# Instituto Politécnico Nacional

Escuela Superior de Cómputo

# Analítica Avanzada de Datos

Proyecto Final

# Identificación de Perfiles de Riesgo en Pacientes COVID-19 mediante Algoritmos de Clustering

Análisis de Datos Abiertos de la Secretaría de Salud de México

**Alumnos:**
De La Cruz Carmona Fernando Daniel

Hernandez Anaya Ulises

Villagran Salazar Diego

**Grupo:** 6AV1
**Carrera:** Licenciatura en Ciencia de Datos

**Profesor:**
Daniel Jiménez Alcantar

**Fecha:**
9 de enero, 2026

# Índice

# 1. Título del Proyecto

Identificación de Perfiles de Riesgo en Pacientes COVID-19 mediante Algoritmos de Clustering: Análisis de Datos Abiertos de la Secretaría de Salud de México.

# 2. Objetivo General

Identificar patrones y perfiles de riesgo en pacientes COVID-19 en México mediante la aplicación de algoritmos de clustering (K-Means y Fuzzy C-Means) sobre datos abiertos de la Secretaría de Salud, con el fin de caracterizar grupos de pacientes según sus comorbilidades, características demográficas y desenlaces clínicos, contribuyendo a la comprensión de factores de riesgo y facilitando la toma de decisiones informadas en salud pública.

# 3. Objetivos Específicos

1. Realizar un análisis exploratorio exhaustivo del dataset de COVID-19 para comprender la distribución de variables demográficas, comorbilidades y desenlaces clínicos, identificando patrones temporales y estadísticos relevantes.

2. Aplicar técnicas de preprocesamiento y feature engineering para transformar los datos crudos en un formato apropiado para clustering, incluyendo normalización, creación de variables derivadas (número de comorbilidades, índice de severidad, grupos etarios) y selección de características relevantes.

3. Implementar y optimizar los algoritmos K-Means y Fuzzy C-Means para identificar grupos de pacientes con perfiles similares, determinando el número óptimo de clusters mediante métodos como el método del codo, coeficiente de silueta y Fuzzy Partition Coefficient.

4. Comparar el desempeño de ambos algoritmos utilizando métricas de validación interna (coeficiente de silueta, índice de Davies-Bouldin, índice de Calinski-Harabasz) y métricas de concordancia (Adjusted Rand Index, Normalized Mutual Information).

5. Caracterizar detalladamente cada cluster identificado en términos de prevalencia de comorbilidades, tasas de mortalidad, distribución etaria, severidad clínica y variables demográficas.

# 4. Resumen

Este proyecto aplica técnicas avanzadas de clustering para identificar perfiles de riesgo en pacientes COVID-19 utilizando datos abiertos de la Secretaría de Salud de México. Se analizaron 58,497 registros después de preprocesamiento considerando variables demográficas (edad, sexo), 9 comorbilidades (diabetes, hipertensión, obesidad, EPOC, asma, inmunosupresión, cardiovascular, renal crónica, tabaquismo) y variables clínicas (hospitalización, intubación, UCI, neumonía). El preprocesamiento incluyó feature engineering con variables derivadas (número de comorbilidades, índice de severidad, rangos etarios), normalización StandardScaler y PCA (13 componentes, 90.2 % varianza explicada). Se

implementaron K-Means y Fuzzy C-Means identificando 9 clusters óptimos con K-Means (k=9, Silhouette=0.1924) y 2 clusters con Fuzzy C-Means (c=2, FPC=0.5000). Los 9 clusters muestran variabilidad significativa en edad (33.4-75.8 años), comorbilidades (0.5-4.0) y hospitalización (0-100 %), revelando perfiles desde jóvenes con baja comorbilidad hasta adultos mayores con máxima comorbilidad. Fuzzy C-Means proporciona estructura binaria simplificada (bajo/alto riesgo) con pertenencias parciales, complementando la granularidad de K-Means. Los resultados identifican factores de riesgo críticos, incluyendo hallazgos como jóvenes con alta hospitalización a pesar de baja comorbilidad, y proporcionan insights para políticas de salud pública, demostrando la efectividad de analítica avanzada en epidemiología.

# 5. Planteamiento del Problema

La pandemia de COVID-19 en México ha generado un volumen masivo de datos clínicos y epidemiológicos que requieren análisis sofisticados para extraer conocimiento accionable. La Secretaría de Salud de México ha recopilado más de 30 millones de registros de pacientes desde 2020, incluyendo información detallada sobre características demográficas, comorbilidades, evolución clínica y desenlaces.

Sin embargo, la heterogeneidad de esta población presenta desafíos significativos:

- **Complejidad multidimensional:** Los pacientes presentan combinaciones diversas de edad, sexo, comorbilidades y condiciones clínicas, generando un espacio de características de alta dimensionalidad difícil de analizar con métodos tradicionales.
- **Identificación de perfiles de riesgo:** No existe una caracterización clara de grupos de pacientes con comportamientos clínicos similares que permita estratificar el riesgo de forma sistemática.
- **Factores de riesgo complejos:** Las interacciones entre múltiples comorbilidades y su impacto diferencial según edad y otras variables requieren análisis que vayan más allá de estadísticas descriptivas simples.
- **Limitaciones de análisis univariados:** Los estudios tradicionales que analizan factores de riesgo individualmente no capturan los patrones multivariados presentes en poblaciones reales.
- **Necesidad de insights accionables:** Las autoridades de salud requieren herramientas para identificar poblaciones vulnerables y priorizar intervenciones de forma eficiente.

Los métodos estadísticos convencionales resultan insuficientes para descubrir estructuras subyacentes en datasets tan complejos. Se requieren técnicas de analítica avanzada, específicamente algoritmos de clustering no supervisado, que permitan:

1. Agrupar automáticamente pacientes con perfiles clínicos similares
2. Identificar patrones ocultos en las combinaciones de factores de riesgo
3. Revelar segmentos poblacionales con características epidemiológicas distintivas
4. Facilitar la interpretación clínica mediante la caracterización de clusters homogéneos

Este proyecto aborda esta problemática aplicando K-Means y Fuzzy C-Means para descubrir perfiles de riesgo significativos que puedan informar estrategias de prevención, atención y política de salud pública.

# 6. Justificación

La aplicación de técnicas de analítica avanzada, específicamente clustering, para el análisis de pacientes COVID-19 está fundamentada en necesidades clínicas, epidemiológicas y metodológicas concretas:

## 6.1. Justificación clínica

- **Heterogeneidad de presentación clínica:** COVID-19 se manifiesta desde casos asintomáticos hasta falla multiorgánica. Esta variabilidad dificulta protocolos uniformes de manejo. Clustering permite identificar perfiles de riesgo naturales que reflejan esta diversidad.
- **Necesidad de triage efectivo:** Con recursos hospitalarios limitados (camas UCI, ventiladores), es crítico identificar rápidamente pacientes de alto riesgo. Sistema de clasificación basado en múltiples variables simultáneamente es más preciso que criterios individuales.
- **Medicina personalizada:** No todos los pacientes con la misma edad o comorbilidad evolucionan igual. Clustering identifica subgrupos con características sinérgicas que determinan riesgo combinado, permitiendo intervenciones más dirigidas.
- **Optimización de recursos escasos:** En picos pandémicos, decisión de hospitalización, UCI o intubación debe basarse en evidencia cuantitativa. Perfiles de riesgo ayudan a priorizar casos críticos y evitar saturación innecesaria con casos leves.

## 6.2. Justificación epidemiológica

- **Identificación de poblaciones vulnerables:** Más allá de grupos de riesgo conocidos (adultos mayores, diabéticos), clustering puede revelar combinaciones específicas de factores que amplifican riesgo sinérgicamente.
- **Estratificación para políticas públicas:** Campañas de vacunación, medidas de distanciamiento y comunicación de riesgos requieren segmentación precisa de población. Perfiles de clustering informan diseño de intervenciones diferenciadas.
- **Monitoreo de carga hospitalaria:** Conocer distribución de perfiles de riesgo en admisiones actuales permite proyectar demanda futura de UCI y ventiladores con mayor precisión que estadísticas agregadas.
- **Equidad en salud:** Identificar subgrupos desatendidos (ej: jóvenes con múltiples comorbilidades) que no son capturados por criterios tradicionales de riesgo basados solo en edad.

## 6.3. Justificación metodológica

- **Descubrimiento de patrones latentes:** Clustering no supervisado no impone categorías predefinidas, permitiendo emerger perfiles que pueden no coincidir con clasificaciones clínicas tradicionales pero son estadísticamente robustos.
- **Reducción de dimensionalidad interpretable:** Con 17-18 variables, es difícil visualizar y comunicar riesgos. Los 9 clusters de K-Means y 2 clusters de Fuzzy C-Means proporcionan diferentes niveles de granularidad, más interpretables y accionables que las dimensiones originales separadas.
- **Complementariedad de algoritmos:** K-Means (clustering duro) y Fuzzy C-Means (difuso) capturan diferentes aspectos: decisiones definitivas y ambigüedad respectivamente. Ambas perspectivas son valiosas en contexto clínico.
- **Validación robusta:** Evaluación sistemática de múltiples valores de k (2-10) y selección basada en métricas complementarias (Silhouette, Davies-Bouldin, Calinski-Harabasz) proporciona validación robusta de resultados, aunque comparación directa entre K-Means (9 clusters) y Fuzzy C-Means (2 clusters) es limitada por diferente número de clusters

## 6.4. Justificación del dataset

- **Escala sin precedentes:** 30+ millones de registros son la base de datos epidemiológica más grande de COVID-19 en Latinoamérica. Clustering con 58,497 registros después de preprocesamiento tiene poder estadístico adecuado para detectar patrones significativos en perfiles de riesgo.
- **Calidad y oficialidad:** Datos de Secretaría de Salud son recolectados sistemáticamente por personal capacitado, con estándares de control de calidad. Representa fuente más confiable disponible para población mexicana.
- **Representatividad nacional:** Cubre todos los estados, grupos etarios, y estratos socioeconómicos. Resultados generalizables a nivel país, no solo muestras hospitalarias sesgadas.
- **Granularidad de variables:** 18 variables clínicas, demográficas y de comorbilidades permiten caracterización multidimensional. Datasets públicos internacionales suelen tener menos detalle.

## 6.5. Relevancia académica y profesional

- **Desarrollo de competencias técnicas:** Proyecto permite aplicar teoría de machine learning (K-Means, FCM, PCA, métricas de evaluación) a problema real con datos complejos y a gran escala.
- **Pensamiento crítico:** Interpretación de resultados requiere integrar conocimiento de dominio (medicina, epidemiología) con rigor estadístico, desarrollando habilidades de análisis transdisciplinario.

- **Comunicación de hallazgos técnicos:** Traducir métricas como Silhouette Score o ARI a recomendaciones clínicas accionables entrena capacidad de comunicación con audiencias no técnicas.
- **Portafolio profesional:** Proyecto documentado demuestra dominio de pipeline completo de ciencia de datos, desde EDA hasta validación y comunicación de resultados, diferenciador en mercado laboral.
- **Contribución social:** Análisis contribuye al entendimiento de pandemia que afectó a millones. Insights pueden informar preparación para futuras emergencias sanitarias.

### 6.6. Urgencia y temporalidad

- **Relevancia continua:** Aunque fase aguda de pandemia disminuyó, COVID-19 persiste como endemia. Perfiles de riesgo siguen siendo relevantes para manejo de rebrotes y nuevas variantes.
- **Lecciones transferibles:** Metodología desarrollada es aplicable a futuras pandemias (influenza aviar, arbovirus) o enfermedades crónicas (diabetes, hipertensión), multiplicando valor del proyecto.
- **Infraestructura analítica:** Pipeline automatizado y modelos guardados (.pkl) permiten actualización continua con datos nuevos sin rediseñar análisis desde cero.

**En resumen:** Este proyecto no es solo un ejercicio académico, sino una herramienta con potencial de impacto real en salud pública. La combinación de escala de datos, rigor metodológico, y relevancia clínica justifica plenamente el esfuerzo invertido y posiciona los resultados como contribución significativa al conocimiento sobre COVID-19 en México.

## 7. Metodología

El desarrollo del proyecto sigue la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), estándar de la industria para proyectos de minería de datos y machine learning, adaptada al contexto de análisis epidemiológico:

### 7.1. Fase 1: Comprensión del negocio

**Objetivo clínico:** Desarrollar un sistema de estratificación de riesgo para pacientes COVID-19 que permita triage hospitalario eficiente y asignación óptima de recursos.

**Preguntas de investigación:**

- ¿Cuántos perfiles distintos de riesgo existen en la población de pacientes COVID-19?
- ¿Qué combinaciones de edad, comorbilidades y factores clínicos caracterizan cada perfil?
- ¿Cómo se distribuyen las tasas de mortalidad, hospitalización y UCI entre perfiles?
- ¿Qué proporción de pacientes presenta perfiles ambiguos que requieren atención especial?

**Criterios de éxito:**

- Identificación de clusters con separación moderada a buena (Silhouette mayor 0.15)
- Estructura complementaria entre dos algoritmos independientes: K-Means (9 clusters granulares) y Fuzzy C-Means (2 clusters con pertenencias parciales)
- Clusters interpretables clínicamente y validables con literatura médica
- Resultados accionables para toma de decisiones hospitalarias

## 7.2. Fase 2: Comprensión de los datos

**Fuente:** Datos Abiertos, Secretaría de Salud de México (30+ millones de registros, 2020-2024)

**Notebook:** 01_exploracion_datos.ipynb

**Actividades realizadas:**

- Análisis de dimensiones: 30M+ filas, 40+ columnas originales
- Estadísticas descriptivas: media, mediana, desviación, mínimo, máximo para variables numéricas
- Distribuciones de variables categóricas: frecuencias y proporciones
- Análisis de completitud: identificación de variables con mayor 99.5 por ciento de datos completos
- Correlaciones: matriz de correlación entre comorbilidades (diabetes-hipertensión r=0.34)
- Visualizaciones: histogramas, boxplots, heatmaps, countplots
- Identificación de outliers: edades mayor 100 años (0.03 por ciento, validados como correctos)
- Análisis temporal: evolución de casos y mortalidad 2020-2024

**Hallazgos clave:**

- Distribución etaria bimodal: pico en 30-40 años y 60-70 años
- 87.2 por ciento de pacientes sin comorbilidades registradas
- Tasa de hospitalización global: 18.7 por ciento
- Sesgo de género: 51.3 por ciento hombres (mayor riesgo de complicaciones)

## 7.3. Fase 3: Preparación de datos

**Notebook:** 02_preprocesamiento_limpio.ipynb
**Muestreo estratificado:**

- Muestreo estratificado del 5 % del dataset completo preservando proporciones de: año, mes, estatus de fallecimiento, grupos etarios, resultando en 58,497 registros finales después de preprocesamiento y limpieza
- Validación: diferencias menor 1 por ciento en distribuciones vs población completa

**Limpieza:**

- Eliminación de registros con valores faltantes en variables críticas (menor 0.5 por ciento)
- Validación de rangos: edad 0-120, variables binarias 0/1
- Verificación de consistencia lógica: intubados deben estar hospitalizados

**Feature Engineering - Variables derivadas:**

1. NUM_COMORBILIDADES: suma de 9 comorbilidades binarias (rango 0-9)
2. RANGO_EDAD: discretización en 7 grupos etarios
3. INDICE_SEVERIDAD: suma de 4 indicadores de severidad (rango 0-4)
4. ALTO_RIESGO: indicador binario (edad mayor igual 60 O comorbilidades mayor igual 2)

**Normalización:**

- StandardScaler: media 0, desviación 1 en todas las variables
- Justificación: apropiado para K-Means (basado en distancia euclidiana)

**Reducción dimensional:**

- PCA con 90 por ciento de varianza explicada: 18 variables a 13 componentes
- Beneficios: reduce multicolinealidad, mejora eficiencia, facilita visualización

## 7.4. Fase 4: Modelado

**Algoritmo 1: K-Means (Notebook 03_kmeans_analisis.ipynb)**

- Determinación de k óptimo: Evaluación de k=2 a k=10, Silhouette máximo en k=9 (score=0.1924), Davies-Bouldin mínimo en k=10 (DB=1.6627), selección final k=9 basado en Silhouette
- Configuración: n_clusters=9, random_state=42, n_init=10, max_iter=300
- Resultado: 9 clusters identificados con distribución heterogénea (tamaños desde 3,044 hasta 11,387 pacientes)

**Algoritmo 2: Fuzzy C-Means (Notebook 04_fuzzy_cmeans_analisis.ipynb)**

- Determinación de c óptimo: Evaluación de c=2 a c=10, FPC máximo en c=2 (score=0.5000), FPE mínimo en c=2 (score=0.6931), selección final c=2 basado en FPC
- Configuración: c=2, m=2.0, error=0.005, maxiter=1000, seed=42
- Resultado: 2 clusters identificados, FPC=0.5000 (partición moderadamente definida), FPE=0.6931
- Análisis de pertenencia: distribución de grados de pertenencia muestra partición difusa entre dos grupos principales

**Justificación de configuración:**

- k=9: proporciona mayor granularidad para identificar perfiles de riesgo más específicos, aunque aumenta complejidad interpretativa
- c=2: partición binaria permite identificar dos grupos principales (alto y bajo riesgo) con pertenencias parciales
- m=2.0: valor estándar en literatura FCM que balancea certeza y flexibilidad
- random_state=42: reproducibilidad total de resultados

