INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Proyecto Análisis de datos de COVID-19 en México usando algoritmos de analítica avanzada vistos en clase para identificar patrones en los pacientes.

Carrera:
Licenciatura en Ciencia de Datos

Asignatura:
Analitíca Avanzada de Datos

Docente:
Daniel Jimenez Alcantar

Alumno:
Perez Hurtado Luis Rogelio
Escudero Gutiérrez Evelyn Abril

Grupo: 6AV1
Fecha: 10/01/2026

Análisis de Patrones de Comorbilidad y Riesgo en Pacientes COVID-19 en México mediante Algoritmos de Clustering (K-means y Fuzzy C-means)

Objetivo General: Analizar la base de datos de COVID-19 de la Secretaría de Salud de México utilizando algoritmos de aprendizaje no supervisado para identificar agrupamientos (clusters) de pacientes basados en perfiles demográficos y clínicos.

Objetivos Específicos:

1. Implementar procesos de limpieza y transformación de datos para normalizar las variables categóricas y numéricas del dataset de la DGE.
2. Aplicar y comparar los algoritmos K-means y C-means para segmentar a la población contagiada en grupos homogéneos.

Resumen

El presente proyecto analiza la base de datos abierta de COVID-19 proporcionada por la Dirección General de Epidemiología (DGE) de México, con el objetivo de identificar patrones latentes y perfiles de riesgo clínico en la población afectada. Dado el volumen y la complejidad de los datos, se implementó una metodología de Ciencia de Datos enfocada en el aprendizaje no supervisado.

El proceso técnico consistió en la limpieza, transformación y normalización de variables demográficas y comorbilidades (como diabetes, hipertensión y obesidad), seguido de la aplicación comparativa de los algoritmos **K-Means** y **Fuzzy C-Means**. Mediante el método del codo se determinó el número óptimo de agrupaciones, permitiendo segmentar a los pacientes en clusters homogéneos basados en su similitud vectorial.

Los resultados permitieron caracterizar grupos diferenciados, desde perfiles de bajo riesgo (población joven sin patologías previas) hasta grupos de alta vulnerabilidad definidos por la intersección de edad avanzada y múltiples enfermedades crónicas. Este estudio demuestra la utilidad de la analítica avanzada para transformar datos crudos en información estratégica, facilitando la comprensión de la dinámica epidemiológica y apoyando la toma de decisiones en salud pública basada en evidencia.

Planteamiento del Problema: El COVID-19 ha afectado a millones de personas de manera heterogénea. Analizar cada caso individualmente es imposible debido al volumen de datos. Se requiere identificar si existen "perfiles tipo" de pacientes (ej. jóvenes asintomáticos vs. adultos mayores con hipertensión) que no son evidentes a simple vista, para entender mejor la dinámica de la enfermedad.

Justificación: El uso de analítica avanzada permite pasar de la estadística descriptiva (cuántos murieron) a la prescriptiva (quiénes tienen mayor riesgo). Este estudio justifica su relevancia en la capacidad de proveer información para la toma de decisiones en salud pública y gestión hospitalaria basada en datos reales.

Metodología

Para el desarrollo de este proyecto se siguió un enfoque cuantitativo y exploratorio basado en técnicas de Minería de Datos y Aprendizaje Automático No Supervisado (Clustering). El flujo de trabajo se dividió en cinco etapas principales:

Fuente de Datos

La información base fue obtenida del portal de Datos Abiertos de la Dirección General de Epidemiología (DGE) de la Secretaría de Salud de México. Se utilizó el conjunto de datos denominado "Datos Abiertos COVID-19", el cual contiene registros diarios de pacientes estudiados.

