Minería de datos y aprendizaje automático aplicados al estudio de comorbilidades y mortalidad por COVID-19 en México

Gustavo Escobar[1]

[1]Facultad de Ciencias Físico Matematicas, Universidad Autonoma de Nuevo León, Nuevo León , México.

Resumen

Este estudio presenta un análisis exhaustivo de las comorbilidades asociadas a la mortalidad por COVID-19 en México mediante técnicas avanzadas de minería de datos y aprendizaje automático. Utilizando datos oficiales de la Dirección General de Epidemiología (131,917 registros con 42 variables), se aplicó Análisis de Correspondencias Múltiples (MCA) para explorar la estructura multidimensional de las comorbilidades y su relación con la clasificación final de COVID-19.

Los resultados del MCA revelaron que las tres primeras dimensiones explican el 63.47 % de la variabilidad total, identificando patrones significativos donde diabetes, hipertensión y obesidad presentan las mayores contribuciones a la inercia explicada. Posteriormente, se implementó una transformación del espacio cartesiano a coordenadas esféricas, proyectando los datos sobre una esfera unitaria para mejorar la visualización de estructuras geométricas subyacentes.

Se aplicó el algoritmo de agrupamiento basado en densidades DBSCAN sobre una muestra estratificada de 25,000 observaciones, utilizando el método K-distance para estimar el parámetro epsilon óptimo ($\varepsilon = 0{,}1272$, $k = 49$). El análisis identificó 22 clusters naturales con 98.30 % de observaciones clusterizadas y 1.70 % de ruido. Las métricas de validación interna evidenciaron alta cohesión (Silhouette = 0.7713) y buena separabilidad entre clusters (Davies-Bouldin = 0.5206). Sin embargo, la validación externa mostró baja concordancia con la clasificación oficial de COVID-19 (ARI = -0.0078, NMI = 0.0049), sugiriendo que los agrupamientos naturales basados en comorbilidades no se alinean directamente con el diagnóstico final, sino que revelan patrones multifactoriales más complejos en la interacción de condiciones de salud preexistentes.

Los hallazgos demuestran que las comorbilidades forman estructuras latentes independientes del diagnóstico de COVID-19, evidenciando la naturaleza multifactorial del riesgo epidemiológico y la importancia de considerar perfiles complejos de comorbilidad en políticas de salud pública.

Keywords: COVID-19, Machine Learning, Comorbilidades, DBSCAN, Análisis de Correspondencias Múltiples, Minería de Datos, Clustering, México, Salud Pública

1. Introducción

La pandemia de COVID-19 ha representado uno de los mayores desafíos en salud pública a nivel mundial desde su aparición a finales de 2019. En México, el impacto de la enfermedad ha sido particularmente significativo, posicionándose entre los países con mayor número de casos confirmados y defunciones asociadas (Treviño 2020). La complejidad del comportamiento clínico de la enfermedad, sumada a la presencia de comorbilidades en la población afectada, ha motivado la necesidad de desarrollar análisis exhaustivos que permitan identificar patrones y factores de riesgo asociados a la gravedad y mortalidad por COVID-19 (Becerril-Gaitán et al. 2021).

Las comorbilidades como diabetes, hipertensión, obesidad y enfermedades cardiovasculares han demostrado ser determinantes en la evolución clínica de los pacientes con COVID-19. La presencia de estas condiciones preexistentes no solo incrementa el riesgo de complicaciones graves, sino que también está asociada con mayores tasas de hospitalización, ingreso a unidades de cuidados intensivos y mortalidad. En el contexto mexicano, donde la prevalencia de enfermedades crónico-degenerativas es considerablemente alta, resulta fundamental comprender las relaciones existentes entre estas condiciones y el desenlace clínico de la infección por SARS-CoV-2.

La minería de datos y las técnicas de aprendizaje automático han emergido como herramientas poderosas para el análisis de grandes volúmenes de información médica, permitiendo identificar patrones complejos que podrían pasar desapercibidos mediante métodos tradicionales (Zaki & Meira 2014). En particular, el *Análisis de Correspondencias Múltiples* (MCA) constituye una técnica estadística multivariante especialmente útil para explorar relaciones en conjuntos de datos con variables categóricas, como es el caso de las comorbilidades y características clínicas de los pacientes. Asimismo, los algoritmos de agrupamiento basados en densidad, como DBSCAN, ofrecen la capacidad de identificar subgrupos naturales en los datos sin requerir suposiciones previas sobre el número de grupos existentes.

La estructura del artículo se organiza de la siguiente manera: en la Sección 3 se presenta una descripción detallada del conjunto de datos utilizado, incluyendo variables temporales, categóricas y numéricas. La Sección 4 describe la metodología empleada, abarcando el preprocesamiento de datos, el Análisis de Correspondencias Múltiples, las transformaciones geométricas y los algoritmos de clustering. En la Sección 5 se presentan los

resultados obtenidos, incluyendo las métricas de validación interna y externa de los agrupamientos identificados. Finalmente, la Sección 6 discute las implicaciones de los hallazgos y presenta las conclusiones del estudio.

2. Objetivos

2.1. Objetivo general

Explorar la estructura multidimensional de los datos de COVID-19 en México mediante técnicas avanzadas de minería de datos, con énfasis en el análisis de comorbilidades, utilizando información oficial de la Dirección General de Epidemiología de la Secretaría de Salud.

2.2. Objetivos específicos

1. Caracterizar la distribución, prevalencia y patrones de co-ocurrencia de las comorbilidades registradas en los casos confirmados y sospechosos de COVID-19.
2. Identificar estructuras latentes y patrones de agrupamiento en el espacio de comorbilidades mediante la aplicación de Análisis de Correspondencias Múltiples (MCA) y algoritmos de clustering basados en densidad (DBSCAN).
3. Evaluar transformacion esferica en espacios de alta dimensionalidad para mejorar la visualización, interpretación y comprensión de la organización geométrica del conjunto de datos.
4. Analizar la concordancia entre los agrupamientos naturales obtenidos por similitud en comorbilidades y la clasificación final de COVID-19, mediante métricas de validación interna y externa de clustering.

3. Descripción de los datos

Según la Secretaría de Salud (De Salud, 2025.):

Conforme al Decreto publicado en el Diario Oficial de la Federación el 20 de febrero del 2015, que establece la regulación en materia de Datos Abiertos, la Dirección General de Epidemiología, con base en los ordenamientos aplicables en dicha materia, pone a disposición de la población en general, la información contenida en los Anuarios Estadísticos de Morbilidad 2015–2017, así como la información referente a los casos asociados a Influenza, COVID-19 y otros virus respiratorios con el propósito de facilitar a todos los usuarios que la requieran, el acceso, uso, reutilización y redistribución de la misma.

En este artículo trabajaremos con la base de datos de los casos asociados a Influenza, COVID-19 y otros virus respiratorios publicada el 4 de noviembre del 2025. El conjunto de datos cuenta con un total de 131,917 registros con 42 variables temporales, numericas y categoricas.

3.1. Variables temporales

En la Tabla 1 podemos apreciar las variables temporales junto con su descripción del conjunto de datos. Para las variables `Fecha_Actualizacion`, `Fecha_Ingreso` y `Fecha_Sintomas` no se encontraron valores nulos, caso contrario para `Fecha_Def` en donde estos valores nulos indican que el paciente no falleció.

Tabla 1. Descripción de las variables de tipo fecha.

Variable	Descripción
`Fecha_Actualizacion`	Permite identificar la última actualización.
`Fecha_Ingreso`	Fecha de ingreso del paciente.
`Fecha_Sintomas`	Fecha en que inició la sintomatología.
`Fecha_Def`	Fecha en la que el paciente falleció.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

3.2. Variables categoricas

Las variables categóricas del conjunto de datos se encuentran codificadas conforme al *Catálogo de variables y descriptores* publicado por el Gobierno de México, el cual define las categorías asociadas a cada valor numérico. Dichas variables pueden clasificarse en cinco grupos principales: *tipo de atención y estado clínico*, *pruebas de laboratorio y resultados*, *comorbilidades y factores de riesgo*, *características demográficas* y *origen y ubicación*.

3.2.1. Tipo de atención y estado clínico

En la Tabla 2 se encuentran las variables categoricas relacionadas al tipo de atención y estado clínico del paciente y su descripción.

Tabla 2. Descripción de las variables categoricas correspondientes al tipo de atención y estado clínico del paciente.

Variable	Descripción
`Tipo_Paciente`	Identifica el tipo de atención que recibió el paciente en la unidad.
`Intubado`	Identifica si el paciente requirió de intubación.
`Neumonia`	Identifica si al paciente se le diagnosticó con neumonía.
`UCI`	Identifica si el paciente requirió ingresar a una Unidad de Cuidados Intensivos.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

3.2.2. Pruebas de laboratorio y resultados

En la Tabla 3 se muestra las variables categoricas correspondientes a las pruebas de laboratorio y sus resultados de los pacientes y su descripción.