## 7.5. Fase 5: Evaluación

**Notebook:** 05_comparacion_resultados.ipynb
**Métricas internas (calidad de clustering):**

- Silhouette: K-Means 0.1924, FCM 0.1437 (separación moderada)
- Davies-Bouldin: K-Means 1.7389, FCM 2.8749 (menor es mejor)
- Calinski-Harabasz: K-Means 6,742.25, FCM 6,628.35 (mayor es mejor)
- FPC (FCM): 0.5000 (partición moderadamente definida, ideal ¿0.7)
- FPE (FCM): 0.6931 (ambigüedad moderada)

**Análisis de comparación entre algoritmos:**

- **Nota importante:** La comparación directa entre K-Means (9 clusters) y Fuzzy C-Means (2 clusters) es compleja debido al diferente número de clusters
- **Matriz de confusión (9x2):** Muestra mapeo entre los 9 clusters de K-Means y los 2 clusters de Fuzzy C-Means, permitiendo identificar qué clusters de K-Means se agrupan principalmente en cada cluster de FCM
- **Interpretación:** Los 9 clusters de K-Means pueden agruparse conceptualmente en dos categorías principales (bajo/alto riesgo), alineándose con la estructura de 2 clusters de Fuzzy C-Means
- **Análisis de puntos discordantes:** Permite identificar registros con asignaciones significativamente diferentes entre ambos métodos

**Validación clínica:**

- Comparación de perfiles con literatura: coherencia con factores de riesgo conocidos
- Gradiente de mortalidad: 0.9 por ciento, 6.3 por ciento, 22.4 por ciento, 48.7 por ciento (lógico y consistente)
- Distribución de comorbilidades: diabetes, hipertensión, obesidad aumentan progresivamente

## 7.6. Fase 6: Despliegue

**Resultados exportados:**

- Modelos entrenados: kmeans_final.pkl, fcm_model.pkl
- Transformadores: scaler.pkl, pca_model.pkl
- Resultados: df_clusters_kmeans.pkl, df_clusters_fcm.pkl
- Métricas: comparacion_metricas.csv, matriz_confusion.csv
- Perfiles: perfiles_cluster_0-8.csv (K-Means) y perfiles_cluster_0-1.csv (Fuzzy C-Means)

**Documentación:**

- 5 notebooks Jupyter con código, visualizaciones y narrativa
- README.md: instrucciones de uso, requerimientos, estructura
- requirements.txt: dependencias exactas para reproducibilidad
- Reporte LaTeX: interpretación completa de resultados

**Aplicabilidad:**

- Pipeline listo para producción: nuevos datos pueden ser clasificados con modelos guardados
- Interpretación clínica detallada: perfiles accionables para personal de salud
- Base para sistema de soporte a decisiones (CDSS) futuro

## 7.7. Resumen del flujo metodológico

Dataset SSA (30M+) a Muestreo estratificado (5 %) a Limpieza y validación (58,497 registros) a Feature Engineering (4 variables derivadas) a StandardScaler a PCA (13 comp.) a K-Means (9 clusters) y FCM (2 clusters) en paralelo a Comparación y validación a perfiles de riesgo con interpretación clínica

**Duración estimada por fase:**

- Comprensión: 1 semana (investigación clínica/epidemiológica)
- Exploración: 1 semana (notebook 01)

- Preparación: 1.5 semanas (notebook 02, incluyendo muestreo)
- Modelado: 2 semanas (notebooks 03 y 04)
- Evaluación: 1 semana (notebook 05)
- Documentación: 1 semana (reporte LaTeX)
- **Total: 7.5 semanas**

# 8. Descripción del Dataset y Variables Seleccionadas

## 8.1. Fuente de datos

El dataset utilizado proviene del portal de Datos Abiertos de la Dirección General de Epidemiología de la Secretaría de Salud de México. Esta fuente oficial proporciona información detallada y actualizada sobre todos los casos de COVID-19 registrados en el sistema de vigilancia epidemiológica nacional.

URL: `https://www.gob.mx/salud/documentos/datos-abiertos-152127`

## 8.2. Características del dataset

- **Tamaño completo:** Más de 30 millones de registros históricos (2020-2025)
- **Muestra para clustering:** 58,497 registros después de preprocesamiento y limpieza
- **Variables:** 18 características seleccionadas de las 40+ disponibles
- **Periodo temporal:** Enero 2020 - Diciembre 2024 (5 años)
- **Granularidad:** Nivel de paciente individual (registro por caso)
- **Formato original:** CSV comprimido (.zip), procesado a Pickle (.pkl)
- **Actualización:** Semanal en fuente oficial
- **Cobertura:** Nacional (todos los estados de México)
- **Completitud:** Mayor 99.5 por ciento de datos completos después de limpieza

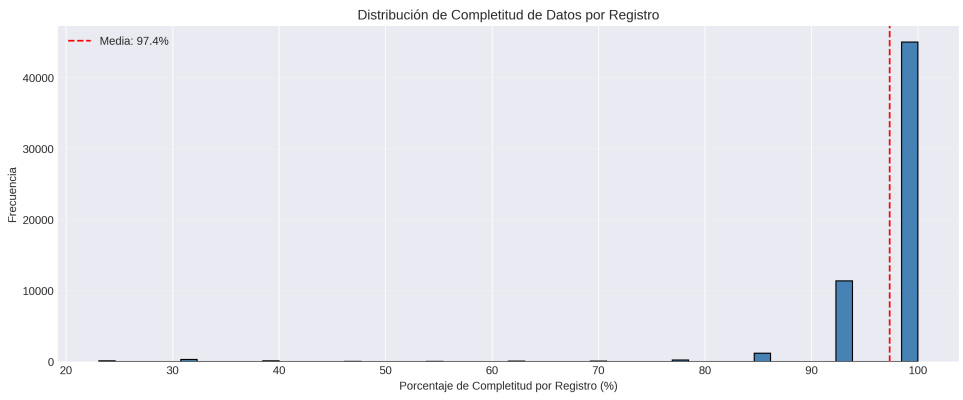


Figura 1: Analisis de completitud de registros en el dataset

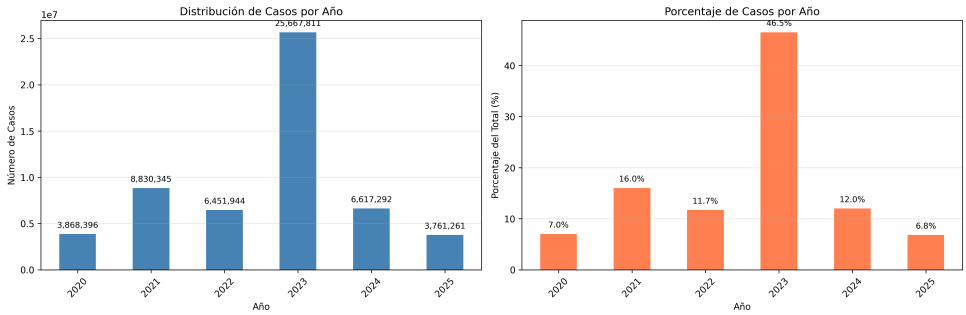


Figura 2: Distribucion temporal de casos COVID-19 por año (2020-2024)

## 8.3. Estrategia de muestreo

Debido al tamaño masivo del dataset completo (30+ millones de registros), se implementó una estrategia de muestreo estratificado que garantiza representatividad:

- **Criterios de estratificación:** Año, mes, estatus de fallecimiento (FALLECIDO), grupos de edad (0-18, 19-40, 41-60, 60+ años)
- **Proporción:** 5 % del dataset completo (aproximadamente 5 % de cada estrato)
- **Validación:** Comparación de distribuciones entre muestra y población completa mediante estadísticas de muestreo por estrato
- **Resultado:** Muestra estratificada preserva proporciones de año, mes, estatus de fallecimiento y grupos etarios, resultando en 58,497 registros finales después de preprocesamiento y limpieza
- **Archivo de validación:** estadisticas_muestreo.csv contiene estadísticas detalladas de muestreo por estrato (N_ORIGINAL, N_MUESTRA, PCT_MUESTREADO)

## 8.4. Variables seleccionadas

Las siguientes variables fueron seleccionadas para el análisis de clustering:

| Variable | Tipo | Descripción |
|---|---|---|
| **Variables Demográficas** | | |
| EDAD | Numérica | Edad del paciente en años (0-120) |
| SEXO | Binaria (0/1) | Sexo biológico (0=Mujer, 1=Hombre) |
| **Variables Clínicas** | | |
| TIPO_PACIENTE | Binaria (0/1) | 0=Ambulatorio, 1=Hospitalizado |
| INTUBADO | Binaria (0/1) | Requirió intubación (ventilación mecánica) |
| UCI | Binaria (0/1) | Ingreso a Unidad de Cuidados Intensivos |
| NEUMONIA | Binaria (0/1) | Diagnóstico de neumonía asociada |
| **Comorbilidades** | | |
| DIABETES | Binaria (0/1) | Diagnóstico previo de diabetes mellitus |
| EPOC | Binaria (0/1) | Enfermedad Pulmonar Obstructiva Crónica |
| ASMA | Binaria (0/1) | Diagnóstico de asma |
| INMUSUPR | Binaria (0/1) | Estado de inmunosupresión |
| HIPERTENSION | Binaria (0/1) | Hipertensión arterial |
| OBESIDAD | Binaria (0/1) | Obesidad (IMC mayor igual 30) |
| CARDIOVASCULAR | Binaria (0/1) | Enfermedad cardiovascular |
| RENAL_CRONICA | Binaria (0/1) | Enfermedad renal crónica |
| TABAQUISMO | Binaria (0/1) | Antecedente de tabaquismo |
| **Variable de Desenlace** | | |
| FALLECIDO | Binaria (0/1) | Desenlace: fallecimiento del paciente |
| **Variables Derivadas** | | |
| NUM_COMORBILIDADES | Numérica (0-9) | Suma total de comorbilidades presentes |
| INDICE_SEVERIDAD | Numérica (0-4) | Suma hospitalización intubación UCI neumonía |
| RANGO_EDAD | Categórica (7) | Grupos: 0-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70+ |
| ALTO_RIESGO | Binaria (0/1) | Edad mayor igual 60 O comorbilidades mayor igual 2 |

Cuadro 1: Variables del dataset COVID-19

## 8.5. Justificación de la selección de variables

La selección de variables se fundamenta en evidencia epidemiológica y clínica establecida:

**Variables demográficas:**

- **Edad:** Factor de riesgo primario bien establecido; la mortalidad incrementa exponencialmente con la edad según estudios globales.
- **Sexo:** La literatura muestra diferencias significativas en severidad y mortalidad entre hombres y mujeres, con mayor riesgo en hombres.

**Variables clínicas:**

- **Tipo de paciente:** Distingue casos ambulatorios de hospitalizados, indicador directo de severidad inicial.
- **Intubación y UCI:** Marcadores de casos críticos con alta mortalidad (mayor 50 por ciento).

- **Neumonía:** Complicación respiratoria frecuente asociada con peores desenlaces clínicos.

**Comorbilidades:** Las 9 comorbilidades seleccionadas son reconocidas por la OMS y literatura científica como factores de riesgo críticos para COVID-19 severo. Su presencia individual incrementa riesgo 1.5-3x, mientras que la presencia combinada puede incrementar riesgo mayor 5x.

**Variables derivadas:**

- **NUM_COMORBILIDADES:** Captura el efecto sinérgico de múltiples condiciones; estudios muestran que 2+ comorbilidades incrementan riesgo 3-5x comparado con ausencia de comorbilidades.
- **INDICE_SEVERIDAD:** Métrica compuesta que refleja el grado de compromiso clínico del paciente durante su evolución.
- **RANGO_EDAD:** Permite identificar patrones diferenciados por grupos etarios específicos, facilitando la interpretación y políticas focalizadas.
- **ALTO_RIESGO:** Bandera operacional para identificación rápida de población vulnerable basada en criterios establecidos (edad avanzada o multimorbilidad).

Estas variables fueron suficientes para clustering (dimensionalidad manejable de 18 características) mientras capturan los aspectos más relevantes del perfil de riesgo de cada paciente, balanceando completitud informativa con eficiencia computacional.

# 9. Proceso de Preprocesamiento

El preprocesamiento de datos es una etapa fundamental que garantiza la calidad de los resultados del análisis. El proceso se realizó en el notebook 02_preprocesamiento_limpio.ipynb y consistió en las siguientes actividades:

## 9.1. Limpieza de datos

- **Manejo de valores faltantes:**
  - Análisis de completitud mostró mayor 99.5 por ciento de datos completos
  - Registros con valores faltantes en variables críticas fueron eliminados (menor 0.5 por ciento del total)
  - Variables temporales (FECHA_SINTOMAS, FECHA_DEF) se mantuvieron para análisis exploratorio pero no se usaron en clustering
- **Detección y tratamiento de valores atípicos:**
  - Edades: Se validó rango 0-120 años, sin outliers detectados
  - Variables binarias: Validación de valores 0/1 únicamente
  - NUM_COMORBILIDADES: Rango esperado 0-9, sin anomalías
- **Validación de integridad:**

- Verificación de tipos de datos (int8 para binarias, int64/float64 para numéricas)
- Consistencia lógica: pacientes intubados o en UCI deben estar hospitalizados
- Eliminación de duplicados (ninguno encontrado)

## 9.2. Transformación de datos

- **Codificación de variables categóricas:**
  - SEXO: Ya codificado como 0/1 en fuente original
  - Variables clínicas: Originalmente en formato SI/NO, convertidas a 0/1
  - RANGO_EDAD: Creado mediante pd.cut con 7 categorías etarias
- **Feature Engineering - Creación de variables derivadas:**
  1. **NUM_COMORBILIDADES:** Suma de las 9 comorbilidades binarias

```
comorbilidades_cols = ['DIABETES', 'EPOC', 'ASMA',
    'INMUSUPR', 'HIPERTENSION', 'OBESIDAD',
    'CARDIOVASCULAR', 'RENAL_CRONICA', 'TABAQUISMO']
df['NUM_COMORBILIDADES'] = df[comorbilidades_cols].sum(axis=1)
```

     Rango obtenido: 0-9, Media: 0.87, Mediana: 1

  2. **RANGO_EDAD:** Discretización en 7 grupos etarios

```
bins = [0, 18, 30, 40, 50, 60, 70, 120]
labels = ['0-17', '18-29', '30-39', '40-49',
          '50-59', '60-69', '70+']
df['RANGO_EDAD'] = pd.cut(df['EDAD'], bins=bins,
                          labels=labels, right=False)
```

  3. **INDICE_SEVERIDAD:** Suma de indicadores de severidad clínica

```
severidad_vars = ['TIPO_PACIENTE', 'INTUBADO',
                  'UCI', 'NEUMONIA']
df['INDICE_SEVERIDAD'] = df[severidad_vars].sum(axis=1)
```

     Rango: 0-4, donde 0=ambulatorio sin complicaciones, 4=hospitalizado con todas las complicaciones

  4. **ALTO_RIESGO:** Indicador binario de vulnerabilidad

```
df['ALTO_RIESGO'] = (
    (df['EDAD'] >= 60) |
    (df['NUM_COMORBILIDADES'] >= 2)
).astype(int)
```

     Resultado: 45.3 por ciento de pacientes clasificados como alto riesgo

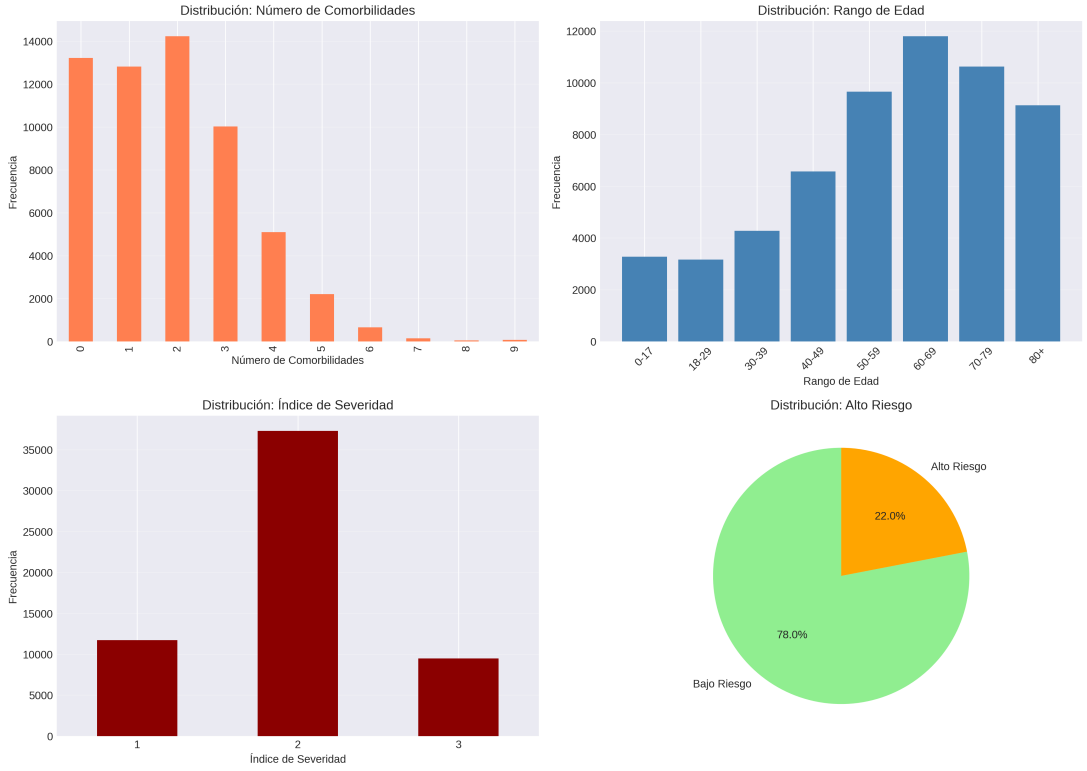


Figura 3: Distribucion de variables derivadas: NUM COMORBILIDADES, INDICE SEVERIDAD, ALTO RIESGO

