UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

MAESTRÍA EN OPTIMIZACIÓN Y CÓMPUTO APLICADO

Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRÍA EN OPTIMIZACIÓN Y CÓMPUTO APLICADO

PRESENTA

CARVANTES BARRERA ALEJANDRO

DIRECTOR DE TESIS

Dra. Lorena Díaz González

CO-DIRECTOR

Dr. Mauricio Rosales Rivera

REVISORES:

Dr. José Alberto Hernández Aguilar

Dra. Blanca Itzelt Taboada Ramírez

Dr. Luis Alberto Chávez Almazán

Dra. Lorena Díaz González

Dr. Outmane Oubram

CUERNAVACA, MORELOS NOVIEMBRE, 2022

Contenido

Resumen

La pandemia por COVID-19 ha tenido un fuerte impacto en la vida cotidiana de la sociedad a nivel global. Por esta razón, surge un fuerte interés por hacer uso de aprendizaje máquina y algoritmos de inteligencia artificial para analizar información de conjuntos de datos con registros de pacientes de COVID-19. En México, para el 15 de abril del 2022, se han reportado 5,737,475 casos positivos, de los cuales, 681,357 fueron hospitalizados y 324,670 fallecieron.

Existen diversos estudios análisis epidemiológicos y aplicación de herramientas de aprendizaje automático a nivel global y nacional, entre ellos se encuentran los que buscan identificar factores de riesgo para pacientes de COVID-19 haciendo uso de la técnica SHAP. El objetivo de esta técnica es explicar el resultado obtenido de un modelo de aprendizaje máquina. Es así como se puede conocer el impacto de cada variable en el resultado obtenido por un modelo, a este impacto se le denomina 'valor de SHAP'. Entonces, analizando estos valores se pretende identificar los factores de riesgo al padecer COVID-19. Sin embargo, trabajos de esta naturaleza en México no son comunes.

Este panorama proporciona una oportunidad para identificar factores de riesgo que pudieran estar asociados a la letalidad. En particular, esta metodología se implementó utilizando la base de datos del gobierno federal de México para alimentar modelos computacionales basados en aprendizaje máquina y finalmente calcular los valores de SHAP de dichos modelos. Además, este análisis se aplicó a cada ola epidemiológica habida en México para abordar la oportunidad de analizar el conjunto de datos por olas epidemiológicas.

En este contexto, este trabajo presenta una serie de modelos de predicción binaria para la defunción en pacientes con COVID-19 basado en XGBoost. Posteriormente, estos modelos son explicados por medio de SHAP, obteniendo así los valores de SHAP para cada variable, con los cuales se puede llevar a cabo la identificación de factores de riesgo de letalidad.

Las variables más importantes en la predicción de la defunción en los pacientes fueron la neumonía y la edad avanzada, las cuales aumentaron el riesgo de fallecer. Con una considerable menor importancia se encontró que los casos registrados en el Instituto Mexicano del Seguro Social (IMSS) presentaron un mayor riesgo de fallecer durante las primeras cuatro olas epidemiológicas de la pandemia. Por el contrario, casos registrados en la Secretaría de Salud de México (SSA) presentaron un menor riesgo de fallecer. Del mismo modo, el pertenecer al género femenino y tener entre 18 a 29 años redujo el riesgo de fallecer al padecer COVID-19. La comorbilidad más notable fue la diabetes y el haber sido intubado en etapas tempranas de la pandemia elevó el riesgo de defunción de manera importante.

Índice de figuras

Índice de tablas

1 Introducción

El mundo presenció el 6 de enero de 2020 la primera muerte por la enfermedad por coronavirus COVID-19 [1]. Posteriormente, el 30 de enero de 2020 la Organización Mundial de la Salud (OMS) declaró la epidemia por COVID-19 una emergencia de salud pública de preocupación internacional. Finalmente, el 11 de marzo de 2020, la OMS declaró que la enfermedad por coronavirus 2019 puede caracterizarse como pandemia [2]. A la fecha (18 de Agosto de 2022) no se ha emitido una declaración distinta con respecto a esta enfermedad por lo que aún es caracterizada como pandemia.

La enfermedad por coronavirus COVID-19 es resultado de la infección por el virus SARS-CoV-2. La cual, puede provocar fallas respiratorias agudas en los casos más graves [3]. La OMS reportó de manera global, hasta el 18 de agosto del 2022 un total de 389,680,368 casos confirmados, los cuales incluyen 6,436,519 muertes [1]. De las evidencias anteriores, es seguro afirmar que la pandemia ha afectado severamente a la población mundial.

Esta pandemia ha motivado la investigación en diversos campos de la ciencia y tecnología para hacerle frente. En particular, el uso de técnicas de *Machine Learning* (ML) ha jugado un rol importante en el desarrollo de métodos para el diagnóstico, detección y predicción de casos de COVID-19, ya que estas técnicas funcionan bien en circunstancias donde se cuenta con datos numéricos correctamente estructurados y preferentemente, en gran cantidad [4].

En México, el gobierno a través del Sistema de Vigilancia Epidemiológica de Enfermedades Respiratorias Virales [5] provee una base de datos (BD) con casos sospechosos y confirmados de COVID-19, así como también, difuntos que padecieron dicha enfermedad. Esta base de datos es actualizada diariamente y contiene información de cada uno de los pacientes, tal como su historial médico, información demográfica e información médica reciente. Para los fines específicos del presente trabajo, se consideró esta base de datos desde sus inicios hasta el 15 de abril del 2022, tomando en cuenta un aproximado de 15 millones de registros, de los cuales 5,737,475 fueron clasificados como positivos o casos confirmados, 681,357 hospitalizados y, en general, 324,670 fallecieron por consecuencia de esta enfermedad. Dada la gran cantidad de datos recopilados, esta base de datos es adecuada para llevar a cabo experimentos de ML. Particularmente, la finalidad del presente trabajo es dilucidar cuáles fueron aquellos factores de riesgo que estuvieron asociados a la letalidad por COVID-19 considerando las primeras cuatro olas de la pandemia en México. A continuación, se mencionan algunos trabajos relacionados, en los cuales se usaron técnicas de ML para analizar bases de datos de COVID-19 a nivel mundial y nacional.

Antecedentes

Se han realizado estudios previos con el fin de destacar algunos factores de riesgo tales como obesidad, diabetes, hipertensión y enfermedades cardiovasculares (p. ej., [6]–[12]). Así como también, un riesgo mayor al contar con más de una comorbilidad a la vez (p. ej., [7], [11]). Además, se han realizado diferentes estudios sobre el efecto de esta enfermedad, no sólo por género sino también por edades. Con respecto a la edad, se observó desde el inicio de la pandemia, que tener más de 60 años es uno de los principales factores de riesgo de muerte (p. ej., [8], [12]), lo cual se debe principalmente a una inmunidad ya desgastada y una mayor prevalencia de enfermedades crónicas en la población de la tercera edad. De manera similar, se ha reportado que tener 40 años y ser varón [6] está asociado a otros posibles factores de riesgo.

Por otro lado, se han evaluado las influencias de los factores socioeconómicos y demográficos en municipios y estados de México por medio diversos métodos estadísticos [13], concluyendo que la mayor transmisión ocurre en las zonas con alto y muy alto desarrollo económico, mientras que aquellas donde existe más pobreza y carencias sociales resultan más afectadas por la letalidad de la enfermedad.

Por otra parte, Quiroz-Juárez et al. [14] desarrollaron modelos de ML basados en redes neuronales artificiales para predecir la defunción en pacientes con base en comorbilidades, información demográfica y médica reciente. En dicho estudio se utilizó la BD de México, con fecha de acceso 31 de enero del 2021, a la cual le aplicaron un balance de clases (sobrevivientes y fallecidos), descartando aleatoriamente casos de sobrevivientes hasta obtener un balance perfecto (50% sobrevivientes y 50% fallecidos). De este modo mejoraron los resultados de sus modelos y concluyeron que las variables que presentan una correlación más fuerte con el resultado del modelo fueron la edad, el estado de hospitalización, la intubación y la unidad de cuidados intensivos.

Por otro lado, a nivel mundial se han realizado estudios de las distintas olas de la pandemia (p. ej., [15] y [16]), no obstante, en México no se ha abordado un análisis desde ese enfoque que cubra las primeras cuatro olas sucedidas en México (hasta el 15 de abril del 2022).

Planteamiento del problema y justificación

El COVID-19 es un importante problema emergente de salud pública en México, por lo que ha sido importante identificar los factores de riesgo que están asociados con la letalidad por esta enfermedad, lo cual considera las primeras cuatro olas de la pandemia.

En este contexto, el cálculo de los valores de SHAP constituye una técnica efectiva para conocer los factores de riesgo asociados a un modelo predictivo (p. ej., [11], [17]–[22]). Este método ha tomado popularidad en la literatura, ya que propone una manera objetiva de conocer el nivel de contribución de una característica en un resultado binario [23]. Particularmente en esta tesis, la técnica de SHAP permite comprender las relaciones entre las características de los pacientes con COVID-19 registrados en la BD oficial de México.

Además, el hecho de llevar a cabo el análisis por oleadas permitirá realizar un análisis retrospectivo para entender las vulnerabilidades que sufrió la población en México a través de las etapas de la pandemia.

1.1 Objetivo general

Identificar los factores de riesgo más importantes que están asociados con la letalidad en pacientes con COVID-19 durante las primeras cuatro olas epidemiológicas en México, mediante la aplicación de técnicas de aprendizaje automático.

1.2 Objetivos específicos

1. **Preprocesar el conjunto de datos:** Realizar un análisis de datos exploratorio, incluyendo limpieza de datos, partición por periodos de tiempo e ingeniería de características para desarrollar los modelos de ML.
2. **Entrenar modelos predictivos para la defunción:** Desarrollar modelos clasificadores binarios que pronostiquen la letalidad de los pacientes.
3. **Analizar los factores de riesgo haciendo uso de los valores de SHAP:** Implementar el método SHAP con el fin de obtener la contribución de cada variable con respecto a la letalidad de los pacientes. De este modo, se identificarán las variables que representan los principales factores de riesgo asociados con la letalidad. Así como también, cuáles son los factores que se asocian negativamente a ella, es decir, los factores de protección.

2 Marco teórico

El Aprendizaje máquina o *Machine Learning* (ML) se define como un conjunto de métodos que permiten a los modelos aprender a partir de una base de datos, para posteriormente realizar predicciones [4]. En este capítulo se describen las técnicas utilizadas en este trabajo, así como las métricas de evaluación de desempeño de los modelos.

2.1 Valores de Shapley

Los valores de Shapley son un método para distribuir las riquezas dentro de la teoría de juegos cooperativos, donde dos o más jugadores no compiten entre sí, sino que colaboran para conseguir el mismo objetivo. Por lo tanto, ganan o pierden en conjunto. Lo que busca este método es determinar el impacto de cada jugador (variable o característica) en la victoria, para que sea remunerado según la importancia de su contribución [17].

En los valores de Shapley se hace uso de la 'función característica' la cual, es un modelado matemático del problema en cuestión. De modo que, dicha función evalúa la combinación de características contenidas en un vector y determina si devolverá un resultado igual a 1 (victoria) o igual a 0 (derrota) [23].

A continuación, se describe a manera de ejemplo el cálculo de los valores de Shapley para un jugador [24]. Es de suma importancia recordar que un jugador (variable o característica) hace referencia a un elemento de un vector (registro o renglón individual de la base de datos).

i) Seleccionar una variable (jugador). Por ej., se tiene un vector con los siguientes jugadores [1, 2, 3] y se selecciona al jugador 1.

ii) Generar la lista de permutaciones de todas las variables (jugadores) comenzando la lista con la variable seleccionada en el costado izquierdo. En este ejemplo, la lista de permutaciones sería la siguiente: [**1**,2,3]; [**1**, 3, 2]; [2, **1**, 3]; [2, 3, **1**]; [3, **1**, 2]; [3, 2, **1**].

iii) Generar dos subconjuntos para cada permutación. El primero, contiene todos los jugadores (variables/características) que se encuentren a la izquierda del jugador seleccionado más el mismo jugador. El segundo, es igual al primero, pero sin el jugador seleccionado.

iv) Evaluar la función característica en cada subconjunto. Ejemplo, definiremos la función característica en la ecuación 1.

$$v(S) = \{1 \; si \; S \in \{\{1,3\},\{2,3\},\{1,2,3\}\};\; 0 \; de \; otro \; modo. \qquad (1)$$

donde S es un subconjunto de jugadores, entonces $v(S)$ es el valor que representa la suma total de las ganancias a obtener de manera cooperativa por los miembros del subconjunto S.

v) Restar el resultado de haber evaluado la función característica en el primer subconjunto menos el resultado de la evaluación en el segundo. De este modo se obtienen las

contribuciones marginales del jugador seleccionado. En la Tabla 1 se ejemplifican los pasos hasta el momento.

Tabla 1 Ejemplo de cálculo de valores de Shapley.

Permutaciones para evaluar al jugador 1	Subconjuntos evaluados	Contribuciones marginal del jugador 1
1,2,3	$v(\{1\}) - v(\emptyset)$	0 – 0 = 0
1,3,2	$v(\{1\}) - v(\emptyset)$	0 – 0 = 0
2,1,3	$v(\{1,2\}) - v(\{2\})$	0 – 0 = 0
2,3,1	$v(\{1,2,3\}) - v(\{2,3\})$	1 – 1 = 0
3,1,2	$v(\{1,3\}) - v(\{3\})$	1 – 0 = 1
3,2,1	$v(\{1,3,2\}) - v(\{3,2\})$	1 – 1 = 0

vi) Sumar todas las contribuciones marginales y posteriormente dividir entre el número total de permutaciones. Este resultado es el valor de Shapley para el jugador evaluado. En el ejemplo, el valor de Shapley para el jugador 1 es igual a 1/6.

Es conveniente enfatizar que la suma de todos los valores de Shapley de todos los jugadores es igual a 1. En resumen, este cálculo hace referencia a cuántas veces la variable (jugador) evaluada fue determinante en el resultado de la función característica.

De acuerdo con los valores de Shapley, la contribución que el jugador i recibe de un juego cooperativo (v, N) se describen en la ecuación 2.

$$\rho_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (2)$$

Donde N es un conjunto de n jugadores y la suma se extiende a todos los subconjuntos S de N que no contienen al jugador i [23].

2.2 SHAP (*SHapley Additive exPlanations*)

SHAP es una librería o *framework* que se basa en los valores de Shapley para averiguar qué variables o características son más relevantes en las predicciones obtenidas por un modelo. La finalidad de la técnica de SHAP es explicar la instancia x de un modelo predictivo calculando la contribución de cada variable a la predicción misma.

