayunas tiene un valor medio de 114 mg/dL, mientras que la HbA1c tiene una media de 6.1%. Los triglicéridos presentan una alta variabilidad (desviación estándar de 45.7), con valores que en algunos casos superan ampliamente el umbral clínico deseable, alcanzando hasta 345 mg/dL. El HDL, conocido como el colesterol bueno muestra un valor promedio de 44.7 mg/dL. Finalmente, los niveles de LDL se ubican en un promedio de 140.6 mg/dL, indicando riesgo moderado.

Tabla 4

Resumen Estadístico de las Variables Clínicas

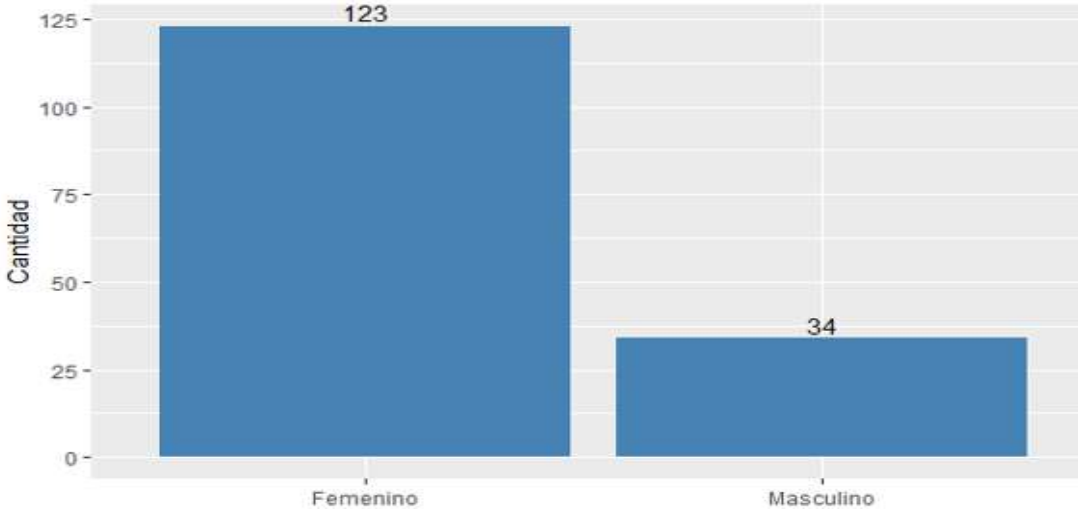
Variable	Promedio	Mediana	Varianza	Desviacón estándar	Mínimo	Máximo
Peso (kg)	80.21	77.00	156.65	12.52	55.0	117.0
Talla (cm)	155.69	156.00	104.87	10.24	94.0	182.0
Edad (años)	63.38	64.00	127.38	11.29	27.0	97.0
IMC (kg/m²)	33.33	31.92	46.64	6.83	25.3	96.2
Glucosa ayuna (mg/dL)	114.05	111.00	230.95	15.20	87.0	159.0
HbA1c (%)	6.12	6.00	0.38	0.62	5.2	8.4
Triglicéridos (mg/dL)	181.75	180.00	2086.06	45.67	121.0	345.0
HDL (mg/dL)	44.68	45.00	26.91	5.19	30.0	55.0
LDL (mg/dL)	140.64	140.00	486.93	22.07	100.0	199.0

Nota. Resumen estadístico de las variables clínicas reales y simuladas seleccionadas en el estudio.

Por otro lado, al examinar la variable género, se observa una clara predominancia del sexo femenino entre los pacientes con obesidad, con un total de 123 casos registrados. En contraste, únicamente 23 hombres presentan esta condición, lo que evidencia una distribución desigual entre ambos géneros dentro de la muestra analizada (véase figura 2).

Figura 2

Cantidad de Pacientes por Genero



No obstante, se realiza un análisis con el objetivo de explorar los factores sociales asociados a los niveles de obesidad, utilizando variables categóricas como el sexo, el nivel educativo, la ocupación y el nivel de obesidad.

La Figura 3 muestra las relaciones entre las categorías de sexo, ocupación, nivel educativo y nivel de obesidad, donde las categorías cercanas entre sí están más asociadas. La dimensión 1 (eje X) explica el 10.1% de la variabilidad y la dimensión 2 (eje Y), el 7.8%.

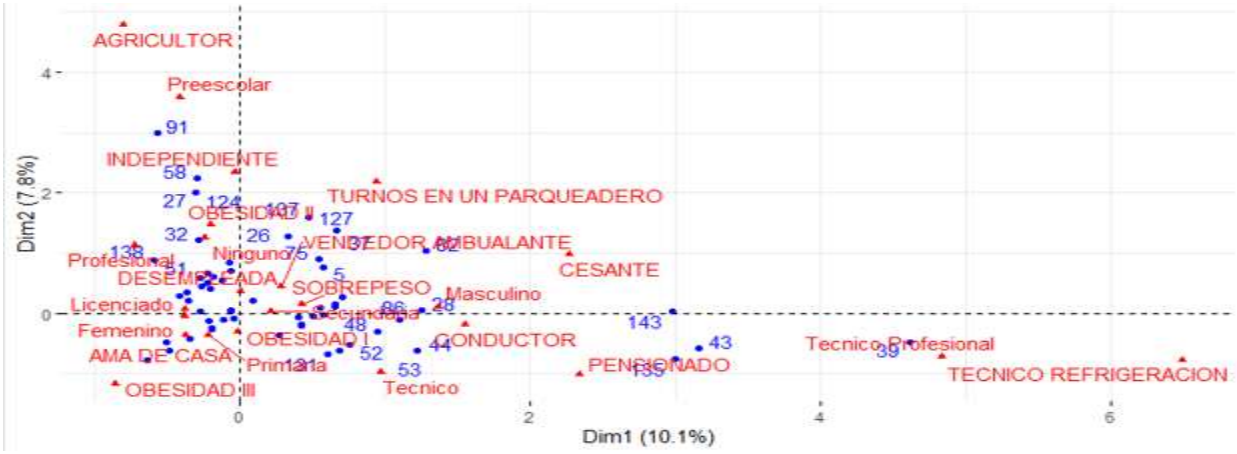
Como se puede observar, el género femenino se asocia estrechamente con los niveles educativos primaria y secundaria, así como con las ocupaciones de ama de casa, desempleada y vendedora ambulante. Este conjunto también se ubica cerca de las categorías Obesidad I y II, lo cual sugiere una posible tendencia de obesidad moderada en este perfil sociodemográfico. El tipo de Obesidad III se posiciona más aislada, hacia el cuadrante inferior izquierdo, lo que podría indicar un perfil menos común, posiblemente vinculado con características específicas no compartidas por la mayoría del grupo. El género masculino se relaciona más con las

