Resumen

Este trabajo presenta un análisis descriptivo del comportamiento del COVID-19 durante las fiestas decembrinas en el Estado de México, considerando datos de los años 2020 a 2023. Se utilizaron técnicas de ciencia de datos y aprendizaje automático para analizar 36,427 registros provenientes del repositorio oficial de la Secretaría de Salud a través de su portal de datos abiertos.

Se realizó un preprocesamiento que incluye imputación de datos faltantes y uso de la codificación OHE para normalizar los valores de los atributos. Mediante el método del codo se determinó un número de clústeres óptimos y no óptimos para el algoritmo de agrupamiento K-means, los cuales fueron posteriormente evaluados con clasificadores como Árboles de Decisión (AD), Máquinas de Vectores de Soporte (SVM), Perceptrón Multicapa (MLP) y Naive Bayes (NB).

Los modelos AD y MLP alcanzaron un F1-score de 1.0 en ambos escenarios, los resultados para SVM demuestran una ligera variación en su F1-score para el escenario de seis grupos, mientras que NB mostró desempeño inferior en ambos casos. Para comprender la importancia de cada atributo en las decisiones de los clasificadores, se utilizó XAI con el método SHAP, destacando la edad, el género, presencia de neumonía y resultados de laboratorio como factores clave.

Los resultados permiten identificar perfiles de alto y bajo riesgo y a su vez apoyar a la toma de decisiones en políticas públicas de salud, así como la posibilidad de expandir este trabajo a otras regiones, periodos y enfermedades con comportamiento similar al COVID-19.

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

Facultad De Ingeniería

“Ciencia de Datos Aplicada al análisis de Comportamiento Del Covid-19 en el Estado de México en fiestas decembrinas”

Artículo Especializado para Publicar en Revista Indizada

Qué Para Obtener El Título De

Ingeniero En Computación

Presenta

Gustavo Alberto Díaz Rojas

Asesoras:

Dra. Rosa María Valdovinos Rosas

Dra. Angélica Guzmán Ponce

Toluca, Estado de México, Junio de 2025

Contenido

Pág.

Resumen

La pandemia de COVID-19 ha planteado desafíos globales sin precedentes, volviéndose particularmente crítica durante las temporadas invernales, cuando las condiciones climáticas pueden favorecer la propagación del virus. Este fenómeno fue especialmente notable en el Estado de México entre 2020 y 2023, un período marcado por un aumento significativo en la transmisión del SARS-CoV-2. La efectividad de los modelos predictivos desempeña un papel crucial en la gestión de la salud pública, ya que permite anticipar brotes y adaptar de manera efectiva las medidas preventivas. En este contexto, surge la necesidad de evaluar y mejorar las herramientas de clasificación que predicen los riesgos asociados con el COVID-19, adaptándolas a las complejidades impuestas por la diversidad de datos y la variabilidad estacional. Por esta razón, en este artículo se realiza un estudio descriptivo de los casos de COVID-19 presentados durante las fiestas decembrinas de los años 2020 al 2022, con la intención de conocer el comportamiento del virus utilizando técnicas de ciencia de datos. Los resultados de la inteligencia artificial explicable destacan la influencia de la edad, el género y las pruebas de laboratorio como las características que más influyen en los modelos.

Palabras clave: COVID-19, *K*-medias, Agrupamiento, XAI.

Abstract

The COVID-19 pandemic has posed unprecedented global challenges, becoming particularly critical during winter seasons when weather conditions can favor the spread of the virus. This phenomenon was especially notable in the State of Mexico between 2020 and 2023, a period marked by a significant increase in SARS-CoV-2 transmission. The effectiveness of predictive models plays a crucial role in public health management, as it enables the anticipation of outbreaks and the effective adaptation of preventive measures. In this context, the need arises to evaluate and improve classification tools that predict risks associated with COVID-19, tailoring them to the complexities imposed by data diversity and seasonal variability. For this reason, this article presents a descriptive study of COVID-19 cases reported during the holiday seasons from 2020 to 2022, with the aim of understanding the virus's behavior using data science techniques. The results of explainable artificial intelligence highlight the influence of age, gender, and laboratory tests as the main features which most influence the models.

Key words: COVID-19, K-medias, Predicción, Agrupamiento, XAI.

I. INTRODUCCIÓN

En diciembre de 2019, en Wuhan, provincia de Hubei, China, tuvieron lugar los primeros acontecimientos de la última catástrofe mundial en tema sanitario. Se habían diagnosticado los primeros casos del Síndrome Respiratorio Agudo Severo Coronavirus 2 denominado SARS-CoV-2 por su similitud con el SARS-CoV descubierto en 2003 y hoy conocido por la denominación otorgada por la Organización Mundial de la Salud, COVID-19 [1]. Actualmente, a poco más de 4 años del primer caso de COVID-19 reportado en China nivel mundial se tienen cifras de más de 776 millones de casos confirmados y más de 7 millones han muerto a causa de sus complicaciones[1].

En México, la Dirección General de Epidemiología se encarga de recopilar todos estos datos en un repositorio de acceso abierto, que hoy en día se ha convertido en una mina de datos de acceso abierto con millones de registros, entre los que se pueden realizar investigación descriptiva o predictiva [2].

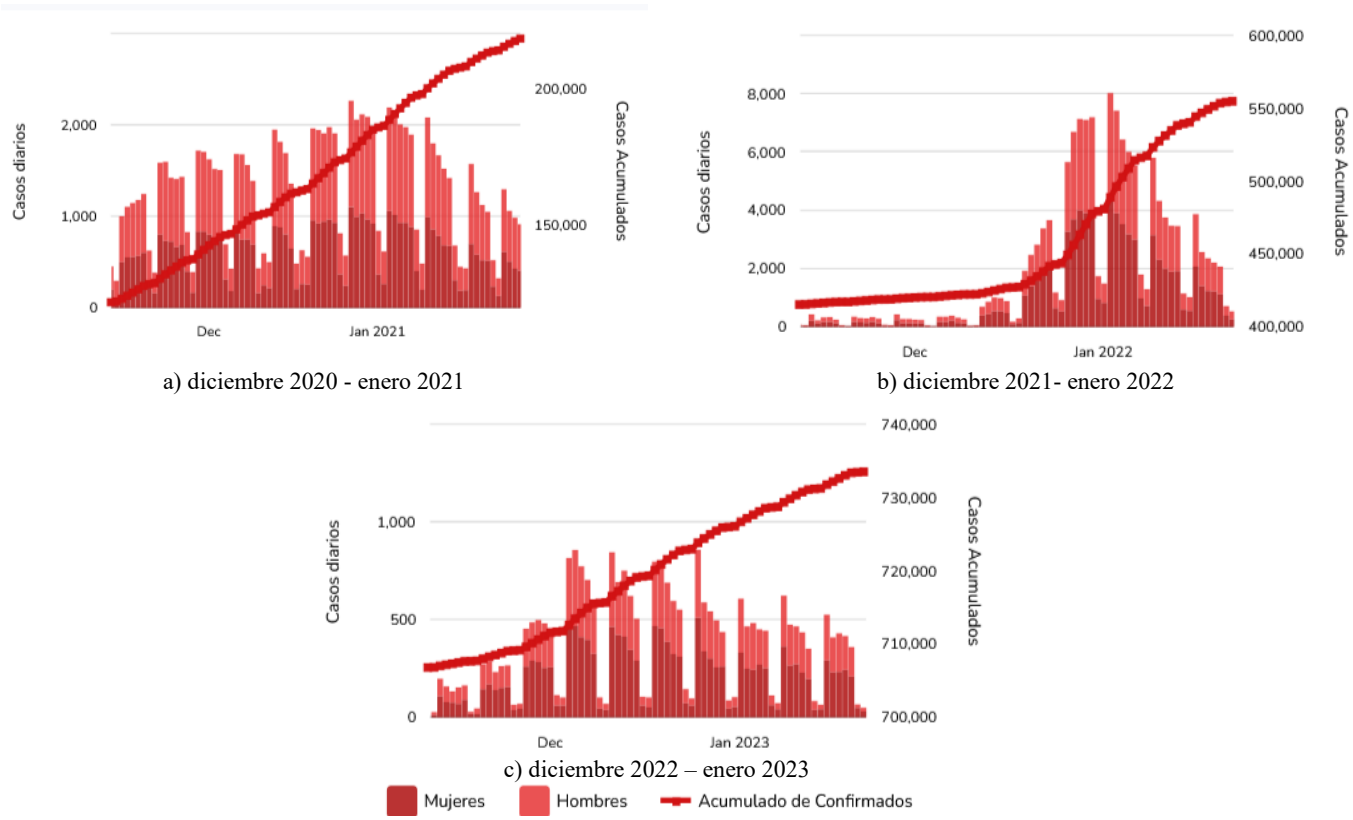
Para el 20 de octubre de 2024 en México las cifras señalan que han ocurrido más de 334 mil muertes y más de 7.6 millones de casos positivos se han registrado por esta enfermedad [3]. A lo largo de los más de tres años de pandemia, se observó una variación importante tanto en los casos nuevos, los hospitalizados y las defunciones. Afortunadamente, con la

[1] https://www.who.int/publications/m/item/covid-19-epidemiological-update-edition-173

administración de vacunas de emergencia iniciada en 2021, la cifra de muertes disminuyó considerablemente según datos del mapa de Johns Hopkins University & Medicine (https://coronavirus.jhu.edu/map.html).