SHAP calcula los valores de Shapley mediante un método lineal de atribución de características aditivas usando la ecuación 3.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \quad (3)$$

Donde g es el modelo explicativo, $z' \in \{0,1\}^M$ es el vector de coalición (subconjunto de jugadores), M es el tamaño máximo de coalición y $\phi_j \in R$ es la contribución por característica o variable j (los valores de Shapley). En el vector de coalición un valor de 1 significa que la variable está presente y 0 que está ausente [25]. Es importante entender que en esta técnica la función característica que utilizan los valores de Shapley es reemplazada por un modelo predictivo binario.

2.2.1 Representación esquemática de SHAP

SHAP genera una representación gráfica de la importancia de cada variable o característica, en donde cada punto representa un valor de SHAP para una característica y una instancia. La posición del eje-y está determinada por la característica y la del eje-x por el valor de SHAP. El color representa los valores de la característica de bajo a alto. En una representación de características binarias el color azul representaría 0 y el rojo 1. Los puntos que se sobreponen se amontonan con respecto al eje-y para dar un mejor entendimiento de la distribución de los valores de SHAP por característica. Finalmente, el orden de las características es descendente, partiendo de mayor importancia a menor importancia. A continuación, en la Fig. 1 se muestra un ejemplo de la gráfica de resumen de SHAP.

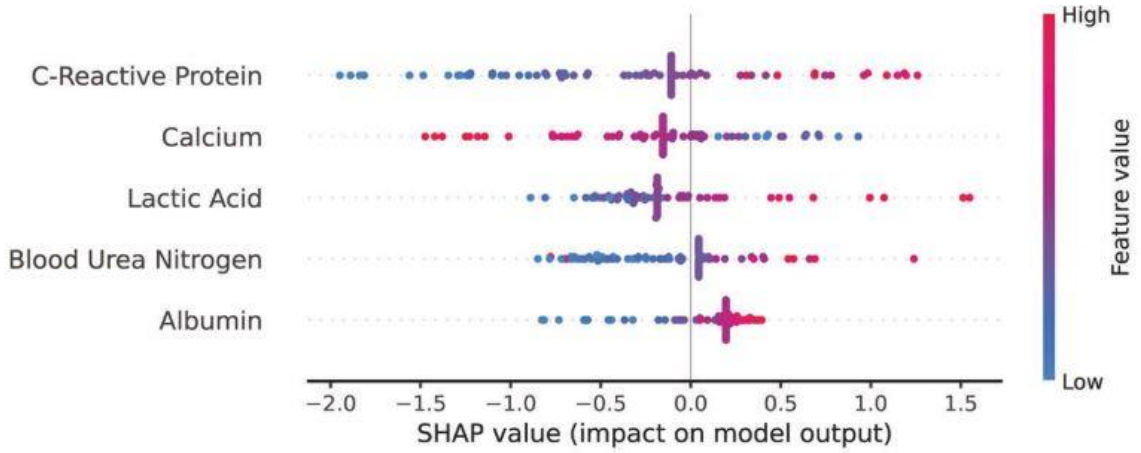

Fig. 1 Ejemplo del resumen gráfico generado por SHAP para un modelo predictivo de la mortalidad de pacientes con COVID-19 a partir de cinco parámetros de laboratorio de química sérica (proteína C reactiva-PCR, calcio sérico, ácido láctico, nitrógeno ureico en sangre y albúmina sérica). Esta figura fue tomada de [18].

2.3 XGBoost (*Extreme Gradient Boosting*)

XGBoost es un método escalable de extremo a extremo que usa árboles de decisión aumentados (*tree boosting*). A continuación, se describen de manera breve los siguientes conceptos: (i) Métodos basados en árboles; y, (ii) Aumento (*Boosting*).

2.3.1 Métodos basados en árboles

Estos métodos permiten desarrollar modelos sencillos e interpretables, que visualmente son similares a una estructura de árbol como se muestra en la Fig. 2. En donde, el nodo ubicado en la

cima es llamado **nodo raíz**, el cual tiene ramificaciones hacía otros nodos. Los nodos que tienen ramificaciones son llamados **nodos internos** (*splits*) y los que no, son llamados **nodos terminales** u **hojas**. Este tipo de árboles suelen ser binarios, lo que significa que cada nodo interno solo tiene dos ramificaciones.

Los nodos en estos árboles representan datos. Cada ramificación contiene un conjunto de atributos o reglas de clasificación (expresables como 'si… entonces…') vinculadas a una etiqueta que se encuentra al final de la ramificación; reglas condicionales. En el ámbito de ML, un algoritmo se encarga de construir el árbol de decisión según los datos de entrada. Posteriormente, para hacer una predicción o clasificación, el dato de entrada seguirá el flujo desde la raíz hasta una hoja, para obtener un resultado de salida.

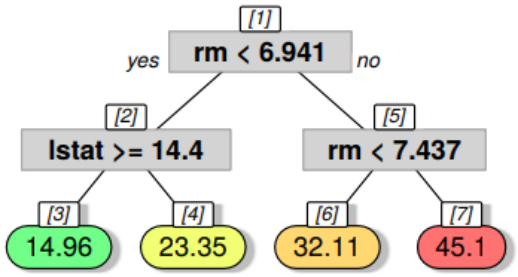

Fig. 2 Visualización de un árbol de decisión. Esta figura fue tomada de [26].

En general, los árboles de decisión sencillos tienen capacidades predictivas limitadas. Sin embargo, sus capacidades predictivas aumentan cuando múltiples modelos de árboles son combinados en uno global [26].

2.3.2 Aumento (*Boosting*)

Se utiliza el término Boosting cuando se construye un modelo por medio de la combinación o ensamble de modelos más simples [24, 25], los cuales son comúnmente llamados **modelos base** y son aprendidos utilizando un **aprendedor base** o **débil** (*base learner o weak learner*).

2.3.3 Árboles de gradiente aumentado (*Gradient Tree Boosting*)

Una vez teniendo conocimiento de lo que es el aumento y los métodos de árboles de decisión es intuitivo entender que los árboles aumentados (*boosted trees*) son algoritmos que hacen uso de múltiples árboles de decisión como modelos base para ensamblar un modelo más robusto. En la Fig. 3 se muestra un sencillo ejemplo con dos árboles de decisión ensamblados. La predicción final para la ejemplificación ilustrada es la suma de predicciones de cada árbol.

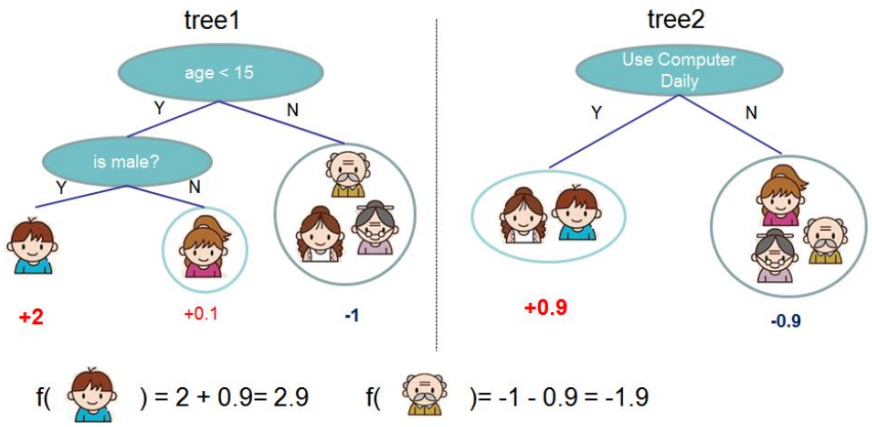

Fig. 3 Ejemplo de un modelo de ensamble de árboles. Esta figura fue tomada de [28].

Debido a que en estos modelos de ensamble incluyen funciones como parámetros estos no pueden ser optimizados utilizando métodos de optimización tradicional en un espacio Euclidiano. Entonces, el modelo se entrena de manera aditiva. De modo que, de manera voraz agrega la función (**f**) que represente un mejor rendimiento [29].

2.3.3.1 Parámetros de los árboles de gradiente aumentado.

Algunos de los parámetros de estos modelos son listados a continuación [30].

1. ***learning_rate*:** es la tasa de aprendizaje, la cual reduce la contribución de cada árbol de decisión que se agrega, de modo que impacta en el tiempo de aprendizaje del modelo.
2. ***n_estimators*:** es el número de estimadores o cantidad de árboles a emplear.
3. ***max_depth*:** es la profundidad máxima, la cual limita el número de nodos en el árbol.
4. ***min_child_weight*:** es el peso mínimo necesario para cada nodo.
5. ***gamma*:** es el costo de complejidad por introducir una hoja nueva al modelo.
6. ***subsample*:** es la tasa de submuestreo en las instancias de entrenamiento; al reducir este parámetro se pueden prevenir problemas de sobreajuste.
7. ***alpha*:** es el término de regularización L1 en los pesos; el aumento de este valor hará que el modelo sea más conservador.
8. ***colsample_bytree*:** es la tasa de submuestreo de columnas cuando se construye cada árbol; el submuestreo ocurre una vez cada que se construye un árbol.

2.4 Métricas de evaluación

En este apartado se explicarán las métricas de evaluación utilizadas en el presente trabajo. Primeramente, en la Tabla 2 se muestra la matriz de confusión, la cual engloba los cuatro posibles resultados de un modelo predictivo. Entonces, el modelo puede acertar de dos maneras, al predecir correctamente un resultado positivo (verdadero positivo; TP por sus siglas en inglés - True Positive) y al predecir correctamente un negativo (verdadero negativo; TN por sus siglas en inglés - True Negative). De manera semejante, puede errar de dos maneras, al predecir un resultado positivo erróneamente (falso positivo; False Positive FP) y al predecir un negativo erróneamente (falso

negativo; False Negative FN). Entender estos cuatro escenarios resulta importante para la interpretación de las métricas de evaluación.

Tabla 2 Matriz de confusión.

	Caso real positivo	**Caso real negativo**
Predicción positiva	Verdadero positivo (TP)	Falso positivo (FP)
Predicción negativa	Falso negativo (FN)	Verdadero negativo (TN)

2.4.1 Exactitud (Accuracy)

La exactitud es una métrica que nos indica la proporción de las instancias verdaderas pronosticadas correctamente entre todas las instancias recibidas por el modelo, la cual se calcula usando la ecuación 4.

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

2.4.2 Precisión (Precision)

La precisión es una métrica que mide basada totalmente en los casos positivos, ya sean falsos o verdaderos. En consecuencia, cada error del modelo en predecir un falso positivo afecta de modo importante esta métrica, la cual se calcula usando la ecuación 5.

$$Precisión = \frac{TP}{(TP + FP)} \quad (5)$$

2.4.3 Recuerdo (Recall)

Esta métrica también es conocida como **sensibilidad** o tasa de verdaderos positivos. Su función es medir el número de instancias verdaderas correctamente pronosticadas entre todas las predicciones resultantes en las instancias positivas. Esta métrica es sensible cuando el modelo se equivoca al no predecir correctamente un caso real positivo y se calcula por medio de la ecuación 6.

$$recall = \frac{TP}{(TP + FN)} \quad (6)$$

2.4.4 Especificidad (Specificity)

Esta métrica se basa en la tasa de verdaderos negativos y es la proporción de casos negativos que fueron pronosticados correctamente. Esta métrica es sensible cuando el modelo predice demasiado casos negativos que no son realmente negativos y se calcula por medio de la ecuación 7.

$$especificidad = \frac{TN}{(TN + FP)} \quad (7)$$

2.4.5 AUC ROC

AUC ROC es la abreviación de los términos en inglés ***'area under the curve'*** **(AUC)** y ***'receiver operating characteristic'*** **(ROC)**. Esta métrica resume la compensación entre sensibilidad y especificidad. Para dar una breve explicación geométrica en el panel **a)** de la Fig. 4 se muestra el punto de sensibilidad/especificidad conectado por la función de paso. El área bajo la curva (AUC) asociada es representada en el área sombreada de azul la cual representa el **área bajo la curva por definición o estricta**. Por otra parte, en el panel **b)** se muestran zonas sombreadas adicionales de color naranja y rojo, que representan el **AUC con lazos.** De lo anterior, cabe destacar que el **AUC con lazos** es el utilizado comúnmente en la literatura [31]. Esta métrica es calculada por librerías tales como **scikit-learn** [32]. La fórmula para el cálculo del AUC con lazos se representa en la ecuación 8 [31].

$$AUC = \frac{1}{2}(sensibilidad + especificidad) \quad (8)$$

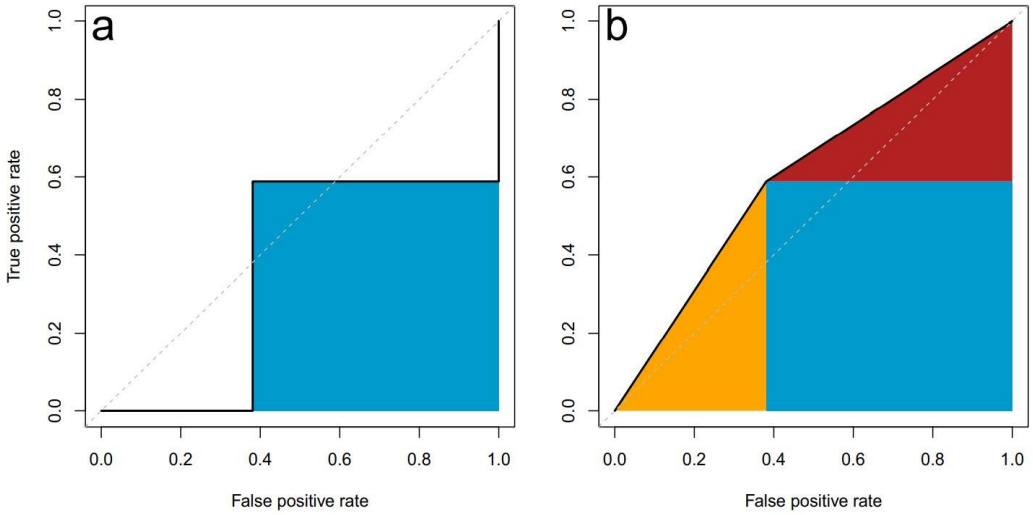

Fig. 4 Definición estricta y con lazos de la curva ROC [31].

2.5 Validación cruzada

La validación cruzada consiste en tomar un conjunto de datos y dividirlo en *K* secciones, a cada una de ellas se le conoce como doblez. Posteriormente, se utiliza el primer doblez como conjunto de validación utilizado para probar un modelo entrenado. El resto de los dobleces conformarán el conjunto de entrenamiento. Después, se repite el mismo proceso utilizando el siguiente doblez como validación y el resto como entrenamiento, esto se repite sucesivamente *K* número de veces, hasta que cada doblez haya sido conjunto de validación una vez. En la Fig. 5 se muestra un ejemplo gráfico de la validación cruzada con *K*=10 (diez dobleces) donde D_{val} hace referencia al conjunto de validación y D_{train}al conjunto de entrenamiento [33].