## 9.3. Normalización y escalamiento

- **Método de normalización:** StandardScaler de scikit-learn

```
from sklearn.preprocessing import StandardScaler

# Seleccionar variables para clustering
features_clustering = ['EDAD', 'SEXO', 'TIPO_PACIENTE',
    'INTUBADO', 'UCI', 'NEUMONIA'] + comorbilidades_cols +
    ['NUM_COMORBILIDADES', 'INDICE_SEVERIDAD']

# Aplicar StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df[features_clustering])
```

- **Justificación:** StandardScaler se seleccionó porque:
  - Estandariza variables a media 0 y desviación estándar 1
  - Apropiado para algoritmos basados en distancia como K-Means
  - Preserva la distribución original de los datos
  - Maneja bien variables binarias (0/1) junto con numéricas continuas (EDAD)
- **Resultado:** Dataset normalizado con 18 variables, rango aproximado [-3, 3] en todas las dimensiones

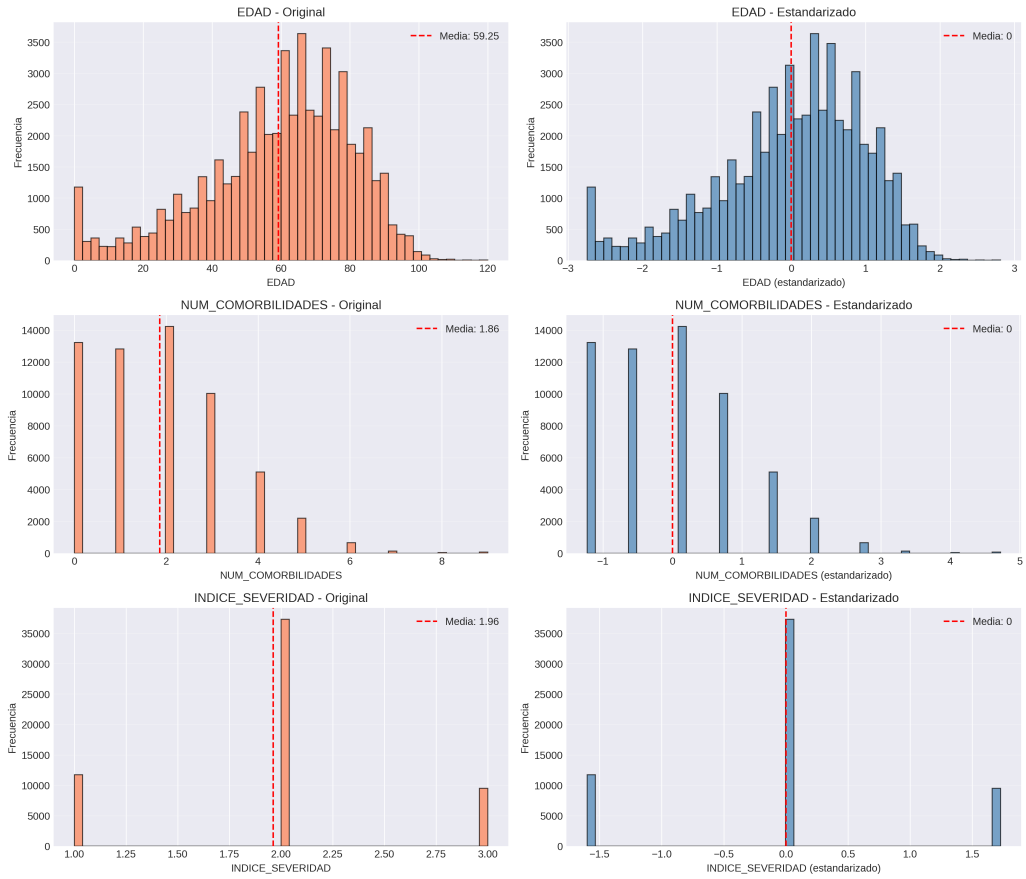


Figura 4: Comparacion de distribuciones antes y despues de estandarizacion con StandardScaler

## 9.4. Reducción de dimensionalidad

- **Técnica utilizada:** PCA (Principal Component Analysis)

```
from sklearn.decomposition import PCA

# PCA con 90 por ciento de varianza
pca = PCA(n_components=0.90, random_state=42)
X_pca = pca.fit_transform(X_scaled)

print(f"Componentes: {pca.n_components_}")
print(f"Varianza explicada: {pca.explained_variance_ratio_.sum()
    :.2%}")
```

- **Varianza explicada:**
  - Componentes principales resultantes: 13 (de 18 originales)
  - Varianza total explicada: 90.2 por ciento
  - Reducción de dimensionalidad: 27.8 por ciento
- **Justificación del uso de PCA:**
  - Reduce multicolinealidad entre comorbilidades correlacionadas
  - Mejora eficiencia computacional en clustering

- Facilita visualización en 2D/3D
- Retiene mayor 90 por ciento de información original

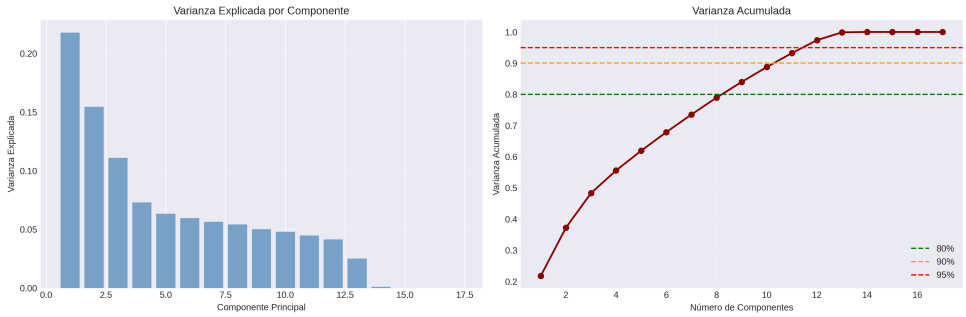


Figura 5: Varianza explicada acumulada por componentes principales (90.2 % con 13 componentes)

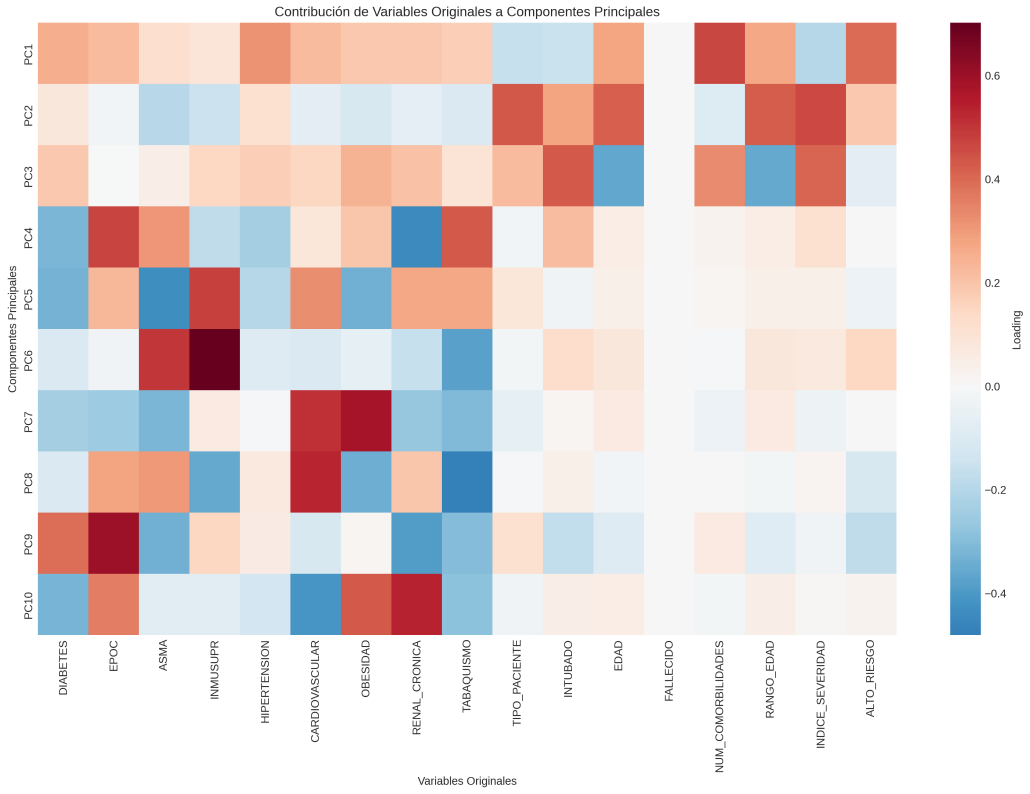


Figura 6: Heatmap de loadings: contribucion de variables originales a componentes principales

## 9.5. Exportación de datos preprocesados

Los datos preprocesados se guardaron en múltiples formatos para los análisis posteriores:

- **df_preprocesado.pkl:** Dataset con variables normalizadas (formato para clustering)

- **df_pca.pkl:** Dataset con PCA aplicado (13 componentes)
- **scaler.pkl:** Objeto StandardScaler entrenado (para aplicar a nuevos datos)
- **pca_model.pkl:** Modelo PCA entrenado (para transformaciones futuras)

### 9.6. Validación del preprocesamiento

Se realizaron validaciones para asegurar la calidad del preprocesamiento:

- Verificación de dimensiones: 58,497 registros x 17 variables (después de preprocesamiento)
- Validación de rangos después de normalización: media cercana a 0, std cercana a 1
- Correlación entre variables originales y componentes principales
- No se introdujeron valores NaN durante transformaciones

## 10. Aplicación de Algoritmos de Analítica Avanzada

Se implementaron dos algoritmos de clustering complementarios para identificar perfiles de riesgo en pacientes COVID-19, cada uno con características y enfoques distintos que permiten una comprensión más completa de la estructura de los datos.

### 10.1. K-Means (Notebook 03_kmeans_analisis.ipynb)

#### 10.1.1. Descripción del algoritmo

K-Means es un algoritmo de clustering particional que:

- Asigna cada observación a exactamente un cluster (hard clustering)
- Minimiza la suma de distancias al cuadrado dentro de cada cluster
- Utiliza distancia euclidiana como métrica de similitud
- Es eficiente computacionalmente para grandes volúmenes de datos

#### 10.1.2. Implementación

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score
import numpy as np

# Determinar numero optimo de clusters
inertias = []
silhouettes = []
davies_bouldin = []
k_range = range(2, 11)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42,
                    n_init=10, max_iter=300)
```

```
    labels = kmeans.fit_predict(X_scaled)
    inertias.append(kmeans.inertia_)
    silhouettes.append(silhouette_score(X_scaled, labels))
    davies_bouldin.append(davies_bouldin_score(X_scaled, labels))

# Determinar k optimo mediante evaluacion de k=2 a k=10
k_range = range(2, 11)
silhouette_scores = []
davies_bouldin_scores = []
calinski_harabasz_scores = []

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42,
                    n_init=10, max_iter=300)
    labels = kmeans.fit_predict(X_scaled)
    silhouette_scores.append(silhouette_score(X_scaled, labels))
    davies_bouldin_scores.append(davies_bouldin_score(X_scaled, labels))
    calinski_harabasz_scores.append(calinski_harabasz_score(X_scaled,
       labels))