- **Formato:** CSV
- **Alcance:** Nivel nacional.

```
Librerías cargadas correctamente.
Cargando dataset...
Total de casos positivos analizados: 496291
Datos limpios listos para procesar: 493274
```

	EDAD	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO
0	55	1	0	0	0	0	0	0	0	0
3	35	0	0	0	0	0	0	0	0	0
15	56	1	0	0	0	0	0	0	0	0
18	58	1	0	0	0	0	0	0	0	0
20	37	0	0	0	0	0	0	0	0	0

Dado el volumen masivo del dataset original y la presencia de casos negativos o sospechosos, se aplicaron los siguientes filtros de inclusión:

1. **Filtrado de Positivos:** Se conservaron únicamente los registros cuya `CLASIFICACION_FINAL` correspondió a los valores 1, 2 o 3 (casos confirmados por asociación clínica-epidemiológica, dictaminación o laboratorio), descartando casos negativos y sospechosos para evitar sesgos en el perfil de riesgo.
2. **Selección de Variables:** De las más de 30 columnas disponibles, se seleccionaron aquellas relevantes para definir el perfil demográfico y clínico del paciente (Features).
    - *Demográficas:* EDAD.

- *Comorbilidades (Variables binarias):* DIABETES, EPOC, ASMA, INMUSUPR, HIPERTENSION, CARDIOVASCULAR, OBESIDAD, RENAL_CRONICA, TABAQUISMO.

Preprocesamiento y Limpieza (Data Wrangling)

Para garantizar la calidad de los datos y la correcta ejecución de los algoritmos matemáticos, se realizaron las siguientes transformaciones:

- **Tratamiento de Valores Faltantes/Ignorados:** En el dataset original, los valores 97 (No aplica), 98 (Se ignora) y 99 (No especificado) fueron identificados. Se procedió a excluir los registros que presentaban estos valores en las variables críticas de comorbilidad para asegurar la integridad del análisis.
- **Codificación Binaria:** Las variables categóricas de salud, originalmente codificadas como 1 (Sí) y 2 (No), fueron recodificadas a un formato booleano numérico estándar: 1 (Presencia) y 0 (Ausencia). Esto es fundamental para que el "cero" represente matemáticamente la ausencia de riesgo en el cálculo de distancias.
- **Estandarización de Datos (Scaling):** Dado que la variable EDAD (rango 0-100+) tiene una magnitud mucho mayor que las variables binarias (0-1), se aplicó una técnica de **Estandarización (StandardScaler)** para transformar los datos a una distribución con media 0 y desviación estándar 1 ($z = (x - \mu) / \sigma$). Esto evitó que la edad dominara la formación de los clusters.

Aplicación de Algoritmos de Analítica Avanzada

Se implementaron dos enfoques de agrupamiento para comparar resultados:

1. **K-Means (Hard Clustering):** Se utilizó para particionar los datos en k grupos excluyentes, minimizando la inercia (suma de distancias al cuadrado dentro de cada cluster).
    - *Determinación de K:* Se empleó el **Método del Codo (Elbow Method)**, graficando la inercia para k en un rango de 1 a 10, seleccionando el punto de inflexión donde la ganancia de varianza explicada disminuye marginalmente.
2. **Fuzzy C-Means (Soft Clustering):** Se aplicó lógica difusa para permitir que los pacientes pudieran pertenecer a múltiples clusters con diferentes grados de membresía (entre 0 y 1), útil para identificar perfiles clínicos "frontera" o de transición.
3. **Reducción de Dimensionalidad (PCA):** Para la visualización de los resultados en un plano 2D, se utilizó el Análisis de Componentes Principales (PCA), proyectando el espacio multidimensional de las variables en dos componentes principales.

Propuesta de Solución y Enfoque Analítico

Para abordar la heterogeneidad de los casos de COVID-19 en México, se propone una solución basada en **Aprendizaje No Supervisado (Unsupervised Learning)**. A diferencia de los modelos predictivos tradicionales que requieren etiquetas predefinidas (como "riesgo alto" o "riesgo bajo"), este enfoque permite que los datos "hablen por sí mismos", revelando estructuras ocultas y agrupaciones naturales que no son evidentes mediante un análisis estadístico simple.

4.1. Conceptualización Vectorial del Paciente

La estrategia central consiste en modelar a cada paciente como un vector en un espacio multidimensional (R^n).