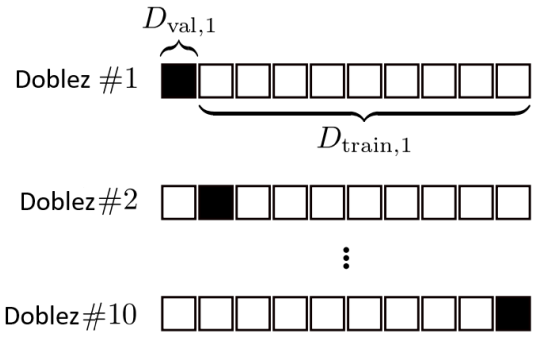

Fig. 5 Diagrama esquemático para representar la validación cruzada de diez dobleces. Esta figura fue tomada de [33].

2.6 Coeficiente de correlación punto-biserial

La correlación punto-biserial es una variante del coeficiente de correlación de Pearson o *r* de Pearson, el cual describe la relación lineal entre dos variables, midiendo la fuerza, dirección y probabilidad de la asociación entre dos variables de intervalo o relación. Los valores que puede tomar oscilan entre -1 y 1, siendo los valores negativos una relación lineal negativa, el 0 una relación nula y los valores positivos una relación lineal positiva. La correlación punto-biserial mide específicamente la relación entre una variable cualitativa dicotómica (si o no, hombre o mujer, se tiene o no algún padecimiento etc.) y una variable de relación o de intervalo [34]. En la ecuación 9 se muestra como se calcula el coeficiente de relación de Pearson descrito de manera analógica [35].

$$r = \frac{(covarianza\ entre\ la\ variable\ A\ y\ B)}{(Desviación\ estandar\ de\ la\ variable\ A)\ \times\ (Desviación\ estandar\ de\ la\ variable\ B)} \quad (9)$$

3 Metodología

La metodología implementada fue constituida por las siguientes etapas principales (ver Fig. 6), las cuales son explicadas en detalle en las siguientes subsecciones:

i) Se descargó la base de datos con última fecha de acceso el 30 de abril del 2022, la cual cuenta con 15,453,958 registros y 39 variables, de las cuales solo se seleccionaron las variables de interés para el estudio.

ii) Se realizó un preprocesamiento que consistió en lo siguiente: a) se convirtieron todas las variables a formato binario; b) se seleccionaron únicamente los registros correspondientes a pacientes positivos en los cuales se confirmó COVID-19; c) se eliminaron registros de pacientes con información incompleta; d) se aplicó ingeniería de características para generar nuevas variables de interés (intervalos de edades, agrupaciones de unidades médicas e intervalos de índice de desarrollo humano); y, finalmente, e) se delimitaron los registros pertenecientes a cada ola.

iii) Se generó un conjunto por cada ola de SARS-CoV-2 del país y un conjunto general que abarcó desde el inicio de la primera ola (2020/05/24) hasta el final de la cuarta (2022/04/15).

iv) Se generaron divisiones manualmente para cada conjunto con el fin de combatir el alto desbalance de clases particular de cada conjunto. Se buscó que estas divisiones mantuvieran un balance aproximado al 50% de casos positivos y 50% de casos fallecidos, con el objetivo de facilitar el entrenamiento de los modelos. Debido a que cada ola epidemiológica cuenta con un desbalance de clases singular, el número de divisiones es establecido individualmente para cada subconjunto. Las divisiones constan de todos los pacientes fallecidos, que se repiten en todas las divisiones del conjunto, y un número muy cercano de sobrevivientes, los cuales no se repiten en ninguna otra división.

v) Se realizó para cada conjunto la búsqueda de hiperparámetros para los modelos de XGBoost en la primera división, posteriormente, con esos mismos parámetros se entrenó y evaluó un modelo para cada división.

vi) Finalmente, se interpretaron los modelos de cada división con el algoritmo SHAP y obtuvieron promedios y medianas de estos valores. A continuación, se detalla cada uno de estos pasos metodológicos.

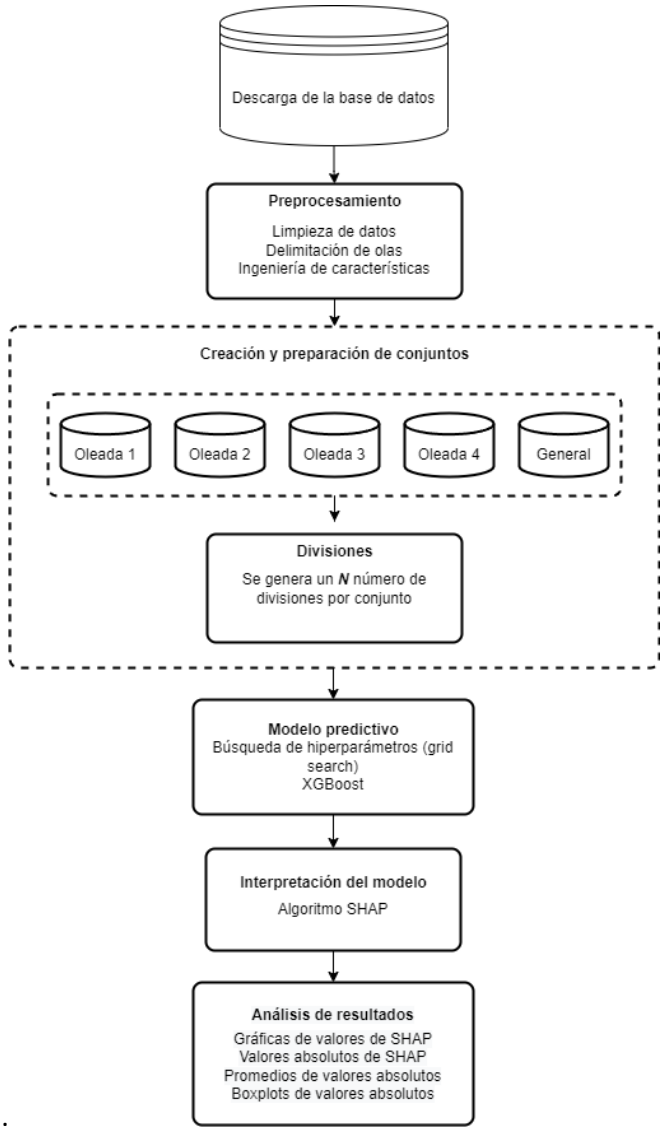

Fig. 6 Diagrama de flujo de la metodología.

3.1 Descarga de la base de datos

Se realizó la descarga del archivo de datos abiertos de casos de COVID-19 desde el portal de la Dirección General de Epidemiología de la Secretaría de Salud del gobierno federal (http://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip) con fecha de acceso 30 de abril del 2022 [5]. La base de datos contenía 39 variables de las cuales se extrajeron las de interés para el presente estudio (Tabla 3).

Tabla 3 Variables extraídas de la base de datos del gobierno de México.

Variable	Descripción
FECHA_DEF	Fecha de defunción del paciente, en caso de no ser un caso fallecido se utiliza el valor 9999-99-99.
EDAD	Edad del paciente.
SEXO	Si el paciente es mujer.

INTUBADO	Si el paciente fue intubado.
NEUMONIA	Si el paciente presentó neumonía.
DIABETES	Comorbilidad.
HIPERTENSION	
OBESIDAD	
CARDIOVASCULAR	
RENAL_CRONICA	
TABAQUISMO	
EPOC (enfermedad pulmonar obstructiva crónica)	
ASMA	
INMUNUSUPRESORA	
ENTIDAD_RES	Estado de residencia del paciente.
MUNICIPIO_RES	Municipio de residencia del paciente.
FECHA_SINTOMAS	Fecha en que el paciente presentó síntomas.
SECTOR	Unidad médica que registró el caso. Se utilizó para generar variables de agrupaciones de unidades médicas. Se describe en la Tabla 5.
TIPO_PACIENTE	Si el paciente es ambulatorio u hospitalizado.
CLASIFICACIÓN_FINAL	Determina si el paciente es o no un caso de COVID-19, las claves que puede tomar esta variable se describen en la Tabla 4.

3.2 Preprocesamiento

Una vez obtenido el conjunto de datos de COVID-19, se realizó un preprocesamiento que consistió en las siguientes tareas.

3.2.1 Limpieza de datos

Es importante denotar que en el presente estudio se decidió trabajar únicamente con casos de COVID-19 confirmados, por lo que se hizo un filtro usando la variable llamada **'CLASIFICACIÓN_FINAL',** donde se removieron los registros con una clave mayor a 3, ya que estos casos son negativos, sospechosos o no se tiene suficiente información de ellos. Las claves de la variable son descritas en la Tabla 4, la cual fue extraída directamente del catálogo de variables de la base de datos de COVID-19 [36].

Tabla 4 Descripción de variable 'CLASIFICACIÓN_FINAL'. Las claves 1, 2 y 3 representan los registros utilizados en el presente trabajo.*

CLAVE	CLASIFICACIÓN	DESCRIPCIÓN
1*	CASO DE COVID-19 CONFIRMADO POR ASOCIACIÓN CLÍNICA EPIDEMIOLÓGICA	Confirmado por asociación: aplica cuando el paciente informó estar en contacto con un caso positivo a COVID-19, que se encuentra registrado en el SISVER. A este paciente no se le tomó muestra o la muestra resultó no válida.

2*	CASO DE COVID-19 CONFIRMADO POR COMITÉ DE DICTAMINACIÓN	Confirmado por dictaminación solo aplica para defunciones bajo las siguientes condiciones: Al caso no se le tomó muestra o sí se tomó muestra, pero resultó no válida.
3*	CASO DE SARS-COV-2 CONFIRMADO POR LABORATORIO	Confirmado por laboratorio aplica cuando: El caso tiene muestra y resultó positiva a SARS-CoV-2, sin importar si el caso tiene asociación clínica epidemiológica.
4	INVÁLIDO POR LABORATORIO	Inválido, aplica cuando el caso no tiene asociación clínico epidemiológica, ni dictaminación a COVID-19. Se le tomó muestra y esta resultó no válida.
5	NO REALIZADO POR LABORATORIO	No realizado aplica cuando el caso no tiene asociación clínico-epidemiológica, ni dictaminación a COVID-19 y se le tomó muestra, la cual no se procesó.
6	CASO SOSPECHOSO	Sospechoso aplica cuando el caso no tiene asociación clínico-epidemiológica, ni dictaminación a COVID-19 y no se le tomó muestra, o se le tomó muestra y está pendiente de resultado, sin importar otra condición.
7	NEGATIVO A SARS-COV-2 POR LABORATORIO	Negativo, aplica cuando al caso se le tomó muestra y resultó negativa a SARS-COV-2 o positiva a cualquier otro virus respiratorio (Influenza, VSR, Bocavirus, otros) sin importar que este caso tenga asociación clínico-epidemiológica o dictaminación a COVID-19.

Así también, se descartaron los registros incompletos, por lo que fueron eliminados aquellos registros con valores 97 (no aplica), 98 (se ignora) o 99 (no especificado) en alguna de las variables seleccionadas. La única excepción a esta regla fue la variable '**INTUBADO'**, ya que no aplica si el paciente no fue hospitalizado, por lo que simplemente se reemplazó el valor de 97 por 0 (no intubado).

3.2.2 Delimitación de olas epidemiológicas de contagios

Haciendo uso de la base de datos del gobierno de México (http://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip) con fecha de acceso 30 de abril del 2022 [5], se realizó la gráfica que se muestra en la Fig. 7, la cual muestra el número de casos positivos de COVID-19 por día de la fecha 23/02/2020 a 15/04/2022. Se escogió este intervalo ya que, aunque la base de datos se actualice diariamente, se reporta un retraso de dos semanas aproximadamente en los registros [14]. Se utilizaron las variables **'FECHA_SINTOMAS'**, para delimitar la fecha del caso de COVID-19, **'FECHA_DEF'** para la fecha de fallecimiento, **'INTUBADO',** para identificar si el caso presentó intubación y **'TIPO_PACIENTE'** para

determinar si el paciente se encontró hospitalizado. Además, para suavizar las líneas de casos positivos, hospitalizados, intubados y fallecidos, se calculó un promedio semanal de estas incidencias. Por otra parte, es importante notar que los datos se graficaron en escalas distintas. De 0 a 80,000 para los casos confirmados y de 0 a 4,000 para los otros casos (hospitalizados, intubados y muertes). Al mismo tiempo, los datos parten de la novena semana epidemiológica (**SE9**) del 2020, lo cual equivale al 23 de febrero de ese año y terminan el 15 de abril del 2022, fecha que forma parte de la **SE15** de dicho año. Finalmente, las líneas punteadas de colores hacen referencia a los intervalos de las olas definidas. De modo que los intervalos serían:

- Primera ola (negro): **SE22** del 2020 (2020/05/24) – **SE32** del 2020 (2020/08/08)
- Segunda ola (rosa): **SE48** del 2020 (2020/11/22) – **SE06** del 2021 (2021/02/13)
- Tercera ola (verde): **SE22** del 2021 (2021/05/30) – **SE42** del 2021 (2021/10/23)
- Cuarta ola (azul): **SE51** del 2021 (2021/12/19) – **SE15** del 2022 (2022/04/15)

Estos intervalos se obtuvieron a partir de los reportes de la Secretaría de Salud de México, los cuales fueron generados por la dirección de epidemiología y la dirección de información epidemiológica [37].

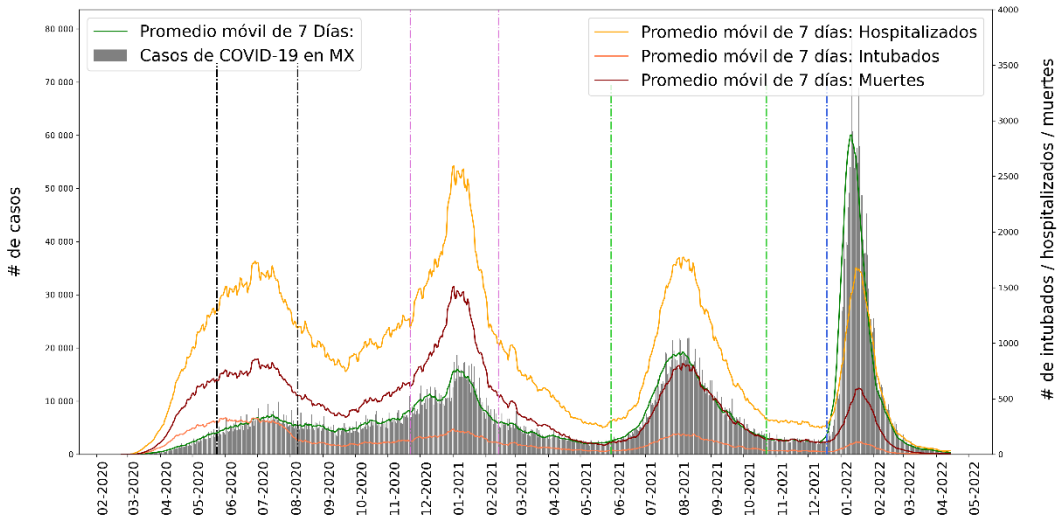

Fig. 7 Casos de intubados, hospitalizados y fallecidos por COVID-19 en México, abarca desde el 23/02/2020 hasta el 15/04/2022.