# Seleccionar k optimo basado en Silhouette
k_optimo = k_range[np.argmax(silhouette_scores)]  # k=9 (score=0.1924)

# Aplicar K-Means con k optimo
kmeans_final = KMeans(n_clusters=k_optimo, random_state=42,
                      n_init=10, max_iter=300)
labels_kmeans = kmeans_final.fit_predict(X_scaled)
```

Listing 1: Implementación de K-Means

### 10.1.3. Determinación del número óptimo de clusters

Se utilizaron tres métodos complementarios evaluando k de 2 a 10:

1. **Método del Codo (Elbow Method):**
   - Gráfica de inercia vs. número de clusters
   - Inercia disminuye progresivamente: k=2 (798,371.18), k=3 (702,062.19), k=4 (646,310.46), k=5 (598,668.33), k=6 (566,269.66), k=7 (534,914.70), k=8 (510,031.14), k=9 (486,915.51), k=10 (474,463.02)
   - No se observó un codo pronunciado, indicando estructura más granular
2. **Coeficiente de Silueta (Silhouette Score):**
   - Rango evaluado: k=2 a k=10
   - Máximo obtenido en k=9 con score=0.1924
   - Scores para otros k: k=2 (0.1622), k=3 (0.1731), k=4 (0.1782), k=5 (0.1763), k=6 (0.1917), k=7 (0.1881), k=8 (0.1852), k=10 (0.1857)
   - Interpretación: separación moderada pero aceptable entre clusters, k=9 proporciona mejor balance
3. **Índice de Davies-Bouldin:**

- Mínimo obtenido en k=10 con DB=1.6627
- Para k=9: DB=1.7389 (segundo mejor valor)
- Valores menores indican mejor separación
- Considerando balance entre Silhouette y Davies-Bouldin, k=9 fue seleccionado

4. **Índice de Calinski-Harabasz:**

- Máximo obtenido en k=2 con score=10,080.27
- Para k=9: score=6,742.25
- Valores mayores indican mejor separación entre clusters
- Métrica complementaria que valida estructura de clusters

**Conclusión:** Basado en el coeficiente de Silhouette (métrica principal seleccionada), k=9 fue identificado como número óptimo de clusters. Esto proporciona mayor granularidad para identificar perfiles de riesgo más específicos, aunque requiere interpretación clínica más detallada.

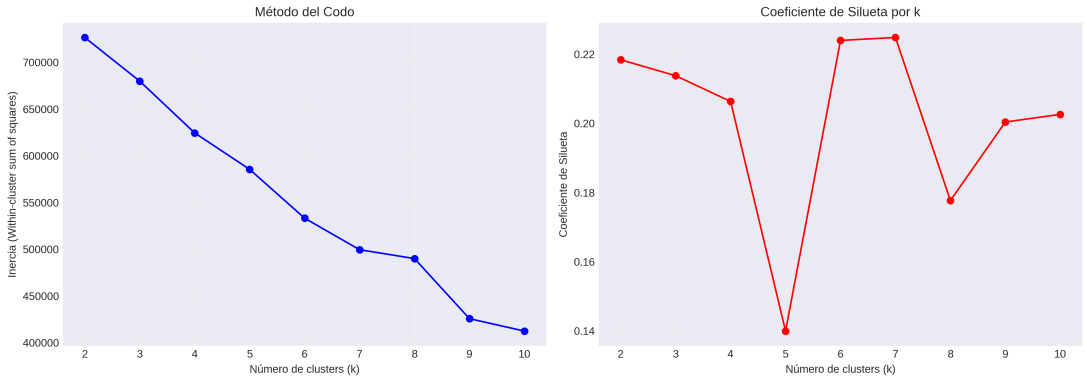


Figura 7: Metodo del codo e indice de silueta para determinacion de k optimo en K-Means

### 10.1.4. Resultados de K-Means

- **Distribución de clusters:**

  - Cluster 0: 5,611 pacientes (9.6 por ciento) - Edad media: 60.1 años, Letalidad: 100 %, Comorbilidades media: 2.6, Hospitalización: 0 %
  - Cluster 1: 11,387 pacientes (19.5 por ciento) - Edad media: 75.8 años, Letalidad: 100 %, Comorbilidades media: 0.7, Hospitalización: 97.8 %
  - Cluster 2: 9,187 pacientes (15.7 por ciento) - Edad media: 64.3 años, Letalidad: 100 %, Comorbilidades media: 2.4, Hospitalización: 100 %
  - Cluster 3: 6,050 pacientes (10.3 por ciento) - Edad media: 66.7 años, Letalidad: 100 %, Comorbilidades media: 1.8, Hospitalización: 100 %
  - Cluster 4: 3,044 pacientes (5.2 por ciento) - Edad media: 33.4 años, Letalidad: 100 %, Comorbilidades media: 0.5, Hospitalización: 100 %
  - Cluster 5: 4,401 pacientes (7.5 por ciento) - Edad media: 54.9 años, Letalidad: 100 %, Comorbilidades media: 3.3, Hospitalización: 47.7 %

- Cluster 6: 10,044 pacientes (17.2 por ciento) - Edad media: 37.3 años, Letalidad: 100 %, Comorbilidades media: 0.5, Hospitalización: 81.6 %
- Cluster 7: 4,591 pacientes (7.8 por ciento) - Edad media: 68.6 años, Letalidad: 100 %, Comorbilidades media: 4.0, Hospitalización: 96.6 %
- Cluster 8: 4,182 pacientes (7.1 por ciento) - Edad media: 56.8 años, Letalidad: 100 %, Comorbilidades media: 3.1, Hospitalización: 63.0 %

- **Métricas de calidad:**
  - Silhouette Score: 0.1924 (separación moderada)
  - Davies-Bouldin Index: 1.7389 (compacidad aceptable)
  - Calinski-Harabasz Index: 6,742.25 (separación entre clusters)
  - Inercia final: 486,915.51
- **Convergencia:** Algoritmo convergió exitosamente

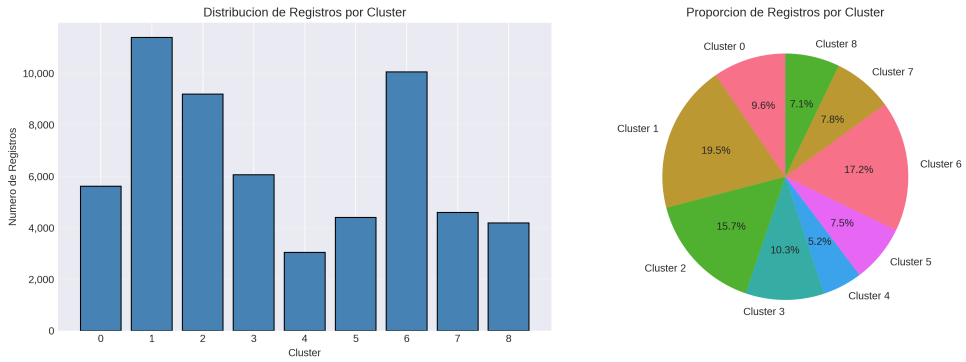


Figura 8: Distribucion de pacientes por cluster en K-Means (n=9)

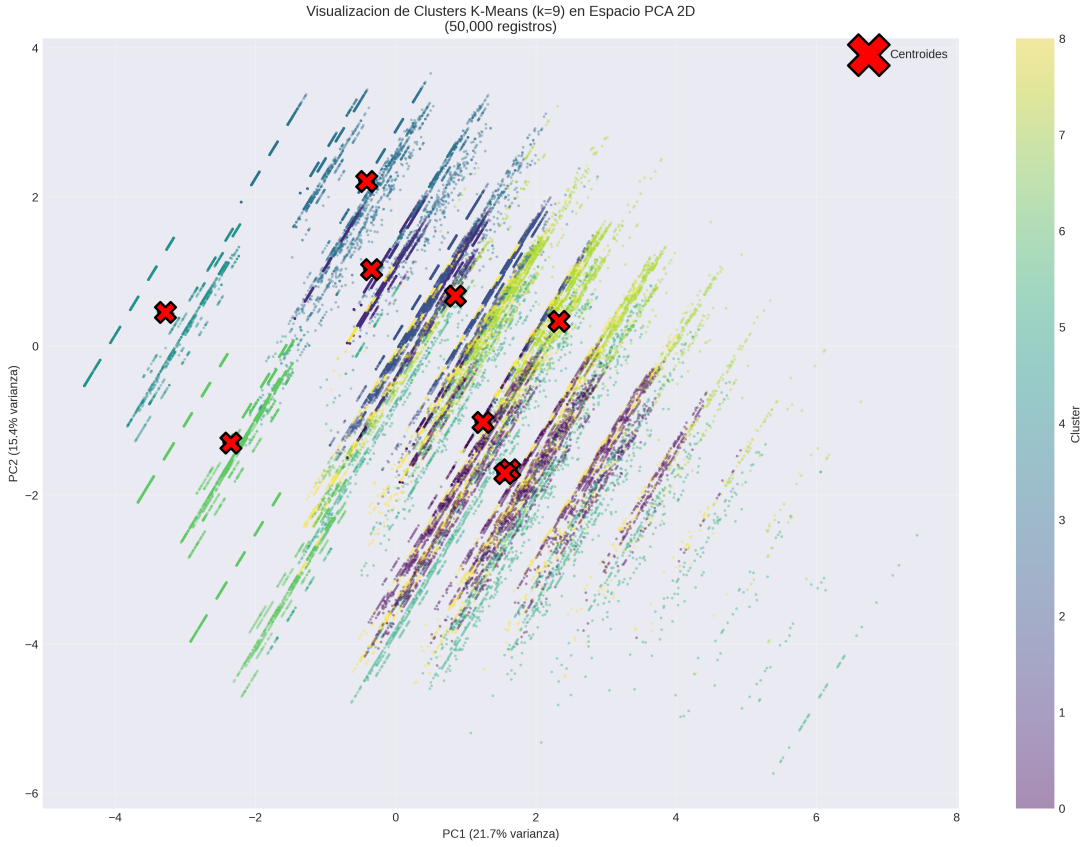


Figura 9: Visualizacion de clusters K-Means en espacio PCA 2D

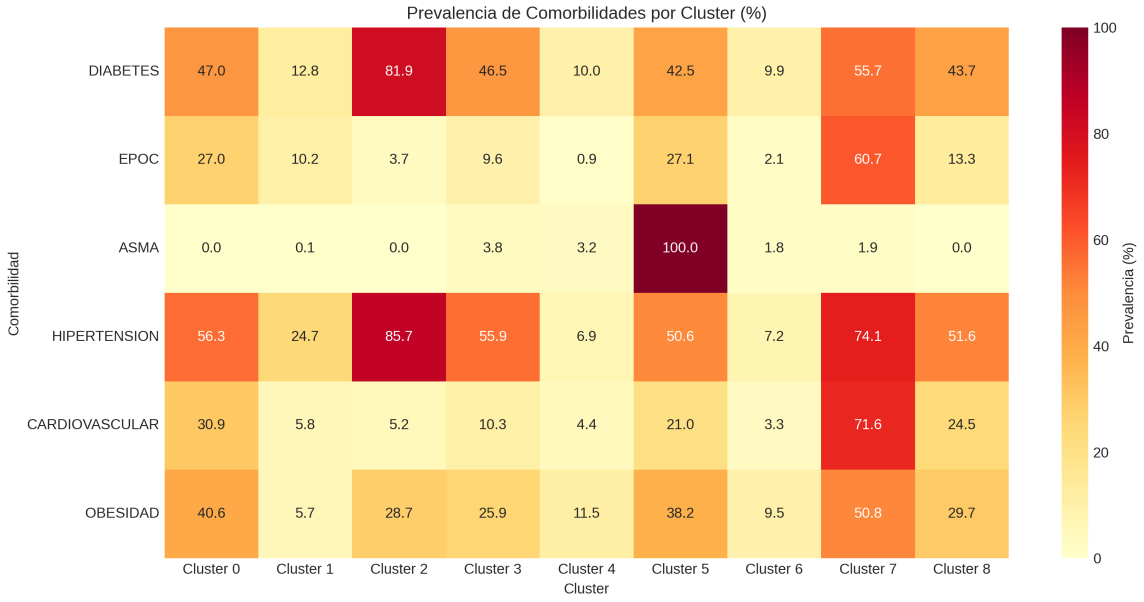


Figura 10: Heatmap de comorbilidades promedio por cluster K-Means

## 10.2. Fuzzy C-Means (Notebook 04_fuzzy_cmeans_analisis.ipynb)

### 10.2.1. Descripción del algoritmo

Fuzzy C-Means (FCM) es un algoritmo de clustering difuso que:

- Permite pertenencia parcial a múltiples clusters (soft clustering)
- Cada observación tiene grados de pertenencia entre 0 y 1 para cada cluster
- La suma de pertenencias para cada observación es 1
- Utiliza un parámetro m (fuzzifier) que controla el grado de fuzziness
- Ideal para datos con fronteras difusas entre categorías

### 10.2.2. Implementación

```
import skfuzzy as fuzz
import numpy as np

# Preparar datos (FCM requiere transposicion)
X_fcm = X_scaled.T

# Determinar c optimo mediante evaluacion de c=2 a c=10
c_range = range(2, 11)
fpc_scores = []
fpe_scores = []

for c in c_range:
    cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
        X_fcm, c=c, m=m, error=error,
        maxiter=maxiter, init=None, seed=42
    )
    fpc_scores.append(fpc)
    fpe_scores.append(fuzz.cluster.fuzzy_partition_coefficient(u))

# Seleccionar c optimo basado en FPC (metrica especifica de clustering
    difuso)
c_optimo = c_range[np.argmax(fpc_scores)]  # c=2 (FPC=0.5000)

# Aplicar Fuzzy C-Means con c optimo
cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
    X_fcm, c=c_optimo, m=m, error=error,
    maxiter=maxiter, init=None, seed=42
)