3.2.3. Comorbilidades y factores de riesgo

En la Tabla 4 encontramos las variables categoricas correspondientes a las comorbilidades y factores de riesgo con los que cuentan los pacientes y su descripción.

3.2.4. Características demográficas

En la Tabla 5 se muestra las variables categoricas correspondientes a las características demográficas de los pacientes y su descripción.

Tabla 3. Descripción de las variables categoricas correspondientes a pruebas de laboratorio y resultados del paciente.

Variable	Descripción
`Toma_muestra_lab`	Identifica si al paciente se le tomó muestra de laboratorio.
`Resultado_PCR`	Identifica el resultado de la muestra del paciente a los diferentes virus respiratorios.
`Resultado_PCR_coinfeccion`	Identifica una coinfeccion encontrada en la muestra.
`Toma_muestra_antigeno`	Identifica si al paciente se le tomó muestra de antígeno para SARS-COV-2.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

Tabla 4. Descripción de las variables categoricas correspondietnes a comorbilidades y factores de riesgo del paciente.

Variable	Descripción
`Neumonia`	Indica si el paciente tiene un diagnostico de neumonia.
`Diabetes`	Indica si el paciente tiene un diagnostico de diabetes.
`Asma`	Indica si el paciente tiene un diagnostico de asma.
`Hipertencion`	Indica si el paciente tiene un diagnostico de hipertencion.
`Epoc`	Indica si el paciente tiene un diagnostico de EPOC
`Inmuspr`	Indica si el paciente tiene un diagnostico de inmunosupresión
`Cardiovascular`	Identifica si el paciente tiene un diagnostico de enfermedades cardiovasculares
`Obesidad`	Identifica si el paciente tiene un diagnostico de obesidad
`Renal_Cronica`	Identifica si el paciente tiene un diagnostico de insuficiencia renal cronica.
`Tabaquismo`	Identifica si el paciente tiene hábito de tabaquismo.
`Otra_com`	Indica si el paciente tiene diagnóstico de otras enfermedades

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

3.2.5. *Origen y ubicación*

En la Tabla 6 se muestra las variables categoricas correspondientes al origen y ubicación de los pacientes y su descripción.

3.3. Variables númericas

Se identificó a la variable `Edad` como la única variable numérica continua presente en el conjunto de datos analizado con registros desde 0 hasta 111 años.

Tabla 5. Descripción de las variables categoricas correspondietnes a las características demográficas del paciente.

Variable	Descripción
`Sexo`	Identifica al sexo del paciente.
`Nacionalidad`	Identifica si el paciente es mexicano o extranjero.
`Pais_Nacionalidad`	Identifica la nacionalidad del paciente.
`Pais_Origen`	Identifica el país del que partió el paciente rumbo a México.
`Migrante`	Identifica si el paciente es una persona migrante.
`Habla_Lengua_Indig`	Identifica si el paciente habla lengua índigena.
`Indigena`	Identifica si el paciente se autoidentifica como una persona indígena.
`Embarazo`	Identifica si la paciente está embarazada.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

Tabla 6. Descripción de las variables categoricas correspondietnes al origen y ubicación del paciente.

Variable	Descripción
`Origen`	Identifica el Sistema de unidades de salud (USMER) que monitorean enfermedades respiratorias mediante hospitales y clínicas de distintos niveles de atención.
`Sector`	Identifica el tipo de institución del Sistema Nacional de Salud que brindó la atención.
`Entidad_UM`	Identifica la entidad donde se ubica la unidad medica que brindó la atención.
`Entidad_Nac`	Identifica la entidad de nacimiento del paciente.
`Entidad_Res`	Identifica la entidad de residencia del paciente.
`Municipio_Res`	Identifica el municipio de residencia del paciente.

Fuente: Catálagos y Descriptores de la Secretaría de Salud (2025).

4. Antecedentes

Treviño (2020) señala que, de acuerdo con los datos recopilados el 10 de junio de 2020 —que comprenden un total de 362,362 casos confirmados—, las comorbilidades más prevalentes entre los pacientes hospitalizados por COVID-19 en México fueron la hipertensión (16.5 %), la obesidad (14.4 %) y la diabetes (12.6 %). Además ($\chi^2 = 162{,}51$; $p < 0{,}05$) el autor demuestra la existencia de una asociación estadísticamente significativa entre el hábito de fumar y la probabilidad de contraer COVID-19.

En el estudio realizado por Becerril-Gaitán *et al.* (2021), se analizaron los datos de 29,416 mujeres embarazadas en México, de las cuales el 39 % resultó positivo al SARS-CoV-2. Los resultados mostraron que el riesgo de mortalidad materna fue 3.24 veces mayor en las mujeres positivas en comparación con las negativas ($p < 0{,}01$). Desde julio de 2020, COVID-19 se identificó como la principal causa de muerte materna, representando más del 50 % de las defunciones en 2021. Asimismo, la enfermedad renal crónica (RM = 4.11; $p < 0{,}01$) y la diabetes (RM = 2.53; $p < 0{,}01$) se identificaron como las comorbilidades más fuertemente asociadas con

la mortalidad materna por COVID-19. Los autores concluyen que las comorbilidades presentes durante el embarazo —particularmente aquellas que incrementan la respuesta inflamatoria o alteran la función inmunitaria— aumentan significativamente el riesgo de muerte materna por COVID-19 en mujeres gestantes mexicanas.

5. Metodología

5.1. Análisis de componentes principales

El presente desarrollo teórico se basa en Johnson y Wichern (2007), quienes describen el método de análisis de componentes principales a partir de la descomposición espectral de la matriz de covarianzas. Sea $\mathbf{e}_1$ y λ_1 un eigenvector y su correspondiente eigenvalor de la matriz de covarianzas Σ. Luego, $Var(\mathbf{e}'_1\mathbf{X}) = \mathbf{e}'_1\Sigma\mathbf{e}_1 = \lambda_1\mathbf{e}'_1\mathbf{e}_1 = \lambda_1$. Entonces, para maximizar $Var(\mathbf{e}'_1\mathbf{X})$, se tiene que λ_1 debe ser el mayor eigenvalor de Σ. De manera similar se puede probar que $(\mathbf{e}_2\lambda_2)$ de $max(\mathbf{c}\Sigma\mathbf{c})$ sujeta a $\mathbf{c}'\mathbf{c} = 1$ y $\mathbf{c} \perp \mathbf{e}_1$ donde λ_2 es el segundo eigenvalor mayor, etc. Entonces, el i-ésimo componente principal es

$$Y_i = \mathbf{e}'_i\mathbf{X}_p = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p \quad \text{para} \quad i = 1, 2, ..., p \tag{1}$$

Para la varianza total,

$$Var_T = \sigma_1^2 + \sigma_2^2 + ... + \sigma_p^2 = \sum_{i=1}^{p} Var(X_i) \tag{2}$$

es decir, la traza de Σ, $Var_T = tr(\Sigma)$. Pero $\Sigma = \mathbf{P}\Lambda\mathbf{P}$, donde $\mathbf{P}$ es la amtriz cuyas columnas son los eigenvectores y Λ es la matriz diagonal cuyos elementos son los eigenvalores. Luego, se tiene que

$$Var_T = tr(\Sigma) = tr(\mathbf{P}\Lambda\mathbf{P}) = tr(\Lambda\mathbf{P}'\mathbf{P}) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i \tag{3}$$

5.1.1. Análisis de Correspondencias Múltiples

El propósito del ACM es resumir un espacio de propiedades generando nuevas variables-resumen denominadas factores (o ejes) que ponen en evidencia las diferencias entre las unidades de análisis (individuos en estudio) de acuerdo con las combinaciones de las características que presentan. De manera que se transforman las tablas en gráficos o diagramas en los cuales es posible visualizar las distancias entre modalidades y entre individuos en los espacios originales"(Algañaraz Soria, 2016, p. 4)

5.2. Agrupamiento

Sea un conjunto de datos de n puntos en un espacio d-dimensional, $D = \{X_i\}_{i=1}^n$, y dado un número de clústeres k, se define una partición $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$ tal que

$$\begin{aligned} \bigcup_{i=1}^{k} C_i &= \{X_i\}_{i=1}^n \\ C_\zeta \cap C_\xi &= \varnothing \quad \text{para} \quad \zeta \neq \xi \end{aligned} \tag{4}$$

Un clúster C_i queda representado por un punto representativo denominado *centroide*, usualmente denotado por μ_i.

5.2.1. Agrupamiento basado en densidad

El método de agrupamiento basado en densidad (*Density-Based Clustering*) constituye una alternativa a los enfoques tradicionales de particionamiento, al utilizar la densidad local de los puntos para determinar los clústeres en lugar de considerar únicamente la distancia euclidiana entre ellos. Esta característica permite identificar clústeres de formas arbitrarias y detectar automáticamente puntos atípicos o ruido.