3.2.3 Ingeniería de características

En esta sección se describe la selección y transformación de variables con el objetivo de ser usadas posteriormente para desarrollar un modelo de ML. Las variables generadas por medio de ingeniería de características se muestran en la Tabla 6 y se describen con más a detalle en esta sección.

La primera variable generada fue la **defunción** con base en la fecha de fallecimiento de los pacientes. Si un paciente no presentó fecha de defunción (mostrando el valor '9999-99-99' en la base de datos) se le asignó un valor de 0 a esta variable. Por el contrario, si presentó alguna fecha se le asigna un valor de 1. La variable de defunción es la variable objetivo del presente trabajo.

Una problemática con la naturaleza de las variables de nuestra base de datos se ejemplifica a continuación. La variable **'EDAD'** en la base de datos contiene un valor cuantitativo discreto. Por otro lado, una parte importante de las variables utilizadas de la base de datos son cualitativas o categóricas (ej. todas las comorbilidades). Es necesario resaltar que, en este contexto, una variable cuantitativa expresa un valor numérico. Mientras tanto, una variable cualitativa es una etiqueta o código para describir si se tiene o no cierta condición. El problema de esta situación se debe a que en los modelos predictivos es recomendable usar solo un tipo de entrada de datos. Es por esta razón, que en el presente trabajo se utilizó el enfoque tradicional de definir intervalos para posteriormente crear variables indicadoras binarias [38]. Los intervalos generados para la variable edad se basaron en la estrategia de vacunación del gobierno de México [39] y se describen en la Tabla 6.

La variable **'SECTOR'** Identifica el tipo de institución del Sistema Nacional de Salud que brindó la atención. Es una variable categórica con 14 posibles valores descritos en la Tabla 5. Se agruparon las unidades médicas con una baja frecuencia de registros para generar nuevas variables, pero evitando generar demasiadas. Todo sector con menos de 100,000 registros (1.7% de los casos positivos totales aproximadamente) es agrupado junto a otras unidades médicas en una variable nueva, en la Tabla 6 se describen las variables generadas.

Tabla 5 Catálogo para variable 'SECTOR' [36].

CLAVE	DESCRIPCIÓN
1	CRUZ ROJA
2	DIF
3	ESTATAL
4	IMSS
5	IMSS-BIENESTAR
6	ISSSTE
7	MUNICIPAL
8	PEMEX
9	PRIVADA
10	SEDENA
11	SEMAR
12	SSA
13	UNIVERSITARIO
99	NO ESPECIFICADO

Además, se anexa la variable de IDH del municipio de residencia del paciente, la cual está basada en el IDH municipal reportado por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) en el 2015 [40]. Para ello, se hizo uso de la clave de estado y municipio de residencia de cada individuo. Posteriormente, con base en esta información se unen los dos conjuntos de datos, obteniendo así el valor del IDH del municipio de residencia de cada individuo. Es importante

mencionar que en el conjunto de datos de IDH se rellenaron los valores faltantes en municipios con el IDH del estado. Finalmente, se divide la variable en intervalos para mantenerla en formato binario, siendo estos los siguientes:

- **IDH_MUY_ALTO:** [0.85, 1]
- **IDH_ALTO:** [0.825, 0.85)
- **IDH_MEDIO_ALTO:** [0.8, 0.825)
- **IDH_MEDIO:** [0.75, 0.8)
- **IDH_BAJO:** [0.7, 0.75)
- **IDH_MUY_BAJO:** [0, 0.7)

Tabla 6 Variables generadas a partir de métodos de ingeniería de características.

Variable	Descripción
DEFUNCIÓN	Indica si el paciente falleció y se genera a partir de la variable **'FECHA_DEF'**. Es la variable objetivo en el presente trabajo.
INTERVALOS DE EDAD	La variable '**EDAD'** se dividió en intervalos, generando las siguientes nuevas variables: **'E0_17', 'E18_29', 'E_30_39', 'E40_49', 'E50_59', 'E60'.** Esta última representa el intervalo de 60 años en adelante. En las demás el dígito previo al guion bajo representa el límite inferior de años y el derecho el superior.
SECTOR	Indica la unidad médica (UM) que registró el caso del individuo. Esta variable se dividió en las siguientes variables: **'H_IMSS', 'H_ISSSTE', 'H_MILITAR', 'H_PRIVADA', 'H_SSA' y 'H_OTRO'.** La variable de hospital militar (**H_MILITAR**) abarca las UM de SEDENA y SEMAR. La variable de otro hospital (**H_OTRO**) combina las UM con menor incidencia, las cuales son: PEMEX, IMSS Bienestar, hospital estatal, cruz roja, hospital universitario y DIF.
ÍNDICE DE DESARROLLO HUMANO (IDH)	El IDH se obtuvo de la base de datos del CONEVAL del 2015 [40] y se dividió en intervalos, generando las siguientes variables adicionales: **'IDH_MUY_BAJO', 'IDH_BAJO', 'IDH_MEDIO', 'IDH_MED_ALTO', 'IDH_ALTO', 'IDH_MUY_ALTO'.**

3.3 Creación de conjuntos por ola y general

Una vez realizado el preprocesamiento, se generaron los cuatro conjuntos de datos que fueron delimitados por los intervalos de fechas de cada oleada, agregando un conjunto general con todos los individuos desde el inicio de la primera ola hasta el final de la cuarta ola (incluyendo el periodo inter-ola).

3.4 Preparación de subconjuntos

La base de datos presentó un desbalance de casos fallecidos y sobrevivientes, solo un 5.7% de los registros totales son defunciones. Un desbalance tan importante no es un tema fácil de manejar. En [14], propusieron realizar un balance al conjunto de datos de manera aleatoria, esto es, forzar el balance a través de descartar casos de sobrevivientes hasta tener el 50% de casos tanto de sobrevivientes como fallecidos. Con esta metodología, los resultados de exactitud, especificidad y sensibilidad de los modelos predictivos binarios fueron superiores al 90%. Se adaptó esta idea para el presente trabajo, sin embargo, para no perder registros de pacientes sobrevivientes se realizaron **divisiones** para cada conjunto de datos. Cada una de estas divisiones contiene los mismos registros de difuntos, pero distintos registros de sobrevivientes, esto con la finalidad de tomar en cuenta la totalidad de los casos de sobrevivientes. Ya que cada conjunto (ola epidemiológica) presentó diferentes desbalances de clases, la cantidad de divisiones generadas fue distinta para cada uno. En general, se generaron divisiones hasta que cada una de estas presentara un balance muy cercano al 50/50 (sobrevivientes/defunciones).

3.5 Modelo predictivo

El lenguaje de programación utilizado para implementar los modelos fue Python (ver. 3.8.3). Se utilizó la paquetería de *XGBoost* (ver. 1.2.1) para entrenar los modelos predictivos debido a que XGBoost es el método más utilizado junto a SHAP reportado en trabajos anteriores de la misma naturaleza (p. ej., [17], [20]–[22]). A continuación, se describen los pasos realizados para entrenar los modelos y realizar la interpretación de estos haciendo uso de la paquetería *SHAP* (ver. 0.37.0). Las variables utilizadas para el entrenamiento de los modelos se muestran en la Fig. 8, siendo un total de 30 además de la defunción que fue la variable a predecir por los modelos.

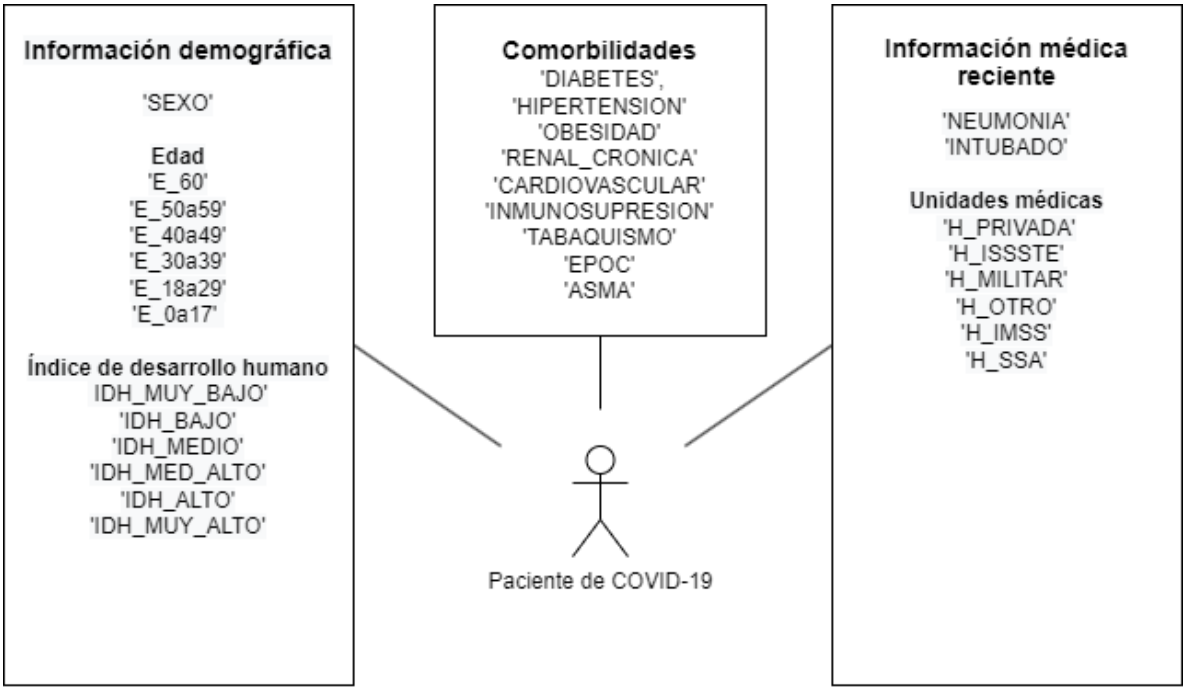

Fig. 8 Variables utilizadas para entrenar los modelos de XGBoost.

3.5.1 Búsqueda de hiperparámetros

Se realizó una búsqueda de cuadrícula (*grid search*) para encontrar las mejores combinaciones de parámetros para los modelos predictivos. Para ello se aplicó una validación cruzada de 2 dobleces y la semilla de aleatoriedad 196. El objetivo de aprendizaje fue la regresión logística para clasificación binaria. La métrica de evaluación fue ROC AUC. Debido a la alta exigencia computacional, la búsqueda de hiperparámetros se realizó solo en el modelo de la primera división y la búsqueda se realizó por segmentos. Para el primer segmento, se optimizaron los parámetros ***max_depth*** y ***min_child_weight***. Para el segundo, se optimizó el parámetro ***gamma*****.** Para el tercero, se optimizaron los parámetros ***subsample*** y ***colsample_bytree*****.** Finalmente, para el cuarto se optimizaron los parámetros ***reg_lambda*** y ***n_estimators*****.** El listado completo de los parámetros a optimizar para XGBoost se muestra en la Tabla 7.

Tabla 7 Listado de parámetros a optimizar para XGBoost.

Parámetros	**Inicial**	**Rango**
max_depth	3	[3, 15, 2]
min_child_weight	1	[1, 6, 1]
gamma	0	[0, 0.5, 0.1]
subsample	0.8	[0.1, 1, 0.1]
colsample_bytree	0.8	[0.2, 1, 0.1]
reg_lambda	0.5	[0.1, 1, 0.1]
n_estimators	500	[50, 500, 50]

3.5.2 Entrenamiento del modelo

Se reservaron 20% de los registros para la etapa de prueba del modelo y el 80% restante para la etapa de entrenamiento. Esta partición de los datos se realizó en cada división junto a una validación cruzada de 3 dobleces. Se mantuvo la misma semilla de aleatoriedad y objetivo de aprendizaje con respecto a la búsqueda de hiperparámetros.

3.6 Interpretación del modelo

Haciendo uso del Algoritmo SHAP se realizó la explicación del modelo y se obtuvieron los valores de SHAP, los cuáles indican las características más importantes del modelo para predecir la defunción en pacientes de COVID-19. Además, los valores de SHAP de cada variable indican si la variable contribuye positiva o negativamente a la predicción de la defunción.

3.6.1 Algoritmo SHAP

En esta etapa, se interpretaron los modelos previamente entrenados y se calcularon los valores de SHAP para cada característica. Además, se calcularon los promedios y medianas (de los valores de SHAP) de todas las divisiones de cada conjunto y se generaron representaciones visuales con el fin de dar una mejor interpretación. El análisis de resultados se muestra en el capítulo siguiente.

4 Resultados

En este capítulo se presentan los resultados de la aplicación de la metodología anteriormente descrita, además de una discusión de los hallazgos obtenidos. Adicionalmente, se describen algunos métodos elegidos para ayudar a la claridad, orden y análisis de los resultados.

4.1 Resultados de los modelos predictivos

En la Tabla 8 se muestra la cantidad de registros de cada conjunto, así como sus divisiones y la cantidad de registros por división. La cantidad de difuntos y sobrevivientes en cada conjunto refleja un desbalance de clases distinto. Por lo tanto, el número de divisiones necesario es distinto para cada conjunto, con la finalidad de mantener un balance aproximado de 50/50 en cada división. En la Tabla 9 se muestran los promedios de los resultados de prueba y entrenamientos de los modelos predictivos. Es importante recordar que para cada modelo se generó una distinta cantidad de divisiones, por lo tanto, estos valores resultan de promediar el número total de divisiones de cada conjunto. Los experimentos realizados en el presente trabajo se llevaron a cabo en una computadora personal con un procesador Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz y con 24GB de RAM.

Tabla 8 Descripción de cada conjunto de datos y sus respectivas divisiones.

Conjunto	Difuntos	Sobrevivientes	Divisiones**	Difuntos por división	Sobrevivientes por división
Oleada 1	55,363 (*14%)	394,086	8	55,363 (52.9%)	49,261 (47.1%)
Oleada 2	82,439 (*10%)	834,589	11	82,439 (52%)	75,872 (48%)
Oleada 3	58,111 (*4%)	1,317,297	22	58,111 (49.3%)	59,878 (50.7%)
Oleada 4	20,718 (*1%)	1,705,858	80	20,718 (49.3%)	21,324 (50.7%)
Conjunto general	297,719 (*6%)	5,269,013	16	297,719 (47.5%)	329,314 (52.5%)

**Letalidad = (fallecidos⁄sobrevivientes) × 100; ** Es el número de submuestreos realizados en sobrevivientes para balancear las clases.*

Tabla 9 Promedios de resultados de los modelos predictivos entrenados para cada conjunto.