Una estrategia promovida por los ministerios de salud en varios países es la integración de repositorios de acceso abierto puestos a disposición de la comunidad científica para su estudio, y de este modo llegar a conclusiones de forma más rápida. En estos repositorios, se recolectan los registros diarios de los pacientes hospitalizados, identificando las comorbilidades asociadas, los factores de riesgo y otros aspectos de la estancia hospitalaria, como estudios de laboratorio, rayos X, tomografías, entre otros.

Dentro de la estadística realizada por el gobierno mexicano, se pudo identificar con cierta precisión los periodos más críticos en los que las olas de contagio se incrementaban. Tal como se observa en las gráficas de la Fig. 1, estas olas de contagio cobraron la mayor cantidad de vidas en el periodo de diciembre 2020 a enero 2021, disminuyendo de forma considerable durante el 2021 a causa de las campañas de vacunación [3].



a) diciembre 2020 - enero 2021

b) diciembre 2021- enero 2022

c) diciembre 2022 – enero 2023

Fig. 1**.** Casos acumulados de COVID-19.

De lo antes mencionado, dos aspectos importantes se pueden rescatar: El primero se refiere al incremento sostenido de casos en el periodo de fiestas decembrinas (navidad, fin de año, día de reyes), principalmente por lo tradicionalista de la población mexicana y debido a que tanto el sector público, como el privado otorgan días de descanso a los trabajadores. En tanto, que el segundo a una disminución de la cantidad de casos gracias la distribución de vacunas.

Ahora bien, según datos oficiales, el Estado de México ocupa el segundo lugar en actividad de contagios acumulados (760,699) y defunciones (48,353) a causa del SARS-CoV-2 [3], lo que hace atractivo su estudio para comprender un poco más el comportamiento de la enfermedad. Adicional a lo anterior, y considerando el incremento histórico de enfermedades respiratorias durante el invierno (que coincide con las fiestas decembrinas), resulta interesante realizar un estudio descriptivo con el objetivo de comprender el comportamiento del COVID-19 en el Estado de México durante las vacaciones decembrinas

II. TRABAJOS RELACIONADOS

La pandemia del Covid-19 ha generado una gran cantidad de datos, así como una necesidad del desarrollo de nuevos modelos predictivos, para abordar y comprender la de la enfermedad, identificación de factores de riesgo, y optimización de estrategias de intervención a futuro.

Debido a esto, trabajos como el de Alyasseri et al. [4] presentan una investigación en la cual aborda la importancia de identificar los modelos de diagnósticos más precisos en el ámbito de DL (Deep Learning) y ML (Machine Learning). En su estudio, se identifican los modelos de Máquinas de Vectores de Soporte (SVM en sus siglas en inglés) y Redes Convolucionales como los más adecuados para el diagnóstico y predicción de brotes del COVID-19.

Por otro lado, en el trabajo de Kurniawan et al. [5], se implementan técnicas de agrupamiento como k-medias para la predicción de brotes del COVID-19, encontrando una fuerte correlación entre la mortalidad y los pacientes críticos.

En un enfoque similar, Nicholson et al. [6] analizan y categorizan a nivel nacional los condados de EE.UU. según su vulnerabilidad al COVID-19 utilizando métodos de ML y k-medias. Usando datos proporcionados por la universidad de John Hopkins y la oficina de censos de EE.UU., identificaron grupos de condados con características similares en términos de vulnerabilidad a la enfermedad y otros factores sociodemográficos.

De manera similar, Gohari et al. [7] realizan un análisis de agrupamiento de países bajo tasas de incidencia y mortalidad, utilizando datos de la OMS en 216 países. Los resultados, obtenidos mediante k-medias, identificaron 3 trayectorias bien definidas, destacando que México presentó una tendencia de incidencia y mortalidad "moderada", un dato relevante para el contexto de este estudio.

El trabajo de Pérez-Ortega et al. [8] aborda la problemática de la clusterización de mortalidad por COVID-19 a nivel municipal en México. Utilizando técnicas de Ciencia de Datos y datos de diversas dependencias oficiales del país, los autores encontraron que los clústeres con alta densidad poblacional pero baja pobreza muestra una mayor tasa de mortalidad en el país.

Finalmente, el trabajo de Escudero et al. [9] analiza la situación del COVID-19 y las implicaciones que tiene para México en el año 2021, utilizando estrategias médicas tales como el análisis de diagnóstico médico, identificando así los grupos de alto y bajo riesgo en México, siendo el grupo de alto riesgo conformado por adultos de 60+ años, además de incluir pacientes con comorbilidades relacionadas al sobrepeso.

III. ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Con el fin de analizar y prever el comportamiento de la COVID-19 durante el periodo de fiestas decembrinas en el Estado de México, se seleccionaron métodos de aprendizaje automático, análisis estadístico e inteligencia artificial explicable que permiten una interpretación detallada de los datos epidemiológicos estudiados.

A. Imputación de datos faltantes

Para abordar tratar los datos con información faltante, se implementó el algoritmo Hot Deck Vecino más cercano [10] para realizar la imputación de datos.

Algoritmo 1. Hot Deck Vecino más cercano

Entradas:
M = Conjunto de datos
Comienzo
Dividir el conjunto de datos considerando los registros con información completa y los que tienen datos faltantes
Para cada registro con información faltante
a. Identifica el vecino más cercano entre los datos con información completa
b. Reemplazar el valor faltante con el del vecino más cercano identificado
Fin

El algoritmo de Hot Deck vecino más cercano identifica el registro análogo al registro con dato faltante o que requiere corrección, utilizando como métrica de similaridad la distancia euclídea, optando por aquel registro que presente la menor distancia para reemplazar el valor ausente o incorrecto.

B. Agrupamiento

Con el fin de segmentar a la población en grupos homogéneos según características epidemiológicas y demográficas. En este estudio se utilizó el algoritmo de k-medias [11] (Algoritmo 2).

Algoritmo 1. Hot Deck Vecino más cercano

Entradas:
M = Conjunto de datos
Comienzo
Dividir el conjunto de datos considerando los registros con información completa y los que tienen datos faltantes
Para cada registro con información faltante
a. Identifica el vecino más cercano entre los datos con información completa
b. Reemplazar el valor faltante con el del vecino más cercano identificado
Fin

C. Clasificación

Los clasificadores utilizados en la experimentación son: el árbol de decisión (AD), la máquina de vector soporte (SVM), el perceptrón multicapa (MLP) y Naive Bayes (NB).

Árbol de decisión. Es un método que utiliza una estructura de árbol, donde cada nodo representa una característica del conjunto de datos y las aristas o enlaces a los nodos denotan las decisiones basadas en estos atributos, las hojas del árbol son las clases [12].

Algoritmo 3. Árbol de decisión

Entradas:
M = conjunto de datos
C = clase objetivo
A = conjunto de atributos

Salidas:
Árbol de decisión
Comienzo
Si todos los ejemplos de M pertenecen a C
Crear nodo hoja con etiqueta C
Si A esta vacío
Crear nodo hoja con el valor de C más común dentro de M
Caso contrario
Seleccionar el mejor atributo para dividir
Para cada atributo α dentro de A, calcular la ganancia de información (entropía) con respecto a M
Seleccionar el atributo α_{mejor} con la mayor ganancia de información
Crear un nodo de decisión
Crear un nodo para el atributo α_{mejor}
Dividir M en subconjuntos M_v (M_v = subconjunto de M donde A = v), uno para cada valor de v de α_{mejor}
Construir recursivamente los subárboles
Para cada valor v de α_{mejor}
Si S_v está vacío
Crear un nodo hoja con el valor de C más frecuente en M
Caso contrario
Llamar recursivamente al algoritmo usando M_vA-{α_{mejor}}, y C
Retornar árbol de decisión final
Fin

Máquina de Vector Soporte. Por otro lado, la máquina de vector soporte (SVM, por sus siglas en inglés) construye un hiperplano o un conjunto de hiperplanos en un espacio multidimensional, con el fin de separar las diferentes clases con un margen lo más grande posible entre ellas [13].