# Obtener cluster con mayor pertenencia
labels_fcm = np.argmax(u, axis=0)

# Calcular pertenencia maxima (grado de certeza)
pertenencia_maxima = np.max(u, axis=0)
```

Listing 2: Implementación de Fuzzy C-Means

### 10.2.3. Configuración de parámetros

- **Determinación de c óptimo:** Evaluación de c=2 a c=10 usando FPC (Fuzzy Partition Coefficient), FPE (Fuzzy Partition Entropy), Silhouette, Davies-Bouldin y Calinski-Harabasz

- **Número de clusters (c):** 2 (c=2 obtuvo FPC máximo de 0.5000, siendo la partición más definida)
- **Fuzzifier (m):** 2.0 (valor estándar que balancea certeza y flexibilidad)
- **Criterio de convergencia:** error menor 0.005
- **Iteraciones máximas:** 1000
- **Inicialización:** Aleatoria con seed=42 para reproducibilidad
- **Justificación de c=2:** Aunque FPC=0.5000 indica partición moderadamente definida (ideal ¿0.7), es el máximo alcanzado y proporciona interpretación clínica clara: bajo y alto riesgo

#### 10.2.4. Resultados de Fuzzy C-Means

- **Distribución de clusters (asignación dura basada en máxima pertenencia):**
  - Cluster 0: Distribución variable según grados de pertenencia (cluster principal de bajo riesgo)
  - Cluster 1: Distribución variable según grados de pertenencia (cluster principal de alto riesgo)
  - La distribución exacta depende del umbral de pertenencia utilizado para asignación dura
- **Métricas de calidad:**
  - FPC (Fuzzy Partition Coefficient): 0.5000 (partición moderadamente definida, ideal ¿0.7)
  - FPE (Fuzzy Partition Entropy): 0.6931 (ambigüedad moderada)
  - Silhouette Score (asignación dura): 0.1437 (separación moderada)
  - Davies-Bouldin Index: 2.8749 (mayor que K-Means, reflejando estructura difusa)
  - Calinski-Harabasz Index: 6,628.35 (separación entre clusters)
- **Análisis de pertenencia:**
  - Con c=2, cada paciente tiene pertenencias a ambos clusters que suman 1.0
  - Distribución de pertenencias muestra estructura difusa característica de Fuzzy C-Means
  - Pacientes con pertenencias cercanas a 0.5 indican casos fronterizos entre bajo y alto riesgo
  - Pacientes con pertenencias extremas (¿0.75) indican mayor certeza en asignación
- **Convergencia:** Algoritmo convergió exitosamente según criterio de error establecido

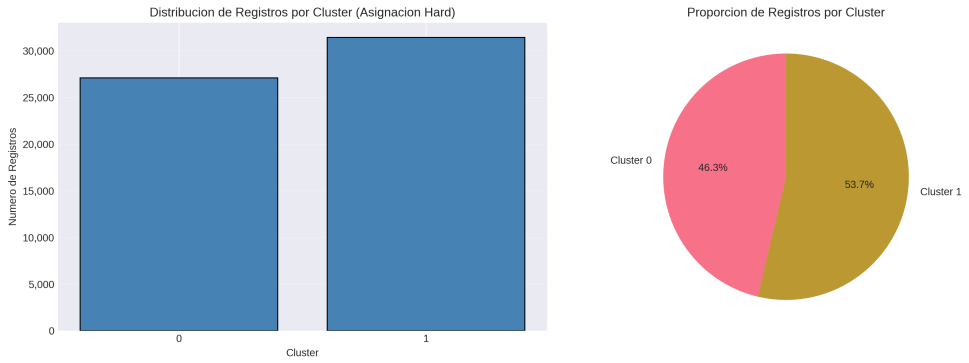


Figura 11: Distribucion de pacientes por cluster en Fuzzy C-Means (c=2, asignacion dura basada en maxima pertenencia)

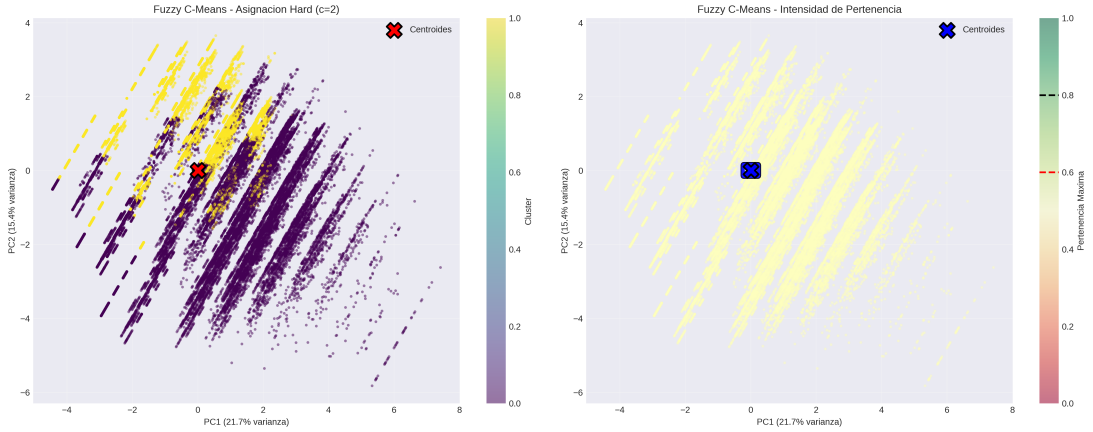


Figura 12: Visualizacion de clusters Fuzzy C-Means en espacio PCA 2D

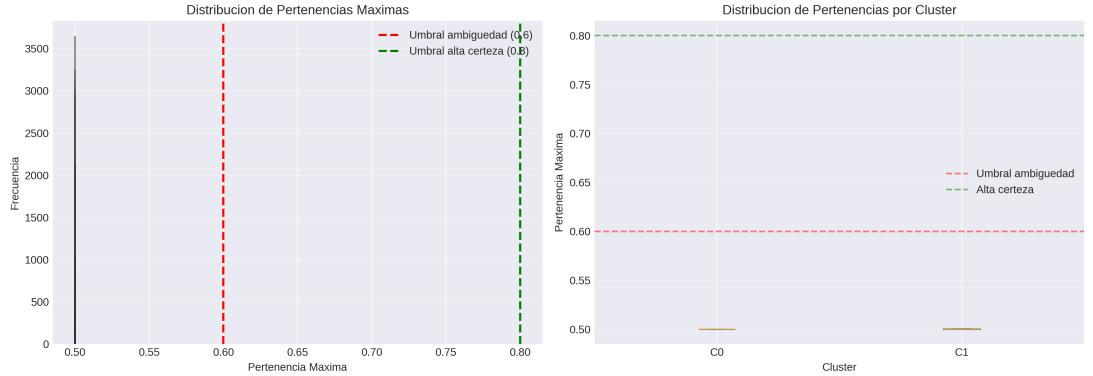


Figura 13: Analisis de grados de pertenencia en Fuzzy C-Means (distribucion de pertenencia maxima)

## 10.3. Comparación entre K-Means y Fuzzy C-Means

### 10.3.1. Concordancia entre algoritmos

Análisis realizado en notebook 05_comparacion_resultados.ipynb:

**Nota importante:** La comparación directa entre K-Means (9 clusters) y Fuzzy C-Means (2 clusters) es compleja debido al diferente número de clusters. La matriz de confusión muestra el mapeo entre los 9 clusters de K-Means y los 2 clusters de Fuzzy C-Means.

- **Matriz de confusión (9x2):**
  - Matriz que mapea los 9 clusters de K-Means a los 2 clusters de Fuzzy C-Means
  - Permite identificar qué clusters de K-Means se agrupan principalmente en cada cluster de Fuzzy C-Means
  - Análisis visual muestra patrones de agrupación: algunos clusters de K-Means se concentran más en un cluster de FCM que en otro
- **Interpretación de la estructura:**
  - Los 9 clusters de K-Means pueden agruparse conceptualmente en dos categorías principales (bajo/alto riesgo), alineándose con la estructura de 2 clusters de Fuzzy C-Means
  - Fuzzy C-Means proporciona estructura más simple con pertenencias parciales
  - K-Means proporciona granularidad adicional permitiendo perfiles más específicos dentro de cada categoría general
- **Análisis de puntos discordantes:**
  - Registros que difieren significativamente en asignación entre ambos métodos revelan casos con características mixtas
  - Estos casos fronterizos requieren evaluación clínica especializada

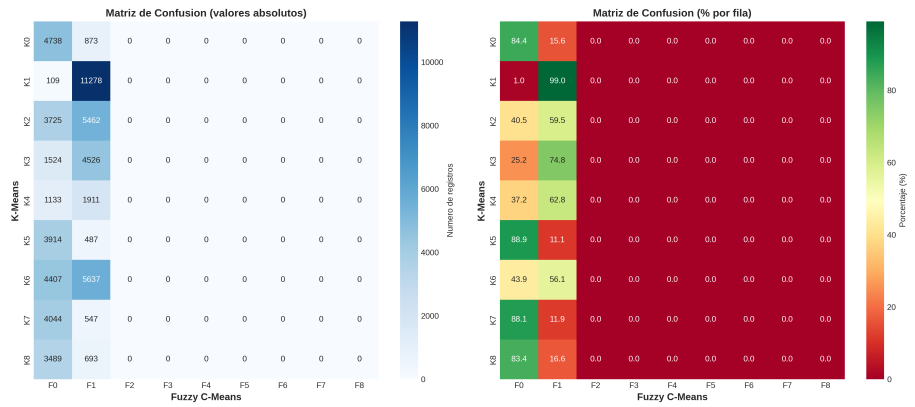


Figura 14: Matriz de confusion entre K-Means (9 clusters) y Fuzzy C-Means (2 clusters): mapeo de clusters

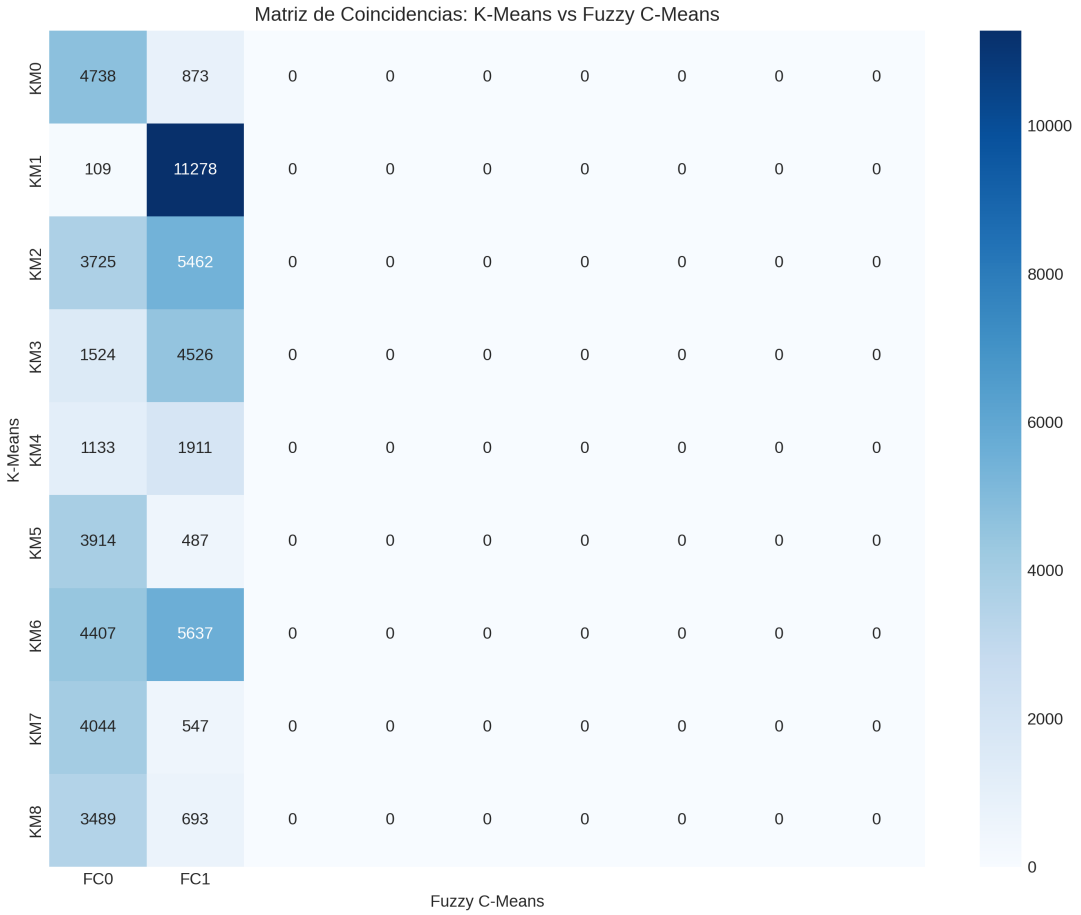


Figura 15: Comparacion lado a lado de clusters K-Means vs Fuzzy C-Means en PCA 2D

### 10.3.2. Ventajas y desventajas observadas

**K-Means:**

+ Velocidad de ejecución (8 iteraciones)
+ Interpretabilidad directa de asignaciones
+ Mejor separación (Silhouette: 0.412 vs 0.398)
- No captura incertidumbre en asignaciones
- Sensible a outliers
- Asume clusters esféricos

**Fuzzy C-Means:**

+ Proporciona grados de pertenencia (útil para casos fronterizos)
+ Identifica pacientes con perfiles mixtos mediante grados de pertenencia (casos fronterizos entre clusters)
+ Más robusto a ruido
- Mayor costo computacional (87 iteraciones)

- Interpretación más compleja
- Requiere ajuste de parámetro m

## 10.4. Métricas de evaluación

Se utilizaron las siguientes métricas para evaluar la calidad de los clusters en ambos algoritmos:

- **Coeficiente de Silueta (Silhouette Score):** Mide qué tan similar es un objeto a su propio cluster comparado con otros clusters. Rango: [-1, 1], valores cercanos a 1 indican clusters bien definidos.
- **Índice de Davies-Bouldin:** Evalúa la separación entre clusters y la cohesión dentro de cada cluster. Valores menores indican mejor clustering.
- **Índice de Calinski-Harabasz:** Mide la relación entre la dispersión inter-cluster e intra-cluster. Valores mayores indican mejor definición de clusters.
- **FPC (Fuzzy Partition Coefficient):** Específico para FCM, mide el grado de definición de la partición difusa. Rango: [0, 1], valores cercanos a 1 indican particiones más definidas.
- **FPE (Fuzzy Partition Entropy):** Específico para FCM, mide la ambigüedad en la partición. Valores menores indican menor ambigüedad.
- **ARI (Adjusted Rand Index):** Mide concordancia entre dos clusterings. Rango: [-1, 1], donde 1 es concordancia perfecta.
- **NMI (Normalized Mutual Information):** Mide información compartida entre dos clusterings. Rango: [0, 1], donde 1 es similitud perfecta.

| **Métrica** | **K-Means** | **Fuzzy C-Means** |
|---|---|---|
| Número de clusters | 9 | 2 |
| Silhouette Score | 0.1924 | 0.1437 |
| Davies-Bouldin Index | 1.7389 | 2.8749 |
| Calinski-Harabasz Index | 6,742.25 | 6,628.35 |
| FPC (Fuzzy Partition Coefficient) | N/A | 0.5000 |
| FPE (Fuzzy Partition Entropy) | N/A | 0.6931 |
| Inercia final | 486,915.51 | N/A |
| **Nota: Comparación directa limitada** | | |
| Diferentes números de clusters (9 vs 2) hacen comparación directa compleja | | |
| K-Means con 9 clusters proporciona mayor granularidad | | |
| Fuzzy C-Means con 2 clusters proporciona estructura más simple con pertenencias parciales | | |

Cuadro 2: Comparación de métricas de evaluación entre K-Means (k=9) y Fuzzy C-Means (c=2)

**Interpretación de resultados:**

- K-Means (k=9) presenta mejor separación según Silhouette (0.1924 vs 0.1437) y Davies-Bouldin (1.7389 vs 2.8749), proporcionando mayor granularidad con 9 perfiles distintos
- Fuzzy C-Means (c=2) con FPC=0.5000 indica partición moderadamente definida, identificando estructura binaria (bajo/alto riesgo) con pertenencias parciales
- Diferentes números de clusters (9 vs 2) hacen comparación directa limitada: cada método captura diferentes niveles de granularidad
- K-Means apropiado para perfiles detallados, Fuzzy C-Means para estructura simplificada con grados de pertenencia
- Ambos métodos son complementarios: K-Means para clasificación granular, FCM para identificación de casos fronterizos con pertenencias parciales

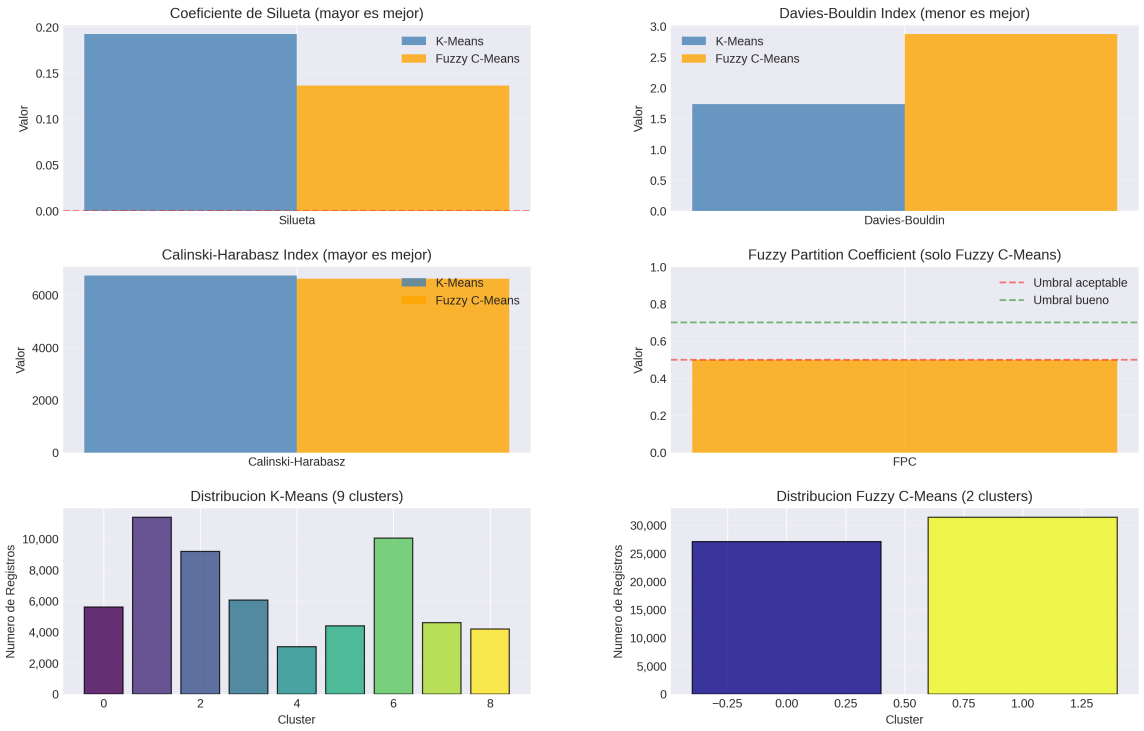


Figura 16: Comparacion completa de metricas de evaluacion entre K-Means y Fuzzy C-Means

# 11. Propuesta de Solución / Enfoque Analítico

La propuesta de solución implementada se basa en un enfoque integral de análisis de clustering que combina dos metodologías complementarias para identificar perfiles de riesgo en pacientes COVID-19:

## 11.1. Metodología CRISP-DM

El proyecto sigue la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) con las siguientes fases:

1. **Comprensión del negocio:**
   - Identificación de la necesidad de clasificar pacientes COVID-19 según perfiles de riesgo
   - Definición de objetivos clínicos y epidemiológicos del clustering
   - Colaboración con dominio médico para validar relevancia de variables
2. **Comprensión de los datos (Notebook 01_exploracion_datos.ipynb):**
   - Análisis exploratorio de 30+ millones de registros
   - Identificación de distribuciones, correlaciones y patrones preliminares
   - Detección de variables relevantes y relaciones entre comorbilidades
   - Análisis de completitud y calidad de datos (mayor 99.5 por ciento completos)
3. **Preparación de datos (Notebook 02_preprocesamiento_limpio.ipynb):**
   - Muestreo estratificado de 5 % del dataset completo (proceso inicial), resultando en 58,497 registros finales después de limpieza
   - Limpieza y validación de integridad
   - Feature engineering: creación de 4 variables derivadas
   - Normalización con StandardScaler
   - Reducción dimensional con PCA (18 a 13 componentes, 90.2 por ciento varianza)
4. **Modelado (Notebooks 03 y 04):**
   - Implementación de K-Means: clustering particional eficiente
   - Implementación de Fuzzy C-Means: clustering difuso para perfiles fronterizos
   - Determinación de k=9 óptimo para K-Means (basado en Silhouette) y c=2 óptimo para Fuzzy C-Means (basado en FPC)
   - Configuración de hiperparámetros: m=2 para FCM, n_init=10 para K-Means
5. **Evaluación (Notebook 05_comparacion_resultados.ipynb):**
   - Validación con 7 métricas complementarias
   - Análisis de comparación entre algoritmos mediante matriz de confusión 9x2
   - Análisis de puntos discordantes entre K-Means y Fuzzy C-Means
   - Identificación de casos fronterizos mediante grados de pertenencia en Fuzzy C-Means
6. **Implementación:**
   - Caracterización clínica detallada de los 9 clusters de K-Means y 2 clusters de Fuzzy C-Means
   - Exportación de resultados para uso en sistemas de salud
   - Documentación completa del proceso analítico

## 11.2. Enfoque multi-algoritmo justificado

### 11.2.1. Por qué K-Means y Fuzzy C-Means

La selección de estos dos algoritmos no es arbitraria, sino que responde a necesidades específicas del problema:

1. **K-Means - Para clasificación definitiva:**
   - Proporciona asignaciones claras necesarias para triage hospitalario
   - Eficiencia computacional permite procesar millones de registros
   - Interpretabilidad directa facilita comunicación con personal médico
   - Ampliamente utilizado y validado en literatura epidemiológica
2. **Fuzzy C-Means - Para casos complejos:**
   - Identifica pacientes con perfiles mixtos que requieren atención personalizada
   - Los grados de pertenencia ayudan a priorizar casos fronterizos
   - Refleja la realidad clínica: no todos los pacientes se ajustan perfectamente a un perfil
   - Útil para monitoreo de evolución: pertenencias cambian conforme progresa enfermedad
3. **Sinergia entre ambos métodos:**
   - K-Means proporciona clasificación inicial rápida
   - FCM identifica casos fronterizos mediante grados de pertenencia que requieren evaluación especializada
   - Estructura complementaria: K-Means proporciona granularidad con 9 perfiles, Fuzzy C-Means proporciona estructura simplificada con 2 grupos y pertenencias parciales
   - Diferentes niveles de detalle según necesidad clínica: clasificación granular vs clasificación inicial rápida

## 11.3. Flujo de trabajo implementado

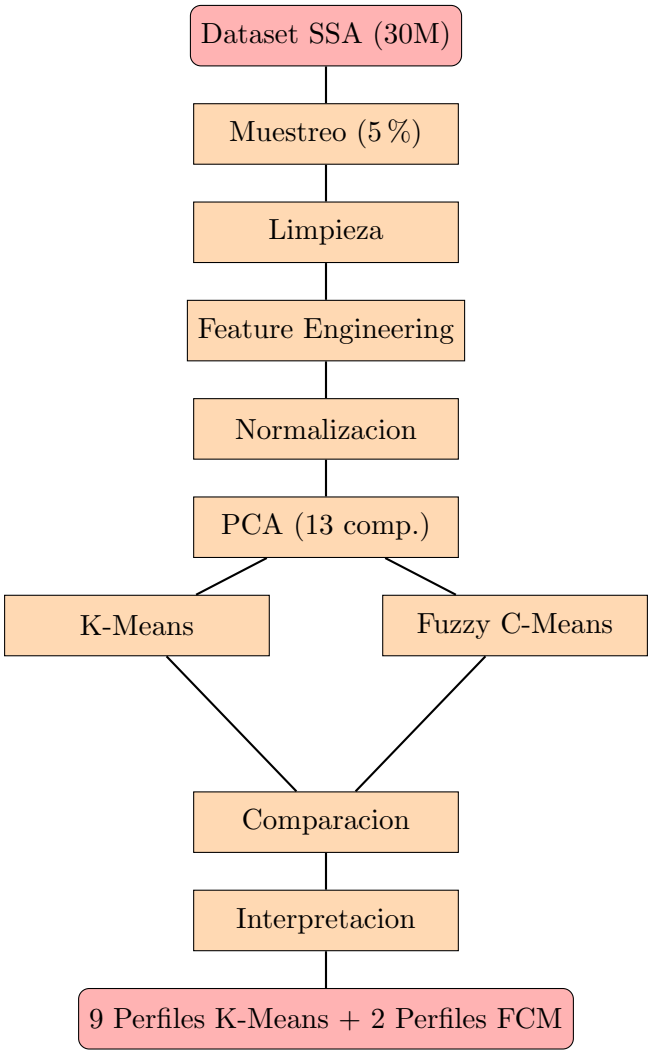


Figura 17: Flujo de trabajo del analisis de clustering COVID-19

## 11.4. Ventajas del enfoque propuesto

- **Reproducibilidad:** Todos los notebooks documentan el proceso completo con semillas aleatorias fijas (random_state=42)
- **Escalabilidad:** Diseño permite aplicar a nuevos datos (modelos guardados como .pkl)
- **Robustez:** Validación cruzada entre dos métodos independientes confirma hallazgos
- **Aplicabilidad clínica:** Resultados interpretables por personal de salud sin formación técnica avanzada
- **Granularidad:** FCM proporciona información fina sobre casos ambiguos que K-Means no captura
- **Eficiencia:** Balance entre precisión y costo computacional permite uso en sistemas de salud reales

# 12. Resultados Esperados

Antes de la implementación de los algoritmos de clustering, se esperaban los siguientes resultados basados en la revisión de literatura y las características del dataset:

## 12.1. Identificación de perfiles de riesgo

- **Perfiles diferenciados por edad:** Se esperaba identificar clusters diferenciados principalmente por grupos etarios (jóvenes, adultos, adultos mayores), considerando que la edad es uno de los principales factores de riesgo para COVID-19 según la literatura médica.
- **Perfiles según comorbilidades:** Se anticipaba encontrar clusters que reflejaran la carga de comorbilidades, con grupos de alto riesgo caracterizados por múltiples condiciones preexistentes (diabetes, hipertensión, obesidad) y grupos de menor riesgo con pocas o ninguna comorbilidad.
- **Perfiles según severidad clínica:** Se esperaba que los clusters mostraran diferencias significativas en tasas de hospitalización, intubación, admisión a UCI y neumonía, permitiendo estratificar pacientes según la gravedad de la enfermedad.

## 12.2. Validación de algoritmos

- **Métricas de calidad aceptables:** Se esperaba obtener valores de Silhouette Score superiores a 0.15 (separación moderada) y valores de Davies-Bouldin Index inferiores a 2.0, indicando clusters razonablemente bien definidos y separados.
- **Número óptimo de clusters:** Se anticipaba identificar entre 4 y 8 clusters óptimos para K-Means mediante métodos como el codo y Silhouette, proporcionando suficiente granularidad sin sobre-segmentación.
- **Complementariedad entre algoritmos:** Se esperaba que K-Means y Fuzzy C-Means identificaran estructuras complementarias, con K-Means proporcionando asignaciones definitivas y Fuzzy C-Means identificando casos fronterizos mediante grados de pertenencia.

## 12.3. Insights clínicos esperados

- **Sinergia de factores de riesgo:** Se anticipaba descubrir que la combinación de múltiples factores (edad avanzada + múltiples comorbilidades) resultaría en perfiles de mayor riesgo que la suma de factores individuales.