Subconjunto	Exactitud	Precisión	*Recall*	ROC AUC
Oleada 1 entrenamiento	0.90	0.89	0.92	0.96
Oleada 1 prueba	0.87	0.86	0.90	0.93
Oleada 2 entrenamiento	0.92	0.91	0.94	0.97
Oleada 2 prueba	0.84	0.83	0.93	0.88
Oleada 3 entrenamiento	0.93	0.93	0.93	0.98
Oleada 3 prueba	0.84	0.83	0.88	0.91
Oleada 4 entrenamiento	0.96	0.96	0.96	0.99
Oleada 4 prueba	0.87	0.87	0.93	0.94
Conjunto general entrenamiento	0.93	0.92	0.93	0.97
Conjunto general prueba	0.85	0.82	0.91	0.90

4.2 Valores de SHAP

Para esta sección es importante tomar en cuenta que los resultados obtenidos hacen referencia al comportamiento de los modelos predictivos entrenados. Además, se denomina **factor de riesgo** a una característica cuando aumenta positivamente el riesgo del paciente de fallecer según el modelo predictivo. Por el contrario, al denominar una característica como **factor de protección** es en referencia a una disminución en el riesgo de fallecer. Para definir cuáles son las variables que son factores de riesgo y cuáles de protección generalmente se observan las gráficas de resumen de SHAP. En la Fig. 9 se muestra la gráfica de resumen para una división del conjunto general. Sin embargo, ya que en el presente trabajo se entrenaron demasiados modelos, se propuso el siguiente método para determinar que variables presentaron un comportamiento de factor de riesgo o protección. No obstante, se anexa un enlace para consultar las 16 gráficas de resumen generadas para el conjunto general [41].

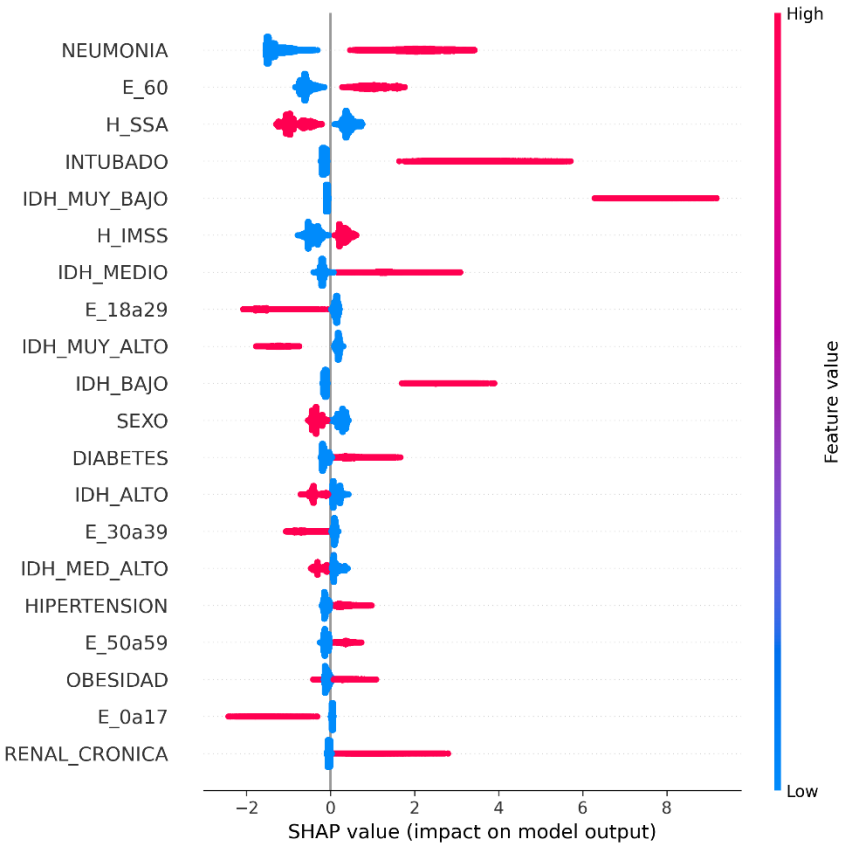

Fig. 9 Gráfica de resumen de SHAP de una división del conjunto general. El color rojo hace referencia al valor 1 y el azul al valor 0 en la variable.

Para llevar a cabo una mejor visualización las variables fueron asignadas a un color único, este color fue determinado según si su comportamiento era el de un **factor de riesgo**, donde el valor 1 de la variable provoca que el modelo aumente la probabilidad de predicción de defunción, o un **factor de protección**, donde el valor 1 de la variable provoca que el modelo reduzca la probabilidad de predicción de defunción. Para determinar estos comportamientos se utilizó el conjunto general que engloba las 4 olas epidemiológicas. Debido a que este conjunto cuenta con 16 divisiones determinar el comportamiento de cada variable no es trivial puesto que los valores de SHAP calculados son distintos para cada registro de cada división. En la Fig. 10 y Fig. 11 se muestran dos gráficas de fuerza, las cuales permiten observar como las variables contribuyen a la predicción del modelo en un registro específico. Los valores de SHAP explican como el modelo toma la decisión de predecir si el paciente fallece o no. La manera en que esto se lleva acabo es partiendo de un valor base (0) y a partir de ahí sumar o restar los valores de SHAP de cada característica hasta llegar a un resultado. Si el valor resultante es positivo el modelo predice que el paciente falleció y viceversa. En la Fig. 10 se observa que las tres variables que el modelo consideró más importantes para predecir la defunción (en color rojo) de este paciente en particular fueron la neumonía, la diabetes y el ser hombre, mientras que las dos variables que contribuyeron a la predicción contraria (en color azul) fueron el no pertenecer a la tercera edad ni haber sido intubado. Sin embargo, se puede observar como la suma de las variables en color rojo fue mayor por lo que la predicción del modelo fue que el paciente falleció. Es importante denotar que todas las variables contribuyen en cierta medida (con la

excepción de que su valor de SHAP sea exactamente 0), pero las que se muestran explícitamente en las Fig. 10 y Fig. 11, fueron las de mayor valor para sumar o restar al valor base.

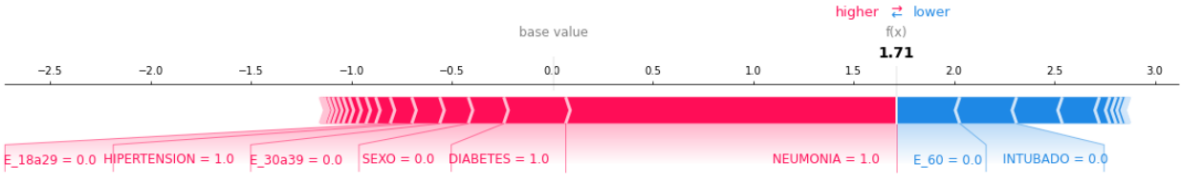

Fig. 10 Gráfica de fuerza para un individuo aleatorio del conjunto general (1).

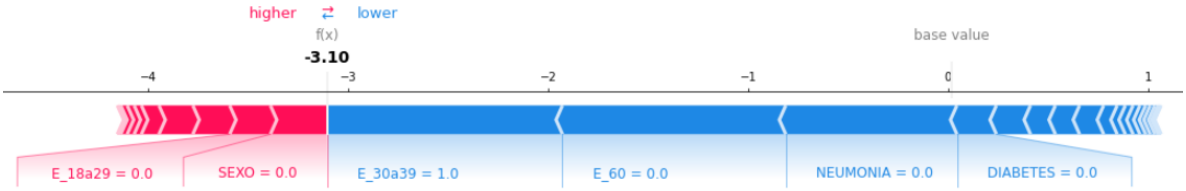

Fig. 11 Gráfica de fuerza para un individuo aleatorio del conjunto general (2).

De lo anterior se puede entender que las variables no siempre tendrán el mismo impacto ni necesariamente la misma dirección (valor de SHAP positivo o negativo) pues los valores de SHAP varían dependiendo del registro. Además, esta variación aumenta en este trabajo en particular porque se tienen distintos modelos por cada conjunto de datos. Por lo tanto, para clasificar las variables se calculó la **correlación punto biserial** entre el valor de SHAP y el valor de entrada de cada variable. Posteriormente, si la relación fue positiva, entonces se le asignó un valor de 1 a esa variable, de lo contrario, se le asignó 0. Esto se realizó para cada división del conjunto general (16 divisiones) y se realizó una sumatoria que se muestra en la Fig. 12. Se establecieron cuatro colores para los siguientes escenarios: (i) Rojo, variables con una sumatoria igual o mayor a 14 (aproximadamente el 10% más alto); (ii) Morado, variables con una sumatoria menor a 14 y mayor o igual a 8; (iii) Verde, variables con una sumatoria menor que 8 y mayor que 2; y, (iv) Azul, variables con una sumatoria igual o menor a 2 (aproximadamente el 10% más bajo).

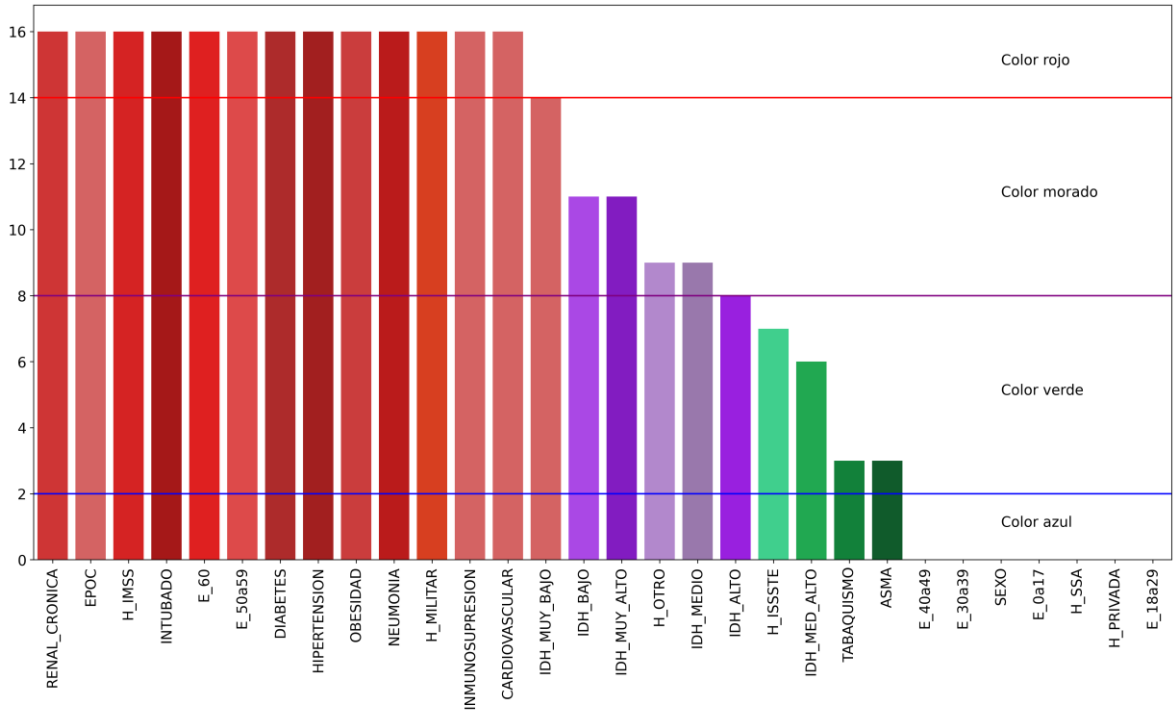

Fig. 12 Sumatoria de las veces que se encontró una correlación positiva entre una variable y sus valores de SHAP en las 16 divisiones del conjunto general.

En la Fig. 14 se muestran los promedios de las medianas de valores de SHAP para las 20 características más importantes para el conjunto de datos general, mientras que en la Fig. 14 se muestra una gráfica de cajas indicando la mediana de los valores de SHAP para las 16 divisiones del conjunto general. Para las gráficas de cajas se establecieron tres umbrales de manera arbitraria para clasificar la importancia de las variables, los umbrales (con base en las medianas de los valores de SHAP) se definen del siguiente modo: (i) Rojo, con un valor de 0.1, las variables que se encuentran debajo de este umbral no fueron consideradas significantes, aproximadamente el 50% de las variables se encuentran debajo de este; (ii) Azul turquesa, con un valor de 0.25, las variables que se encuentran en este intervalo [0.1 – 0.25] se consideran de moderada importancia; y, (iii) Purpura, con un valor de 0.5, las variables dentro de este intervalo (0.25 – 0.5] son consideradas como variables de alta importancia y las variables sobre este umbral son consideradas como las variables más importantes para la predicción de la defunción. En la Tabla 10 se muestran las variables de moderada y alta importancia para cada conjunto de datos.

Como se aprecia en la Fig. 14 y Fig. 14, para el conjunto general la **neumonía** fue la variable de mayor importancia, la cual presentó un claro comportamiento de factor de riesgo al estar clasificada en color rojo. La siguiente variable más importante fue pertenecer a la **tercera edad**, clasificada también como un factor de riesgo. Posteriormente, en el intervalo de variables de alta importancia se encontró el ser hospitalizado en el IMSS, la cual fue clasificada consistentemente como factor de riesgo. Posteriormente, entre las variables de moderada importancia destacaron el pertenecer al género femenino (**'SEXO'**) y el haber sido un caso registrado por la Secretaría de Salud (**'H_SSA'**). Se puede observar en la Fig. 12 que este par de variables presentaron un comportamiento de factor de protección, además, la variable **'SEXO'** tuvo una variabilidad casi nula a través de todas las divisiones

del conjunto general. Por el contrario, la variable '**H_SSA**' presentó una variabilidad muy elevada. Con respecto a las variables relacionadas al IDH, estas presentaron una variabilidad muy elevada y no tuvieron un comportamiento consistente como factor de riesgo o protección, sin embargo, no quedaron bajo el umbral de la irrelevancia. Por otra parte, el haber sido intubado y estar entre los 50 y 59 años figuraron consistentemente como factores de riesgo entre las variables de moderada importancia. Con respecto a comorbilidades, la diabetes y la hipertensión se mantuvieron entre las variables de moderada importancia con un comportamiento consistente de factor de riesgo. Finalmente, haber tenido una edad entre 18 y 29 años o entre 30 y 39 años se consideró un consistente factor de protección entre las variables de moderada importancia.