Algoritmo 4. Algoritmo de SVM

```
Entradas:
M = conjunto de datos
Y = etiquetas de clase
intI = interceptos iniciales
PI = pesos iniciales
Max_iter = máximas iteraciones
salidas:
pesos óptimos
intercepto optimo
representación gráfica (opcional)

Comienzo
   Inicializar los valores de PI e intI
   Repetir hasta el Max_iter o hasta alcanzar la convergencia
      Para cada clase repetir
         Calcular la predicción ajustada utilizando pesos, interceptos y características de entrada
         Determinar el error las predicciones con las etiquetas reales
      Evaluar la medida de optimización basada en los pesos actuales
      Para cada clase repetir:
         Si los pesos actuales minimizan la medida de optimización
            Actualizar los pesos óptimos y el intercepto optimo
   Retornar los pesos óptimos y el intercepto óptimo
Fin
```

Perceptrón Multicapa. Este algoritmo funciona procesando los datos de entrada a través de múltiples capas de neuronas interconectadas, cada neurona recibe una entrada, la pondera y aplica la función de activación para generar una salida, estas salidas se pasan a la siguiente capa y así sucesivamente hasta llegar a la capa de salida [14].

Algoritmo 5. Algoritmo de MLP

```
Entradas:
M = conjunto de datos
Tasa_aprendizaje = parámetro que determina la magnitud de ajustes de pesos
Max_iter = máximas iteraciones
Arquitectura_red = número de neuronas por capa
Función_activacion = función de activación (e.g. sigmoide)
U = umbral

Comienzo
   Inicializar la red
      Crear la estructura de la red según Arquitectura_red
      Asignar pesos iniciales aleatorios
      Establecer bias inicial para cada neurona
   Para cada iteración desde 1 hasta Max_iter
      Para cada par de entrada y salida dentro de M
         Propagación hacia adelante
            Establecer las entradas de la capa de entrada con los valores del patrón actual
            Para cada capa de la red
               Calcular el coste de activación de cada neurona
               Aplicar la Funcion_activacion para obtener la salida
         Cálculo de error
            Comparar el valor de salida con el valor esperado
         Propagación hacia atrás
            Calcular el gradiente del error con respecto a las salidas de las neuronas de la capa de salida
            Propagar el gradiente hacia atrás en cada una de las capas ocultas ajustando pesos y umbrales en función de Tasa_aprendizaje y los gradientes calculados
         Actualización de pesos y umbrales
            Ajustar pesos y umbrales de cada neurona en la dirección que minimice el error
         Fin de iteración
            Evaluar el rendimiento de la red en el conjunto de entrenamiento
            Si el error alcanza un valor de U aceptable
               Detener el entrenamiento
Fin
```

Naive Bayes. En su funcionamiento, este algoritmo calcula la probabilidad de que un dato pertenezca a cierta clase, considerando la independencia entre las características. En otras palabras, asume que la presencia de una característica en una clase no está relacionada con la presencia de otra característica [15].

Algoritmo 5. Algoritmo de NB

Entradas:
M = conjunto de datos
F = (f_1, f_2, f_3, …, f_n) valores de las variables predictoras del conjunto de datos de prueba

Salidas:
clase predicha para el conjunto de pruebas

Comienzo
  Leer el conjunto de datos de entrenamiento M
  Calcular la media y desviación estándar de las variables predictoras para cada clase en M
  Repetir:
    Calcular la probabilidad de f_i utilizando la ecuación de densidad gaussiana para cada clase
  Hasta que la probabilidad de todas las variables predictoras (f_1, f_2, f_3, …, f_n) hayan sido calculadas
  Calcular la verosimilitud para cada clase
  Obtener la clase con mayor verosimilitud
Fin

D. Inteligencia Artificial Explicable

Con el fin de ampliar las predicciones de clasificación para dilucidar la fiabilidad y el razonamiento que sostienen las predicciones de los modelos se utilizan técnicas de Inteligencia Artificial Explicable (XAI, por sus siglas en inglés). SHAP (*SHapley Additive exPlanation*) [16] es una técnica que distribuye de manera equitativa el resultado del modelo entre sus diversas características de entrada. Cada característica se conceptualiza como un participante en un juego colaborativo, y su impacto individual en la predicción final determina su importancia relativa. Este proceso implica calcular la contribución marginal de cada característica examinando todas las combinaciones posibles de características, lo cual permite entender mejor cómo cada una influye en el resultado del modelo.

IV. METODOLOGÍA

La metodología utilizada es la de CRISP-DM [17], todo el proceso de la metodología se presenta de manera gráfica en la fig. 3, a continuación, se especifica el proceso para cada uno de los pasos.

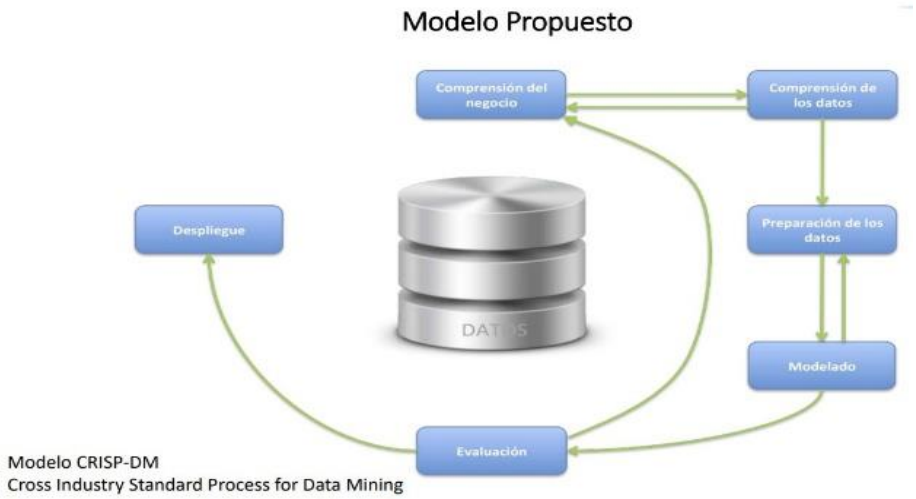


Fig. 2 Diagrama del proceso de la metodología CRISP-DM

E. Adquisición de datos

El registro de datos sobre el COVID-19 a nivel nacional se encuentra disponible en el repositorio nacional (https://www.gob.mx/salud/documentos/datos-abiertos-152127), este repositorio es mantenido y actualizado por la Dirección General de Epidemiología de la secretaria de salud desde el 2020 hasta la fecha.

Para fines experimentales en este artículo se consideró el periodo de diciembre a enero en los años 2020-2023 (36,427 registros), debido a que en estas fechas se presenta un aumento de casos positivos hospitalizados, respecto a los demás años.

Las características estudiadas son: sexo, edad, intubación, neumonía, embarazo, comorbilidades (diabetes, enfermedad pulmonar obstructiva crónica, asma, inmunosupresión, hipertensión, enfermedades cardiovasculares, obesidad y enfermedad renal crónica), tabaquismo, las pruebas diagnósticas, incluyendo los resultados de laboratorio, antígenos y si estuvo en cuidados intensivos.

F. Preprocesado de datos

Varias tareas de preprocesado se aplicaron a los datos a fin de mejorar su calidad tales como: Imputación de datos faltantes, Categorización de los datos y la codificación one-hot encoding, y de este modo, el aprendizaje por parte de los algoritmos utilizados.

Imputación de datos faltantes: Debido a la presencia de registros con datos faltantes, se aplicó el método de imputación Hot Deck Vecino más cercano (Tabla 1).

TABLA 1

CANTIDAD DE DATOS IMPUTADOS

Atributo	Datos imputados	Atributo	Datos imputados
Intubado	164 (0.45%)	Cardiovascular	41 (0.11%)
Neumonía	2008 (5.51%)	Inmunosuprimido	37 (0.1%)
Diabetes	56 (0.15%)	Insuficiencia renal	35 (0.1%)
EPOC	34 (0.09%)	Tabaquismo	45 (0.12%)
Obesidad	36 (0.1%)	Hipertensión	46 (0.13%)
Asma	37 (0.1%)	Otras comorbilidades	1899 (5.21%)
UCI	166 (0.46%)		

Categorización: Para las edades, se establecieron los siguientes rangos [18]: Grupo 1: individuos menores de 14 años. Grupo 2: individuos entre 15 y 47 años. Grupo 3, adultos de 48 a 63 años. Grupo 4, personas de 64 años o más.

Codificación de datos: Asimismo, para los atributos categóricos como estado de embarazo, resultados de laboratorio y resultados de antígenos, se realizó una codificación *one-hot encoding*. Esta técnica, genera una nueva columna para cada categoría posible dentro de las variables, asignando 0 o 1 que indican la presencia o ausencia de dicha categoría.