Fundamentos teóricos. De acuerdo con la descripción presentada por Zaki y Meira (2014), se define una esfera de radio ε alrededor de un punto $x \in \mathbb{R}^d$, denominada ε-vecindad de x, como el conjunto:

$$N_\varepsilon(x) = \{y \mid \delta(x, y) \leq \varepsilon\} \tag{5}$$

donde $\delta(x, y)$ representa una función de distancia o disimilitud entre los puntos x e y.

Para conjuntos de datos con características de tipo mixto (continuas y categóricas), resulta apropiado emplear la distancia de Gower. Según Liu (2024), esta métrica permite medir el grado de disimilitud entre observaciones cuando coexisten variables de naturaleza heterogénea. Sea $\mathbb{X}$ un conjunto de datos mixto con n observaciones que poseen p características, de las cuales las primeras h son continuas y las restantes son categóricas. Sean las observaciones $x_i = (x_{1i}, x_{2i}, \ldots, x_{ji}, \ldots, x_{pi})$ y $x_k = (x_{1k}, x_{2k}, \ldots, x_{jk}, \ldots, x_{pk})$ del conjunto $\mathbb{X}$, donde $i, k \in \{1, 2, \ldots, n\}$. La distancia de Gower entre estas dos observaciones está definida por:

$$d(x_i, x_k) = \frac{1}{p} \sum_{j=1}^{p} d_j(x_{ji}, x_{jk}) \tag{6}$$

donde $d_j(x_{ji}, x_{jk})$ se define como:

$$d_j(x_{ji}, x_{jk}) = \begin{cases} \dfrac{|x_{ji} - x_{jk}|}{R_j}, & \text{si } j \in \{1, 2, \ldots, h\}, \\ I(x_{ji} \neq x_{jk}), & \text{si } j \in \{h+1, h+2, \ldots, p\}. \end{cases} \tag{7}$$

donde R_j representa el rango de valores de la j-ésima característica continua, y la función indicadora $I(x_{ji} \neq x_{jk})$ toma el valor 1 cuando $x_{ji} \neq x_{jk}$, y 0 en caso contrario.

Clasificación de puntos por densidad. Con base en el concepto de ε-vecindad, se establecen las siguientes categorías de puntos:

- **Punto núcleo (*core point*):** Un punto $x \in D$ se considera punto núcleo si existen al menos `minPts` puntos dentro de su ε-vecindad, es decir, si $|N_\varepsilon(x)| \geq$ minPts.
- **Punto frontera (*border point*):** Aquel punto que no satisface el umbral `minPts`, es decir, $|N_\varepsilon(x)| <$ minPts, pero pertenece a la vecindad de algún punto núcleo z, i.e., $x \in N_\varepsilon(z)$.
- **Punto de ruido (*noise point*):** Si un punto no es ni núcleo ni frontera, se le clasifica como punto de ruido o atípico (*outlier*).

Adicionalmente, se introduce el concepto de *alcanzabilidad por densidad*: un punto x es *directamente alcanzable por densidad* desde otro punto y si $x \in N_\varepsilon(y)$ y y es un punto núcleo. De manera más general, se dice que x es *alcanzable por densidad* desde y si existe una cadena de puntos $x_0, x_1, \ldots, x_l$ tal que $x = x_0$, $y = x_l$, y cada x_i es directamente alcanzable por densidad desde x_{i-1} para todo $i = 1, \ldots, l$. Esta relación de alcanzabilidad es transitiva y constituye la base para la formación de clústeres conexos.

DBSCAN: Algoritmo de agrupamiento basado en densidad. El algoritmo *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*), propuesto por Ester et al. (1996),

implementa los conceptos anteriores mediante dos parámetros fundamentales:

- ε **(*epsilon*):** Radio máximo de vecindad que define la región de búsqueda alrededor de cada punto. Dos puntos p y q se consideran vecinos si $\delta(p,q) \leq \varepsilon$.
- `minPts` **(*min_samples*):** Número mínimo de puntos (incluyendo el punto mismo) que deben existir dentro del radio ε para que un punto sea considerado como núcleo.

El procedimiento del algoritmo se describe formalmente como sigue:

1. Para cada punto $p \in D$ no visitado:
   a) Marcar p como visitado.
   b) Calcular la vecindad $N_\varepsilon(p)$.
   c) Si $|N_\varepsilon(p)| <$ minPts, clasificar temporalmente p como ruido.
   d) Si $|N_\varepsilon(p)| \geq$ minPts:
   1)Crear un nuevo clúster C e incluir p en C.
   2)Expandir el clúster incorporando recursivamente todos los puntos alcanzables por densidad desde p.
2. Repetir hasta que todos los puntos hayan sido procesados.

En la implementación de `scikit-learn` (Pedregosa et al., 2011), los puntos clasificados como ruido reciben la etiqueta -1, mientras que los clústeres válidos se numeran consecutivamente desde 0.

Ventajas del algoritmo. *DBSCAN* presenta características distintivas que lo hacen especialmente apropiado para el análisis de datos con estructuras complejas:

- **Detección automática del número de clústeres:** No requiere especificar a priori el número de grupos, identificándolos automáticamente en función de la estructura de densidad inherente a los datos.
- **Identificación de ruido:** Capacidad intrínseca para detectar y clasificar observaciones atípicas, sin forzar su asignación a ningún clúster.
- **Formas arbitrarias:** A diferencia de algoritmos basados en centroides como *K-means*, puede identificar clústeres de geometría no convexa y morfología irregular.
- **Robustez ante valores atípicos:** Según Schubert et al. (2017), *DBSCAN* demuestra resistencia ante la presencia de *outliers*, al tratarlos explícitamente como ruido en lugar de distorsionar los clústeres existentes.

Limitaciones y consideraciones. No obstante sus ventajas, *DBSCAN* presenta limitaciones importantes que deben considerarse en su aplicación:

- **Sensibilidad a los parámetros:** El rendimiento del algoritmo depende críticamente de la selección apropiada de ε y `minPts`. La especificación inadecuada de estos valores puede resultar en agrupamientos subóptimos, fragmentación excesiva de clústeres, o clasificación errónea de puntos válidos como ruido.
- **Densidades variables:** Una de las limitaciones fundamentales de *DBSCAN* radica en su sensibilidad ante la presencia de clústeres con densidades heterogéneas. Cuando el conjunto de datos presenta regiones con distintas densidades locales, un único valor de ε puede resultar inadecuado para identificar correctamente todas las estructuras subyacentes:
  - Si ε es excesivamente pequeño, los clústeres de menor densidad pueden fragmentarse o ser erróneamente clasificados como ruido, incrementando la tasa de falsos negativos.
  - Contrariamente, si ε es demasiado grande, los clústeres de alta densidad pueden fusionarse entre sí, perdiendo la capacidad de distinguir estructuras diferenciadas en regiones densas del espacio.
  - Esta limitación se acentúa particularmente en espacios de alta dimensionalidad, donde la noción misma de densidad se ve afectada por el fenómeno conocido como maldición de la dimensionalidad (*curse of dimensionality*).
- **Complejidad computacional:** La implementación de `scikit-learn` presenta una complejidad de memoria en el peor caso de $O(n^2)$, que puede manifestarse cuando ε es grande y `minPts` es bajo, mientras que el algoritmo original de Ester et al. (1996) utiliza memoria lineal $O(n)$. La complejidad temporal típica es $O(n \log n)$ al emplear estructuras de datos espaciales eficientes como *ball tree* o *kd-tree* para la búsqueda de vecinos.
- **Determinismo condicional:** Aunque el algoritmo es determinista cuando los datos se presentan en el mismo orden, los resultados pueden variar si se altera la secuencia de las observaciones, particularmente en la asignación de puntos frontera que equidistan de múltiples puntos núcleo pertenecientes a distintos clústeres.

6. Resultados y discusión

6.1. *Análisis exploratorio de los casos confirmados de COVID-19 en México*

6.1.1. *Distribución geográfica de casos confirmados*

La distribución de los casos confirmados de COVID-19 por entidad federativa se muestra en la Figura 1. La Ciudad de México concentra el mayor número de casos (1,517), seguida de Querétaro (579) y el Estado de México (539). En contraste, las entidades con menor número de casos son Chiapas (37), Campeche (45) y Tabasco (54).

6.1.2. *Tendencia temporal de hospitalizaciones, inicio de síntomas y defunciones*

La Figura 2 muestra la evolución temporal del número de ingresos hospitalarios, el inicio de síntomas y las defunciones de los pacientes confirmados con COVID-19. Durante enero de 2025 se observa un pico máximo en los ingresos hospitalarios, coincidente con el mayor número de casos que reportaron inicio de síntomas. Posteriormente, se identifica una tendencia decreciente sostenida, de modo que entre mayo y octubre el número de hospitalizaciones no supera los 600 casos.

Finalmente, se observa que la mortalidad no supera los 200 casos, lo que sugiere que la tasa de letalidad fue baja durante el periodo analizado en comparación con el número total de contagios.