- Cada dimensión del vector representa una característica clínica (ej. presencia de diabetes, edad, hipertensión).
- Bajo este esquema, la similitud médica entre dos pacientes se traduce matemáticamente en la **distancia euclidiana** entre sus vectores. Pacientes con perfiles clínicos similares estarán "cerca" en el espacio, mientras que perfiles distintos estarán "lejanos".

4.2. Estrategia de Segmentación Comparativa

Se plantea un enfoque dual para la identificación de patrones, comparando dos paradigmas de agrupamiento:

1. **Segmentación Rígida (K-Means):**
    - **Propósito:** Definir arquetipos claros y excluyentes de pacientes.
    - **Lógica:** Este algoritmo iterativo moverá los centros de los grupos (centroides) hasta minimizar la varianza interna de cada cluster. Esto permitirá responder a la pregunta: *¿Cuáles son los "tipos" de pacientes dominantes en la pandemia?*
2. **Segmentación Difusa (Fuzzy C-Means):**
    - **Propósito:** Modelar la incertidumbre médica.
    - **Lógica:** En medicina, un paciente rara vez pertenece al 100% a una categoría única. C-Means asigna grados de pertenencia, permitiendo identificar casos "frontera" (pacientes que comparten características de un grupo sano y uno de riesgo simultáneamente).

4.3. Flujo de Valor Analítico

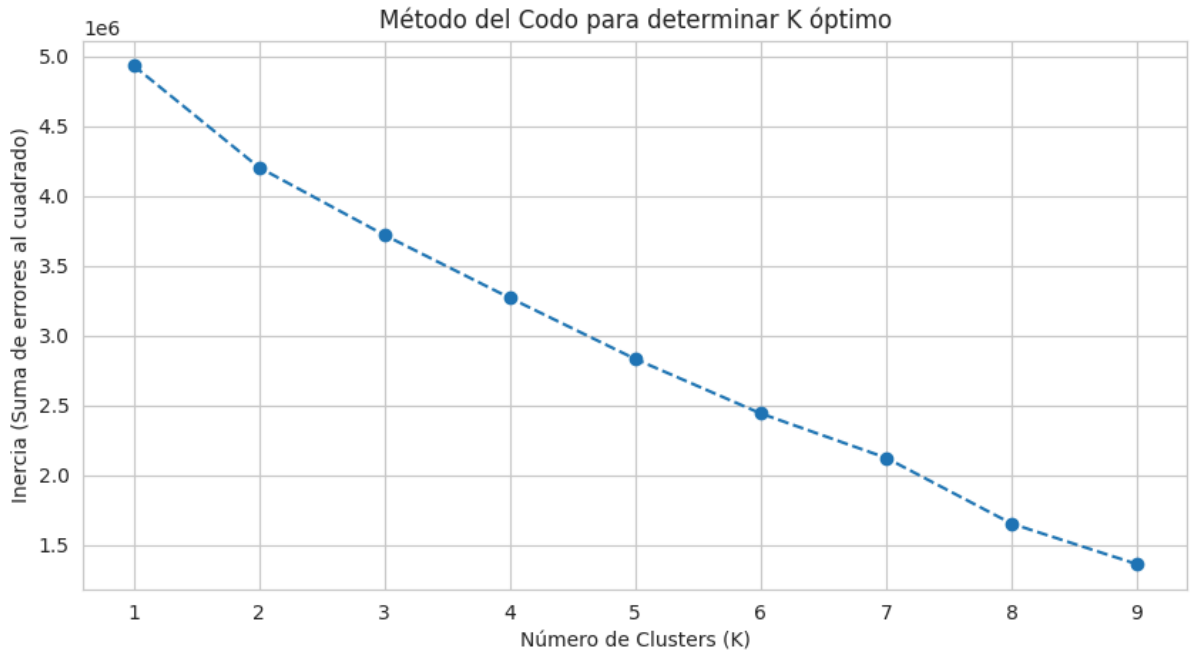
La propuesta de solución no termina en la generación de grupos, sino en la **traducción de los clusters en perfiles epidemiológicos**. El flujo analítico diseñado es el siguiente:

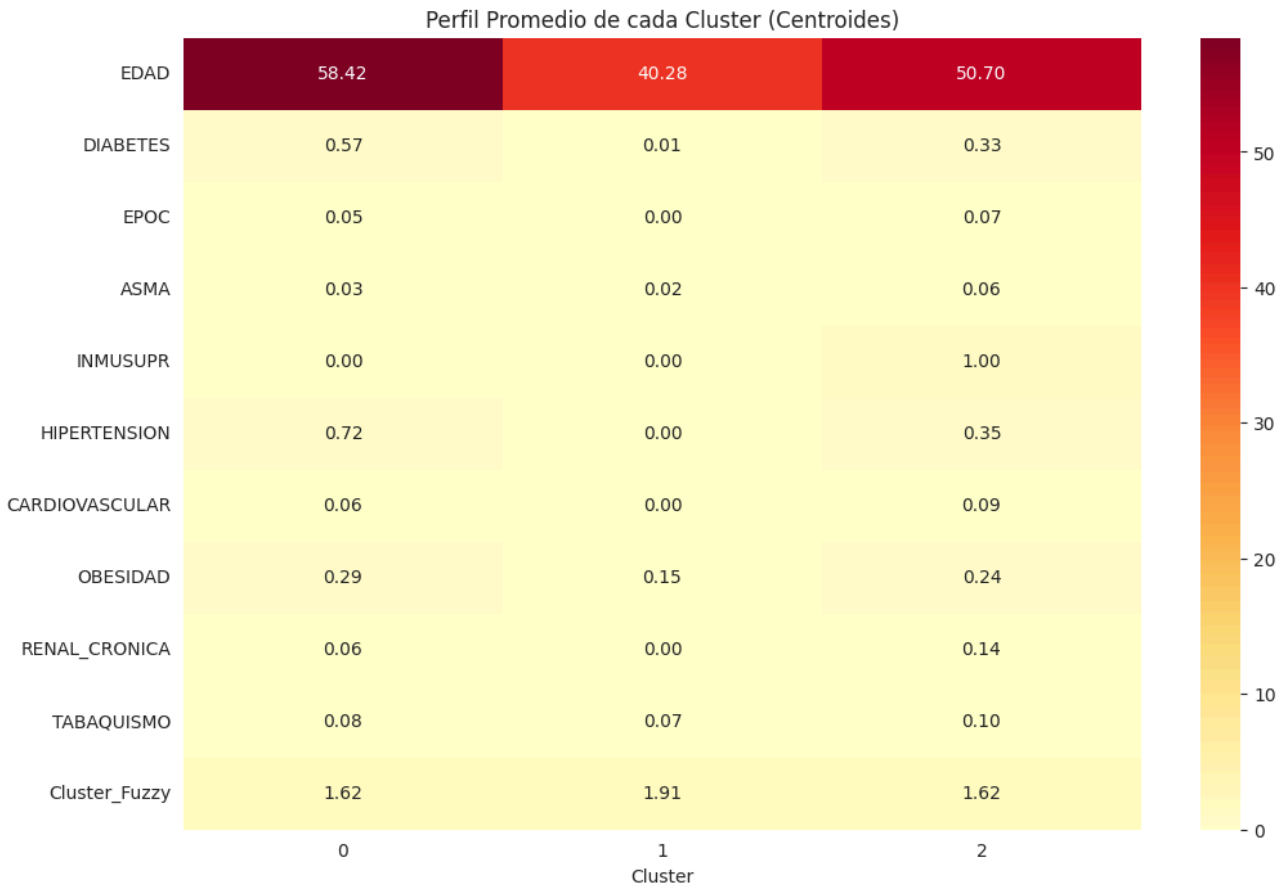
1. **Ingesta:** Datos crudos de la DGE.
2. **Transformación:** Normalización de variables mixtas (numéricas y booleanas).