Como resultado de este proceso, se contó con un total de 15 características descriptivas. (SEXO, INTUBADO, NEUMONÍA, DIABETES, EPOC, ASMA, INMUNOSUPR, HIPERTENSIÓN, OTRA_COM, CARDIOVASCULAR, OBESIDAD, RENAL_CRÓNICA, TABAQUISMO, UCI, JUVENIL, ADULTOS_JOVENES, ADULTOS_MEDIA, MAYORES, EMBARAZO_SI, EMBARAZO_NO, EMBARAZO_NA, RESULTADO_LAB_POS, RESULTADO_LAB_NEG, RESULTADO_LAB_NoAdeq, RESULTADO_LAB_NA, RESULTADO_ANTIGENO_POS, RESULTADO_ANTIGENO_NEG, RESULTADO_ANTIGENO_)

G. *Agrupamiento*

Para determinar el valor óptimo de k se utilizó el método del codo variando el valor de k de 2-12. La Fig. 3 muestra los resultados del agrupamiento con los diferentes valores de k. El punto de inflexión marcado en rojo indica el número óptimo de grupos, en nuestro caso es 6.

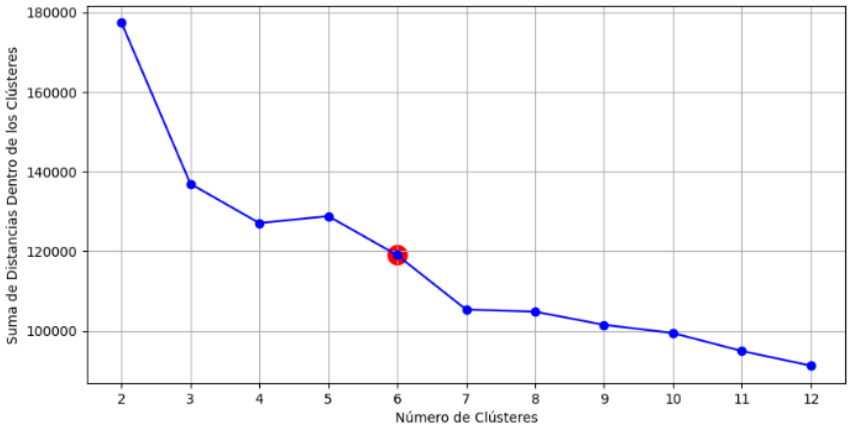


Fig. 3. Determinación del número óptimo de grupos.

H. *Métricas de evaluación*

Para analizar la calidad de agrupamiento se emplearon índices de validación, los cuales miden tanto la cohesión como la separación entre los clústeres.

Dado C_e el centroide del conjunto de datos, K número de clústeres y con M muestras, el índice Davies-Bouldin [19] mide la dispersión y la separación de los clústeres.

$$DB(C_e) = \frac{1}{K} \sum_{c_k=1} \max_{c_l \in C_e \setminus c_k} \left(\frac{\Delta(X_k + \Delta(X_l))}{\delta(X_k, X_l)} \right) \qquad (1)$$

Donde:

ΔX_k = es la distancia promedio intra-clúster

$\delta(X_k, X_l)$ = es la distancia promedio entre los clústeres X_k yX_l

Por otro lado, el índice Calinski-Harabasz considera la suma de los cuadrados entre grupos (BSS) y la suma de los cuadrados intra-grupo (WSS):

$$CH(C_e) = \frac{M-K}{K-1} \frac{\sum_{c_k \in C_e} |c_k| \, d_E(\overline{c_k}, \bar{X})}{\sum_{c_k \in C_e} \sum_{x_i \in c_k} d_E(x_i, \overline{c_k})} \qquad (2)$$

Donde:

$|c_k|$ = Número de observaciones en el clúster k, C_k = Centroide del clúster k, $\bar{X}$ = Centroide de todos los puntos en el conjunto de datos, d_E = Computa la distancia entre dos puntos

B. Rendimiento de los modelos

Con la ayuda de la matriz de confusión se analizó la efectividad del clasificador. Se trata de una matriz cuadrada de $n \times n$ donde cada entrada N_{ij} con $i \neq j$ representa el número de instancias que fueron clasificadas como pertenecientes a la clase g_j, pero que en realidad pertenecen a la clase g_i [20].

TABLA 1

MATRIZ DE CONFUSIÓN PARA MÁS DE DOS CLASES

		Predichas			
		g_1	g_2	⋯	g_n
Reales	g_1	N_{11}	N_{12}	⋯	N_{1n}
	g_2	N_{21}	N_{22}	⋯	N_{2n}
	⋯	⋯	⋯	⋯	⋯
	g_n	N_{n1}	N_{n2}	⋯	N_{nn}

A partir de la matriz de confusión de la Tabla 2, se pueden obtener las siguientes métricas:

***Recall* o Sensibilidad.** Mide el número de instancias positivas correctamente clasificadas como clase g (Ec. 1).

$$Recall_g = \sum_{i=1}^{n} \frac{N_{ii}}{\sum_{j=1}^{n} N_{ij}} \qquad (1)$$

Precisión. Es una medida del grado de acuerdo entre las clases asociadas a los datos positivos con los establecidos como tales por el clasificador (Ec. 2).

$$Precision_g = \sum_{i=1}^{n} \frac{N_{ii}}{\sum_{j=1}^{n} N_{ji}} \qquad (2)$$

F1-score (Ec. 3). Promedio armónico de ambas métricas [21].

$$F1 - score_g = 2 * \frac{Precision_g \times Recall_g}{Precision_g + Recall_g} \qquad (3)$$

Por lo que, para un problema multiclase la métrica global es la media de las F1-score por clase Ec. 4.

$$F1_{macro} = \frac{1}{|G|} \sum_{g=1}^{|G|} F1 - score_g \qquad (4)$$

I. Implementación

Para la implementación de esta investigación, se utilizará la plataforma GitHub como repositorio central, donde se almacenarán todos los elementos esenciales del proyecto. Esto incluirá el conjunto de datos utilizados, los algoritmos empleados y una documentación detallada sobre la configuración y parametrización de cada modelo.

El propósito de esta implementación es garantizar la reproducibilidad del experimento, permitiendo que futuros investigadores o profesionales puedan replicar y adaptar el análisis a la misma enfermedad o a diferentes patologías. Para ello, el repositorio contará con el conjunto de datos, códigos de los algoritmos utilizados, guía de uso para la configuración de los parámetros utilizados en los algoritmos.

V. RESULTADOS

En la experimentación se utilizó validación cruzada con 10 repeticiones, considerando 80% de los datos para entrenamiento y el 20% restante para las pruebas. Los modelos analizados son Árboles de decisión (AD), Máquinas de vector soporte (SVM), Naive Bayes (NB), Perceptrón Multicapa (MLP), considerando los parámetros predeterminados de la biblioteca scikit-learn[2] de Python, excepto la semilla de los valores aleatorio, la cual se estableció en 42.

J. Resultados del agrupamiento

Se compararon dos esquemas de agrupamiento, k=2 y k=6, cuya disposición de los centros se muestra en las Figuras 3 y 4 se muestra. *True* indica la presencia de una característica específica y el valor *False*, su ausencia.