ocupaciones como conductor y pensionado, lo que sugiere que en este grupo podría predominar la obesidad leve o el sobrepeso, ya que también se encuentra próximo a esas categorías.

En la parte superior de la figura se observan ocupaciones como agricultor, independiente y trabajo por turnos en parqueadero, junto a niveles bajos de escolaridad como ninguno o preescolar, asociados a la categoría de Obesidad II, lo que puede señalar un grupo vulnerable por su baja escolaridad y condiciones laborales informales.

Figura 3

Análisis de Correspondencia Multiple



Después de examinar los factores mediante el Análisis de Correspondencias Múltiple (ACM), se procede a explorar la relación entre las variables cuantitativas del estudio mediante un análisis de correlación. Este análisis permite identificar la intensidad y dirección de las asociaciones entre variables como el peso, la talla, la edad, el índice de masa corporal (IMC), y los exámenes médicos tales como Glucosa ayuno HbA1 c trigicelidos, colesterol.

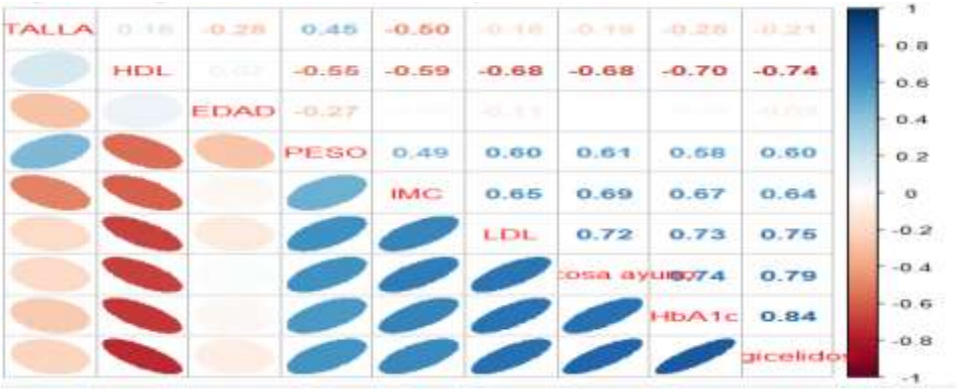
De acuerdo a la Figura 4, se puede evidenciar que la relación entre varias de las variables es alta, ya que las esferas representadas en la matriz están bien definidas, con un tamaño considerable y una coloración intensa a excepción de la edad cuyas esferas son opacas indicando

una correlación baja (<0.3). En dicha figura se destaca una fuerte asociación positiva entre HbA1c y triglicéridos, reflejada con una esfera de color azul, correspondiente a un coeficiente de correlación de 0.84, lo que indica que a mayor HbA1c, también tienden a aumentar los triglicéridos. También se puede observar un correlación positiva y moderada entre variables como IMC, colesterol ldl, glucosa, peso y la talla, lo cual era de esperarse puesto que a mayor talla y peso mayor IMC, a mayor peso, mayor colesterol del malo y niveles de glucosa.

Por otro lado, también se observan correlaciones negativas, como la relación entre el IMC y el HDL (colesterol "bueno"), con un coeficiente de -0.59, lo que indica que, a menor índice de masa corporal, mayores niveles de HDL (colesterol del bueno) se presentan. Esta asociación aparece en el gráfico con una esfera de color rojo intenso. Asimismo, se observa una correlación negativa entre los triglicéridos y el HDL (-0.74), lo cual también es esperable desde un punto de vista clínico, dado que niveles elevados de triglicéridos suelen estar asociados a una disminución del colesterol "bueno".

Figura 4

Correlación entre las Variables



Para corroborar la información visualizada en el gráfico de correlación, se realizó una prueba de correlación entre las variables, considerando un nivel de significancia del 5% (α =

0.05). Los resultados obtenidos muestran que los valores p (p-values) asociados a la mayoría de las correlaciones son 0 o cercanos a cero, lo que permite concluir que existe una relación estadísticamente significativa entre estas variables (véase Tabla 5).

Es importante destacar que la edad de los pacientes no presenta significancia estadística en la mayoría de sus relaciones con otras variables, con valores p superiores al 0.05. Por lo tanto, se considera que la edad es la única variable que no muestra una relación estadísticamente significativa con las demás. No obstante, esto no afecta negativamente el modelo, ya que el objetivo principal es analizar la estructura de correlación entre las variables para posteriormente definir una medida de distancia adecuada**.**

Dado que la mayoría de las variables están correlacionadas entre sí, y según lo sugerido por (Hastie et al., 2009) se procederá a utilizar una medida de distancia basada en correlaciones, lo cual resulta apropiado para técnicas como el fuzzy clustering, donde la similitud entre pacientes se puede definir a partir del grado de asociación entre sus características.