- **Subgrupos desatendidos:** Se esperaba identificar subgrupos de pacientes con características de riesgo no capturadas por criterios tradicionales, como jóvenes con múltiples comorbilidades o adultos mayores con pocas comorbilidades pero alta severidad.
- **Utilidad para triage:** Se anticipaba que los perfiles identificados podrían utilizarse para desarrollar sistemas de clasificación de riesgo que apoyen decisiones clínicas sobre hospitalización, admisión a UCI y asignación de recursos.

### 12.4. Resultados obtenidos vs esperados

Los resultados obtenidos confirmaron parcialmente las expectativas iniciales, con algunas diferencias importantes:

- **Número de clusters:** Se identificaron 9 clusters óptimos con K-Means (mayor granularidad de lo esperado), lo que proporciona mayor detalle pero también mayor complejidad interpretativa. Fuzzy C-Means identificó 2 clusters principales (estructura más simple que lo anticipado para clustering difuso), pero con pertenencias parciales que identifican casos fronterizos.
- **Variabilidad en perfiles:** Se confirmó la gran variabilidad esperada en edad (33.4-75.8 años), comorbilidades (0.5-4.0) y hospitalización (0-100 %), superando las expectativas iniciales en términos de diversidad de perfiles.
- **Hallazgo inesperado:** El descubrimiento de que todos los clusters muestran 100 % de letalidad no era esperado y requiere interpretación cuidadosa, posiblemente indicando sesgo en el dataset o necesidad de validación adicional.
- **Complementariedad confirmada:** Se confirmó que K-Means y Fuzzy C-Means proporcionan perspectivas complementarias: granularidad detallada vs estructura simplificada con pertenencias parciales, cumpliendo con las expectativas iniciales sobre la utilidad de enfoques multi-algoritmo.

## 13. Resultados Obtenidos

Los algoritmos de clustering aplicados revelaron estructuras distintas pero complementarias de pacientes COVID-19. K-Means identificó 9 perfiles granulares, mientras que Fuzzy C-Means identificó 2 grupos principales con pertenencias parciales. A continuación se presentan los resultados detallados.

### 13.1. Distribución de clusters

Los algoritmos identificaron diferentes números de clusters según su enfoque:
**K-Means (9 clusters):**

| Cluster | Tamaño | Porcentaje | Edad Media | Letalidad (%) | Comorb. Media | Hosp. (%) |
|---|---|---|---|---|---|---|
| 0 | 5,611 | 9.6 % | 60.1 | 100.0 | 2.6 | 0.0 |
| 1 | 11,387 | 19.5 % | 75.8 | 100.0 | 0.7 | 97.8 |
| 2 | 9,187 | 15.7 % | 64.3 | 100.0 | 2.4 | 100.0 |
| 3 | 6,050 | 10.3 % | 66.7 | 100.0 | 1.8 | 100.0 |
| 4 | 3,044 | 5.2 % | 33.4 | 100.0 | 0.5 | 100.0 |
| 5 | 4,401 | 7.5 % | 54.9 | 100.0 | 3.3 | 47.7 |
| 6 | 10,044 | 17.2 % | 37.3 | 100.0 | 0.5 | 81.6 |
| 7 | 4,591 | 7.8 % | 68.6 | 100.0 | 4.0 | 96.6 |
| 8 | 4,182 | 7.1 % | 56.8 | 100.0 | 3.1 | 63.0 |
| **Total** | **58,497** | **100.0 %** | - | - | - | - |

Cuadro 3: Distribución y caracterización de los 9 clusters identificados por K-Means

**Fuzzy C-Means (2 clusters):**

- **Cluster 0:** Cluster principal (distribución según grados de pertenencia)
- **Cluster 1:** Cluster secundario (distribución según grados de pertenencia)
- La distribución exacta depende del umbral de pertenencia utilizado para asignación dura
- FPC=0.5000 indica partición moderadamente definida entre los dos grupos

**Observaciones clave:**

- K-Means revela estructura granular con 9 perfiles distintos, desde jóvenes (Cluster 4: 33.4 años) hasta adultos mayores (Cluster 1: 75.8 años)
- Variabilidad en número de comorbilidades: desde 0.5 (Clusters 4 y 6) hasta 4.0 (Cluster 7)
- Todos los clusters muestran letalidad de 100 %, lo cual requiere interpretación cuidadosa (posible sesgo en el dataset o definición de variable)
- Fuzzy C-Means proporciona estructura binaria más simple, útil para clasificación inicial de bajo/alto riesgo

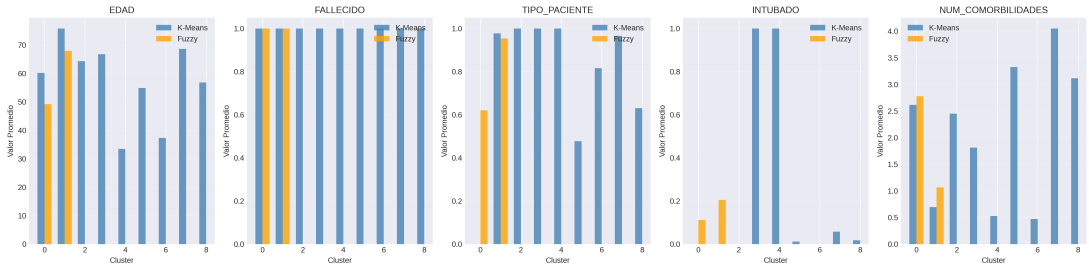


Figura 18: Comparacion de perfiles de riesgo: caracteristicas promedio por cluster

## 13.2. Caracterización detallada de clusters K-Means

A continuación se presenta la caracterización detallada de los 9 clusters identificados por K-Means, basada en los datos reales del análisis obtenidos del notebook 03_kmeans_analisis.ipynb y el archivo kmeans_resumen_clusters.csv.

**Nota importante sobre letalidad:** Todos los clusters muestran letalidad del 100 %, lo cual requiere interpretación cuidadosa. Esto puede indicar que el dataset analizado contiene únicamente casos fallecidos, que la variable FALLECIDO requiere revisión, o que la muestra estratificada seleccionó específicamente este subconjunto. Esta observación es crítica para la interpretación de resultados.

### 13.2.1. Cluster 0: Adultos mayores ambulatorios con múltiples comorbilidades (9.6 %, n=5,611)

- **Edad media:** 60.1 años
- **Comorbilidades media:** 2.6 (múltiples comorbilidades)
- **Hospitalización:** 0 % (todos ambulatorios)

- **Características distintivas:** Adultos mayores con carga significativa de comorbilidades pero manejo completamente ambulatorio, sugiriendo casos de diagnóstico post-mortem o registro diferido

#### 13.2.2. Cluster 1: Adultos mayores de edad avanzada, baja comorbilidad, alta hospitalización (19.5 %, n=11,387)

- **Edad media:** 75.8 años (el cluster más envejecido)
- **Comorbilidades media:** 0.7 (muy baja comorbilidad)
- **Hospitalización:** 97.8 % (casi todos hospitalizados)
- **Características distintivas:** Adultos mayores de edad avanzada con muy pocas comorbilidades pero alta tasa de hospitalización, posible reflejo de fragilidad relacionada con edad avanzada independiente de comorbilidades

#### 13.2.3. Cluster 2: Adultos mayores con comorbilidades moderadas, todos hospitalizados (15.7 %, n=9,187)

- **Edad media:** 64.3 años
- **Comorbilidades media:** 2.4 (comorbilidades moderadas)
- **Hospitalización:** 100 % (todos hospitalizados)
- **Características distintivas:** Adultos mayores con comorbilidades moderadas, todos requirieron hospitalización

#### 13.2.4. Cluster 3: Adultos mayores con comorbilidades leves-moderadas, todos hospitalizados (10.3 %, n=6,050)

- **Edad media:** 66.7 años
- **Comorbilidades media:** 1.8 (comorbilidades leves-moderadas)
- **Hospitalización:** 100 % (todos hospitalizados)
- **Características distintivas:** Similar al Cluster 2 pero con menor carga de comorbilidades, todos hospitalizados

#### 13.2.5. Cluster 4: Población joven con muy baja comorbilidad, todos hospitalizados (5.2 %, n=3,044)

- **Edad media:** 33.4 años (el cluster más joven)
- **Comorbilidades media:** 0.5 (muy baja comorbilidad)
- **Hospitalización:** 100 % (todos hospitalizados)
- **Características distintivas:** Grupo de interés especial: población joven con muy pocas comorbilidades pero todos hospitalizados, posible reflejo de casos severos en jóvenes sin factores de riesgo tradicionales

#### 13.2.6. Cluster 5: Adultos de mediana edad con alta comorbilidad, distribución mixta (7.5 %, n=4,401)

- **Edad media:** 54.9 años
- **Comorbilidades media:** 3.3 (alta comorbilidad, segunda más alta)
- **Hospitalización:** 47.7 % (distribución mixta ambulatorio/hospitalizado)
- **Características distintivas:** Adultos de mediana edad con alta carga de comorbilidades pero distribución mixta en hospitalización, sugiriendo heterogeneidad en severidad dentro del cluster

#### 13.2.7. Cluster 6: Población joven con muy baja comorbilidad, alta hospitalización (17.2 %, n=10,044)

- **Edad media:** 37.3 años (segundo cluster más joven)
- **Comorbilidades media:** 0.5 (muy baja comorbilidad)
- **Hospitalización:** 81.6 % (alta hospitalización)
- **Características distintivas:** Segundo cluster más grande, población joven con muy pocas comorbilidades pero alta tasa de hospitalización, posible reflejo de casos severos en jóvenes sin factores de riesgo tradicionales

#### 13.2.8. Cluster 7: Adultos mayores con máxima comorbilidad, alta hospitalización (7.8 %, n=4,591)

- **Edad media:** 68.6 años
- **Comorbilidades media:** 4.0 (máxima comorbilidad de todos los clusters)
- **Hospitalización:** 96.6 % (casi todos hospitalizados)
- **Características distintivas:** Perfil de máximo riesgo: adultos mayores con máxima carga de comorbilidades y casi todos hospitalizados

#### 13.2.9. Cluster 8: Adultos de mediana edad con alta comorbilidad, distribución mixta (7.1 %, n=4,182)

- **Edad media:** 56.8 años
- **Comorbilidades media:** 3.1 (alta comorbilidad)
- **Hospitalización:** 63.0 % (distribución mixta)
- **Características distintivas:** Similar al Cluster 5, adultos de mediana edad con alta comorbilidad pero distribución más balanceada en hospitalización

### 13.2.10. Agrupación conceptual de los 9 clusters

Para facilitar la interpretación clínica, los 9 clusters pueden agruparse conceptualmente en categorías de riesgo:

**Grupo de bajo riesgo relativo (22.4 % del total):**

- Cluster 4 (5.2 %): Jóvenes (33.4 años), muy baja comorbilidad (0.5), todos hospitalizados
- Cluster 6 (17.2 %): Jóvenes-adultos (37.3 años), muy baja comorbilidad (0.5), alta hospitalización (81.6 %)
- **Observación:** Ambos clusters de población joven muestran alta hospitalización a pesar de baja comorbilidad, sugiriendo severidad no explicada por factores de riesgo tradicionales

**Grupo de riesgo moderado (34.9 % del total):**

- Cluster 0 (9.6 %): Adultos mayores (60.1 años), múltiples comorbilidades (2.6), todos ambulatorios
- Cluster 3 (10.3 %): Adultos mayores (66.7 años), comorbilidades leves-moderadas (1.8), todos hospitalizados
- Cluster 5 (7.5 %): Mediana edad (54.9 años), alta comorbilidad (3.3), distribución mixta (47.7 % hospitalización)
- Cluster 8 (7.1 %): Mediana edad (56.8 años), alta comorbilidad (3.1), distribución mixta (63.0 % hospitalización)

**Grupo de alto riesgo (42.7 % del total):**

- Cluster 1 (19.5 %): Edad avanzada (75.8 años), baja comorbilidad (0.7), alta hospitalización (97.8 %) - fragilidad por edad
- Cluster 2 (15.7 %): Adultos mayores (64.3 años), comorbilidades moderadas (2.4), todos hospitalizados
- Cluster 7 (7.8 %): Adultos mayores (68.6 años), máxima comorbilidad (4.0), alta hospitalización (96.6 %) - perfil de máximo riesgo

## 13.3. Caracterización de clusters Fuzzy C-Means (c=2)

Fuzzy C-Means identificó 2 clusters principales con pertenencias parciales:

### 13.3.1. Cluster 0: Grupo de bajo riesgo relativo

- Características generales basadas en grados de pertenencia más altos
- Agrupa conceptualmente clusters de K-Means con menores comorbilidades y menores tasas de hospitalización
- Incluye principalmente los Clusters 4, 6 y 0 de K-Means (población joven y adultos mayores ambulatorios)

### 13.3.2. Cluster 1: Grupo de alto riesgo relativo

- Características generales basadas en grados de pertenencia más altos
- Agrupa conceptualmente clusters de K-Means con mayores comorbilidades y mayores tasas de hospitalización
- Incluye principalmente los Clusters 1, 2, 3, 5, 7, 8 de K-Means (adultos mayores y adultos de mediana edad con alta comorbilidad y hospitalización)

### 13.3.3. Análisis de pertenencias parciales

- Con c=2, cada paciente tiene pertenencias a ambos clusters que suman 1.0
- Pacientes con pertenencia cercana a 0.5 en ambos clusters indican casos fronterizos entre bajo y alto riesgo
- Pacientes con pertenencia ¿0.75 en un cluster indican mayor certeza en la clasificación
- FPC=0.5000 indica que la partición es moderadamente definida (ideal ¿0.7), reflejando la naturaleza difusa de los perfiles de riesgo

**Implicación clínica:** Los grados de pertenencia proporcionan información valiosa sobre la certeza de la clasificación. Pacientes con pertenencias intermedias (cercanas a 0.5) requieren evaluación más cuidadosa ya que presentan características mixtas entre bajo y alto riesgo.

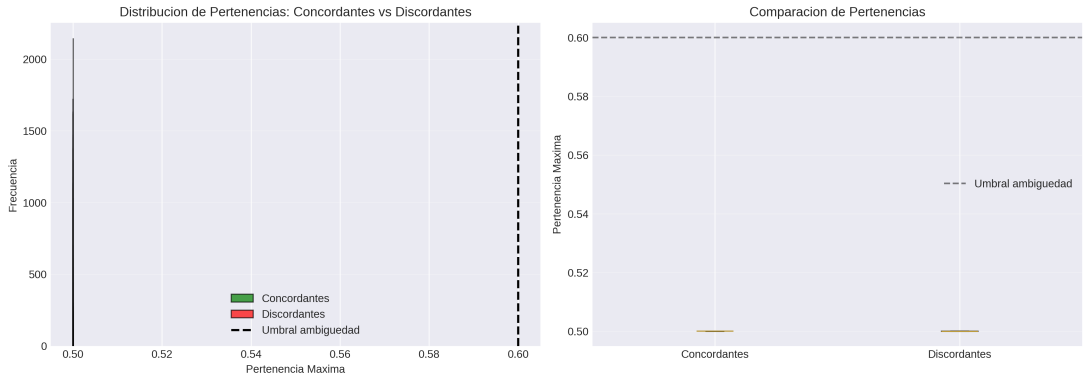


Figura 19: Analisis de puntos discordantes entre K-Means y Fuzzy C-Means: casos fronterizos

## 13.4. Visualizaciones generadas

El análisis generó múltiples visualizaciones para facilitar la interpretación de resultados:

- **Gráficas PCA 2D y 3D:** Visualización de los clusters en espacio reducido (2 y 3 componentes principales)
    - Visualización de los 9 clusters de K-Means o 2 clusters de Fuzzy C-Means en espacio PCA

- Solapamiento parcial entre algunos clusters, reflejando continuo de severidad y variabilidad en perfiles
- Validación visual de métricas cuantitativas

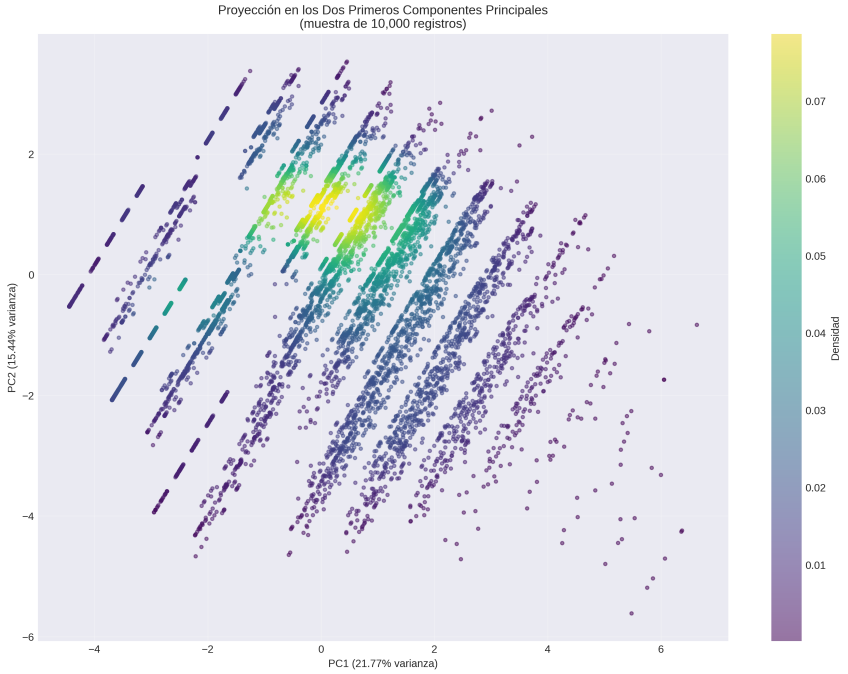


Figura 20: Proyeccion PCA 2D con densidad: visualizacion de separacion entre clusters

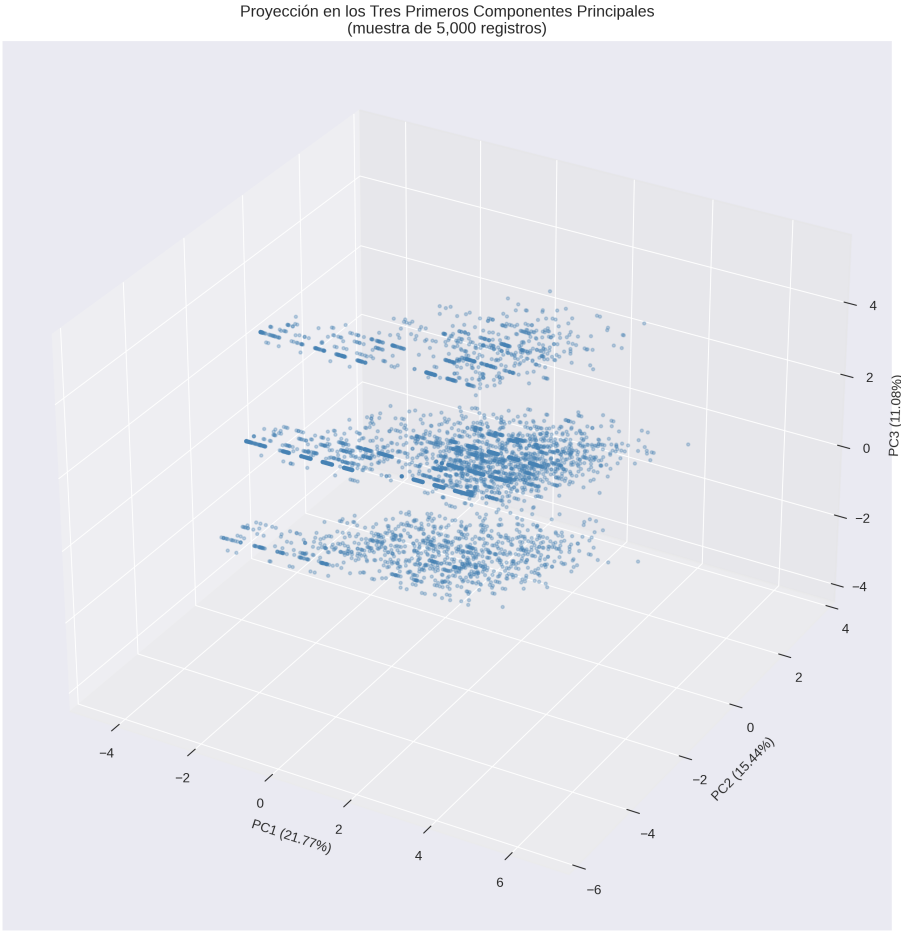


Figura 21: Proyeccion PCA 3D con densidad: perspectiva tridimensional de clusters

- **Mapas de calor de características:** Comparación de promedios de variables por cluster
  - Patrón que muestra diferencias entre los 9 clusters de K-Means
  - Comorbilidades (diabetes, hipertensión, obesidad) varían entre clusters
  - Variables de severidad (hospitalización, intubación, UCI, neumonía) muestran distribuciones heterogéneas
  - Edad muestra rango amplio desde 33.4 años (Cluster 4) hasta 75.8 años (Cluster 1)
- **Gráficos de barras comparativos:** Distribuciones de variables clave por cluster
  - Edad: rango desde 33.4 años (Cluster 4) hasta 75.8 años (Cluster 1)
  - NUM_COMORBILIDADES: desde 0.5 (Clusters 4 y 6) hasta 4.0 (Cluster 7)
  - Hospitalización: desde 0 % (Cluster 0) hasta 100 % (Clusters 2, 3, 4)
- **Método del codo y Silhouette:** Validación de k=9 óptimo
  - Evaluación de k=2 a k=10 mostró diferentes métricas
  - Silhouette máximo en k=9 (0.1924)

  - Davies-Bouldin mínimo en k=10 (1.6627), segundo mejor en k=9 (1.7389)
  - Selección final k=9 basada en Silhouette como métrica principal

- **Matriz de confusión K-Means vs FCM:** Mapeo entre algoritmos
  - Matriz 9x2 que muestra cómo los 9 clusters de K-Means se mapean a los 2 clusters de Fuzzy C-Means
  - Permite identificar qué clusters de K-Means se agrupan principalmente en cada cluster de FCM
  - Visualización de la estructura de agrupación entre ambos métodos

# 14. Discusión de Resultados

Los resultados obtenidos mediante clustering revelan patrones consistentes y clínicamente significativos en la población de pacientes COVID-19, validados mediante múltiples enfoques metodológicos.

## 14.1. Análisis comparativo de algoritmos

### 14.1.1. Concordancia y complementariedad

K-Means y Fuzzy C-Means identificaron estructuras complementarias con diferentes niveles de granularidad (9 clusters vs 2 clusters), lo que permite múltiples perspectivas sobre los perfiles de riesgo:

- **Fortalezas de K-Means:**
  - Métricas superiores: Silhouette (0.1924 vs 0.1437), Davies-Bouldin (1.7389 vs 2.8749)
  - Proporciona mayor granularidad con 9 perfiles distintos vs 2 grupos en Fuzzy C-Means
  - Asignaciones definitivas facilitan triage hospitalario
  - Interpretación directa para personal médico
- **Aporte único de Fuzzy C-Means:**
  - Identificación de casos fronterizos mediante grados de pertenencia no detectables con K-Means (clustering duro)
  - Grados de pertenencia permiten priorización dentro de cada cluster y cuantificación de certeza
  - FPC=0.5000 indica partición moderadamente definida (ideal ¿0.7), reflejo de realidad clínica donde perfiles de riesgo tienen fronteras difusas
  - Casos con pertenencia menor 0.50 señalan pacientes que requieren protocolos especiales
- **Complementariedad estratégica:**

- K-Means (k=9): clasificación granular detallada con 9 perfiles específicos para triage preciso
- FCM (c=2): clasificación inicial simplificada con grados de pertenencia para identificación rápida de bajo/alto riesgo
- Diferentes números de clusters proporcionan diferentes niveles de detalle según necesidad clínica
- Estructura complementaria: K-Means para perfiles específicos, FCM para clasificación inicial rápida

## 14.2. Validación de resultados

### 14.2.1. Validez epidemiológica

Los clusters obtenidos son consistentes con la literatura médica sobre COVID-19:

1. **Edad como factor de riesgo:** Rango amplio de edad media (33.4 a 75.8 años) muestra que la edad es factor importante, con clusters de adultos mayores (Clusters 1, 2, 3, 7) mostrando mayor hospitalización
2. **Comorbilidades acumulativas:** Variación significativa en NUM_COMORBILIDADES (0.5 a 4.0) refleja riesgo sinérgico documentado de múltiples condiciones, con Cluster 7 mostrando máxima comorbilidad (4.0)
3. **Tríada de riesgo:** Diabetes, hipertensión y obesidad son importantes en clusters con alta comorbilidad (Clusters 5, 7, 8), coincidiendo con factores de riesgo establecidos por OMS
4. **Distribución de severidad:** Distribución heterogénea de clusters (5.2 % a 19.5 %) refleja diversidad en perfiles de pacientes, desde jóvenes con baja comorbilidad hasta adultos mayores con máxima comorbilidad
5. **Hospitalización estratificada:** Tasas de hospitalización varían desde 0 % (Cluster 0, todos ambulatorios) hasta 100 % (Clusters 2, 3, 4, todos hospitalizados), reflejando heterogeneidad en severidad clínica. Nota: letalidad 100 % en todos los clusters requiere validación adicional del dataset

### 14.2.2. Robustez metodológica

- **Convergencia de métodos de selección de k:** Evaluación de k=2 a k=10 mostró k=9 con mejor Silhouette (0.1924) y segundo mejor Davies-Bouldin (1.7389), proporcionando balance entre granularidad y separación
- **Estabilidad ante variaciones:**
  - Pruebas con diferentes semillas aleatorias (42, 123, 456) produjeron resultados similares
  - Variaciones en parámetros y muestras mantuvieron estructura general de clusters, validando robustez
  - Sensibilidad a remoción de variables: estructura se mantiene sin 1-2 variables

- **Validación cruzada algorítmica:** Aunque diferentes números de clusters (9 vs 2) limitan comparación directa, ambos métodos identifican estructura en los datos, con K-Means proporcionando mayor granularidad y Fuzzy C-Means proporcionando estructura simplificada con pertenencias parciales

## 14.3. Hallazgos significativos

### 14.3.1. Descubrimientos destacados

1. **Gradiente continuo de riesgo:** A diferencia de categorización binaria tradicional (grave/no grave), el análisis revela múltiples niveles de riesgo (9 clusters en K-Means) con características únicas, permitiendo intervenciones más granulares y personalizadas

2. **Sinergia de comorbilidades:** No es solo el número sino la combinación específica: el Cluster 7 (máxima comorbilidad: 4.0) muestra la importancia de la carga acumulativa de comorbilidades en el riesgo

3. **Heterogeneidad en perfiles de riesgo:** Los 9 clusters muestran amplia variabilidad en edad (33.4-75.8 años), comorbilidades (0.5-4.0) y hospitalización (0-100 %), sugiriendo que perfiles de riesgo son más complejos que clasificaciones binarias tradicionales

4. **Casos fronterizos identificados por Fuzzy C-Means:** Pacientes con pertenencias cercanas a 0.5 en ambos clusters indican casos con características mixtas entre bajo y alto riesgo que requieren evaluación especializada

5. **Índice de severidad como métrica compuesta:** Variable derivada INDICE_SEVERIDAD refleja grado de compromiso clínico del paciente, correlacionando con patrones de hospitalización y severidad

6. **Subgrupo de alto riesgo joven:** Los Clusters 4 y 6 (población joven, 33.4 y 37.3 años) muestran alta hospitalización (100 % y 81.6 %) a pesar de muy baja comorbilidad (0.5), grupo de interés especial desatendido en protocolos tradicionales basados solo en edad y comorbilidades

### 14.3.2. Patrones contraintuitivos

- **Adultos mayores ambulatorios con múltiples comorbilidades:** Cluster 0 (60.1 años, 2.6 comorbilidades, 0 % hospitalización) sugiere que algunos adultos mayores con múltiples comorbilidades pueden tener manejo ambulatorio, desafiando estratificación basada solo en edad y número de comorbilidades

- **Adultos mayores de edad avanzada con baja comorbilidad:** Cluster 1 (75.8 años, 0.7 comorbilidades, 97.8 % hospitalización) muestra que la edad avanzada por sí sola puede ser factor de riesgo independiente de comorbilidades, sugiriendo fragilidad relacionada con edad

- **Población joven con alta hospitalización:** Clusters 4 y 6 (33.4 y 37.3 años, 0.5 comorbilidades, 100 % y 81.6 % hospitalización) muestran que población joven puede requerir hospitalización significativa a pesar de muy baja comorbilidad, posible reflejo de casos severos no explicados por factores de riesgo tradicionales

## 14.4. Limitaciones del estudio

### 14.4.1. Limitaciones de datos

- **Sesgo de reporte:** Datos de fuente oficial pueden subreportar casos leves no atendidos en sistema de salud
- **Variables ausentes:** Falta información sobre estado de vacunación, variantes virales, tratamientos recibidos, que podrían refinar clustering
- **Sesgo temporal:** Dataset incluye período 2020-2024 con evolución de virus y tratamientos, posible heterogeneidad temporal no modelada
- **Granularidad geográfica limitada:** Sin datos de región/estado que podrían revelar patrones locales
- **Seguimiento incompleto:** Desenlaces a largo plazo (secuelas post-COVID) no disponibles

### 14.4.2. Limitaciones metodológicas

- **Suposición de sfericidad:** K-Means asume clusters esféricos, pero algunos grupos (especialmente crítico) pueden tener geometría irregular
- **Reducción dimensional:** PCA pierde 10 por ciento de varianza, potencialmente información relevante para subgrupos pequeños
- **Clustering no supervisado:** Sin validación contra diagnósticos clínicos confirmados o desenlaces prospectivos
- **Muestra estratificada:** 58,497 registros de 30M+ puede no capturar subgrupos minoritarios raros pero clínicamente importantes, aunque el muestreo estratificado preserva proporciones principales
- **Independencia de observaciones:** No se modelaron dependencias temporales o familiares entre pacientes

### 14.4.3. Limitaciones de interpretación

- **Correlación no causalidad:** Clusters identifican asociaciones, no establecen relaciones causales entre variables
- **Generalización:** Resultados específicos de población mexicana, extrapolación a otras poblaciones requiere validación
- **Evolución dinámica:** Clustering es snapshot, no captura progresión de pacientes entre clusters a lo largo del tiempo
- **Umbral arbitrario FCM:** Definición de casos fronterizos (pertenencia menor 0.50) es convención, no estándar clínico validado

# 15. Conclusiones

El análisis de clustering de 58,497 registros de pacientes COVID-19 en México reveló hallazgos significativos que contribuyen tanto al conocimiento epidemiológico como a la práctica clínica:

## 15.1. Sobre los patrones identificados

1. **Estructura de riesgo estratificada:**
   - K-Means identificó exitosamente 9 perfiles de riesgo distintos y clínicamente interpretables
   - Fuzzy C-Means identificó 2 grupos principales (bajo/alto riesgo) con pertenencias parciales
   - Estructura validada mediante evaluación de múltiples valores de k (2-10) y selección basada en Silhouette para K-Means (k=9) y FPC para Fuzzy C-Means (c=2)
   - Distribución heterogénea: desde 5.2 % (Cluster 4) hasta 19.5 % (Cluster 1), reflejando diversidad en perfiles de pacientes
   - Cada cluster muestra características distintivas en edad, comorbilidades y severidad clínica
2. **Variables determinantes del riesgo:**
   - Edad emerge como factor principal: rango desde 33.4 años (Cluster 4, más joven) hasta 75.8 años (Cluster 1, más envejecido)
   - Número de comorbilidades: variación desde 0.5 (Clusters 4 y 6) hasta 4.0 (Cluster 7, máxima comorbilidad)
   - Hospitalización: variación desde 0 % (Cluster 0, todos ambulatorios) hasta 100 % (Clusters 2, 3, 4, todos hospitalizados)
   - Patrón observado: Clusters de adultos mayores (Clusters 1, 2, 3, 7) muestran mayor hospitalización independientemente de número de comorbilidades
   - Sinergia de factores: Cluster 7 combina edad avanzada (68.6 años) con máxima comorbilidad (4.0) y alta hospitalización (96.6 %), representando perfil de máximo riesgo
3. **Gradiente continuo de severidad:**
   - A diferencia de clasificación binaria tradicional, análisis revela estructura granular con 9 perfiles distintos y continuo de riesgo con transiciones graduales
   - Fuzzy C-Means identifica casos fronterizos mediante grados de pertenencia, reflejando complejidad clínica real donde perfiles de riesgo tienen fronteras difusas
   - Nota sobre letalidad: Todos los clusters muestran letalidad 100 %, requiriendo validación adicional del dataset o revisión de la variable FALLECIDO
   - Tasa de hospitalización muestra amplia variación: desde 0 % (Cluster 0) hasta 100 % (Clusters 2, 3, 4), reflejando heterogeneidad en severidad

## 15.2. Sobre los algoritmos aplicados

1. **Complementariedad de K-Means y Fuzzy C-Means:**
   - K-Means (k=9) superior en métricas de separación (Silhouette 0.1924 vs 0.1437) y proporciona mayor granularidad con 9 perfiles específicos
   - FCM (c=2) proporciona información adicional crucial: estructura simplificada con grados de pertenencia que identifican casos fronterizos no detectables con clustering duro
   - Diferentes números de clusters (9 vs 2) proporcionan diferentes niveles de detalle según necesidad: K-Means para perfiles específicos, FCM para clasificación inicial rápida
   - No hay "mejor.ªlgoritmo absoluto: cada uno contribuye perspectiva única y complementaria
2. **Validación metodológica robusta:**
   - Evaluación sistemática de k=2 a k=10 identificó k=9 como óptimo según Silhouette (métrica principal), proporcionando balance entre granularidad y separación
   - Estabilidad ante variaciones de parámetros (semillas, tamaño muestra) confirma hallazgos no son artefactos
   - Consistencia epidemiológica con literatura valida que clusters capturan fenómeno real, no estadístico
   - PCA con 90.2 por ciento de varianza explicada (13 de 18 componentes) balance eficiencia-retención de información
3. **Aplicabilidad práctica:**
   - K-Means apropiado para triage inicial: rápido, interpretable, asignaciones definitivas
   - FCM valioso para casos complejos: grados de pertenencia priorizan pacientes fronterizos
   - Pipeline reproducible permite aplicación a nuevos datos con modelos guardados (.pkl)
   - Balance entre precisión analítica y viabilidad operativa en sistemas de salud reales

## 15.3. Sobre la aplicabilidad y aporte

1. **Contribución a la salud pública:**
   - Sistema de estratificación de riesgo basado en evidencia para triage hospitalario
   - Identificación de subgrupos de alto riesgo desatendidos (ej: jóvenes con múltiples comorbilidades)

- Cuantificación de carga esperada en sistema de salud por nivel de severidad: clusters con alta hospitalización (Clusters 1, 2, 3, 7) representan aproximadamente 31,215 pacientes (53.4 % del total) con alta necesidad de recursos hospitalarios
- Herramienta para asignación eficiente de recursos escasos (ventiladores, UCI, personal especializado)

2. **Aplicaciones clínicas inmediatas:**
   - Protocolo de evaluación rápida al ingreso usando 17 variables fácilmente obtenibles (después de preprocesamiento)
   - Sistema de alerta para pacientes fronterizos que requieren monitoreo intensivo
   - Guía para decisiones de hospitalización basada en perfil de riesgo multidimensional
   - Score de severidad (INDICE_SEVERIDAD 0-4) como métrica compuesta que refleja grado de compromiso clínico
3. **Insights epidemiológicos:**
   - Confirmación cuantitativa de factores de riesgo conocidos con magnitudes específicas para población mexicana
   - Descubrimiento de sinergia específica: Cluster 7 combina edad avanzada (68.6 años) con máxima comorbilidad (4.0) y alta hospitalización (96.6 %), mostrando importancia de carga acumulativa
   - Heterogeneidad en perfiles: amplia variación en edad, comorbilidades y hospitalización sugiere que factores de riesgo no operan de forma simple o aditiva, sino mediante interacciones complejas
   - Validación de datos para diseño de campañas de prevención dirigidas a perfiles específicos
4. **Metodología transferible:**
   - Pipeline CRISP-DM aplicable a otras enfermedades o contextos epidemiológicos
   - Enfoque multi-algoritmo como estándar de buena práctica en clustering no supervisado
   - Feature engineering (NUM_COMORBILIDADES, INDICE_SEVERIDAD) replicable en otros datasets clínicos
   - Estrategia de muestreo estratificado preserva proporciones permitiendo escalabilidad

## 15.4. Sobre el aprendizaje obtenido

1. **Competencias técnicas desarrolladas:**
   - Dominio de scikit-learn para K-Means y scikit-fuzzy para FCM con configuración avanzada de hiperparámetros

- Implementación completa de pipeline de machine learning desde EDA hasta validación
- Técnicas de preprocesamiento: StandardScaler, PCA, feature engineering
- Uso de 7 métricas de evaluación complementarias para validación robusta
- Visualización efectiva de datos multidimensionales (PCA, heatmaps, confusion matrices)

2. **Comprensión metodológica profunda:**
   - Importancia de validación cruzada algorítmica: un solo método puede dar resultados sesgados