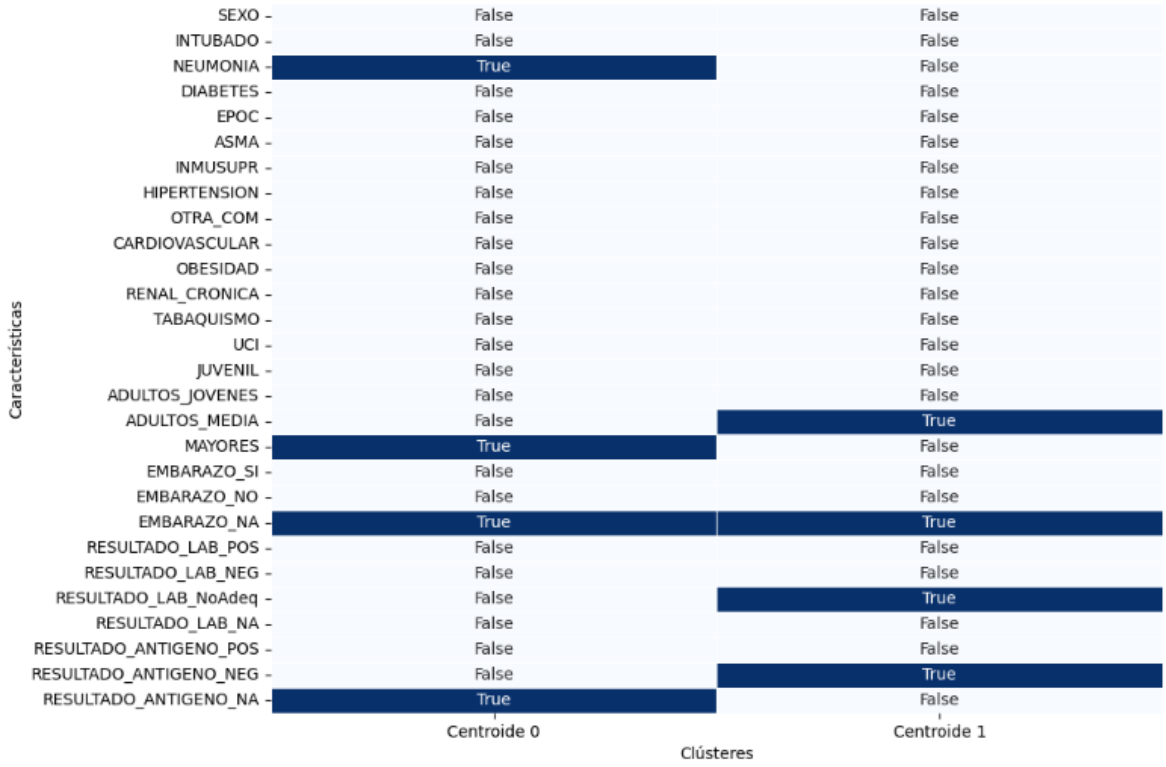


Fig. 4 Centroides para dos grupos

De la Fig. 4 se observan que el clúster 0 se caracteriza por la presencia de neumonía, personas mayores de edad y embarazo no aplicable, esto último sugiere que son hombres. En tanto que el resultado de antígeno no aplicable sugiere que se trata de personas mayores de edad con neumonía, lo que sugiere una asociación específica de esta condición con hombres de este grupo etario.

Por otro lado, el clúster 1 destaca por incluir principalmente a adultos de mediana edad. En este grupo, la falta de toma de muestras de laboratorio y un resultado negativo en pruebas de antígeno sugieren que estos individuos probablemente no padecen COVID-19. La composición de estos clústeres sugiere patrones epidemiológicos diferenciados durante los periodos invernales en el Estado de México, posiblemente influenciados por factores demográficos y de salud específicos de cada grupo.

En la Fig. 5 presenta los centroides correspondientes a seis grupos distintos. En el clúster 0, el sexo *True* sugiere que son mujeres quienes han presentado neumonía, no están embarazadas y no se les ha realizado una prueba de antígeno. Por otro lado, el clúster 1 agrupa mujeres adultas de mediana edad con resultados negativos en las pruebas de laboratorio, y sin prueba de antígeno aplicada.

[2] https://scikit-learn.org/stable/index.html

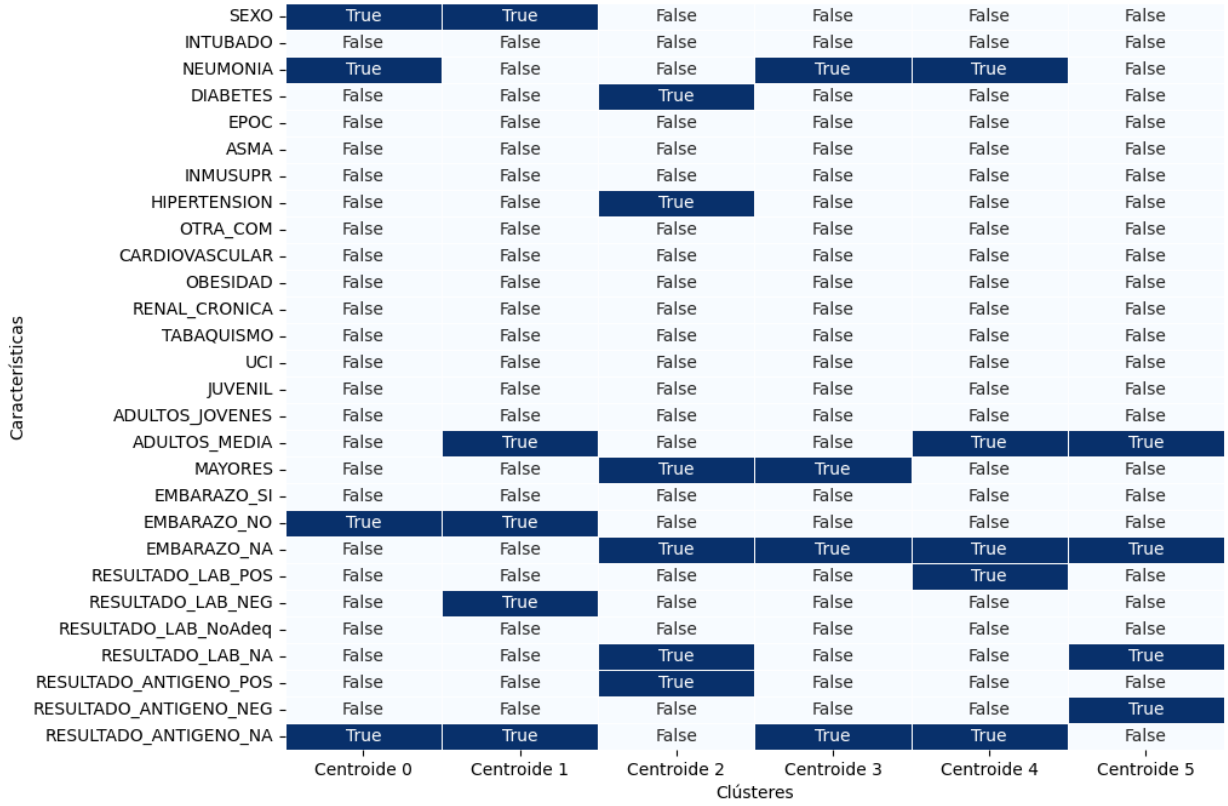


Fig. 5. Centroides para 6 grupos

El clúster 3 parece representar a hombres mayores con neumonía, mientras que el clúster 4, al estar asociado a resultados positivos en las pruebas de laboratorio, sugiere referirse a hombres de mediana edad con COVID-19 confirmado.

Finalmente, el clúster 5 se caracteriza por hombres de mediana edad con resultados negativos en la prueba de antígeno. En conjunto, estos grupos se diferencian en gran medida según las variables de sexo y edad, que son factores clave en la segmentación realizada por *K*-medias.

La Tabla 3 compara los índices de validación obtenidos mediante los métodos de Davies-Bouldin y Calinski-Harabasz para evaluar la calidad de la agrupación en las dos configuraciones de *K*-medias.

TABLA 2

ÍNDICES DE VALIDACIÓN INTERNOS

	Davies Bouldin	**Calinski-Harabasz**
2 grupos	3.3078	3269.1326
6 grupos	2.6426	3873.7733

De la Tabla 3, el índice Davies-Bouldin mide la dispersión y la separación de los clústeres, donde valores más bajos indican una mejor separación entre grupos. En tanto que el índice Calinski-Harabasz mide la relación entre la dispersión intra-clúster e inter-clúster, valores altos indican una mejor calidad de agrupación. Ambos índices sugieren que la estructura de seis clústeres logra una mejor partición del conjunto de datos.

K. Clasificación

La Tabla 4 muestra los F1-scores obtenidos con los modelos utilizados al ser entrenados con el conjunto de datos agrupado

TABLA 3

F1-SCORE OBTENIDO POR CLASIFICADOR

	AD	**SVM**	**MLP**	**NB**
2 grupos	1.0	1.0	1.0	0.9033
6 grupos	1.0	0.9999	1.0	0.7398

Para los resultados con dos grupos los modelos AD, SVM y MLP han alcanzado un F1-score de 1.0. Esto sugiere que, para este caso particular, los modelos son efectivos en separar las dos clases, probablemente porque las características que diferencian a las dos clases son muy distintivas y bien definidas. Mientras que el modelo NB tiene un F1-score de 0.9033, que, aunque es alto, es notablemente inferior en comparación con los otros clasificadores. Esto puede indicar que las suposiciones de independencia de características de NB no se sostienen en este contexto, afectando su capacidad para distinguir entre las dos clases tan efectivamente como los otros modelos.

Por parte de los seis grupos, los modelos AD y MLP mantienen un F1-score de 1.0, indicando que estos modelos manejan muy bien la complejidad añadida de tener más clases. Esto sugiere que ambos modelos pueden captar y modelar la complejidad y las relaciones no lineales entre las características. Mientras que el SVM tiene una ligera disminución en el F1-score a 0.9999. Aunque sigue siendo alto, esta pequeña reducción podría ser resultado de la dificultad de establecer márgenes claros de separación en un espacio con mayor número de clases, o podría ser debido los hiperparámetros que no están optimizados para el escenario de múltiples clases.