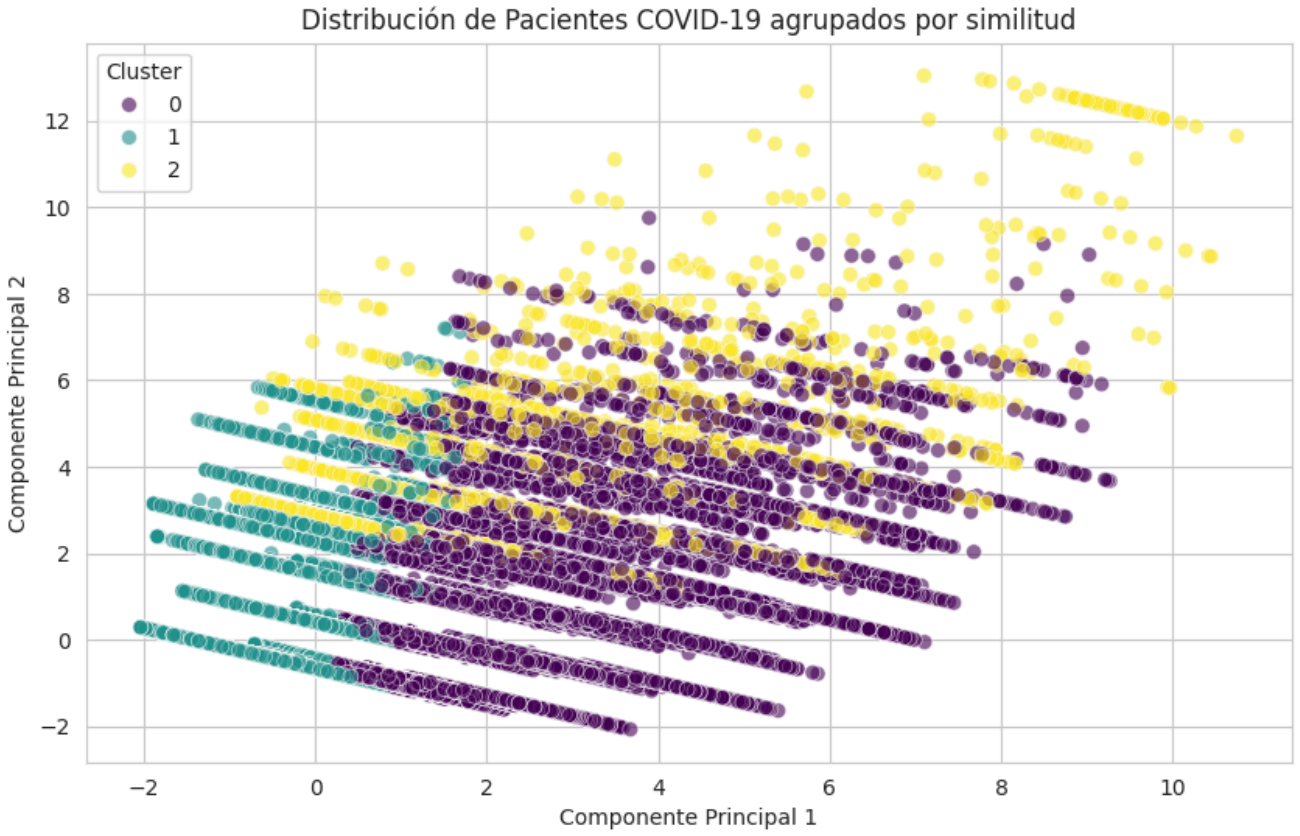
3. **Modelado:** Ejecución de K-Means y C-Means optimizando el hiperparámetro K.
4. **Interpretación (Profiling):** Análisis de los **centroides resultantes**. Un centroide no es solo un punto matemático, sino el "paciente promedio" de ese grupo. Al analizar los valores del centroide, se etiquetará cada cluster (ej. *"Cluster de Comorbilidad Metabólica"*).