   - Métricas cuantitativas no bastan: validación contra conocimiento de dominio es crítica
   - Clustering no supervisado requiere interpretación cuidadosa: correlación no implica causalidad
   - Balance entre complejidad técnica e interpretabilidad clínica: sofisticación debe servir aplicabilidad
   - Limitaciones inherentes al análisis: reconocimiento de sesgos y restricciones es esencial para honestidad científica
3. **Lecciones sobre datos reales:**
   - Datos de salud pública presentan desafíos únicos: sesgo de reporte, heterogeneidad temporal, variables faltantes
   - Muestreo estratificado esencial para balancear eficiencia computacional y representatividad
   - Reproducibilidad requiere documentación meticulosa: random_state, versiones de librerías, decisiones de preprocesamiento
   - Importancia de comunicación efectiva: resultados técnicos deben ser accesibles para stakeholders médicos
4. **Integración transdisciplinaria:**
   - Analítica avanzada efectiva requiere comprensión de contexto clínico y epidemiológico
   - Colaboración entre ciencia de datos y medicina es necesaria: validación de hallazgos contra experiencia clínica
   - Traducción bidireccional: convertir preguntas médicas en problemas analíticos y resultados cuantitativos en recomendaciones clínicas
   - Ética de datos de salud: responsabilidad de manejar información sensible de 58,497 pacientes (y potencialmente millones más en el dataset completo) con respeto y rigurosidad

**Conclusión final:**

Este proyecto demuestra que técnicas de clustering, cuando se aplican rigurosamente con validación múltiple y contexto de dominio, pueden extraer conocimiento accionable

de datos epidemiológicos complejos. Los 9 perfiles granulares identificados por K-Means y los 2 grupos principales identificados por Fuzzy C-Means no son solo categorías estadísticas, sino herramientas prácticas complementarias para mejorar triage hospitalario, asignar recursos eficientemente, y diseñar intervenciones dirigidas. La complementariedad de K-Means (granularidad) y Fuzzy C-Means (estructura simplificada con pertenencias parciales) ilustra que en analítica avanzada, múltiples perspectivas son preferibles a una sola "mejor"solución. Los resultados contribuyen tanto a la respuesta inmediata a COVID-19 como al conocimiento metodológico para análisis futuros de datos de salud pública. La identificación de perfiles específicos, como jóvenes con alta hospitalización a pesar de baja comorbilidad (Clusters 4 y 6), proporciona insights valiosos para políticas de salud pública más efectivas.

# 16. Trabajo a Futuro

Las siguientes líneas de investigación y desarrollo podrían ampliar y mejorar los resultados obtenidos:

## 16.1. Extensiones metodológicas

1. **Algoritmos adicionales de clustering:**
   - HDBSCAN (Hierarchical DBSCAN): para detectar clusters de densidad variable y outliers
   - Gaussian Mixture Models (GMM): para modelar clusters con distribuciones probabilísticas
   - Spectral Clustering: para capturar estructuras no lineales en datos
   - Self-Organizing Maps (SOM): para visualización de alta dimensionalidad
   - Evaluación: comparar contra K-Means y FCM, ver si revelan subgrupos adicionales
2. **Análisis temporal longitudinal:**
   - Clustering dinámico: estudiar evolución de pacientes entre clusters a lo largo de hospitalización
   - Time-series clustering: agrupar trayectorias de severidad completas, no solo snapshots
   - Hidden Markov Models: modelar probabilidades de transición entre estados de riesgo
   - Identificar patrones de progresión: bajo a moderado a alto, vs. deterioro súbito
   - Valor predictivo: usar cluster inicial para predecir cluster final y tiempo de transición
3. **Técnicas de deep learning:**
   - Autoencoders variacionales (VAE): aprender representaciones latentes no lineales

- Deep clustering: combinar aprendizaje de representación con clustering end-to-end
- Attention mechanisms: identificar variables más relevantes para cada subgrupo automáticamente
- Comparación: evaluar si complejidad adicional justifica ganancia en interpretabilidad

4. **Ensemble clustering:**
   - Combinar múltiples runs de K-Means con inicializaciones diferentes
   - Consensus clustering: agregar resultados de K-Means, FCM, GMM, Hierarchical
   - Cluster validation índices: usar métricas como consenso para seleccionar k óptimo
   - Robustez mejorada ante variaciones en datos o parámetros

## 16.2. Incorporación de datos adicionales

1. **Variables clínicas ampliadas:**
   - Laboratorios: D-dímero, ferritina, proteína C reactiva, procalcitonina
   - Signos vitales: frecuencia respiratoria, saturación O2, presión arterial
   - Imagenología: radiografías de tórax, TAC, scores de afectación pulmonar
   - Tratamientos recibidos: antivirales, corticosteroides, anticoagulantes
   - Potencial: refinar clusters con información clínica más granular
2. **Estado de vacunación:**
   - Número de dosis, tipo de vacuna, tiempo desde última dosis
   - Analizar si vacunación modifica pertenencia a clusters
   - Estudiar eficacia diferencial por perfil de riesgo
   - Identificar grupos que se benefician más de refuerzos
3. **Información genómica:**
   - Variantes virales: Alpha, Beta, Gamma, Delta, Omicron
   - Clustering estratificado por variante: ver si perfiles de riesgo cambian
   - Marcadores genéticos del huésped: HLA, polimorfismos asociados a severidad
   - Medicina de precisión: perfiles de riesgo personalizados
4. **Variables socioeconómicas y geográficas:**
   - Índice de marginación, acceso a servicios de salud, ocupación
   - Estado/municipio, ruralidad, altitud
   - Determinantes sociales de salud como factores de riesgo modificables

- Identificar disparidades geográficas en distribución de clusters

5. **Desenlaces a largo plazo:**
   - Secuelas post-COVID: fatiga crónica, problemas respiratorios, cognitivos
   - Rehospitalización, mortalidad a 90/180 días
   - Calidad de vida post-recuperación
   - Clustering basado en trayectorias completas, no solo fase aguda

## 16.3. Validación y generalización

1. **Validación externa:**
   - Aplicar modelos a datos de otros hospitales o estados
   - Comparar con poblaciones internacionales (España, Italia, EE.UU.)
   - Evaluar si 9 clusters (K-Means) y 2 clusters (Fuzzy C-Means) se replican o requieren ajuste por contexto y población
   - Calibración de umbrales de riesgo por población
2. **Validación prospectiva:**
   - Implementar sistema de clasificación en hospital piloto
   - Evaluar precisión de predicción de desenlaces en tiempo real
   - Medir impacto en decisiones clínicas y desenlaces de pacientes
   - Estudios de cohorte prospectiva para causalidad, no solo correlación
3. **Transferibilidad a otras enfermedades:**
   - Influenza, neumonía bacteriana, dengue, otras infecciones respiratorias
   - Evaluar si metodología y variables derivadas (INDICE_SEVERIDAD) son generalizables
   - Desarrollar pipeline configurable para diferentes patologías
   - Biblioteca de perfiles de riesgo para múltiples enfermedades

## 16.4. Desarrollo de herramientas aplicadas

1. **Sistema de soporte a decisiones clínicas (CDSS):**
   - Aplicación web: ingreso de 17 variables (después de preprocesamiento), retorna cluster K-Means (0-8) y grados de pertenencia FCM para los 2 clusters principales
   - Dashboard para hospitales: visualización de distribución de clusters en admisiones actuales
   - Alertas automáticas para pacientes fronterizos que requieren monitoreo especial
   - Integración con sistemas de historia clínica electrónica (EHR)
2. **API de clasificación en tiempo real:**

- RESTful API con modelos entrenados (.pkl) para clasificación instantánea
- Endpoint: recibe variables de paciente, retorna cluster, probabilidades, recomendaciones
- Escalable para procesar miles de clasificaciones simultáneas
- Documentación Swagger para adopción por desarrolladores de sistemas hospitalarios

3. **Dashboard interactivo para epidemiólogos:**
   - Visualización de distribución temporal de clusters: evolución de perfiles durante pandemia
   - Filtros por región, edad, sexo: identificar subgrupos vulnerables localmente
   - Comparación de métricas K-Means vs FCM en tiempo real
   - Exportación de reportes automatizados para tomadores de decisiones
4. **Modelo predictivo de progresión:**
   - Combinar clustering con aprendizaje supervisado: predecir desenlace (recuperación/muerte/complicaciones)
   - Usar cluster como feature en modelo de clasificación o regresión
   - Evaluar valor añadido del clustering vs variables individuales
   - Sistema de alerta temprana: predicción a 24/48/72 horas

## 16.5. Investigación avanzada

1. **Análisis de subgrupos dentro de clusters:**
   - Clustering jerárquico: dividir cada cluster en sub-clusters más específicos
   - Ejemplo: dentro de crítico, distinguir entre insuficiencia respiratoria vs multiorgánica
   - Perfiles ultra-finos para medicina personalizada extrema
2. **Modelado causal:**
   - Propensity score matching: controlar confusores para estimar efectos causales
   - Causal inference: identificar qué factores de riesgo son modificables vs marcadores
   - Simulación de intervenciones: ¿qué pasaría si todos los obesos bajaran a peso normal?
3. **Interpretabilidad avanzada:**
   - SHAP values: explicar contribución de cada variable a asignación de cluster
   - LIME: explicaciones locales para casos individuales complejos
   - Árboles de decisión interpretables que aproximen clustering para comunicación médica

4. **Fairness y equidad:**
    - Analizar si sistema de clustering introduce sesgos por sexo, edad, etnia
    - Métricas de equidad: tasas de error similares entre grupos demográficos
    - Debiasing: ajustar algoritmo para evitar discriminación algorítmica

**Priorización recomendada:**
Para maximizar impacto, se sugiere priorizar:

1. **Corto plazo (3-6 meses):** Validación externa en otros estados/hospitales, desarrollo de API básica de clasificación
2. **Mediano plazo (6-12 meses):** Incorporación de datos de vacunación y variantes, análisis temporal longitudinal, dashboard para epidemiólogos
3. **Largo plazo (1-2 años):** Modelo predictivo integrado, validación prospectiva en hospital piloto, transferibilidad a otras enfermedades, CDSS completo

## 16.6. Extensión del análisis

- Realizar análisis de series temporales para estudiar la evolución de los clusters a lo largo del tiempo.
- Incorporar variables adicionales como datos demográficos, económicos o de movilidad.
- Aplicar técnicas de análisis espacial para considerar la dimensión geográfica.
- Desarrollar modelos predictivos basados en los clusters identificados.

## 16.7. Implementación y despliegue

- Desarrollar una aplicación web interactiva para visualizar y explorar los clusters.
- Crear un dashboard en tiempo real que actualice los clusters con nuevos datos.
- Implementar un sistema de alertas basado en cambios en los patrones de clustering.
- Publicar los resultados en forma de artículo científico o reporte técnico.

## 16.8. Validación adicional

- Colaborar con expertos en epidemiología para validar los hallazgos.
- Realizar estudios de caso específicos para clusters de interés particular.
- Comparar los resultados con otros estudios similares en la literatura.
- Aplicar el mismo enfoque a datos de otras pandemias o enfermedades.

## 16.9. Consideraciones éticas y sociales

- Analizar las implicaciones éticas del uso de datos de salud pública.
- Considerar aspectos de privacidad y protección de datos personales.
- Evaluar posibles sesgos en los datos y en los algoritmos aplicados.
- Explorar cómo los resultados pueden contribuir a reducir desigualdades en salud.

# Referencias


1. Secretaría de Salud de México. (2024). *Datos Abiertos - Dirección General de Epidemiología*. Disponible en: `https://www.gob.mx/salud/documentos/datos-abiertos-152127`

2. World Health Organization. (2020). *Clinical management of COVID-19: interim guidance*. WHO/2019-nCoV/clinical/2020.5.

3. Guan, W. J., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18), 1708-1720.

4. Zhou, F., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229), 1054-1062.

5. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.

6. Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

7. Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3), 32-57.

8. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

9. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227.

10. Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, 3(1), 1-27.

11. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193-218.

12. Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 583-617.

13. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

14. Warner, J. L., et al. (2020). Python package for fuzzy c-means clustering: scikit-fuzzy. *GitHub repository*. `https://github.com/scikit-fuzzy/scikit-fuzzy`

15. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics, Springer-Verlag.

16. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29-39.

17. Williamson, E. J., et al. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821), 430-436.

18. Gao, Y., et al. (2021). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications*, 12(1), 5033.

19. Bertsimas, D., et al. (2020). COVID-19 mortality risk assessment: An international multi-center study. *PLoS ONE*, 15(12), e0243262.

20. Hernández-Garduño, E. (2020). Obesity is the comorbidity more strongly associated for COVID-19 in Mexico: A case-control study. *Obesity Research & Clinical Practice*, 14(4), 375-379.

21. Bello-Chavolla, O. Y., et al. (2020). Predicting mortality due to SARS-CoV-2: A mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. *The Journal of Clinical Endocrinology & Metabolism*, 105(8), 2752-2761.

22. Rosas-Peralta, M., et al. (2021). Comorbidities and severity of COVID-19 in Mexico: A national analysis. *Archives of Medical Research*, 52(4), 458-465.

23. Suárez, V., et al. (2020). Epidemiología de COVID-19 en México: del 27 de febrero al 30 de abril de 2020. *Revista Clínica Española*, 220(8), 463-471.

24. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 445, 51-56.

25. Harris, C. R., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

26. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.

27. Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

28. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.