Tabla 5

Valor P

Variable	Peso	Talla	Edad	IMC	Glucosa ayuno	HbA1c	Triglicéridos	HDL	LDL
Peso	-	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Talla	0.00	-	0.00	0.00	0.13	0.02	0.08	0.28	0.17
Edad	0.00	0.00	-	0.62	1.00	1.00	1.00	1.00	1.00
IMC	0.00	0.00	0.62	-	0.00	0.00	0.00	0.00	0.00
Glucosa ayuno	0.00	0.01	0.88	0.00	-	0.00	0.00	0.00	0.00
HbA1c	0.00	0.00	0.66	0.00	0.00	-	0.00	0.00	0.00

Triglicéridos	0.00	0.01	0.28	0.00	0.00	0.00	-	0.00	0.00
HDL	0.00	0.04	0.36	0.00	0.00	0.00	0.00	-	0.00
LDL	0.00	0.02	0.17	0.00	0.00	0.00	0.00	0.00	-

Nota. Valor P de la prueba de correlación entre las variables

Posterior al análisis de clustering difuso, se evaluó la multicolinealidad entre las variables numéricas mediante el cálculo del Factor de Inflación de la Varianza (VIF). Los resultados obtenidos en laTabla 6 indicaron que todos los VIF se mantuvieron por debajo del umbral crítico de 5, lo que sugiere una colinealidad aceptable. Por tanto, no fue necesario eliminar ninguna variable, permitiendo conservar toda la información relevante del conjunto de datos original para el análisis de agrupamiento. En consecuencia, se reducio el ruido de los datos con función z-score, detectetando solo un valor atipico correspondiente al individuo 98 siendo este eliminado.

Tabla 6

Análisis de Multicolinealidad (VIF)

Variable	VIF
Peso	3.2
Talla	2.9
Edad	1.8
IMC	4.5
Glucosa ayuno	2.6
HBA1c	2.3
Triglicéridos	3.9
HDL	2.7
LDL	3.5

Nota. Valor del vif para las variables reales y simuladas.

Aplicación del Método C-means Clustering (FCM)

Se estandarizaron las variables cuantitativas, obteniendo una media cercana a cero y una varianza unitaria en todas ellas, lo que permitió su comparabilidad y evitó sesgos en el cálculo de distancias.

Tabla 7

Encabezado Variables Estandarizadas

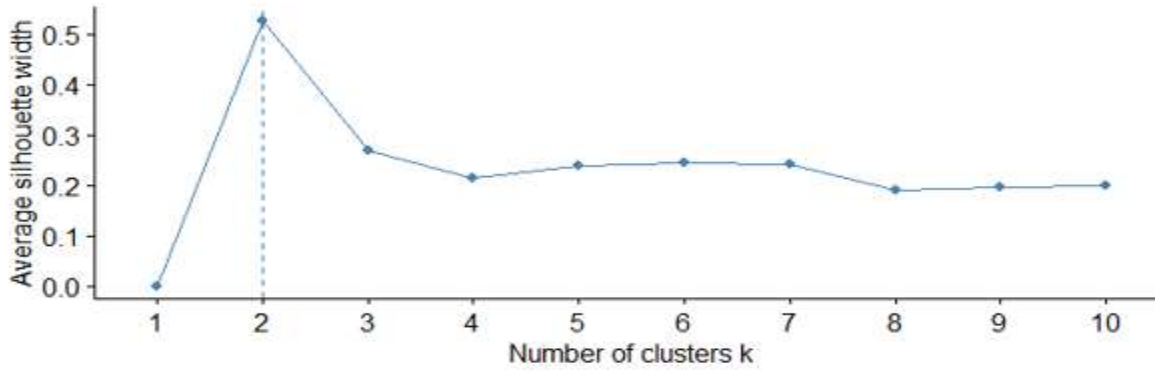
Variable	Media	Desviación Estándar
Peso	0.00	1.00
Talla	0.00	1.00
IMC	0.00	1.00
Triglicéridos	0.00	1.00

Nota. Resultados de las primeras 5 variables estandarizadas

Previo a la estandarización de las variables, se determina el número óptimo de clústeres utilizando el método de la silueta. Como resultado, se identificó que la mejor partición de los datos corresponde a 2 clústeres, lo cual sugiere la existencia de dos grupos diferenciados en la estructura original de los datos.

Figura 5

Numero Optimo de Cluster



Antes de ejecutar el algoritmo de agrupamiento fuzzy, se optó por utilizar el valor por defecto del parámetro de difusividad m = 2, , el cual es ampliamente recomendado en la literatura por (Bezdek, 1981) como estándar para estudios con incertidumbre. Este valor ofrece un equilibrio razonable entre la definición clara de los clústeres y la representación de ambigüedad característica del enfoque fuzzy. Además, estudios recientes como el de Huang et al. (2024) señalan que el valor m = 2 no solo permite suavizar adecuadamente las pertenencias difusas, sino que también contribuye a mitigar el impacto del ruido en los datos. Los autores argumentan que valores muy cercanos o muy lejanos a 2 pueden afectar negativamente la estabilidad del modelo y la interpretabilidad de los clústeres, por lo que recomiendan m = 2 como una opción robusta y balanceada.