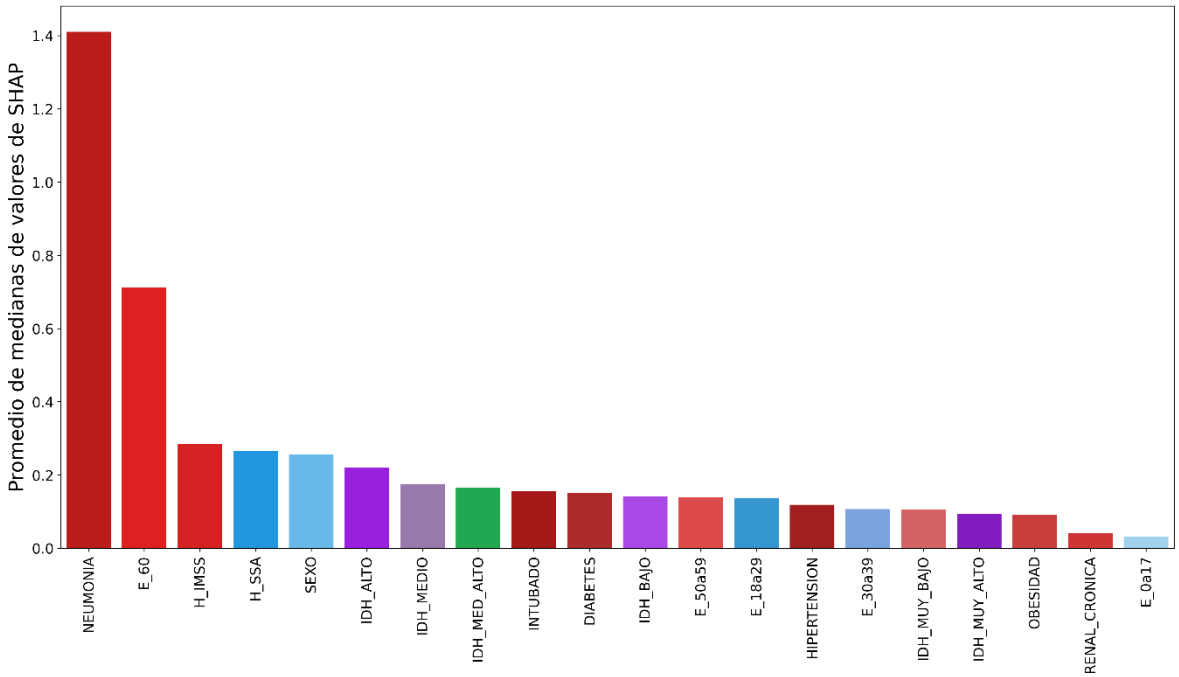

Fig. 13 Promedio de medianas de valores de SHAP para el conjunto de datos general.

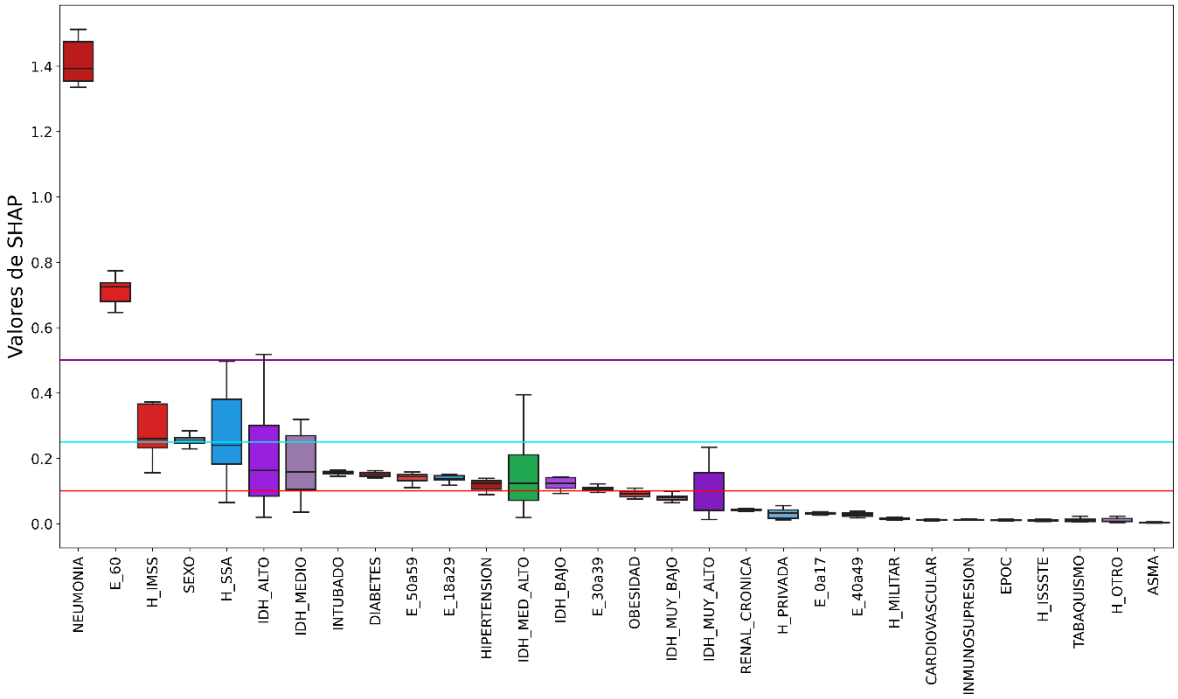

Fig. 14 Gráfica de cajas (boxplot) de las 16 divisiones del conjunto general.

4.3 Análisis de olas epidemiológicas

En esta sección se analizan las gráficas de los resultados para los conjuntos por ola epidemiológica de México. En la Fig. 15 se muestran las 15 variables más relevantes para los cuatro conjuntos y en la Fig. 16 las gráficas de caja para estos mismos conjuntos.

Primeramente, se puede observar que el padecer **neumonía** fue el mayor factor de riesgo en todas las olas excepto en la cuarta, donde pertenecer a la **tercera edad** fue el mayor factor de riesgo. Sin embargo, es importante destacar que la neumonía generalmente es una condición que surge como consecuencia del COVID-19, incluso en un trabajo anterior [14]. se utiliza la variable de neumonía para determinar si el caso de COVID-19 es más grave y así generar distintos subconjuntos, separándola de las comorbilidades que el paciente padecía desde antes de ser un caso confirmado COVID-19. No obstante, en dicho trabajo se obtuvieron mejores resultados al utilizar el subconjunto que incluía la variable de neumonía en el modelo predictivo. En segundo lugar, se encuentra el tener una edad mayor a 60 años para todas las olas, exceptuando en la cuarta. Estas dos variables representan los dos factores de riesgo más importantes encontrados por los modelos. Posteriormente, en la Tabla 10 se encuentran las variables que son clasificadas como de moderada y alta importancia según los umbrales explicados anteriormente.

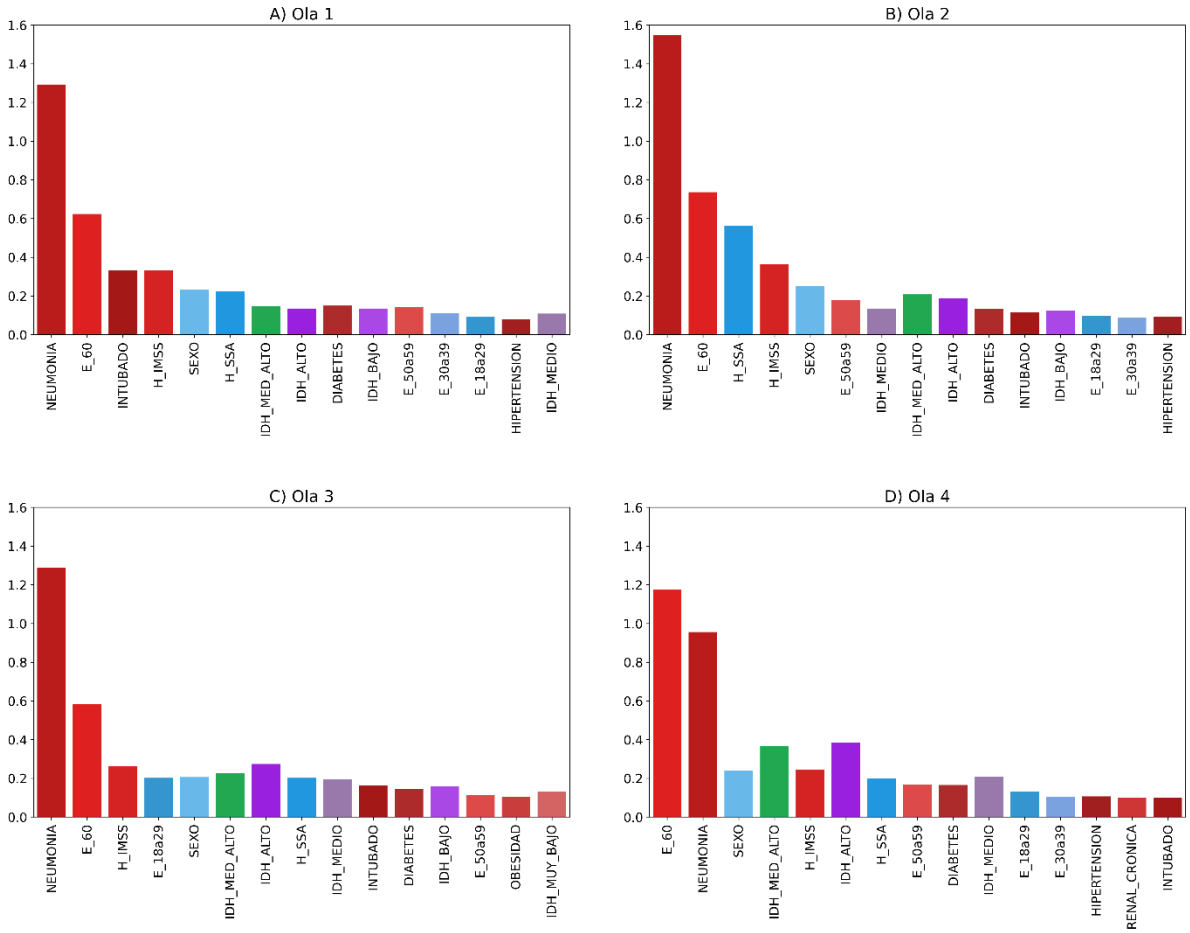

Fig. 15 Promedios de medianas de valores de SHAP para las 15 variables más relevantes de cada oleada.

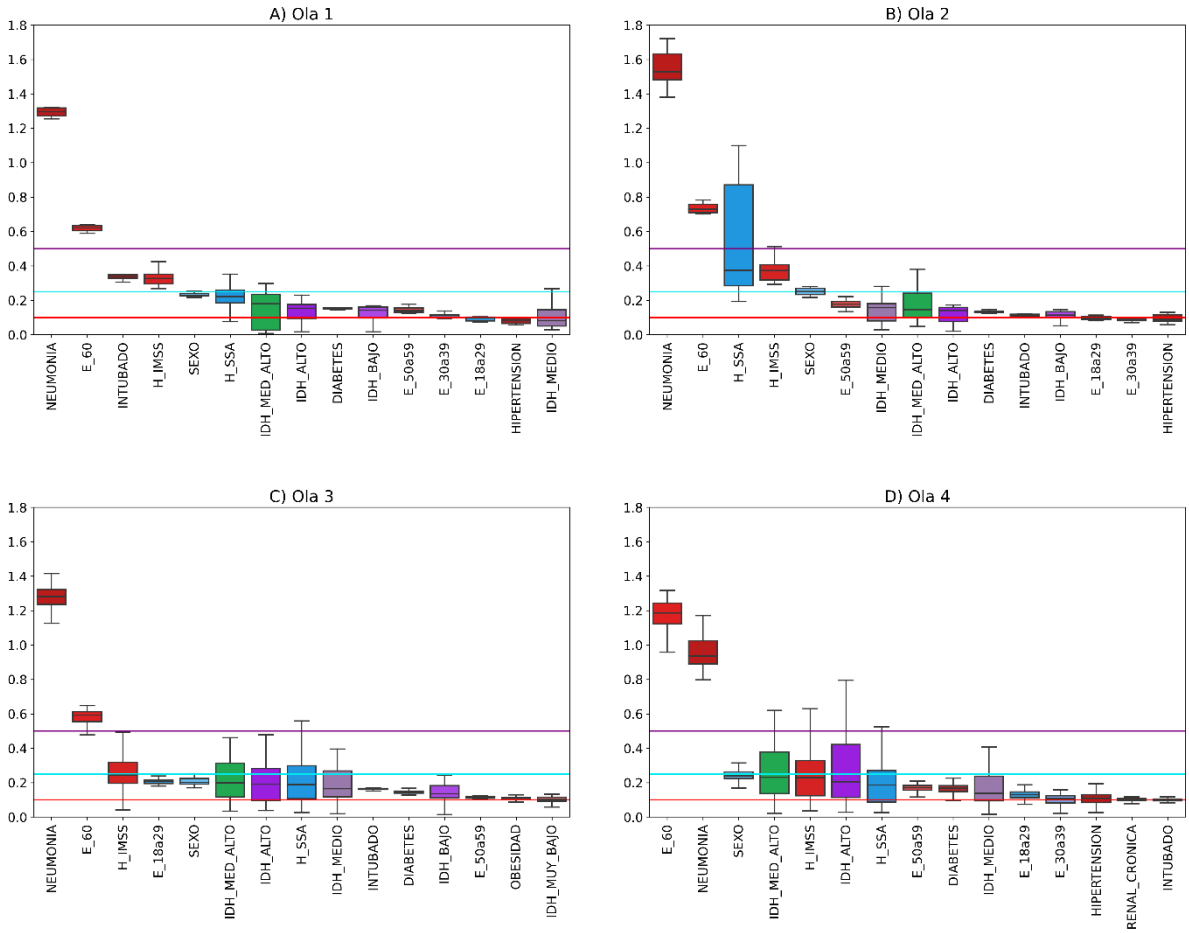

Fig. 16 Gráfica de cajas de valores de SHAP para las 15 variables más relevantes por oleada

Tabla 10 Variables de moderada y alta importancia en los conjuntos de datos. Intervalos definidos con base en las medias de los valores de SHAP. Las variables listadas se encuentran en orden descendente con respecto a su media de valor de SHAP.

Ola	Variables de moderada importancia [0.1 – 0.25)	Variables de alta importancia [0.25 – 0.5)	Variables de más alta importancia [0.5, ∞)
1	SEXO, H_SSA, IDH_MED_ALTO, IDH_ALTO, DIABETES, IDH_BAJO, E_50a59, E_30a39, E_18a29.	INTUBADO, H_IMSS.	NEUMONÍA, EDAD_60.
2	E_50a59, IDH_MEDIO, IDH_MED_ALTO, IDH_ALTO, DIABETES, INTUBADO, IDH_BAJO, E_18a29.	H_SSA, H_IMSS, SEXO.	NEUMONÍA, EDAD_60.
3	E_18a29, SEXO, IDH_MED_ALTO, IDH_ALTO, H_SSA, IDH_MEDIO, INTUBADO, DIABETES, IDH_BAJO, E_50a59, OBESIDAD.	H_IMSS.	NEUMONÍA, EDAD_60.
4	SEXO , IDH_MED_ALTO, H_IMSS, IDH_ALTO, H_SSA, E_50a59, DIABETES, IDH_MEDIO, E_18a29, E_30a39, HIPERTENSIÓN, RENAL_CRÓNICA, INTUBADO, IDH_BAJO.		EDAD_60, NEUMONÍA.
General	H_SSA , IDH_ALTO , IDH_MEDIO, INTUBADO, DIABETES, E_50a59, E_18a29, HIPERTENSIÓN, IDH_MED_ALTO , IDH_BAJO, E_30a39.	H_IMSS, SEXO.	NEUMONÍA, EDAD_60.