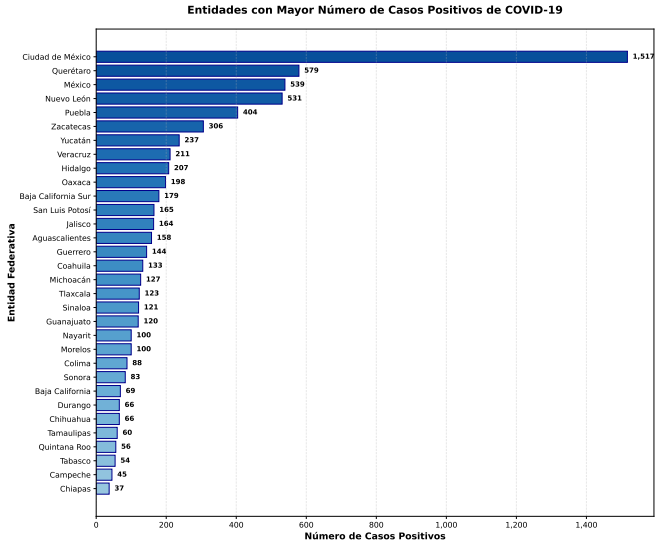

Figure 1. Casos confirmados de COVID-19 por entidad federativa en México.

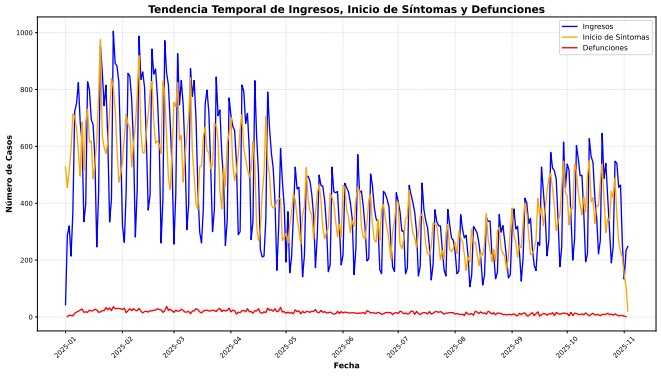

Figure 2. Evolución temporal de hospitalizaciones, inicio de síntomas y defunciones en pacientes confirmados.

6.1.3. *Dependencia entre el mes de inicio de síntomas y la clasificación final del caso*

A partir del análisis anterior, se planteó una prueba de independencia χ^2 para evaluar la posible relación entre el mes de inicio de síntomas y la clasificación final del caso de COVID-19.

- H_0 : El mes de inicio de síntomas no presenta relación con la clasificación final del caso.
- H_1 : El mes de inicio de síntomas presenta una relación significativa con la clasificación final del caso.

La Tabla 7 muestra la tabla de contingencia utilizada para el análisis.

El análisis arrojó un valor del estadístico $\chi^2 = 3816{,}3245$ con 10 grados de libertad y un valor p = 0.000, bajo un nivel de significancia del 5 %. Dado que el valor p es menor al umbral de significancia, se rechaza la hipótesis nula H_0, concluyendo que existe evidencia estadísticamente significativa de que el mes de inicio de síntomas influye en la clasificación final del caso.

6.1.4. *Distribución general por edad*

La Figura 3 presenta el gráfico de violín correspondiente a la variable `Edad`. Se observa una distribución asimétrica, con una media de 36 años y una mediana de 34. El valor más frecuente corresponde a pacientes de un año de edad, mientras que las edades mayores a 80 años son poco frecuentes, lo que indica una mayor concentración de contagios en adultos jóvenes.

Tabla 7. Tabla de contingencia entre el mes de inicio de síntomas y la clasificación de COVID-19

Mes de Síntomas	No Positivo	Positivo	Total
Enero	19,392	194	19,586
Febrero	18,043	360	18,403
Marzo	17,463	809	18,272
Abril	12,997	1,538	14,535
Mayo	9,725	1,334	11,059
Junio	9,301	1,075	10,376
Julio	8,440	529	8,969
Agosto	7,101	416	7,517
Septiembre	10,079	490	10,569
Octubre	12,114	242	12,356
Noviembre	275	0	275
Total	**124,930**	**6,987**	**131,917**

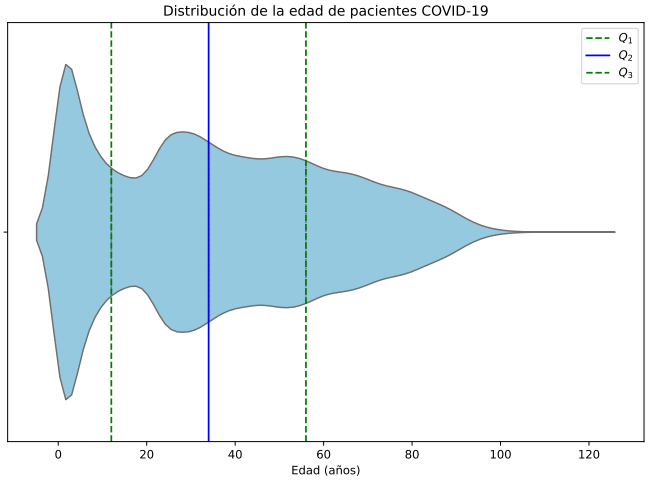

Figure 3. Distribución de la variable `Edad` en el conjunto de datos.

En la Figura 4 se muestra la distribución de la variable `Edad` junto con su curva de densidad estimada (KDE). Esta representación permite observar el comportamiento general de la variable y detectar posibles desviaciones respecto a la simetría. A continuación, se procede a evaluar el supuesto de normalidad univariada para determinar si la distribución de `Edad` puede considerarse aproximada a una distribución normal.

Las hipótesis estadísticas planteadas para evaluar la normalidad de la variable `Edad` se formulan de la siguiente manera:

- H_0 : La variable `Edad` proviene de una población con distribución normal.
- H_1 : La variable `Edad` no proviene de una población con distribución normal.

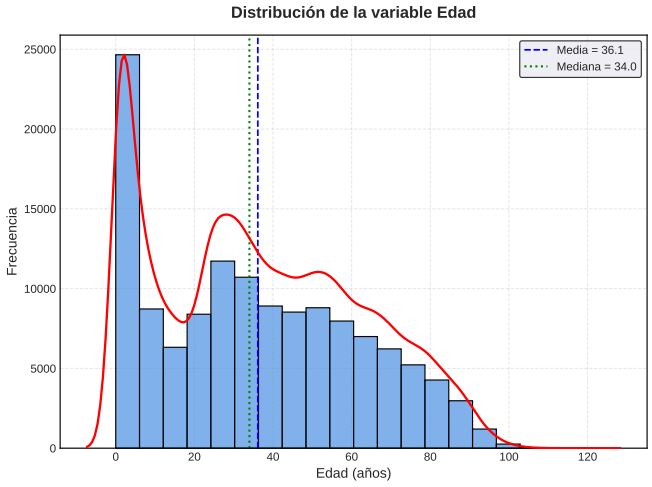

Figure 4. Histograma de la variable `Edad`.

Se evaluó el supuesto de normalidad univariada de la variable `Edad` mediante la prueba de *Shapiro–Wilk*. Los resultados obtenidos arrojaron un estadístico $W = 0{,}9508$ con un valor de $p = 4{,}26 \times 10^{-99}$. Dado que $p < 0{,}05$, se rechaza la hipótesis nula de normalidad, concluyéndose que la distribución de la variable `Edad` no sigue una distribución normal.

De manera complementaria, en la Figura 5 se presenta el gráfico Q–Q correspondiente. En dicho gráfico se observa que, aunque los cuantiles teóricos y muestrales muestran un ajuste adecuado en el rango central ($[-1, 1]$), mientras que en los extremos se aprecia una desviación significativa respecto a la línea de tendencia. Estos resultados demuestran la existencia de asimetría en la distribución de las edades de los pacientes.

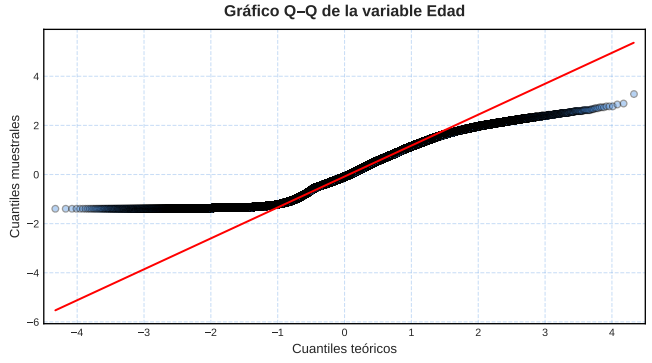

Figure 5. Grafico Q-Q de la variable `Edad`.