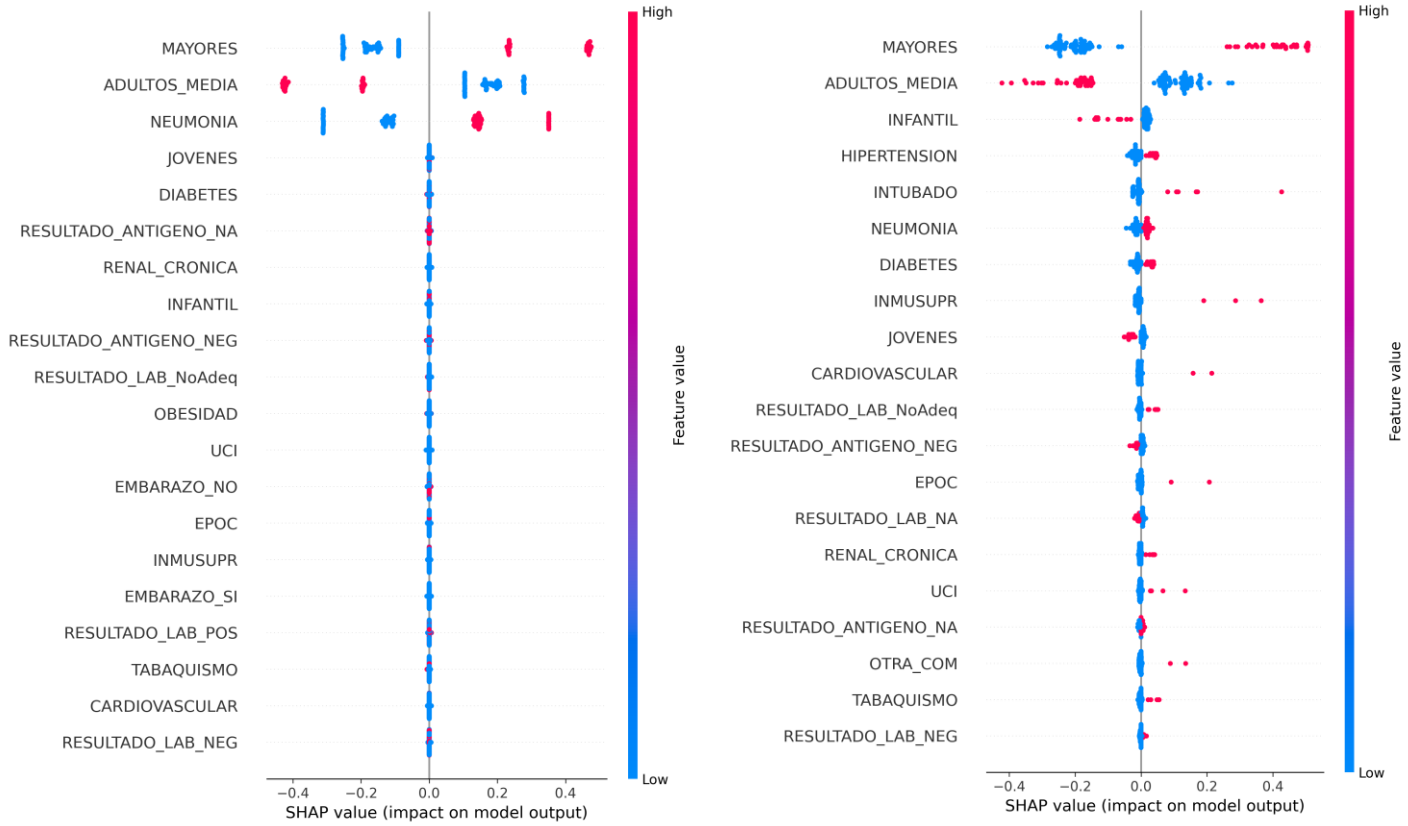
El modelo NB continua con una puntuación baja 0.7398, que es menor en comparación con el escenario de dos clases. Este declive es más pronunciado en comparación con otros clasificadores, reflejando posiblemente la debilidad de NB en manejar la correlación entre características y la complejidad aumentada en un espacio de características más diverso y probablemente interdependiente.

L. Inteligencia Artificial Explicable

Con la finalidad de examinar los mecanismos internos de los modelos predictivos que exhibieron un rendimiento bajo en comparación del resto, se implementa XAI. Con el método SHAP, se han identifican las nueve características más influyentes que determinan la discriminación entre clases en los conjuntos de datos (ADULTOS_MEDIA, SEXO, EMBARAZO_NA, NEUMONÍA, RESULTADO_LAB_NA, RESULTADO_LAB_POS, HIPERTENSIÓN, OBESIDAD, RESULTADO_ANTÍGENO_NA). Estos hallazgos se visualizan en las Figuras 6-8 donde cada punto indica el impacto de una característica específica en la predicción del modelo, con colores que varían desde azul para los valores más bajos hasta rojo para los más altos. El eje horizontal indica la magnitud y la dirección de la contribución de la característica a la predicción del modelo.

Para la evaluación comparativa de dos grupos específicos, la Fig. 6 ilustra los valores SHAP de los modelos SVM y NB. En el caso de SVM se observa, la edad ejerce una influencia notable en la predicción perteneciente al segundo grupo analizado. Este fenómeno se relaciona con la neumonía, ya que ambas características se destacan con valores coloreados en rojo y situados al lado derecho de la línea base. En contraste, con la diabetes, enfermedades renales crónicas y la enfermedad pulmonar obstructiva crónica, entre otras, que muestran valores de SHAP cercanos a cero, lo que indica que no constituyen predictores fuertes o consistentes dentro del modelo.

En contraste, con NB predominantemente, se identifica la presencia de puntos rojos a la derecha de la línea base, lo cual sugiere un impacto significativo de estas características en la clasificación hacia la segunda clase por parte del modelo. Al igual que en análisis previos, los individuos de mayor edad son un factor influyente para la clasificación hacia esta segunda clase. Además, se observa que la presencia de hipertensión, diabetes, condiciones inmunosupresoras y enfermedades cardiovasculares ejerce una influencia considerable en la decisión del modelo respecto a esta misma clase.



a) SVM b) NB

Fig. 6. Resultados SHAP para dos grupos

Por otro lado, para la primera clase, las edades inferiores a 60 años desempeñan un papel decisivo, en clara consonancia con resultados negativos en pruebas de antígeno.

Estos hallazgos resaltan la capacidad de los modelos para diferenciar entre perfiles de riesgo en temporadas invernales en el Estado de México basados en la edad y comorbilidades, subrayando la relevancia de estas variables en las decisiones predictivas de ambos modelos.

En el caso de NB, al operar bajo supuestos de independencia condicional, puede ser sensible a la forma en que se distribuyen y representan las variables dentro del conjunto de datos. Así, la heterogeneidad en las características puede afectar la capacidad predictiva del modelo.

Para la evaluación comparativa de los seis grupos, en la Fig. 7 se muestran los resultados de los valores SHAP para el modelo SVM, mientras que la Figura 8 muestra los resultados correspondientes al modelo NB. En ambas figuras se puede observar que las características más influyentes tienden a tener puntos dispersos a lo largo de un rango más amplio de valores SHAP, mientras que las características menos influyentes tienen puntos más agrupados cerca del centro.