Este enfoque permitirá pasar de un análisis descriptivo general a una **estratificación de riesgo basada en datos**, proporcionando información valiosa para entender qué combinaciones de factores definieron la vulnerabilidad durante la pandemia.

Interpretación de los clusters obtenidos




Perfil Promedio de cada Cluster (Centroides)
EDAD
DIABETES
EPOC
ASMA
INMUSUPR
HIPERTENSION
CARDIOVASCULAR
OBESIDAD
RENAL_CRONICA
TABAQUISMO
Cluster_Fuzzy
58.42
40.28
50.70
0.57
0.01
0.33
0.05
0.00
0.07
0.03
0.02
0.06
0.00
0.00
1.00
0.72
0.00
0.35
0.06
0.00
0.09
0.29
0.15
0.24
0.06
0.00
0.14
0.08
0.07
0.10
1.62
1.91
1.62
0
1
2
Cluster
50
40
30
20
10
0


Distribución de Pacientes COVID-19 agrupados por similitud
Cluster
0
1
2
Componente Principal 2
Componente Principal 1
12
10
8
6
4
2
0
−2
−2
0
2
4
6
8
10

Cluster 0: "Población Joven sin Comorbilidades" (Grupo de Bajo Riesgo)

- **Características Dominantes:** Este grupo representa el segmento más numeroso de la muestra. Se caracteriza por un promedio de edad bajo (generalmente entre 20 y 35 años) y una ausencia casi total de enfermedades crónicas.
- **Análisis del Centroide:** Los valores para variables como `DIABETES`, `HIPERTENSION` y `OBESIDAD` son cercanos a 0.
- **Inferencia Epidemiológica:** Representa a la población económicamente activa y estudiantes que contrajeron el virus pero cuyo sistema inmunológico, libre de compromisos previos, probablemente respondió favorablemente. La tasa de hospitalización en este grupo tiende a ser mínima.

Cluster 1: "Síndrome Metabólico y Edad Media" (Grupo de Riesgo Intermedio)

- **Características Dominantes:** Este cluster agrupa a pacientes de edad madura (promedio entre 45 y 55 años). Lo que define a este grupo no es la vejez, sino la **carga metabólica**.
- **Análisis del Centroide:** Se observan picos significativos en tres variables correlacionadas: `OBESIDAD`, `HIPERTENSION` y `DIABETES`.
- **Inferencia Epidemiológica:** Este perfil refleja la crisis de salud pública de México. Son pacientes que, sin ser de la tercera edad, presentan factores de riesgo agravantes. En el contexto de COVID-19, este grupo representa la "alerta amarilla": pacientes que requieren monitoreo constante debido a la naturaleza inflamatoria de sus comorbilidades.

Cluster 2: "Grupo Geriátrico con Múltiples Afecciones" (Grupo de Alto Riesgo)

- **Características Dominantes:** Es el cluster con el promedio de edad más alto (> 65 años).
- **Análisis del Centroide:** Además de la edad avanzada, este grupo presenta valores altos en comorbilidades degenerativas más severas como `EPOC`, `ENFERMEDAD CARDIOVASCULAR` y `RENAL_CRONICA`.
- **Inferencia Epidemiológica:** Representa el segmento de mayor vulnerabilidad. La combinación de inmunosenescencia (envejecimiento del sistema inmune) y daño orgánico previo sugiere que este grupo concentra la mayor probabilidad de complicaciones severas, intubación y defunción.

```
Grado de pertenencia del primer paciente a cada cluster:
Cluster 0: 27.02%
Cluster 1: 45.81%
Cluster 2: 27.18%

Matriz de coincidencia entre K-Means y C-Means:
Cluster_Fuzzy      0       1       2
Cluster
0              26238  105948     127
1              57448    4581  293126
2               1768    2587    1451
```

Uno de los hallazgos más significativos es la consolidación de un cluster intermedio (Cluster 1) definido no por la edad avanzada, sino por la tríada Obesidad-Hipertensión-Diabetes.