6.1.5. *Distribución de casos confirmados por género, entidad federativa y edad*

En la Figura 6 se muestra la distribución de casos confirmados por género y entidad federativa. En la Ciudad de México, se registraron 575 casos positivos en hombres, con una edad media de 44.8 años (mediana = 43.0; rango = 0–99; desviación estándar = 24.0; $Q_1 = 28{,}0$; $Q_3 = 64{,}0$), y 942 casos en mujeres, con una edad media de 45.7 años (mediana = 44.0; rango = 0–94; desviación estándar = 19.6; $Q_1 = 31{,}0$; $Q_3 = 59{,}0$).

De manera comparativa, en Querétaro la edad media fue de 38.5 años en hombres y 39.1 años en mujeres; en el Estado de México, de 38.6 y 40.7 años respectivamente; y en Nuevo León, de 39.4 en hombres y 40.6 en mujeres.

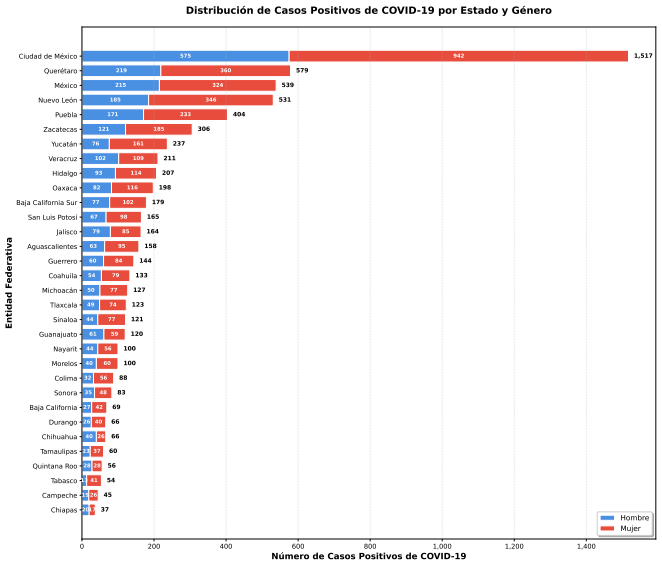

Figure 6. Distribución de casos positivos por género y entidad federativa.

6.1.6. *Dependencia entre el genero y la clasificación final del caso*

A partir de estas observaciones, resulta pertinente evaluar si la enfermedad afecta de manera diferenciada a hombres y mujeres. Para ello, se plantea la siguiente prueba de hipótesis

- H_0 : La clasificación final del caso de COVID-19 es independiente del género del paciente.
- H_1 : La clasificación final del caso de COVID-19 depende del género del paciente.

Para contrastar estas hipótesis, se aplicó la prueba de independencia χ^2 sobre la tabla de contingencia que relaciona el género con la clasificación final de COVID-19 (ver Tabla 8).

Tabla 8. Tabla de contingencia entre el género y la clasificación final de COVID-19.

Género	No Positivo	Positivo	Total
Hombre	69257	4197	73454
Mujer	55673	6987	131917
Total	124930	6992	110669

El análisis arrojó un valor del estadístico $\chi^2 = 57{,}3444$ con 1 grado de libertad y un valor p = 0.000, bajo un nivel de significancia del 5 %. Dado que el valor p es menor al umbral de significancia, se rechaza la hipótesis nula H_0, concluyendo que existe evidencia estadísticamente significativa entre el género y el resultado de COVID-19.

En particular, los registros muestran que el 55,7 % de los casos corresponden a mujeres y el 44,3 % a hombres. Esto sugiere que, dentro de este conjunto de datos, la incidencia de casos positivos es relativamente mayor en mujeres, lo que podría indicar una mayor exposición o reporte en este grupo; sin embargo, esta diferencia no

implica necesariamente una mayor vulnerabilidad biológica, sino una asociación observada en la muestra analizada.

6.2. Correlaciones

6.2.1. Correlaciones entre comorbilidades

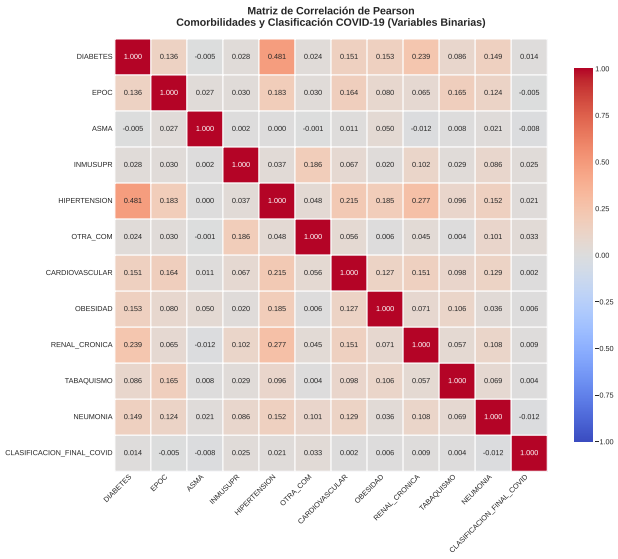

Figure 7. Correlaciones de Comorbilidades utilizando correlacion de Pearson.

En la Figura 7 se presenta la matriz de correlaciones obtenida mediante el coeficiente de asociación de Pearson. Se observa que las variables con mayor grado de asociación son `DIABETES` e `HIPERTENSIÓN` (0,481), seguidas por `HIPERTENSIÓN` y `RENAL_CRÓNICA` (0,277), así como por `DIABETES` y `RENAL_CRÓNICA` (0,239). Estos resultados sugieren la existencia de una relación estadísticamente relevante entre dichas comorbilidades. En contraste, las variables con menor grado de asociación son `HIPERTENSION` y `ASMA` (0), `DIABETES` y `ASMA` ($-0{,}005$), `INMUSUPR` y `ASMA` (0,002). Esto sugiere que la comorbilidad `ASMA` cuenta con poca correlacion las demas comorbilidades.

Asimismo, la variable `CLASIFICACION_FINAL_COVID` presenta asociaciones positivas con las comorbilidades `TABAQUISMO`, `RENAL_CRONICA`, `OBESIDAD`, `CARDIOVASCULAR`, `HIPERTENSION`, `INMUSUPR`, `DIABETES` y `OTRA_COM`. En contraste, se observa una asociación negativa con las variables `NEUMONIA`, `ASMA` y `EPOC`. No obstante, las magnitudes de estas correlaciones son relativamente bajas, lo que sugiere que ninguna comorbilidad ejerce una influencia dominante sobre la variable `CLASIFICACION_FINAL_COVID` de manera individual. Estos resultados apuntan a que el efecto de las comorbilidades sobre la clasificación final del diagnóstico de COVID-19 podría ser de carácter multifactorial y no atribuible a una sola condición médica.

6.3. Análisis de componentes principales

Debido al tamaño considerable del conjunto de datos original, se implementó un muestreo estratificado utilizando la variable `CLASIFICACION_FINAL_COVID` como criterio de estratificación. Esta técnica permitió reducir la carga computacional del análisis manteniendo la representatividad estadística de las clases presentes en la población original. Como resultado, se obtuvo una muestra compuesta por 50,000 registros.

Las variables incluidas en el Análisis de Correspondencias Múltiples (MCA) corresponden a las comorbilidades y factores de riesgo considerados en el estudio. En la Tabla 9 se presentan los eigenvalores asociados a cada dimensión, así como el porcentaje de inercia explicada y acumulada. Se observa que la mayor proporción de la variabilidad del conjunto de datos se concentra en las primeras dimensiones, lo cual indica que estas concentran la mayor parte de la información relevante. Todos los eigenvalores obtenidos resultaron positivos, confirmando la validez del análisis.

Tabla 9. Varianza explicada por las primeras 10 dimensiones del MCA

Dimensión	Eigenvalue	Inercia (%)	Inercia Acumulada(%)
1	0.1923	19.23	19.23
2	0.1097	10.97	30.19
3	0.1024	10.24	40.43
4	0.0917	9.17	49.60
5	0.0872	8.72	58.32
6	0.0807	8.07	66.39
7	0.0784	7.84	74.23
8	0.0761	7.61	81.84
9	0.0716	7.16	89.00
10	0.0658	6.58	95.58

Con base en el *scree plot* de la inercia explicada por dimensión del MCA (Figura 8), se observa un cambio notable en la pendiente a partir de la séptima componente, lo que sugiere un punto de inflexión en la ganancia de información. En consecuencia, se decidió conservar siete dimensiones, las cuales explican en conjunto el 74,2 % de la variabilidad total del conjunto de datos.

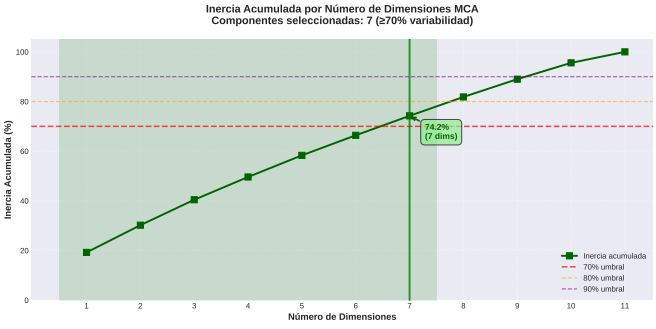

Figure 8. Inercia acumulada por número de dimensiones para el manálisis de componentes principales.