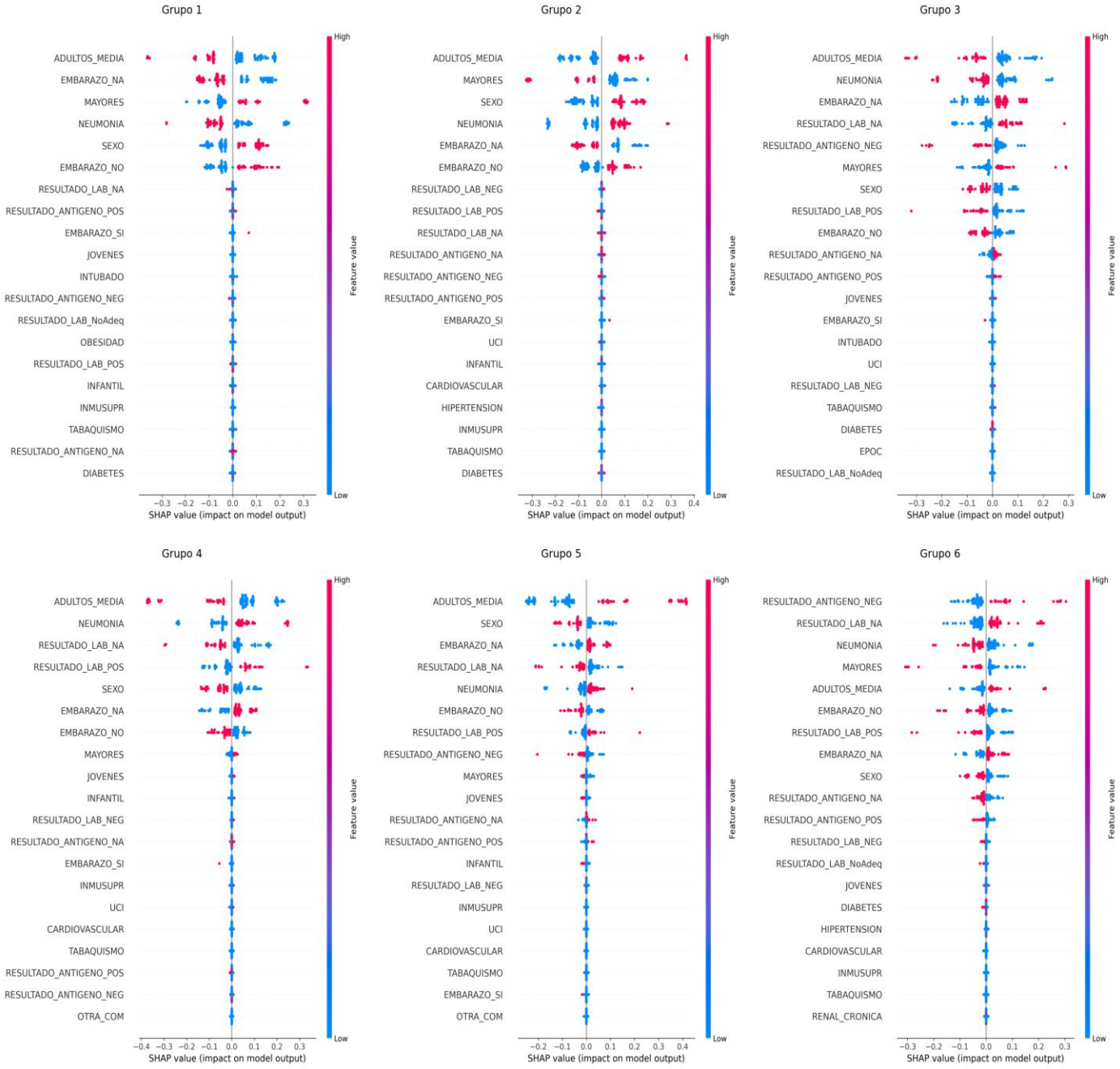


Fig. 7. Resultados SHAP del modelo SVM para seis grupos

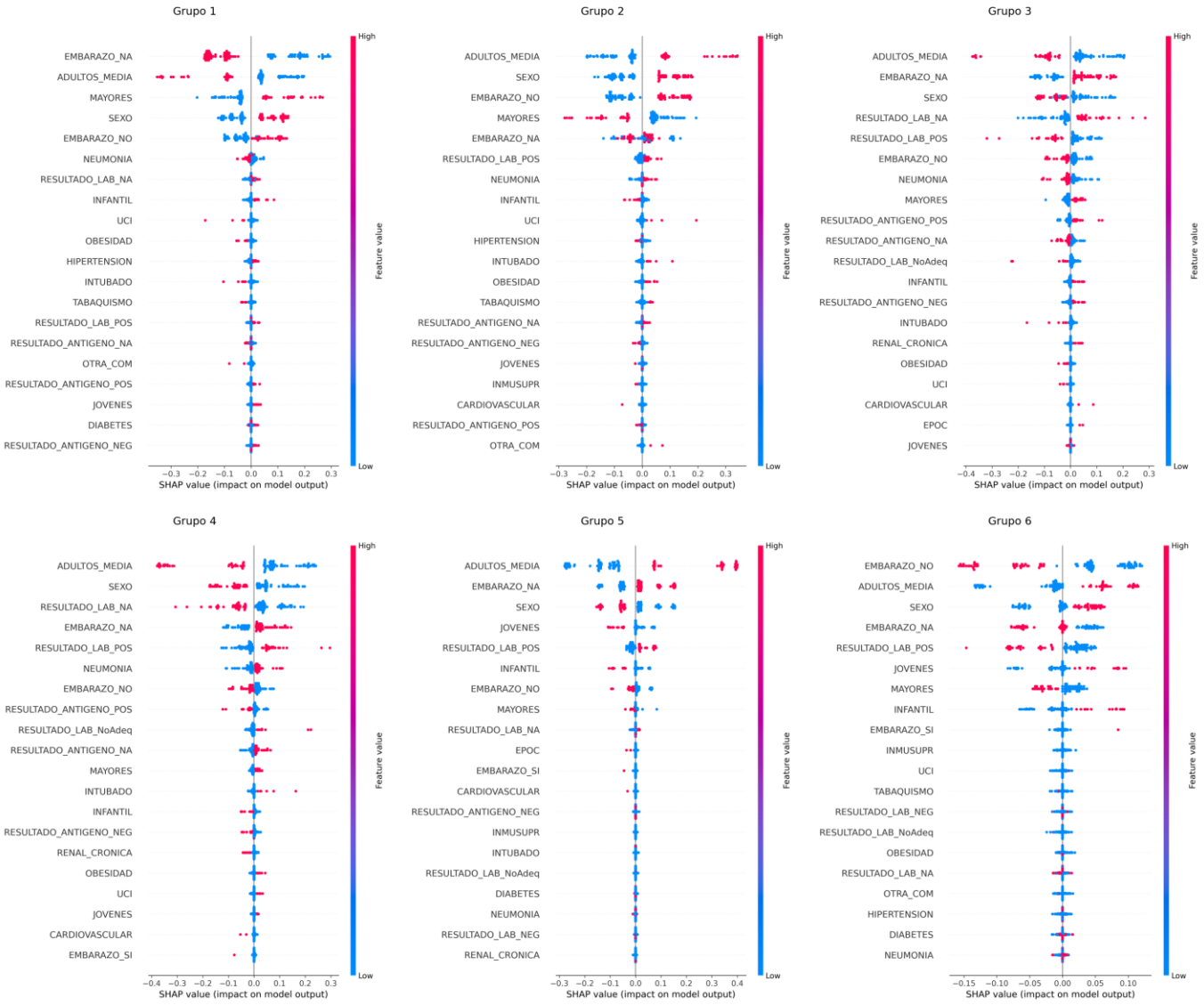


Fig. 8. Resultados SHAP del modelo NB para seis grupos

En la Fig. 7, se destaca que los adultos de mediana edad constituyen una característica significativa para el modelo SVM, ya que aparecen como la variable más influyente en cinco de los grupos estudiados, especialmente en los grupos 2 y 5, donde los valores de SHAP indican un impacto alto y positivo.

A diferencia del análisis con dos grupos, en el escenario de seis grupos se identifica el género como una característica relevante, donde ser mujer tiene un impacto positivo en los grupos 1 y 2 y ser hombre influye en los demás grupos.

Por otro lado, la presencia de neumonía se asocia con un impacto positivo en los grupos 2, 4 y 5. Este hallazgo sugiere un mayor riesgo o la necesidad de intervención específica en estos grupos, dado que las características relacionadas con resultados de laboratorio o la detección de antígenos son negativas a COVID-19 o no fueron aplicadas.

En cuanto a las comorbilidades, los resultados indican que estas no tienen un impacto significativo en la toma de decisiones del modelo, ya que se posicionan en la línea base de los valores SHAP.

Contrastando estos resultados con los obtenidos en el análisis de dos grupos de la Fig. 6 para el mismo modelo, se observa que para los seis grupos tanto el género como los resultados de laboratorio son determinantes en la toma de decisiones, mientras que, en el análisis anterior, los factores decisivos eran principalmente la edad y la presencia de neumonía. Esta

comparación subraya cómo la inclusión de más grupos puede alterar la percepción de las características determinantes en la predicción del modelo SVM.

En el análisis utilizando el modelo de NB y basándose en los valores SHAP para un escenario experimental dividido en seis grupos, se identifica que la edad, particularmente la de los adultos de mediana edad, emerge como la característica de mayor impacto en cuatro de los grupos. En contraste, en dos de los grupos, la característica más influyente es la ausencia de embarazo, aunque la edad media también ocupa un lugar prominente en estos grupos.

Al igual que en el modelo SVM, el género se presenta como una característica significativa. En particular, el ser mujer y no estar embarazada se destacan como atributos de impacto positivo y considerable para la predicción en el grupo 2.

Notablemente, el resultado positivo en las pruebas de laboratorio para COVID-19 también constituye un factor de impacto en este grupo, lo que sugiere la presencia del virus en mujeres no embarazadas. Este hallazgo es de particular interés, ya que no se había contemplado en los centroides (centroide 1) anteriormente. Las características de mayor relevancia son: adulto_media, sexo y neumonía y pueden proporcionar elementos significativos en la estructura y la dinámica de los modelos.