Una vez implementado el algoritmo con los parámetros establecidos se obtuvieron los resultados de la matriz de pertenencia, donde cada fila representa una observación y cada columna indica el grado en que dicha observación pertenece a cada clúster. A continuación, se muestra la matriz de pertenecía.

Tabla 8

Matriz de Pertenencia de la Partición en Dos Clústeres

Obs	Cluster 1	Cluster 2	Obs	Cluster 1	Cluster 2
1	0.4073	0.5927	21	0.0483	0.9517
2	0.9697	0.0303	22	0.9049	0.0951
3	0.9560	0.0440	23	0.5902	0.4098
4	0.5407	0.4593	24	0.8482	0.1518
5	0.7530	0.2470	25	0.0708	0.9292
6	0.9487	0.0513	26	0.2068	0.7932
7	0.6156	0.3844	27	0.1809	0.8191

8	0.8948	0.1052	28	0.8164	0.1836
9	0.5142	0.4858	29	0.7060	0.2940
10	0.6944	0.3056	30	0.8665	0.1335
11	0.4271	0.5729	31	0.8720	0.1280
12	0.5545	0.4455	32	0.0812	0.9188
13	0.9207	0.0793	33	0.4292	0.5708
14	0.8729	0.1271	34	0.9707	0.0293
15	0.9085	0.0915	35	0.5800	0.4200
16	0.9476	0.0524	36	0.9693	0.0307
17	0.8981	0.1019	37	0.1281	0.8719
18	0.5776	0.4224	38	0.8639	0.1361
19	0.8770	0.1230	39	0.9911	0.0089
20	0.6873	0.3127	40	0.9396	0.0604

Nota. Encabezado de la *Matriz de Pertenencias para 20 pacientes.*

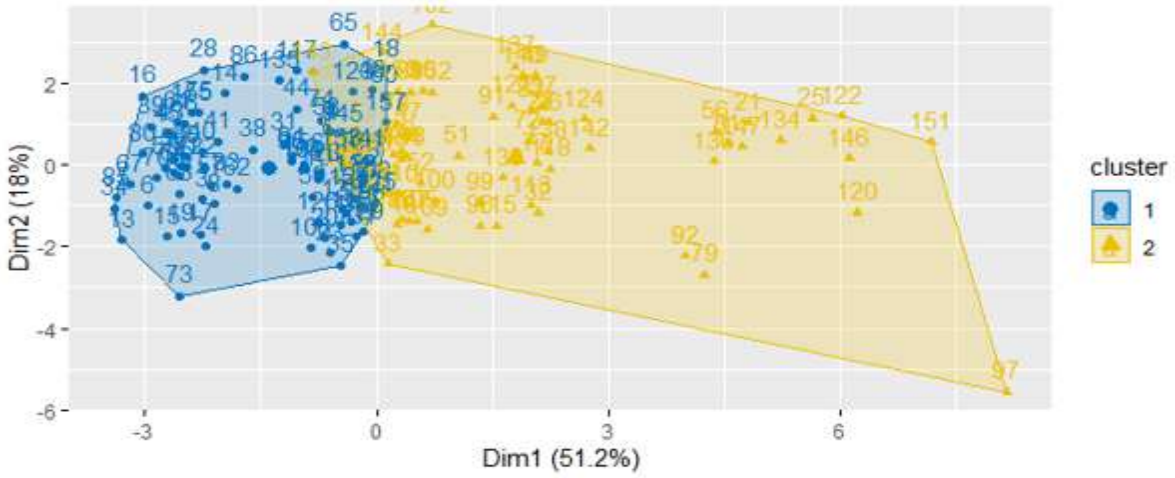
Como se muestra en dicha tabla, algunas observaciones presentan un grado de pertenencia claramente definido. Por ejemplo, las observaciones 2, 3, 6 y 15 con alta pertenencia al clúster 1, mientras que otras presentan grados más equilibrados entre observaciones, como la 1 4 9, lo cual justifica el uso de un enfoque difuso en lugar de un enfoque tradicional jerárquico o aglomerativo.

Gráficamente, como se observa en la Figura 6, el grado de traslapamiento entre los clústeres evidenciando que un individuo puede pertenecer a ambos grupos específicamente se observa un bloque notorio de individuos en el centro del clúster lo que indica grados de pertenencia bastante parecidos a más de un clúster , pero con diferente probabilidad lo que representan mejor la realidad de los datos desde un enfoque difuso. Además, en la gráfica se destaca un valor extremo correspondiente al individuo 97. Este dato podría ser un indicio que

dicho individuo tiene una mayor probabilidad de padecer enfermedades relacionadas con la obesidad y con características diferentes a los demás pacientes.

Figura 6

Visualización de los Clusters



Evaluación del Modelo

La Tabla 9 muestra las métricas de desempeño del modelo de agrupamiento difuso. Se observa que el Índice de Silueta Difuso indica una calidad de partición moderadamente buena, lo que sugiere una separación aceptable entre los clústeres. Por otro lado, el Coeficiente de Partición de 0.69 y el Coeficiente de Partición Modificado 0.38 respaldan una estructura de clústeres definida, pero con cierto grado de traslapamiento, lo cual es de esperarse en métodos donde los datos son inherentes a la incertidumbre. La Entropía de Partición de 0.47 refleja un nivel medio de incertidumbre en la asignación de los datos a los clústeres.En conjunto, estos indicadores sugieren que el modelo es aceptable y proporciona una segmentación válida de los datos para su posterior análisis de caracterización e interpretación de los perfiles.

Tabla 9

Métricas de Desempeño

Métrica	Valor
Índice de Silueta Difuso (Fuzzy SI)	0.81
Entropía de la Partición (PE)	0.4683
Coeficiente de Partición (PC)	0.6909
Coeficiente de Partición Modificado (MPC)	0.3819

Nota. Metricas de desempeño para evaluar el rendimiento del modelo.