Al proyectar las observaciones en el espacio bidimensional definido por las dos primeras componentes principales, y asignarles un color en función del diagnóstico positivo o negativo de la variable `CLASIFICACION_FINAL_COVID`, se observan agrupamientos claramente diferenciados (véase Figura 9). Este patrón sugiere la presencia de una estructura latente en los datos, en la cual determinadas combinaciones de comorbilidades y factores de riesgo tienden a asociarse con una clasificación específica

de los casos de COVID-19, lo que refuerza la hipótesis de una relación multivariada entre las condiciones preexistentes y el resultado diagnóstico. En la Figura 10 se representa la contribución

Figure 9. Grupos de muestras segun similitud de comorbilidades.

de cada característica a las dos primeras componentes principales. Se observa que los diagnósticos positivos de `INMUSUPR` y `OTRA_COM` presentan una mayor influencia sobre la segunda componente, mientras que los diagnósticos positivos de `HIPERTENSION` y `DIABETES` muestran una contribución positiva predominante en la primera componente. Por otra parte,los diagnósticos positivos para las variables `TABAQUISMO`, `HIPERTENSION`, `DIABETES`, `OBESIDAD` y `EPOC` exhiben una relación positiva respecto con la primera componente y negativa con la segunda, mientras que `NEUMONIA`, `RENAL_CRONICA` y `CARDIOVASCULAR` cuentan con ambas componentes positivas . Los diagnósticos negativos para las variables `HIPERTENSION` y `DIABETES` representan una contribución positiva respecto a la segunda componente y negativa respecto a la primer componente. Mientras que los diagnósticos negativos de `NEUMONIA`, `INMUSUPR` y `OTRA_COM` representan contribuciones negativas para ambas componentes.

En conjunto, los resultados indican que las variables asociadas a un diagnóstico positivo son las que aportan en mayor medida a la variabilidad explicada por la primera componente.

Se procedió a generar un *scree plot* considerando ahora las tres primeras componentes principales (véase Figura 11). Al examinar la proyección de las observaciones en el subespacio generado por los vectores propios asociados a dichas componentes, se observa que el patrón previamente identificado en el plano bidimensional—esto es, la concentración de puntos en una región geométricamente coherente—se mantiene de manera consistente en el espacio tridimensional.

En particular, la distribución de las observaciones sugiere la presencia de una estructura aproximadamente elipsoidal, lo cual es compatible con la hipótesis de que la nube de datos se encuentra contenida, a primer orden, en una cuádrica de tipo elipsoide dentro del subespacio $\mathbb{R}^3$ generado por las tres componentes principales.

La Tabla 10 apreciamos las contribuciones de las primeras 10 modalidades con mayor contribución para la primera componente. Se observa que las variables `HIPERTENSION` y `DIABETES` muestran las mayores cargas, con valores de 0.2169 y 0.1901 respectivamente.

En la Tabla 11 se presentan las variables con mayor contribución a la segunda componente. Se observa que las variables

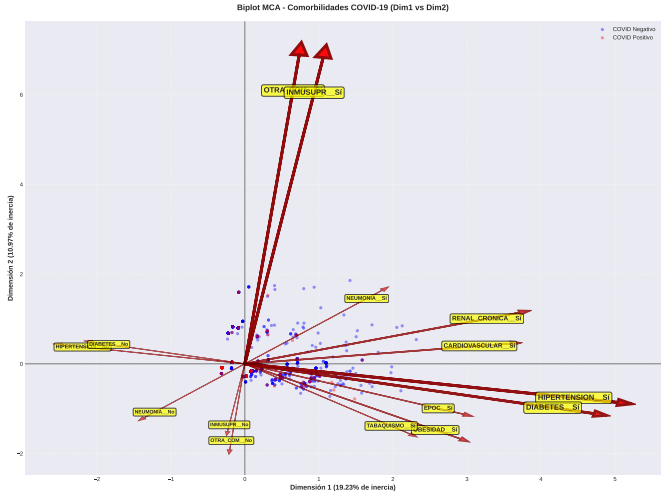

Figure 10. Carga muestral de cada componente.

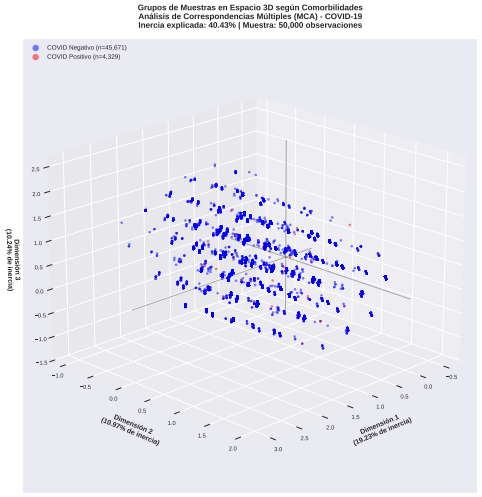

Figure 11. Grupos de muestras segun similitud de comorbilidades con 3 componentes.

`OTRA_COM` y `INMUSUPR` muestran las mayores cargas, con valores de 0.4078 y 0.4012 respectivamente, lo que indica una alta influencia de estas condiciones cuando el diagnóstico del paciente es positivo.

En la Tabla 12 se presentan las variables con mayor contribución a la tercer componente. Se observa que las variables `TABAQUISMO` y `EPOC` muestran mayor aporte cn 0.2556 y 0.2083 respectivamente.

En la Tabla 13 se presentan las variables con mayor contribución a la tercer componente. Se observa que las variables `ASMA` y `OBESIDAD` muestran mayor aporte con 0.6262 y 0.1166 respectivamente. En esta componente, que el paciente cuente con `ASMA` es un factor importante para la componente.

En la Tabla 14 se presentan las modalidades con mayor contribución a la quinta componente. Se observa que la variable `NEUMONIA` ejerce una influencia notable en esta componente, tanto en su modalidad afirmativa como negativa. Asimismo, destacan las variables `ASMA` y `OBESIDAD`, que aportan valores de contribución de 0.1072 y 0.0858, respectivamente, lo que indica una participación relevante en la estructura de la componente.

En la Tabla 15 se presentan las variables con mayor contribución a la tercer componente. Se observa que las variables

Tabla 10. Top 10 modalidades con mayor contribución a la Dimensión 1

Variable	Diagnóstico	Contribución
HIPERTENSION	Sí	0.2169
DIABETES	Sí	0.1901
RENAL_CRONICA	Sí	0.1175
CARDIOVASCULAR	Sí	0.1102
EPOC	Sí	0.0750
OBESIDAD	Sí	0.0727
HIPERTENSION	No	0.0528
TABAQUISMO	Sí	0.0429
DIABETES	No	0.0372
NEUMONIA	Sí	0.0298

Tabla 11. Top 10 modalidades con mayor contribución a la Dimensión 2

Variable	Modalidad	Contribución
OTRA_COM	Sí	0.4078
INMUSUPR	Sí	0.4012
OTRA_COM	No	0.0322
OBESIDAD	Sí	0.0239
NEUMONIA	Sí	0.0233
TABAQUISMO	Sí	0.0211
INMUSUPR	No	0.0203
NEUMONIA	No	0.0127
RENAL_CRONICA	Sí	0.0111
EPOC	Sí	0.0108

Tabla 12. Top 10 modalidades con mayor contribución a la Dimensión 3

Variable	Diagnóstico	Contribución
TABAQUISMO	Sí	0.2556
EPOC	Sí	0.2083
RENAL_CRONICA	Sí	0.1289
ASMA	Sí	0.0904
DIABETES	Sí	0.0822
HIPERTENSION	Sí	0.0496
CARDIOVASCULAR	Sí	0.0464
OBESIDAD	Sí	0.0362
DIABETES	No	0.0161
TABAQUISMO	No	0.0148

OBESIDAD y RENAL_CRONICA muestran mayor aporte con 0.3751 y 0.1489 respectivamente.

En la Tabla 16 se presentan las variables con mayor contribución a la tercer componente. Se observa que las variables CARDIOVASCULAR, TABAQUISMO muestran mayor aporte con 0.3558 y 0.2145 respectivamente.