5 Discusión

En el presente capítulo se comparan los resultados obtenidos con algunos reportados en la literatura. Adicionalmente, en la Tabla 11 y

Tabla 12 se sintetizan los hallazgos con respecto a los factores de riesgo al padecer COVID-19 de dos revisiones literarias publicadas previamente [42], [50]. La discusión se basó principalmente en la Fig. 16 por el uso de la mediana y la separación de los conjuntos por ola.

La **neumonía** es considerada por los modelos como el mayor factor de riesgo de fallecimiento, por ser la consecuencia más grave de la enfermedad por COVID-19. Esta complicación puede conllevar a un estado crítico al paciente [44], por lo que es reportado consistentemente en la revisión de literatura como un alto factor de riesgo [43]. Por otro lado, la **edad avanzada** fue la característica más destacada como factor de riesgo de letalidad para los modelos presentados en el presente trabajo, así como también en un estudio reciente [12].

El hecho de que el pertenecer a la tercera edad no haya tenido un aumento importante si no hasta la cuarta ola epidemiológica se puede deber a lo siguiente: (i) En la primera ola epidemiológica, el índice de letalidad fue el más elevado (14%), la población no estaba preparada para la pandemia y fallecieron personas de todas las edades. De esto mismo se desprende que el ser **intubado** presentó su máxima importancia como factor de riesgo en esta oleada; (ii) En la segunda ola epidemiológica se tiene aún un índice de letalidad elevado (10%) y aunque la tercera edad haya tenido un ligero incremento en importancia, tal vez no fue tan notable por el inicio de la jornada de vacunación o las medidas preventivas especiales para gente de la tercera edad; (iii) Para la tercera ola epidemiológica (con una letalidad del 4%), los adultos mayores ya habían cumplido con un esquema de vacunación completo (2 dosis) [45], por lo que, a pesar de ser población de riesgo, tenían ese estabilizador frente al resto de la población. (iv) Sin embargo, para la cuarta ola epidemiológica, la mayoría de la población estaba terminando su esquema de vacunación por lo que el estabilizador para la tercera edad anteriormente mencionado quedó anulado [46]. Además, la letalidad se redujo en esta ola a 1%. Por lo anterior, se especula que se decrementó la importancia de la neumonía y se aumentó la de la tercera edad, ya que la pandemia se encontraba en un estado más controlado y la mayoría de la población se encontraba con un esquema completo de vacunación para este entonces. Dicho de otro modo, se especula que la edad avanzada tuvo su pico de importancia en la cuarta ola ya que la pandemia estaba en un estado más controlado y la edad avanzada suele ser un factor de riesgo para cualquier enfermedad grave. Por otra parte, esto igual puede deberse a que la tasa de mortalidad en adultos mayores es más elevada [47], por lo que habiendo una letalidad tan baja, puede darse el caso de que múltiples fallecimientos registrados hayan fallecido por otra razón parcialmente ajena al COVID-19 ya que los fallecimientos registrados en la base de datos no necesariamente deben ser directamente por COVID-19.

Los modelos desarrollados en el presente trabajo reflejaron la importancia de la **unidad médica** (UM) donde fue atendido el paciente. Lo cual es congruente con un trabajo reportado [6] en México, en el cual mediante la razón de momios se identificó que el ser **atendido en el IMSS** representó un mayor factor de riesgo con respecto a otras unidades médicas (hasta agosto 15 del 2020). En el presente trabajo el ser atendido en el IMSS se clasificó como una variable de alta importancia en las primeras tres oleadas y el conjunto general, con un consistente comportamiento de factor de riesgo. Su reducción de importancia en la cuarta oleada epidemiológica puede deberse a que el mismo

índice de letalidad bajó a 1% y que la mayoría de la población ya contaba con un esquema de vacunación completo, por lo que debieron complicarse menos los casos. Además, en el mismo trabajo citado [6] se reporta que haber sido atendidos en una **UM de la SSA** o **privada** representa un factor de protección. En el presente trabajo el haber sido atendido en una UM de la SSA o privada consistentemente se clasificó como factor de protección como se observa en la Fig. 12, sin embargo, solo el haber sido atendido en la SSA figuró consistentemente como una variable de moderada importancia durante los periodos de tiempo estudiados en el presente trabajo, destacando únicamente en la segunda oleada epidemiológica como de alta importancia (Tabla 10). Cabe mencionar que esta variable (H_SSA) presentó una alta variabilidad (Fig. 16), esto es probablemente debido a que un caso registrado por una UM no necesariamente es un caso hospitalizado en la UM y en particular esta es la UM que registró más casos (46% del conjunto general).

Los modelos consideran que **pertenecer al género femenino** actúa como factor de protección. Esta variable (**'SEXO'**) se clasifica como de moderada importancia en todas las olas epidemiológicas exceptuando la segunda, donde se clasificó como de alta importancia, así como también, en el conjunto general. Con respecto a la revisiones literarias, en la Tabla 11 se reporta el género masculino como un factor de riesgo de baja consistencia. Sin embargo, en la

Tabla 12 se reporta como de alta consistencia. En un estudio reciente se menciona que la probabilidad de fallecer aumenta con respecto a la edad, pero en una proporción aproximadamente 1.5 veces mayor en hombres que en mujeres [44].

Con respecto a las comorbilidades, **la diabetes** es la única que presenta un comportamiento de factor de riesgo consistente al mismo tiempo que se clasifica como de moderada relevancia en todas las oleadas. Esto es congruente con la literatura, dado que se ha reportado que la diabetes es de las comorbilidades más importantes para agravar la enfermedad por COVID-19 (por ejemplo en: [6], [8]–[10], [12], [42], [43], [48], [49]). También, en la literatura se reporta la obesidad como de alta consistencia y la hipertensión como de consistencia media [42], [43]. Estas últimas dos variables presentaron un consistente comportamiento de factor de riesgo en el presente trabajo (Fig. 12), sin embargo, no figuran en el umbral de la relevancia más que en la tercera y cuarta ola epidemiológica, por lo que su dirección es clara pero la magnitud en que influyen al resultado es baja.

Retomando la variable de la edad, el intervalo que mantuvo un comportamiento consistente de factor de protección y tuvo una relevancia moderada en el conjunto general y en todas las olas epidemiológicas fue el **tener entre 18 a 29 años de edad**. Este valor no tuvo importantes cambios ni variabilidad a través de todas las olas epidemiológicas por lo que se puede considerar como el intervalo de edad que fue menos afectado por el COVID-19 durante el periodo de la pandemia analizado en el presente trabajo.

La inclusión de factores demográficos tales como el **IDH** resultó en un acierto, pues los modelos consideran significativas estas variables (de importancia moderada en su mayoría). Sin embargo, dado que en el presente trabajo se clasificaron las variables como de comportamiento de factor de riesgo o protección con base en el conjunto general que engloba todas las olas epidemiológicas, es probable que el hecho de que estas variables no tuvieran un comportamiento consistente se deba a que, efectivamente, el rol del IDH como factor de riesgo o protección fue cambiando a través de la pandemia. Esto es algo que se podría estudiar más a detalle en un trabajo futuro.

Tabla 11 Síntesis de revisión literaria por Rod et al. (2020) [42]

Factor de Riesgo	Consistencia
Edad avanzada	Alta
Diabetes	Alta
Enfermedad renal crónica	Media
Enfermedad cardiovascular	Media
Género masculino	Baja

Tabla 12 Síntesis de revisión literaria por Gao et al. (2021) [50]

Factor de Riesgo	Consistencia
Edad avanzada	Alta
Diabetes	Alta
Neumonía	Alta
Género masculino	Alta
Obesidad	Alta
Hipertensión	Media
Enfermedad renal crónica	Media
Asma	Baja
Enfermedad Inmunosupresora	Baja

6 Conclusiones

En el presente trabajo se analizan algunas características de pacientes positivos a COVID-19 por medio de una serie de modelos de XGBoost entrenados para predecir la defunción en los pacientes. Se entrenaron múltiples modelos para cada ola epidemiológica en México más un conjunto general que incluye todas las olas y los periodos inter-ola. Se realizaron múltiples divisiones por cada conjunto para utilizar todos los registros positivos de cada uno sin afectar el desempeño de los modelos.

El enfoque principal de este análisis es la interpretación de los modelos de ML. Primeramente, se utilizaron diversas métricas de evaluación para asegurar la calidad de la predicción de los modelos. Posteriormente, se llevó a cabo un análisis de la importancia de las variables haciendo uso de los valores de SHAP. Las variables más importantes en la predicción de la defunción en los pacientes fueron la neumonía y la edad avanzada. Las siguientes variables más importantes fueron las unidades médicas donde haber sido atendido en el IMSS aumentó la probabilidad de fallecer, por el contrario, haber sido atendido en la SSA la redujo, aunque esta variable presentó alta variabilidad, la media de su valor de SHAP la clasificó como una variable de moderada a alta importancia. Con respecto al sexo, el pertenecer al género femenino representó un factor de protección importante. Por otra parte, las variables relacionadas al IDH mostraron una importancia moderada, pero con direcciones inconsistentes (no se pudieron clasificar consistentemente como factores de riesgo o protección), así como también un alto nivel de variabilidad en su impacto en el resultado (valores de SHAP). Con respecto a comorbilidades, se encontró la diabetes como la más importante, siendo esta un factor de riesgo. Por otra parte, el intervalo de edad entre 18 a 29 años fue el que se encontró como un factor de protección relevante y consistente. Finalmente, el haber sido intubado fue un alto factor de riesgo en la primera ola epidemiológica, presentó una reducción de importancia en la segunda y tercera ola hasta que llegó a su nivel de importancia más bajo en la cuarta ola.

7 Limitaciones y perspectivas

El presente trabajo cuenta con importantes limitaciones con respecto al conjunto de datos, ya que este no cuenta con información muy detallada del estado de salud de los pacientes por lo cual entrenar un modelo predictivo para un problema tan complejo presenta complicaciones. Además, presenta un alto desbalance entre los individuos sobrevivientes y fallecidos. En este trabajo se abordó esa problemática realizando divisiones por cada conjunto, sin embargo, en cada división se utilizaron los mismos registros de difuntos, lo que puede no ser el escenario más deseable. Además, la causalidad en la letalidad no puede ser confirmada por los resultados obtenidos en el presente trabajo ya que hay un riesgo de confusión residual por factores desconocidos.

Con respecto a variables, en trabajos futuros se podría explorar más a fondo el IDH en periodos más cortos o específicos, ya que en el presente trabajo se le encontraron comportamientos muy variables, pero no irrelevantes. También se podrían agregar factores como la densidad poblacional, calidad del aire, entre otros.

El presente trabajo fue probado en el contexto específico de la pandemia por COVID-19 en México. Sin embargo, la metodología propuesta es altamente generalizable a otros contextos, incluso a contextos no relacionados a una pandemia. Los resultados obtenidos demuestran que la aplicación de los valores de SHAP a modelos predictivos es un método efectivo para la detección de los factores de riesgo o de protección. Por lo tanto, este trabajo es una importante contribución para estudiar futuros fenómenos, siempre y cuando se cuente con una base de datos fiable. Con respecto a los modelos, se pueden realizar aún distintos ajustes de los hiperparámetros de la técnica XGBoost o incluso probar otras técnicas que puedan dar buenos resultados sin necesidad de tener un conjunto de datos balanceado, esta posibilidad queda abierta a ser explorada pues de lograrse se podrían analizar gráficas de SHAP adicionales que fueron limitadas por la creación de las divisiones.

Finalmente, sería conveniente continuar el análisis haciendo énfasis en los pacientes hospitalizados, esto podría llevarse a cabo teniendo distintos niveles de conjuntos o simplemente haciendo un análisis exclusivo de estos. Sin embargo, también podría solo agregarse la variable hospitalizado junto a otras más a la misma metodología.

8 Referencias

[1] W. H. Organization, “WHO Coronavirus (COVID-19) Dashboard,” 2022. https://covid19.who.int/.

[2] O. P. de la S. (OPS), “La OMS caracteriza a COVID-19 como una pandemia,” 2020.

[3] X. Chen *et al.*, “A systematic review of neurological symptoms and complications of COVID-19,” *J. Neurol.*, vol. 268, no. 2, pp. 392–402, 2021, doi: 10.1007/s00415-020-10067-3.

[4] C. Molnar, B. Boehmke, and B. Greenwell, *Interpretable Machine Learning*. 2022.

[5] “Datos Abiertos de COVID-19.” [Online]. Available: http://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip.

[6] H. Najera and A. G. Ortega-Avila, “Health and Institutional Risk Factors of COVID-19 Mortality in Mexico, 2020,” *Am. J. Prev. Med.*, vol. 60, no. 4, pp. 471–477, 2021, doi: 10.1016/j.amepre.2020.10.015.

[7] G. M. Parra-bracamonte, N. Lopez-villalobos, and F. E. Parra-bracamonte, “Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ’ s public news and information ,” no. January, 2020.

[8] J. E. Salinas-Aguirre, C. Sánchez-García, R. Rodríguez-Sanchez, L. Rodríguez-Muñoz, A. Díaz-Castaño, and R. Bernal-Gómez, “Clinical characteristics and comorbidities associated with mortality in patients with COVID-19 in Coahuila (Mexico),” *Rev. Clin. Esp.*, no. xxxx, pp. 6–10, 2021, doi: 10.1016/j.rce.2020.12.006.

[9] J. E. de la Peña *et al.*, “Hypertension, Diabetes and Obesity, Major Risk Factors for Death in Patients with COVID-19 in Mexico,” *Arch. Med. Res.*, vol. 52, no. 4, pp. 443–449, 2021, doi: 10.1016/j.arcmed.2020.12.002.

[10] A. Ortiz *et al.*, “Chronic kidney disease is a key risk factor for severe COVID-19: A call to action by the ERA-edta,” *Nephrol. Dial. Transplant.*, vol. 36, no. 1, pp. 87–94, 2021, doi: 10.1093/NDT/GFAA314.

[11] K. C. Y. Wong and H.-C. So, “Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK Biobank data,” *medRxiv*, p. 2020.09.18.20197319, 2020, [Online]. Available: https://doi.org/10.1101/2020.09.18.20197319.

[12] S. Tehrani, A. Killander, P. Åstrand, J. Jakobsson, and P. Gille-Johnson, “Risk factors for death in adult COVID-19 patients: Frailty predicts fatal outcome in older patients,” *Int. J. Infect. Dis.*, vol. 102, pp. 415–421, 2021, doi: 10.1016/j.ijid.2020.10.071.