VI. CONCLUSIONES

Ante la crítica propagación del COVID-19 durante las temporadas invernales en el Estado de México entre 2020 y 2023, es fundamental evaluar la eficacia de los modelos de clasificación que predicen los riesgos asociados a la enfermedad. Estos modelos son esenciales para adaptar las intervenciones de salud pública de manera efectiva frente a las variantes del virus y las condiciones ambientales que influyen en su transmisión.

El análisis realizado revela diferencias significativas en la eficacia de dichos modelos, segmentados mediante el algoritmo *k*-medias. Se observa que los modelos de Árboles de Decisión y Redes Neuronales de Perceptrón Multicapa alcanzan una precisión perfecta (F1-score de 1.0) en las configuraciones de dos y seis grupos, demostrando una excelente capacidad para manejar la complejidad añadida por un mayor número de clústeres. Contrariamente, los modelos de Máquinas de Vectores de Soporte y NB mostraron un rendimiento inferior en escenarios con más clústeres, este último alcanzando un F1-score de sólo 0.7398 en la configuración de seis grupos. Esto subraya los desafíos en manejar la correlación entre características y la complejidad creciente en espacios de características más diversificados.

Además, el análisis de inteligencia artificial explicable aplicados a estos modelos proporcionan una visión más profunda sobre cómo características específicas como la edad, género, y resultados de pruebas de laboratorio influyen de manera diferencial en las predicciones de cada modelo. En particular, la edad y la presencia de neumonía son factores de alto impacto en las predicciones del modelo SVM, mientras que en el modelo NB, la capacidad de este último para manejar correctamente la dependencia entre características parece comprometida. SHAP, proporcionó elementos claves sobre las características más influyentes en la predicción de los modelos, subrayando la importancia de factores como la edad, el género y la importancia que tienen tomar muestras de laboratorio.

Estos hallazgos subrayan la importancia de seleccionar y ajustar adecuadamente los modelos de clasificación en estudios epidemiológicos para asegurar que las intervenciones y políticas de salud pública se basen en predicciones precisas y relevantes de los perfiles de riesgo durante las temporadas invernales.

Como líneas abiertas se contempla utilizar otros métodos para determinar las características relevantes, tales como Análisis de Componentes Principales. Además, se busca expandir el análisis a otras regiones y temporadas, así mismo, comparar con otros modelos epidemiológicos y explorar el impacto de las variantes del virus en la dinámica de transmisión.

Así mismo se recomienda a las autoridades poner mayor atención al grupo de personas de 48 a 63 años, poniendo mayor énfasis a los hombres que padezcan comorbilidades relacionadas a la obesidad puesto que algunos de ellos podrían incluso llegar a ser asintomáticos.

VII. BIBLIOGRAFÍA

[1] OMS, «World Health Organization,» OMS, 2019. [En línea]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/. [Último acceso: Agosto 2024].

[2] S. Salud, «Datos Abiertos Bases Historicas,» Gobierno de México, Abril 2020. [En línea]. Available: https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia. [Último acceso: 15 Agosto 2024].

[3] CONAHCYT, «COVID - 19 Tablero México,» CONAHCYT, 27 Febrero 2020. [En línea]. Available: https://datos.covid-19.conacyt.mx/#DOView. [Último acceso: 15 Agosto 2024].

[4] Z. A. A. Alyasseri, «Review on COVID-19 diagnosis models based on machine learning and deep learning approaches,» *Expert Systems,* vol. 39, nº 3, 2022.

[5] R. Kurniawan, S. N. H. Sheikh Abdullah, F. Lestari, M. Z. Ahmad Nazri, A. Mujahidin y N. Adnan, «Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries,» de *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, Pangkal, Indonesia , 2020.

[6] C. Nicholson, L. Beattie, M. Beattie, T. Razzaghi y S. Chen, «A machine learning and clustering-based approach for county-level COVID-19 analysis,» *PLoS ONE,* vol. 17, nº 4, 2022.

[7] K. Gohari, A. Kazemnejad, A. Sheidaei y S. Hajari , «Clustering of countries according to the COVID-19 incidence and mortality rates,» *BMC Public Health,* vol. 22, nº 1, 2022.

[8] J. Pérez-Ortega, N. N. Almanza-Ortega, . K. Torres-Poveda, G. Martínez-González, J. C. Zavala-Díaz y R. Pazos-Rangel, «Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico,» *Mathematics,* vol. 10, nº 13, p. 2167, 2022.

[9] J. Escudero, J. Guarner, A. Galindo-Fraga, M. Escudero-Salamanca, M. A. Alcocer-Gamba y C. Del Río, «La pandemia de Coronavirus SARS-CoV-2 (COVID-19): Situación actual e implicaciones para México,» *Archivos de Cardiología de México,* vol. 90, nº 91, pp. 7-14, 2020.

[10] R. J. A. L. Rebecca R. Andridge, «A Review of Hot Deck Imputation for Survey Non-response,» *Int. Statistical Rev,* vol. 78, nº 1, pp. 40-64, 2010.

[11] M. I. A. S. E. H. M. M. H. I. H. S. Md. Zubair, «An Efficient K-means Clustering Algorithm for Analysing COVID-19,» *Springer International Publishing,* pp. 422-432, 2020.

[12] J. E. Trujillo González y C. Vejerano García, «Clasificación de pacientes con síntomas de COVID-19 mediante árboles de decisión como una aplicación del aprendizaje automático,» *Guacamaya,* vol. 8, nº 1, 2023.

[13] S. Suthaharan, «Support Vector Machine,» *Machine Learning Models and Algorithms for Big Data Classification,* vol. 36, pp. 207-235, 2016.

[14] S. B. Š. N. A. I. L. V. M. Zlatan Car, «Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron,» *Computational and Mathematical Methods in Medicine,* 2020.

[15] J. M. Sugandh Bhatia, «Naïve Bayes Classifier for Predicting the Novel Coronavirus,» de *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, 2021.

[16] E. U. K. M. E. N. M. I. O. G. ,. E. A. Kevser Kübra Kırboğa, «CVD22: Explainable artificial intelligence determination of the relationship of troponin to D-Dimer, mortality, and CK-MB in COVID-19 patients,» *Comput Methods Programs Biomed,* vol. 233, p. 107492, 2023.

[17] J. J. E. Zúñiga, «Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública,» *Ingeniería, investigación y tecnología,* vol. 1, pp. 1-13, 2020.

[18] Z. a. Y. R. a. L. K. a. Y. G. a. L. Z. a. G. J. a. Z. Z. a. J. P. a. L. Y. a. Q. S. a. H. G. Lin, «Establishment of age group classification for risk stratification in glioma patients,» *BMC Neurology,* vol. 20, nº 1, p. 310, 2020.

[19] O. A. a. I. G. a. J. M. a. J. M. P. a. I. Perona, «An extensive comparative study of cluster validity indices,» *Pattern Recognition,* vol. 46, nº 1, pp. 243-256, 2013.

[20] Z. a. Y. R. a. L. K. a. Y. G. a. L. Z. a. G. J. a. Z. Z. a. J. P. a. L. Y. a. Q. S. a. H. G. Lin, «Boosting methods for multi-class imbalanced data classification: an experimental review,» *Journal of Big Data,* vol. 7, nº 1, p. 70, 2020.

[21] C. F. a. J. H.-O. a. R. Modroiu, «An experimental comparison of performance measures for classification,» *Pattern Recognition Letters,* vol. 30, nº 1, pp. 27-38, 2009.

Gustavo Alberto Diaz Rojas is a recently graduated computer engineer with areas of interest in artificial intelligence, data mining, and software development.

Rosa María Valdovinos Rosas holds a Ph.D. in Computer Science and is a Level II member of the National System of Researchers (SNI). She has participated in over 15 research projects with both social and scientific impact. Her academic production includes publications in indexed journals, book chapters, conference papers, books, invited talks, and one granted patent. She has mentored undergraduate, Master's, and Ph.D. students. Dr. Valdovinos is also involved in organizing science outreach events aimed at inspiring early scientific vocations. Her scientific work has received over 1,000 citations, with an H-index of 14, and she is ranked among the 10,000 most influential female scientists in Mexico, according to the International AD Scientific Index 2025.

Angélica Guzmán-Ponce earned a Ph.D. in Computer Science from the Autonomous University of the State of Mexico in 2021. She is a Level I member of the National System of Researchers (CONACYT). Currently, she is a postdoctoral researcher at Jaume I University (Castellón de la Plana, Spain) and the Polytechnic University of Valencia. Her research interests include Machine Learning and Graph Theory.