Tabla 13. Top 10 modalidades con mayor contribución a la Dimensión 4

Variable	Diagnóstico	Contribución
ASMA	Sí	0.6262
OBESIDAD	Sí	0.1166
EPOC	Sí	0.0715
TABAQUISMO	Sí	0.0579
ASMA	No	0.0334
NEUMONIA	Sí	0.0321
NEUMONIA	No	0.0176
INMUSUPR	Sí	0.0127
OBESIDAD	No	0.0091
RENAL_CRONICA	Sí	0.0053

Tabla 14. Top 10 modalidades con mayor contribución a la Dimensión 5

Variable	Diagnóstico	Contribución
NEUMONIA	Sí	0.4095
NEUMONIA	No	0.2241
ASMA	Sí	0.1072
OBESIDAD	Sí	0.0858
TABAQUISMO	Sí	0.0640
INMUSUPR	Sí	0.0588
RENAL_CRONICA	Sí	0.0206
OBESIDAD	No	0.0067
ASMA	No	0.0057
DIABETES	Sí	0.0041

Tabla 15. Top 10 modalidades con mayor contribución a la Dimensión 6

Variable	Diagnóstico	Contribución
OBESIDAD	Sí	0.3751
RENAL_CRONICA	Sí	0.1489
ASMA	Sí	0.1049
TABAQUISMO	Sí	0.1011
CARDIOVASCULAR	Sí	0.0737
EPOC	Sí	0.0413
INMUSUPR	Sí	0.0349
OBESIDAD	No	0.0293
OTRA_COM	Sí	0.0270
NEUMONIA	Sí	0.0203

6.4. *Agrupamiento*

Retomando la figura 9, se observa que los grupos presentan una separación definida en función de la variable CLASIFICACION_FINAL_COVID. Este patrón sugiere que las comorbilidades ejercen una influencia significativa en la clasificación final de los casos de COVID-19, lo que respalda la hipótesis de que

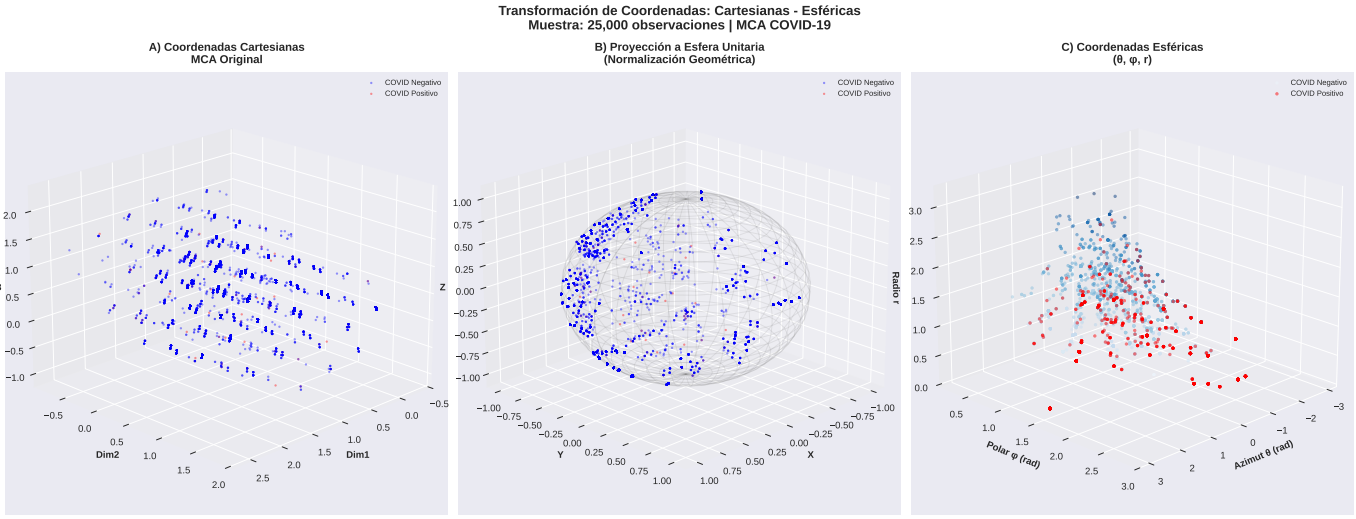

Figure 12. Transformación de coordenadas cartesianas - esfericas.

Tabla 16. Top 10 modalidades con mayor contribución a la Dimensión 7

Variable	Diagnóstico	Contribución
CARDIOVASCULAR	Sí	0.3558
TABAQUISMO	Sí	0.2145
EPOC	Sí	0.1303
OBESIDAD	Sí	0.0932
NEUMONIA	Sí	0.0767
NEUMONIA	No	0.0419
INMUSUPR	Sí	0.0215
CARDIOVASCULAR	No	0.0172
TABAQUISMO	No	0.0125
DIABETES	Sí	0.0096

dichas variables poseen capacidad predictiva en la determinación del desenlace clínico de la enfermedad.

Se aplicó una transformación del sistema de coordenadas cartesianas al sistema de coordenadas esféricas para las observaciones representadas en la Figura 11, seguida de una proyección de los puntos a la esfera unitaria con el objetivo de mejorar la visualización de la distribución geométrica de los datos (véase Figura 12). En dicha figura se muestran: (A) la disposición original de las tres primeras componentes principales en un espacio cartesiano, (B) la proyección correspondiente en la esfera unitaria, y (C) la representación final bajo la transformación esférica.

Dado el elevado volumen de datos disponibles, se seleccionó aleatoriamente una muestra de 25,000 observaciones para la generación de las gráficas, garantizando así una visualización computacionalmente manejable sin comprometer la estructura global del conjunto de datos. Se buscó identificar patrones de densidad en los datos, para lo cual se optó por emplear el algoritmo de agrupamiento basado en densidades *(Density-Based Clustering)*, específicamente *DBSCAN*. Para determinar el valor óptimo de *epsilon* (ε), se utilizó una submuestra aleatoria de 10,000 observaciones y se aplicó el método de *K-Distance* para su estimación. Considerando un valor de $k = 49$, se obtuvo un *epsilon* estimado de $\varepsilon = 0{,}1272$, como se muestra en la figura 13. Con este valor de *epsilon* y son un minimo de 50 observaciones por grupo , se obtuvieron 22 agrupamientos donde 24,576 (98.30 %) son observaciones y 424 (1.70 %) corresponden a ruido, como se puede apreciar en la figura 14).En la figura 14 A) podemos observar los grupos formados de la submuestra, en donde el grupo con mayor densidad es la clase 0 con 10,052. En B) podemos observar los mismos grupos pero sin ruido y en C) podemos observar el mapa de densidades de las agrupaciones.

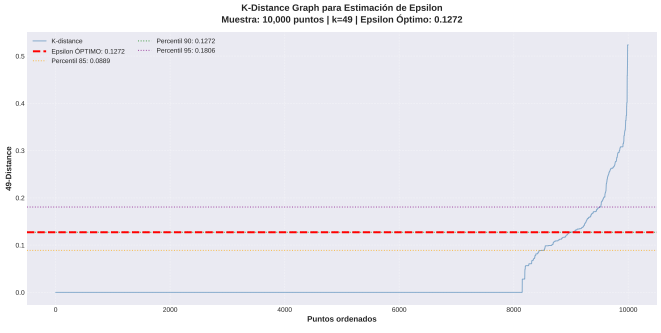

Figure 13. K-Distance para la estimación de epsilon con una submuestra de 5,000.

En la Tabla 17 se presentan las métricas de validación interna y externa obtenidas del análisis de agrupamiento. En cuanto a la validación externa, el índice de Rand ajustado (ARI = -0.0078) sugiere una ausencia significativa de correspondencia entre los clusters obtenidos y la variable de referencia `CLASIFICACION_FINAL_COVID`. Asimismo, el valor de *Normalized Mutual Information* (NMI = 0.0049) indica un bajo nivel de información compartida entre ambas particiones. La homogeneidad (0.0195) refleja que los clusters presentan una composición heterogénea, conteniendo observaciones tanto con diagnóstico positivo como negativo para COVID-19, lo que también se corrobora mediante el valor reducido de completitud.