Robustez de los Clusters

El análisis de robustez mediante remuestreo bootstrap mostró niveles de consistencia moderados en la asignación de clústeres. El promedio de consistencia fue del 71.8%, con un máximo de 82.4% y un mínimo de 60%. Estos resultados reflejan una partición razonablemente estable, aunque con cierta sensibilidad a la variabilidad de la muestra. Dicha fluctuación es coherente con la naturaleza difusa del método, donde se permite la pertenencia parcial de cada observación a múltiples clústeres.

Tabla 10

Consistencia de los Clusters

Estadístico	Valor
Mínimo	0.6000
1er cuartil	0.4932
Mediana	0.7
Media	0.7183
3er cuartil	0.6452

Máximo	0.8241

Nota. Consistencia de los cluster para evaluar la precisión del algoritmo

Con el fin de evaluar la estabilidad de la pertenencia difusa de los individuos a los clústeres, se calculó la desviación estándar (SD) de los valores de pertenencia obtenidos en la muestra Bootstrap . Los resultados mostrados en la tabla 11 indican que un subconjunto de observaciones presentó desviaciones en torno a 0.33, lo cual sugiere una variabilidad moderada en sus niveles de pertenencia. No obstante, la mayoría de las observaciones mantuvieron una desviación estándar inferior a este rango, lo cual respalda una partición globalmente robusta.

Tabla 11

Variabilidad de los Indices Difusos

ID Paciente	Cluster1_SD	Cluster2_SD
1	0.3447538	0.3447538
2	0.3405275	0.3405275
3	0.3400703	0.3400703
4	0.3379344	0.3379344
5	0.3358061	0.3358061

Nota. Encabezado de la Variabilidad de los Indices Difusos para los dos clusters

Validación del Modelo de Agrupamiento Seleccionado

Para validar la idoneidad del modelo Fuzzy C-Means con distancia de correlación, se realizó una comparación con otras variantes del algoritmo, incluyendo Gustafson-Kessel y una aproximación basada en la distancia de Mahalanobis. Esta validación se llevó a cabo mediante métricas de evaluación internas como el índice de silueta difusa, la entropía de partición (PE), el coeficiente de partición (PC) y su versión modificada (MPC). Aunque Gustafson-Kessel muestra

valores extremos, el índice de silueta negativo sugiere agrupaciones artificiales, posiblemente por sobreajuste. El modelo FCM con correlación ofrece un balance más realista entre separación y pertenencia difusa que el modelo con distancia mahalanobis, es decir, los resultados obtenidos permiten inferir que el modelo seleccionado y empleado Fuzzy C-Means con distancia de correlación empleado y seleccionado en este estudio continúa siendo más adecuado.

Tabla 12 .

Comparacion de los Modelos

Modelo	Fuzzy Silhouette	Partition Entropy	Partition Coefficient	MPC
FCM (Correlation)	0.81	0.468	0.691	0.382
FCM (Mahalanobis)*	0.447	0.392	0.744	0.488
Gustafson-Kessel	-0.043	8.00e-15	1.000	1.000

Nota. Comparación de los modelos basados en FCM correlation y Mahalanobis vs Gustafson

Caracterización de los Grupos

La caracterización presentada en la siguiente tabla se fundamenta en el análisis de las variables clínicas y antropométricas evaluadas para los dos clústeres obtenidos mediante el enfoque de agrupamiento difuso (fuzzy clustering). Dicha tabla ofrece una visión detallada de los valores estadísticos principales para variables clave como el peso, talla, edad, índice de masa corporal (IMC), glucosa en ayuno, hemoglobina glicosilada (HbA1c), triglicéridos, HDL y LDL.

Tabla 13

Caracterización de los Clusters

Variable	Estadística	Clúster 1	Clúster 2
Peso	min	55.0	63.5
	max	101	117
	mean	74.93977	86.94058

	sd	9.479046	12.748072
	sesgo	0.4971784	0.4594561
Talla	min	135	94
	max	182	178
	mean	156.8977	154.1449
	sd	9.09849	11.41758
	sesgo	0.3914957	-1.9531664
Edad	min	39	27
	max	97	88
	mean	65.98864	60.04348
	sd	9.678147	12.340077
	sesgo	0.2678980	-0.1674519
IMC	min	25.30	30.41
	max	38.95	96.20
	mean	30.40364	37.06304
	sd	2.527403	8.579526
	sesgo	0.3699365	4.8518620
Glucosa ayuno	min	87	100
	max	125	159
	mean	106.5795	123.5797
	sd	9.770612	15.598271
	sesgo	0.0505869	0.5260445

HbA1c	min	5.2	5.7
	max	6.4	8.4
	mean	5.782955	6.547826
	sd	0.3014225	0.6493455
	sesgo	0.06488087	1.25850563
Triglicéridos	min	121	156
	max	200	345
	mean	157.5682	212.5797
	sd	25.01622	47.57947
	sesgo	0.1651785	1.3697727
HDL	min	40	30
	max	55	48
	mean	48.03409	40.39130
	sd	3.408741	3.695058
	sesgo	-0.04085236	-0.85118525
LDL	min	100	121
	max	159	199
	mean	127.5227	157.3623
	sd	14.32543	18.65710
	Sesgo	0.3679985	0.1171139

Nota. Caracterizacion de clusters para las variables reales y simuladas.