[13] L. A. Chávez-almazán, L. Díaz-gonzález, and M. Rosales-rivera, “Socioeconomic status and its effects on morbidity , mortality , and lethality due to COVID-19 in Mexico,” pp. 1–18.

[14] M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, R. D. J. de León-Montiel, and A. B. U’Ren, “Identification of high-risk COVID-19 patients using machine learning,” *PLoS One*, vol. 16, no. 9 September, pp. 1–21, 2021, doi: 10.1371/journal.pone.0257234.

[15] J. Solis, C. Franco-Paredes, A. F. Henao-Martinez, M. Krsak, and S. M. Zimmer, “Structural vulnerability in the U.S. revealed in three waves of COVID-19,” *Am. J. Trop. Med. Hyg.*, vol. 103, no. 1, pp. 25–278, 2020, doi: 10.4269/ajtmh.20-0391.

[16] A. Di Castelnuovo *et al.*, “Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study,” *Nutr. Metab. Cardiovasc. Dis.*, vol. 30, no. 11, pp. 1899–1913, 2020, doi: 10.1016/j.numecd.2020.07.031.

[17] M. Smith and F. Alvarez, “Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19,” *Expert Syst. Appl.*, vol. 176, no. June 2020, pp. 1–12, 2021, doi: 10.1016/j.eswa.2021.114832.

[18] A. L. Booth, E. Abels, and P. McCaffrey, “Development of a prognostic model for mortality in COVID-19 infection using machine learning,” *Mod. Pathol.*, vol. 34, no. 3, pp. 522–531, 2021, doi: 10.1038/s41379-020-00700-x.

[19] B. Davazdahemami, H. M. Zolbanin, and D. Delen, “An explanatory machine learning framework for studying pandemics: The case of COVID-19 emergency department readmissions,” *Decis. Support Syst.*, no. August 2021, p. 113730, 2022, doi: 10.1016/j.dss.2022.113730.

[20] N. Jing, Z. Shi, Y. Hu, and J. Yuan, “Cross-sectional analysis and data-driven forecasting of confirmed COVID-19 cases,” *Appl. Intell.*, vol. 52, no. 3, pp. 3303–3318, 2022, doi: 10.1007/s10489-021-02616-8.

[21] T. Banerjee, A. Paul, V. Srikanth, and I. Strümke, “Socioeconomic Disparities and COVID-19: The Causal Connections,” *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.4013119.

[22] A. Vaid *et al.*, “Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: Model development and validation,” *J. Med. Internet Res.*, vol. 22, no. 11, pp. 1–19, 2020, doi: 10.2196/24018.

[23] S. Hart, “Shapley Valoe,” pp. 210–216, 1951.

[24] A. Paturel, “Game Theory Game Theory,” *Nav. War Coll. Rev.*, vol. 14, no. 5, pp. 16–42, 2014, [Online]. Available: https://digital-commons.usnwc.edu/nwc-review/vol14/iss5/3.

[25] C. Molnar, “SHAP (SHapley Additive exPlanations),” *Interpretable Machine Learning*, 2022. https://christophm.github.io/interpretable-ml-book/shap.html#fn67.

[26] N. Didrik, “Tree Boosting With XGBoost,” Norwegian University of Science and Technology, 2016.

[27] R. E. Schapire, *Boosting: Foundations and Algorithms*, vol. 42, no. 1. 2013.

[28] Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Assoc. Comput. Mach.*, no. XGBoost: A Scalable Tree Boosting System, pp. 785–794, 216AD.

[29] J. Friedman, R. Tibshirani, and T. Hastie, “Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors),” *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, 2000, doi: 10.1214/aos/1016120463.

[30] “XGBoost Parameters,” 2021. https://xgboost.readthedocs.io/en/stable/parameter.html.

[31] J. Muschelli, “ROC and AUC with a Binary Predictor : a Potentially Misleading Metric,” *J. Classif.*, 2019.

[32] F. Pedragosa, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[33] D. Berrar, “Cross-validation,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. April, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[34] W. Vogt, “Point Biserial Correlation,” *Dict. Stat. Methodol.*, no. 2, pp. 1–3, 2015, doi: 10.4135/9781412983907.n1452.

[35] J. D. Chee and T. Queen, “Pearson ’ s Product-Moment Correlation : Sample Analysis Pearson ’ s Running head : Pearson ’ s Product Moment Correlation Pearson ’ s Product Moment Correlation : Sample Analysis Jennifer Chee University of Hawaii at M ā noa School of Nursing,” *ResearchGate*, no. May 2015, 2016, doi: 10.13140/RG.2.1.1856.2726.

[36] Dirección General de Epidemiología, “Diccionario de datos de COVID-19,” 2021. [Online]. Available: https://epidemiologia.salud.gob.mx/gobmx/salud/datos_abiertos/diccionario_datos_covid19.zip.

[37] Secretaría de Salud de México, “Casos de COVID-19 por olas pandémicas en México,” 2022, [Online]. Available: https://github.com/ItsShinyAlex/IA_Analisis_Datos_Covid19_MX/blob/main/referencias/Casos_de_COVID19_por_olas_pandemicas_en_Mexico.pptx.

[38] F. Pargent, “A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling,” *J. Chem. Inf. Model.*, vol. 53, no. Ludwig-Maximilians-Universität München, pp. 1689–1699, 2019.

[39] R. Cortés Alcalá and H. López-Gatell Ramírez, “Política rectora de vacunación contra covid-19,” 2021. [Online]. Available: https://coronavirus.gob.mx/wp-content/uploads/2021/01/PolVx_COVID_-11Ene2021.pdf.

[40] Consejo Nacional de Evaluación de la Política de Desarrollo Social., “Informe de pobreza en los municipios de México 2015,” Ciudad de México, 2015.

[41] “Summary plots para conjunto general COVID-19 MX.” https://github.com/ItsShinyAlex/IA_Analisis_Datos_Covid19_MX/tree/main/summaries.

[42] J. E. Rod, O. Oviedo-Trespalacios, and J. Cortes-Ramirez, “A brief-review of the risk factors for covid-19 severity,” *Rev. Saude Publica*, vol. 54, pp. 1–11, 2020, doi: 10.11606/S1518-8787.2020054002481.

[43] Y. dong Gao *et al.*, *Risk factors for severe and critically ill COVID-19 patients: A review*, vol. 76, no. 2. 2021.

[44] Q. Zhang *et al.*, “Human genetic and immunological determinants of critical COVID-19 pneumonia,” *Nature*, vol. 603, no. March, pp. 587–598, 2022, doi: 10.1038/s41586-022-04447-0.

[45] Secretaría de Bienestar, “A sus más de cien años, personas adultas mayores reciben la vacuna contra COVID-19 que las protege en nueva batalla,” 2021.

https://www.gob.mx/bienestar/es/articulos/a-sus-mas-de-cien-anos-personas-adultas-mayores-reciben-la-vacuna-contra-covid-19-que-las-protege-en-nueva-batalla?idiom=es#:~:text=El lunes 15 de febrero,por ser las más vulnerables. (accessed Oct. 24, 2022).

[46] Gobierno de México, “Calendario de vacunación,” 2022. https://vacunacovid.gob.mx/calendario-vacunacion/ (accessed Oct. 24, 2022).

[47] G. Soto-Estrada, L. Moreno-Altamirano, and D. Pahua-Díaz, “Epidemiological overview of Mexico’s leading causes of morbidity and mortality,” *Rev. la Fac. Med.*, vol. 59, no. 6, pp. 8–22, 2016, [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0026-17422016000600008.

[48] S. A. Bae, S. R. Kim, M. N. Kim, W. J. Shim, and S. M. Park, “Impact of cardiovascular disease and risk factors on fatal outcomes in patients with COVID-19 according to age: A systematic review and meta-analysis,” *Heart*, vol. 107, no. 5, pp. 373–380, 2021, doi: 10.1136/heartjnl-2020-317901.

[49] W. Guo *et al.*, “Diabetes is a risk factor for the progression and prognosis of COVID-19,” *Diabetes. Metab. Res. Rev.*, vol. 36, no. 7, pp. 1–9, 2020, doi: 10.1002/dmrr.3319.

[50] Y. Gao, M. Ding, and X. Dong, “Risk factors for severe and critically ill COVID-19 patients: A review.”

Cuernavaca, Morelos a 14 de 11 del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, del estudiante **ALEJANDRO CARVANTES BARRERA**, con matricula **10036987**, con el título **Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático,** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. José Alberto Hernández Aguilar
Profesor- investigador
Facultad de Contaduría, Administración e Informática

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JOSE ALBERTO HERNANDEZ AGUILAR | Fecha:2022-11-14 21:37:59 | Firmante

MKeRXwwgS1o6jygd5W8GwCmsBFYATS4MFPF7pLPOQ/t5STFuo1IkQrHsSr6xb5aeeciYmtGMudkLLN3+RvDgktauVsJAFVGqM9Db5bj07jMopMBlqnNoEVdVyIwHXxxNnmbeImIIR4xzvww5y/PmbzMsqoXXG1d9JBt1vLZuOnVg+26z5M5DVTlT1sfcUoxdeKpzqm2hsHuDY5JMR1sb2Y1+EPqcOuwLqn8CPW0/VofAqow9s/x+R/pa0p1pHf4Y26COuyu5HOVpBahDOoz7q3p76EGPShFdbLNte4ov2hejsKaRaAs3ua4XO9dPZoNN7QehP3pX8Vypc7n0XFewTA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

4sVtHZmDA

https://efirma.uaem.mx/noRepudio/Qz2B0NW2FPaIfMgyqG7JLOcD3dHkkyqB

Cuernavaca, Morelos a 14 de 11 del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, del estudiante **ALEJANDRO CARVANTES BARRERA**, con matricula **10036987**, con el título **Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático,** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Blanca Itzelt Taboada Ramírez
Profesor- investigador
Instituto de Biotecnología
Universidad Nacional Autónoma de México

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

BLANCA ITZELT TABOADA RAMÍREZ | Fecha:2022-11-15 00:39:51 | Firmante

qR6EJ+RgfHXoBckaCBTZ7Q4GSeYdsDBUSFIi3yDaw/urGb84ZTz6a/ngVV175u9zjzNLH1iudsahQtlYDe93/PVasYCjDSEC9HtG1yeS43DsetvNwC8B+KV0m9zOfEh7eCnVKDQsw7xyZyVqchrXAr0NAygERPr6YV+iTEaYDakESK05IMbtyZuvi2f1/XgPBj33cizZvv/go+wpa/cYUe0ENu6R1Iej3Ilr4wD+X8i5IQeSzinpV7Mt6TF6/5I+BCvYBh1biExchoS0oja8vuEHD8mlf40VnQS6yk5MFwM23P0USmNXSxMmtIQ4/EK8dphJJtzdtjt2dv7h1blZpQ==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

7ViFuI3Qn

https://efirma.uaem.mx/noRepudio/TJiYQu5wyeLHfIKa6kBTf3Zlkex99EOO

Cuernavaca, Morelos a 14 de 11 del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, del estudiante **ALEJANDRO CARVANTES BARRERA**, con matricula **10036987**, con el título **Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático,** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Luis Alberto Chávez Almazán
Profesor- investigador
Facultad de Ciencias Químico Biológicas
Universidad Autónoma de Guerrero

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LUIS ALBERTO CHÁVEZ ALMAZÁN | Fecha:2022-11-16 23:16:08 | Firmante

yqaw5t8Bk8xOJPPjlX0wRDvJCu5AvDNZFPGjJu/CZKuE4BJTtS3nyBKz8vjtFAMBkVUpoGtiSQhqscIL/JY0+hpqopbvGSIdYDbrgx/w353VC34VC8Has4ySkb/4ZE/HKm1Q4VWsA
TsXR/fsCyjo+9U9Qi5X5cfw5NnwmTkmyxT8YuZD2BnmeQlyJIAnKPO1t8/45DlLrzNQl62MJW9M5dYM70zgxskAKpLJ8PdfOp1dz3Yj6a4beeuN3HGrheKaTQ9DmlK8LMy8mRvdB
+waAdI1wCQ5X9m5Vhr/a0RqgM66PylKPM4EtG2HSBPDTDm+SzJjM5PqmbrOGMZTCMlcPA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

2taqzIYjD

https://efirma.uaem.mx/noRepudio/xronofjWeVIl365FV1Wx92puaBEx5EjQ

Cuernavaca, Morelos a 14 de 11 del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, del estudiante **ALEJANDRO CARVANTES BARRERA**, con matricula **10036987**, con el título **Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático,** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Lorena Díaz González
Profesor- investigador
Centro de Investigación de Ciencias

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LORENA DIAZ GONZALEZ | Fecha:2022-11-15 20:14:40 | Firmante

QovVg02tZxPyEorTbthdsuei8Aig33AOooiftLmAJdTYPpYMzViLJKcrdYg/H3whypnHZMi2gSN4x5Hg/0BcJmiWRKTvYe8U201qo7GAMpQl6cy6CDwSZYweULC637N7m7wujv+ceknm+ek/GHULudse53dPL+7M1mA1caHQ8/C7e3ZPSc5Lh/46S0b0Wc+NorkejNrgTkMezQ7Ke8EPj9jlCDt7y40vlWSq0nuDvMjbOCNXRJ/R5Uoyphoae/I8Oi7/xSL4KR85PXz1VkLfybKYskkHh4DfGCmCJnyOuggRkEiy02eELgcCqokXPNfZkA0uk19t1xzOpMRN45oV+g==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

KgFJEwhyP

https://efirma.uaem.mx/noRepudio/94Rm3ZGA5bZ5VK01VSPweZxqkJkdfuOT

Cuernavaca, Morelos a 14 de 11 del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, del estudiante **ALEJANDRO CARVANTES BARRERA**, con matricula **10036987**, con el título **Identificación de factores asociados a la letalidad por COVID-19 en México mediante el aprendizaje automático,** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Outmane Oubram
Profesor- investigador
Facultad de Ciencias Químicas e Ingeniería

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

OUTMANE OUBRAM | Fecha:2022-11-17 12:46:14 | Firmante

BSLwn7WEXzdLnsotMondP8wDjrPM7oHSlsgajeVp6P6XqwkMseuTox4+G8vnEAqD2jF92+G2jwq0Z0Wco2XJ0d5kk6RKF+a3xkoDW6ak1h+qM1GfjT9yqcAB4bZGYpgKgxSJTf
MLJGTTu5TnfNAochBbIVCCNP+BHW6vfrJCm6uAa/XL9To1vTlkInMjtg3XUQQP/oIBDxxQ3F00PTxocxeyhfwB+DjfcJ2GYs1duCJQzHPW844HVJTMo6c0Nk8vGTGcbLkZVl3Ms
wqlZFvVp46nNWhaQ1pDF/hZ7qdOuWT1dQkYhhz2PGE2L9L0umPvdVzl5zGslkousIECcLENxQ==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

58VxsqPfb

https://efirma.uaem.mx/noRepudio/CoDfSNtePBsOnzTDpIQJkwXfi0kJV7hl