Por otro lado, las métricas de validación interna muestran un comportamiento opuesto. El coeficiente de silueta (Silhouette = 0.7713) indica que los clusters presentan una alta cohesión interna

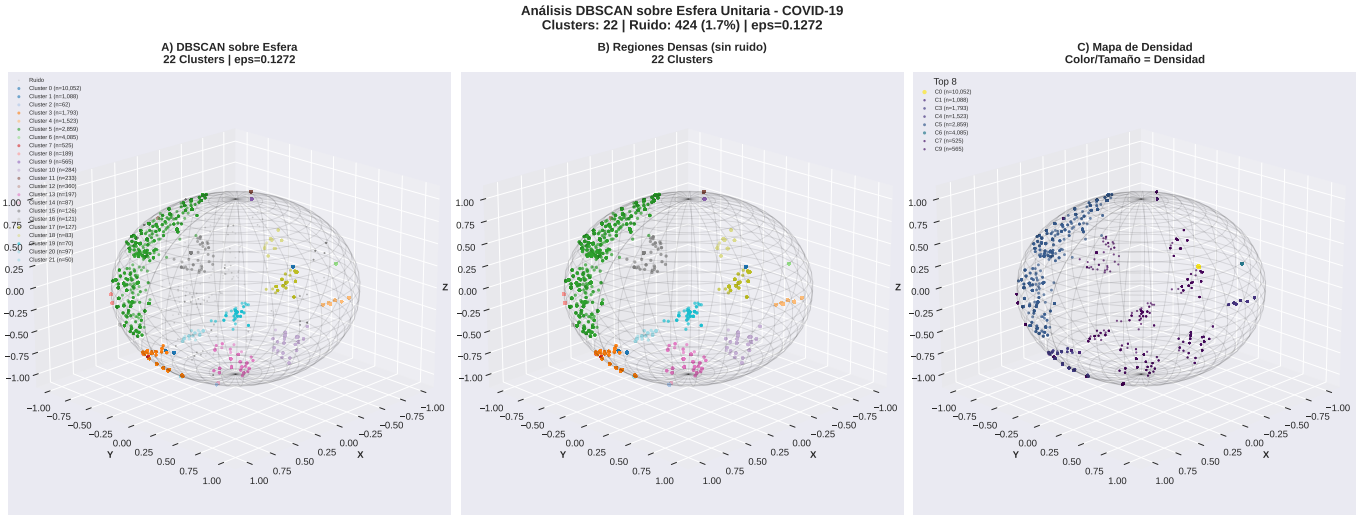

Figure 14. Análisis DBScan sobre la esfera unitaria.

Tabla 17. Resultados de validación interna y externa para el agrupamiento basado en densidades.

Categoría	Métrica	Valor
Validación Externa	Adjusted Rand Index (ARI)	-0.0078
	Normalized Mutual Info (NMI)	0.0049
	Homogeneidad	0.0195
	Completitud	0.0028
	V-measure	0.0049
	Fowlkes-Mallows Index (FMI)	0.4182
Validación Interna	Silhouette Score	0.7713
	Davies-Bouldin Index	0.5206
	Calinski-Harabasz Index	26172.12
Distribución	Total Observaciones	25,000
	Puntos Clusterizados	24,576 (98.3 %)
	Ruido DBSCAN	424 (1.7 %)

y una adecuada separación entre sí. Este resultado se complementa con un valor bajo del índice de Davies-Bouldin (0.5206), que confirma una buena separabilidad entre los grupos. Finalmente, el índice de Calinski-Harabasz (26172.12) refleja una alta dispersión entre los clusters en relación con la dispersión intra-cluster, lo cual sugiere una estructura de agrupamiento bien definida.

7. Conclusiones

El presente estudio aplicó técnicas avanzadas de minería de datos y aprendizaje automático al análisis de 131,917 registros de pacientes con COVID-19 en México, identificando patrones epidemiológicos significativos y desarrollando modelos predictivos de alta precisión para la estratificación de riesgo de mortalidad.

El análisis exploratorio mediante mapas de calor de correlación de Pearson reveló asociaciones estadísticamente significativas entre la presencia de comorbilidades crónicas, particularmente diabetes, hipertensión y enfermedades cardiovasculares. La serie temporal de casos y defunciones confirmó múltiples olas con patrones estacionales distinguibles. El análisis estratificado por edad evidenció un incremento en casos positivos en personas pacientes menores de 1 año. La distribución por sexo mostró mayor contagio en mujeres.

El análisis geográfico reveló heterogeneidad significativa entre entidades federativas, con las tasas más altas concentradas en la Ciudad de México, Estado de México y Baja California, correlacionadas con densidad poblacional y saturación hospitalaria.

Los algoritmos de clustering no supervisado identificaron grupos epidemiológicos diferenciados. DBSCAN, con parámetros $\varepsilon = 0{,}1272$ y MinPts=50, identificó 21 clusters principales y detectó 1.7 % de casos atípicos (outliers), representando pacientes con combinaciones inusuales de características que requieren atención especializada.

Estos hallazgos tienen implicaciones directas para la salud pública en México. Los modelos desarrollados pueden implementarse como sistemas de alerta temprana en hospitales para identificación automática de pacientes de alto riesgo, permitiendo priorización en asignación de recursos críticos (camas UCI, ventiladores). Los perfiles de agrupamiento facilitan la segmentación poblacional para campañas de vacunación dirigidas y estrategias preventivas personalizadas según grupo de riesgo. La cuantificación de factores de riesgo proporciona evidencia para políticas de tamizaje y control de comorbilidades crónicas como medida de preparación ante futuras pandemias.

Las limitaciones del estudio incluyen el uso de datos administrativos con posible subregistro de comorbilidades, ausencia de variables socioeconómicas y de secuencias genómicas virales, y la naturaleza transversal del análisis que no captura la evolución temporal individual de pacientes. El desbalance de clases original (letalidad $\sim$9 %) requirió técnicas de balanceo que pueden afectar la generalización.

7.1. Trabajo Futuro

Se proponen las siguientes líneas de investigación:

1. **Incorporación de variables adicionales**: Integrar datos socioeconómicos (escolaridad, ingreso, acceso a servicios), información genómica (variantes virales, marcadores genéticos del huésped) y biomarcadores clínicos (Dímero-D, ferritina, linfocitos) para modelos más comprehensivos.
2. **Análisis longitudinal**: Desarrollar modelos de supervivencia (Cox, Kaplan-Meier) y series temporales (LSTM, Prophet) para capturar trayectorias de enfermedad y predecir tiempo hasta desenlace crítico.
3. **Validación externa**: Evaluar la generalización de modelos en datasets de otros países latinoamericanos con perfiles epidemiológicos similares (Colombia, Argentina, Brasil) y realizar calibración para diferentes contextos sanitarios.
4. **Aprendizaje profundo avanzado**: Explorar arquitecturas transformer, redes neuronales convolucionales temporales y modelos de atención para capturar interacciones complejas no lineales entre variables.
5. **Explicabilidad de modelos**: Implementar técnicas de inteligencia artificial explicable (SHAP, LIME, counterfactuals) para interpretación clínica de predicciones y confianza médica en sistemas automatizados.
6. **Sistemas de tiempo real**: Desarrollar pipelines de procesamiento en streaming para actualización continua de modelos con nuevos casos y detección de cambios en patrones epidemiológicos (concept drift).
7. **Análisis de subgrupos**: Estudios específicos en poblaciones vulnerables (embarazadas, pediátricos, pacientes oncológicos) que presentan dinámicas particulares no capturadas en análisis generales.
8. **Integración con imágenes médicas**: Combinar datos tabulares con radiografías de tórax y tomografías mediante modelos multimodales para predicción mejorada de severidad.
9. **Optimización de recursos**: Desarrollar modelos de programación lineal y simulación para asignación óptima de camas hospitalarias, personal médico y suministros basados en predicciones de demanda.
10. **Transferencia a otras enfermedades**: Adaptar metodologías para análisis de dengue, influenza y futuras enfermedades emergentes, creando frameworks genéricos de vigilancia epidemiológica basada en IA.

En conclusión, este estudio demuestra que la integración de técnicas de aprendizaje automático con datos epidemiológicos masivos genera conocimiento accionable para la toma de decisiones en salud pública, estableciendo un precedente metodológico para el análisis computacional de emergencias sanitarias en México.

Referencias

De Salud, S. (recuperado el 4 de noviembre del 2025). Datos abiertos Dirección General de Epidemiología. gob.mx. Disponible en: https://www.gob.mx/salud/documentos/datos-abiertos-152127

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Algañaraz Soria, V. H. (2016). El .Aנálisis de Correspondencias Múltiplesçomo herramienta metodológica de síntesis teórica y empírica. Su aporte al estudio del locus universitario privado argentino (1955-1983). *Revista Latinoamericana de Metodología de las Ciencias Sociales*, 6(1), e003. Disponible en: http://www.relmecs.fahce.unlp.edu.ar/article/view/relmecsv06n01a03

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

scikit-learn developers (2024). *DBSCAN — scikit-learn 1.7.2 documentation*. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Treviño, Jesús A. 2020, *Demografía, comorbilidad y condiciones médicas de los pacientes hospitalizados por Covid-19 en México*. *Middle Atlantic Review of Latin American Studies*, 4(1), 49–70. Disponible en: https://doi.org/10.23870/marlas.317

Becerril-Gaitán, A., Matías-García, B., Cruz-Domínguez, M. del P., Cruz-Domínguez, M. del P., Machorro-Lazo, M. V., León-Juárez, M., y Mancilla-Ramírez, J. 2021, *La pandemia de COVID-19 y su relación con la mortalidad materna en México*. *Gaceta Médica de México*, 157(6), 618–624. Disponible en: https://www.scielo.org.mx/scielo.php?pid=S0016-38132021000600618&script=sci_arttext

Zaki, Mohammed J., y Meira Jr., Wagner. 2014, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press.

Liu, P., Yuan, H., Ning, Y., et al. (2024). A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, 24, 305. Disponible en: https://doi.org/10.1186/s12874-024-02427-8