La Tabla 14 muestra los intervalos de confianza al 95% para la media de las variables cuantitativas, separados por clúster. Estos intervalos permiten identificar diferencias significativas entre los grupos formados mediante clustering difuso. Se observa, que el clúster 1 presenta valores promedio más altos en peso, IMC, glucosa en ayuno, HbA1c y triglicéridos, lo que sugiere un perfil metabólico más comprometido en comparación con el clúster 2. Por el contrario, el clúster 2 muestra mayores niveles promedio de HDL, considerado un factor protector. Específicamente El Clúster 1 se caracteriza por presentar un peso promedio entre entre 83.88 y 90.00 kg, con un IMC en el rango de 35.00 a 39.12, lo que indica obesidad tipo II o incluso tipo III. En contraste el cluster 2 presenta peso promedio y IMC menro de 72.93 a 76.95 y 29.87 a 30.94 indicando un riesgo mas bajo que el cluster 1.

Tabla 14

Intervalos de Confianza

Variable	Cluster 1 (IC 95%)	Cluster 2 (IC 95%)
PESO	83.88 – 90.00	72.93 – 76.95
TALLA	151.40 – 156.89	154.97 – 158.83
EDAD	57.08 – 63.01	63.94 – 68.04
IMC	35.00 – 39.12	29.87 – 30.94
Glucosa ayuno	119.83 – 127.33	104.51 – 108.65
HbA1c	6.39 – 6.70	5.72 – 5.85
Triglicéridos	201.15 – 224.01	152.27 – 162.87
HDL	39.50 – 41.28	47.31 – 48.76
LDL	152.88 – 161.84	124.49 – 130.56

Nota. Intervalos de Confianza para la Media de los Clusters

Análisis de Perfiles

Cluster 1. Riesgo moderado de enfermedades metabólica: Agrupa principalmente a mujeres adultas mayores, con una edad media de 65.9 años, ligeramente superior a la del Clúster 2. En cuanto al estado nutricional, este grupo presenta un IMC medio de 30.4, lo que indica un rango de obesidad clase I. El valor máximo de IMC en este clúster es de 38.95, lo que sugiere que, aunque hay presencia de obesidad, no se alcanzan niveles extremos. Además, los niveles promedio de glucosa en ayunas (106.5 mg/dL) y HbA1c (5.78%) sugieren un posible estado de prediabetes en varios casos de acuerdo con los valores establecidos por (Centers for Disease Control and Prevention, 2024) . Este perfil sugiere un riesgo moderado de enfermedades metabólicas, siendo relevante la edad, el IMC, y los laboratorios clínicos como factor contribuyente.

Cluster 2. Riesgo cardiovascular y complicaciones asociadas a obesidad**:** Está compuesto por mujeres y hombres con edad media de 60 años, con una media de IMC de 37.06, lo que las ubica en un rango de obesidad clase II a III, con un valor máximo alarmante de 96.2. Este grupo muestra valores más elevados de glucosa en ayunas (123.6 mg/dL) y HbA1c (6.54%), lo que apunta a una mayor prevalencia de diabetes tipo 2 o riesgo elevado de padecerla (Centers for Disease Control and Prevention, 2024). Además, los triglicéridos y el LDL también son más altos en este clúster. Estadísticamente, este grupo representa un perfil metabólicamente más comprometido debido a que los pacientes presentan mayor grado de obesidad con aumentos severos de exámenes médicos. De acuerdo con la Organización Mundial de la Salud (0MS) y (Gómez & Pérez, 2023) cuando se obtienen resultados como los de esta caracterización se llegan a provocar un aumento del riesgo de diabetes tipo 2 y cardiopatías lo que hace que este grupo

represente un perfil metabólicamente más comprometido, con múltiples indicadores de riesgo cardiovascular y complicaciones asociadas a obesidad severa.

Conclusiónes

En este trabajo se consideraron datos de pacientes con obesidad en Bogotá, Colombia, durante el periodo 2024. Este enfoque incluyo tanto métodos estadísticos descriptivos, tales como el análisis de tendencias centrales, correlaciones y visualización de datos, como técnicas de agrupación difusa, aplicadas a variables relacionadas con indicadores de salud. A lo largo de este análisis, se demostraron una serie de aspectos fundamentales que ayudan a entender la relación entre el nivel de obesidad y ciertos parámetros clínicos, así como también las ventajas y limitaciones de este método, teniendo en cu enta las características específicas de los datos y los objetivos del estudio.

Adicionalmente, a partir de los procesos de modelado realizados, se puede concluir lo siguiente:

Se probo que el método de clustering difuso es eficaz para caracterizar pacientes con obesidad, especialmente en contextos donde existe incertidumbre en los datos o traslape entre grupos. Durante el análisis, se observó que varios pacientes presentaban características comunes a más de un grupo, lo cual justifica el uso de una partición difusa en lugar de un enfoque tradicional.

En particular, para este tipo de datos clínicos con incertidumbre, donde las variables presentan correlaciones entre sí, se demostró que el uso de una distancia basada en correlaciones fue mas adecuado que sus variantes. Esta medida de similitud permitió una mejor representación de las relaciones internas de los datos, facilitando la formación de clústeres bien definidos y coherentes con la realidad clínica.

Se mostró que, si bien los datos de laboratorio fueron simulados a partir de información real, el modelo resultante presentó un desempeño adecuado, evidenciado por métricas de

evaluación que reflejaron buena cohesión y separación entre los clústeres generados. Además, los grupos identificados correspondieron a perfiles clínicos coherentes y con sentido dentro del contexto de la obesidad. Esto refuerza la idea de que, cuando se diseñan correctamente, los datos sintéticos pueden ser una herramienta válida y útil para explorar patrones y validar metodologías en contextos donde el acceso a datos reales es limitado, especialmente en áreas sensibles como la salud.