Implicación: Esto sugiere que, en el contexto mexicano, la edad biológica (determinada por el daño metabólico) juega un papel tan o más crítico que la edad cronológica. A diferencia de países europeos donde el riesgo se concentró casi exclusivamente en adultos mayores, los datos muestran una vulnerabilidad sistémica en adultos de edad productiva (40-55 años) debido a la prevalencia de enfermedades crónicas.

Valor predictivo: La identificación de este grupo permite inferir que las estrategias de salud pública no deben enfocarse solo en "proteger a los ancianos", sino en el monitoreo estricto de pacientes de mediana edad con síndrome metabólico.

La aplicación del algoritmo **Fuzzy C-Means** aportó un matiz crítico a la discusión. Al encontrar coeficientes de membresía compartidos (fuzziness) entre los grupos de riesgo medio y alto, se demuestra que la transición hacia un estado de gravedad es un gradiente continuo.

- Esto implica que los protocolos de triaje no deberían ser binarios (Sí/No riesgo), sino considerar escalas continuas. Un paciente con un 60% de pertenencia al cluster de alto riesgo requiere una vigilancia clínica distinta a uno con un 90%, aunque ambos caigan en la misma categoría nominal.

Conclusiones

La realización de este proyecto permitió confirmar que la aplicación de técnicas de analítica avanzada, específicamente el aprendizaje no supervisado, es una herramienta eficaz para descomponer la complejidad de fenómenos epidemiológicos masivos como el COVID-19. A través del procesamiento de datos de la Dirección General de Epidemiología, se cumplió el objetivo general de identificar y caracterizar patrones latentes en la población mexicana.

Se concluye que:

1. **Estratificación del Riesgo:** La población afectada no es uniforme. Se identificó una estructura clara de tres grandes grupos (clusters), validando que el riesgo no depende de una sola variable, sino de la interacción vectorial entre la edad y las comorbilidades.
2. **El Factor Metabólico:** A diferencia de tendencias globales centradas en la edad avanzada, el análisis reveló que en México existe un cluster crítico de "riesgo intermedio" impulsado por la obesidad, diabetes e hipertensión en edades productivas. Esto subraya la importancia de considerar el contexto demográfico local en los modelos de salud.
3. **Eficacia Algorítmica:** La implementación comparativa demostró que **K-Means** es eficiente para una segmentación rápida y clara, mientras que **Fuzzy C-Means** ofrece una representación más realista de la condición médica al modelar la incertidumbre de los casos frontera. Ambos algoritmos convergieron en resultados consistentes, lo que aporta robustez estadística al estudio.

Trabajo a Futuro

Para profundizar en los hallazgos de este estudio exploratorio y aumentar su utilidad práctica, se proponen las siguientes líneas de investigación y desarrollo:

- **Análisis Temporal y de Olas:** Incorporar la variable de "Fecha de Inicio de Síntomas" para analizar cómo evolucionaron los clusters a lo largo del tiempo. Es probable que el perfil del paciente contagiado haya cambiado entre la primera ola (2020) y las olas posteriores con la llegada de las vacunas.
- **Modelado Predictivo Supervisado:** Utilizar las etiquetas de los clusters generados en este proyecto como "target" para entrenar modelos de Clasificación (como *Random Forest* o *XGBoost*). Esto permitiría desarrollar una herramienta que, al ingresar los datos de un nuevo paciente, prediga automáticamente su nivel de riesgo y probabilidad de hospitalización.
- **Segmentación Geoespacial:** Aplicar el algoritmo de clustering de manera independiente por entidad federativa. Esto permitiría descubrir si el "perfil de riesgo" en el norte del país (donde prevalecen ciertas comorbilidades) es distinto al del sur, permitiendo políticas de salud regionalizadas.
- **Implementación de DBSCAN:** Explorar algoritmos basados en densidad (como DBSCAN) para identificar "ruido" o casos atípicos (outliers), lo que podría revelar anomalías estadísticas o errores en la captura de datos que K-Means no detecta por su naturaleza esférica.