Se identifico patrones clínicos ocultos relevantes para la caracterización de pacientes con obesidad. En particular, se evidenció un primer grupo con riesgo moderado de enfermedades metabólicas, compuesto principalmente por mujeres adultas mayores caracterizdo con patrones asociados a la obesidad clase I . Por otro lado, el segundo grupo reflejó un perfil clínico de mayor severidad, con individuos de ambos sexos que presentaban características asociadas a la obesidad clase II y III, acompañada de alteraciones significativas en los niveles de glucosa y hemoglobina glicosilada, lo que sugiere un riesgo elevado de complicaciones cardiovasculares y metabólicas.

Recomendaciones

Se recomienda que futuros trabajos continúen utilizando datos reales combinados con simulaciones controladas, especialmente en variables clínicas difíciles de obtener. Asimismo, sería pertinente explorar otro algoritmos de clustering difuso, con el fin de reforzar la robustez del análisis en poblaciones diversas. Finalmente, se sugiere ampliar los datos incluyendo nuevas variables clínicas y antropmoetricas que permitan una caracterización aún más detallada de los perfiles de obesidad.

Referencias Bibliográficas

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., & Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Computer Methods and Programs in Biomedicine, 161*, 145-172. https://doi.org/10.1016/j.cmpb.2018.04.013

Alarcón, M. Á. M., Mas, M. T., Morales-Gabardino, J. A., & Buitrago-Ramírez, F. (2021). Prevalencia y grado de control de los factores de riesgo cardiovascular en pacientes con cardiopatía isquémica adscritos a un centro de salud urbano: E202102040. *Revista Española de Salud Pública*, *95*, 6-páginas.

Amaro-López, L., Hernández-González, P. L., Hernández-Blas, A., & Hernández-Arzola, L. I. (2019). La simulación clínica en la adquisición de conocimientos en estudiantes de la Licenciatura de Enfermería. *Enfermería Universitaria*, *16*(4), 402-413. https://doi.org/10.22201/eneo.23958421e.2019.4.543

American Diabetes Association. (2024). *Standards of Medical Care in Diabetes—2024*. https://diabetesjournals.org/care/article/47/Supplement_1/S1/153191/Standards-of-Care-in-Diabetes-2024

Aslan, B., & Hızıroğlu, O. A. (2024). Prediction of Lung Cancer with Fuzzy Logic Methods: A Systematic Review. *Artificial Intelligence Theory and Applications*, *4*(2), 155-192.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer.

Bhatti, P., Mahboob, K., Naeem, S. S., Bhatti, I. H., & Kamran, N. (2023). Enhanced Diabetic Prediction Using Fuzzy C-Means Preprocessing and Random Forest Ensemble Learning. *VFAST Transactions on Software Engineering*, *11*(4), 32-44.

Bijuraj, L. (2013). Clustering and its Applications. *Proceedings of National Conference on New Horizons in IT-NCNHIT*, *169*, 172.

Centers for Disease Control and Prevention. (2024). *Centros para el Control y la Prevención de Enfermedades (CDC)—Página principal en español*. https://www.cdc.gov/spanish/index.html

Clínica Universidad de los Andes. (2021). *7 enfermedades asociadas a la obesidad*. https://www.clinicauandes.cl/noticia/7-enfermedades-asociadas-a-la-obesidad-que-podr%C3%ADan-mejorar-con-cirugia-bariatrica

Congreso de Colombia. (2012). *Ley 1581 de 2012: Por la cual se dictan disposiciones generales para la protección de datos personales*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Constitución Política de Colombia. (1991). *Artículo 15*. https://www.constitucioncolombia.com/titulo-2/capitulo-1/articulo-15

Contreras Contreras, G. F., Medina Delgado, B., Acevedo Jaimes, B. R., & Guevara Ibarra, D. (2022). Metodología de desarrollo de técnicas de agrupamiento de datos usando aprendizaje automático. *Tecnura*, *26*(72), 42-58. https://doi.org/10.14483/22487638.17246

Datos.gov.co. (2023). *Enfermedades Crónicas*. https://www.datos.gov.co/Salud-y-Protecci-n-Social/Enfermedades-Cr-nicas/2uxx-gxp3

Dunn, J. C. (1973). *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters* [Ph.D. dissertation]. University of Illinois.

Edla, D. R., Lone, T., Tapas, N., & Kuppili, V. (2020). Analysis of high dimensional brain data using prototype based fuzzy clustering. *Clinical Epidemiology and Global Health, 8*(4), 1110-1118.

Giordani, P., Ferraro, M. B., Martella, F., Giordani, P., Ferraro, M. B., & Martella, F. (2020). *Introduction to Clustering*. Springer.

Gómez, M., & Pérez, J. (2023). Relación entre obesidad y riesgo cardiovascular en adultos. *Revista Española de Cardiología*, *76*(5), 412-420. https://doi.org/10.1016/j.rec.2023.01.015

Gómez Martínez, V., De la Torre Archundia, E. R., Sánchez Sandoval, A. G., Madin Juárez, B., Dávila Villada, M. S., & Álvarez Orozco, M. E. (2021). Simulación clínica en estudiantes de enfermería ante la pandemia por COVID-19. *Revista Salud y Cuidado*. https://revistasaludycuidado.uaemex.mx/article/view/24452

Hantoli, Y. N. Y. & others. (2021). *Prediction and Classification Analytics of Obesity Datasets Using a Hybrid Model of Clustering and Neuro-Fuzzy Methods* □□□□□ □□□□□□□□ [PhD Thesis]. AAUP.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill Education.

Khamis, G. S. M., Al Qahtani, N. S., Alanazi, S. M., Alruwaili, M. M., Alenazi, M. S., & Alruwaili, M. A. (2024). *Utilizing Fuzzy C-Means Clustering and PCA in Public Health: A Machine Learning Approach to Combat CVD and Obesity*.

MedlinePlus. (2021). *Riesgos de la obesidad para la salud*. https://medlineplus.gov/spanish/ency/patientinstructions/000348.htm

Ministerio de Salud de Colombia. (1999). *Resolución 1995 de 1999: Por la cual se establecen normas para el manejo de la historia clínica*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=2928

Ministerio de Salud y Protección Social. (2023). *Informe sobre la prevalencia del sobrepeso y obesidad en Colombia*. https://www.minsalud.gov.co

Nagase, Y., Satoh, T., Shigetome, K., Tokumaru, N., Matsumoto, E., Yamada, K. D., Imafuku, T., Watanabe, H., Maruyama, T., Ogata, Y., Yoshida, M., Saruwatari, J., & Oniki, K. (2023). Serum Fatty Acid Composition Balance by Fuzzy C-Means Method in Individuals with or without Metabolic Dysfunction-Associated Fatty Liver Disease. *Nutrients*, *15*(4). https://doi.org/10.3390/nu15040809

Nedyalkova, M., Barazorda-Ccahuana, H. L., Sârbu, C., Madurga, S., & Simeonov, V. (2020). Fuzzy partitioning of clinical data for DMT2 patients. *Journal of Environmental Science and Health, Part A*, *55*(12), 1450-1458. https://doi.org/10.1080/10934529.2020.1809925

Obesidad, S. E. para el E. de la. (2023). *Cálculo de IMC*. https://www.seedo.es/index.php/herramientas-seedo/calculo-de-imc

Rojas Diaz, J., Chavarro Porras, J. C., & Moreno Laverde, R. (2009). Tecnicas de logica difusa aplicadas a la mineria de datos. *Scientia et Technica*, *3*(40).

Salud, O. M. de la. (2024). *Obesidad y sobrepeso*. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

Sánchez, L. M. S., Martínez, N. P., Palacios, L. D., & Orozco, K. E. (2020). Prevalencia de sobrepeso, obesidad y factores de riesgo en una cohorte de escolares en Bogotá, Colombia. *Pediatria*, *53*(1), 5-13.

Sato-Ilic, M., & Jain, L. C. (2006). Introduction to Fuzzy Clustering. En *Innovations in Fuzzy Clustering: Theory and Applications*. Springer Berlin Heidelberg.

Sulla-Torres, J., Soto-Paredes, C., Cárdenas-Soria, R., & Huancco-Coila, L. (s. f.). *Sistema Neurodifuso con Optimización por Enjambre de Partículas para la Clasificación de la Obesidad en Niños y Adolescentes*.

Sümbül-Şekerci, B., Pasin, Ö., Egeli, D., Gönenç, S., & Şekerci, A. (2024). Characterizing cognitive phenotypes and clinical correlates in type 2 diabetes using fuzzy clustering and decision tree analysis. *Scientific Reports*, *14*(1), 23965. https://doi.org/10.1038/s41598-024-74741-6

Takeshita, S., Nishioka, Y., Tamaki, Y., Kamitani, F., Mohri, T., Nakajima, H., Kurematsu, Y., Okada, S., Myojin, T., Noda, T., Imamura, T., & Takahashi, Y. (2024). Novel subgroups of obesity and their association with outcomes: A data-driven cluster analysis. *BMC Public Health*, *24*(1), 124. https://doi.org/10.1186/s12889-024-17648-1

Velmurugan, T., & Emayavaramban, K. (2025). Performance Analysis of K-Means and Fuzzy C-Means (FCM) Clustering Algorithms for Diabetic Dataset. En *Intelligent Manufacturing and Cloud Computing* (pp. 130-137). IOS Press.

William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2019). Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Informatics in Medicine Unlocked*, *14*, 23-33. https://doi.org/10.1016/j.imu.2019.02.001

World Health Organization. (2000). *Obesity: Preventing and Managing the Global Epidemic* (Vol. 894). World Health Organization. https://iris.who.int/handle/10665/42330

Zheng, Y., Xu, Z., Wu, T., & Yi, Z. (2024). A systematic survey of fuzzy deep learning for uncertain medical data. *Artificial Intelligence Review*, *57*(9), 230. https://doi.org/10.1007/s10462-024-10871-7

Ápendices

Apéndice A

Sofware Utilizado

Para la implementación del algoritmo de clustering difuso, se utilizó el software R, versión 4.1.1, ejecutado desde el entorno de desarrollo RStudio. Durante el desarrollo del análisis se emplearon diversas librerías especializadas, entre ellas:

ppclust y fclust, utilizadas para la aplicación del algoritmo Fuzzy C-Means y sus variantes (como Gustafson-Kessel);

cluster y factoextra, para determinar el número óptimo de clústeres y realizar visualizaciones de agrupamiento;

dplyr y tidyr, para la manipulación y transformación eficiente de los datos;

ggplot2, utilizada en la generación de gráficos personalizados;y boot, en el caso de realizar remuestreo con bootstrap.

Apéndice B

Recursos Tecnicos

El procesamiento del análisis se realizó en un equipo con características computacionales de 4 núcleos de procesamiento y 12 GB de memoria RAM. La ejecución promedio de cada bloque de código fue de aproximadamente 30 segundos, con excepción del procedimiento de remuestreo mediante bootstrap, el cual presentó una duración promedio de 5 minutos debido a la intensidad computacional requerida por el número de iteraciones y la replicación del modelo.