






Inferencia probabilística de eventos asociados al COVID-19 en México Probabilistic inference of events associated with COVID-19 in México

C. Tino-Salgado ^{a,*}, M. Martínez-Arroyo ^a, M. Hernández-Hernández ^b, E. de la Cruz-Gómez ^a,
J. S. Noguera-Bautista ^c

^a Tecnológico Nacional de México campus Acapulco. Carretera Cayaco - Puerto Marqués, Del P.r.i. s/n, Acapulco de Juárez, Gro.

^b Tecnológico Nacional de México campus Chilpancingo. Av. José F. Ruiz Massieu No. 5, Fracc. Villa Moderna, 39090 Chilpancingo de los Bravo, Gro.

^c Hospital General del ISSSTE. Av. Adolfo Ruiz Cortines 124, CTM, 39601 Acapulco de Juárez, Gro.

Resumen

Actualmente, la población mexicana desconoce la probabilidad de presentar eventos agravantes (intubación, ingreso a la unidad de cuidados intensivos y defunción) derivados del COVID-19. Diversos autores han propuesto modelos gráficos probabilísticos para identificar los factores asociados a esta enfermedad. En este documento, se propone utilizar redes bayesianas para identificar las relaciones de dependencia probabilísticas en 23 variables de estudio del conjunto de datos abiertos de COVID-19, proporcionado por la Dirección General de Epidemiología en México durante el periodo 2020 y 2021. Se generaron modelos de redes bayesianas a través de los algoritmos de aprendizaje estructural: PC y Hill Climb Search. Los resultados permitieron determinar que la diabetes, hipertensión y obesidad son los principales factores que inciden en eventos agravantes de COVID-19, así mismo, la probabilidad de defunción depende del grupo de edad del paciente y si fue o no intubado. La red bayesiana como clasificador obtiene al menos un 94% de precisión y exactitud al clasificar eventos agravantes de COVID-19.

Palabras Clave: Redes bayesianas, Aprendizaje estructural, Inferencia probabilística, COVID-19.

Abstract

Currently, the Mexican population is unknown of the probability of presenting aggravating events (intubation, admission to the intensive care unit, and death) derived from COVID-19. Several authors has proposed probabilistic graphical models for identify the factors associated to this disease. In this document, we propose to use bayesian networks to identify probabilistic dependency relationships in 23 study variables from the COVID-19 open data set, provided by the Dirección General de Epidemiología in Mexico during the period 2020 and 2021. Bayesian network models were generated through structural learning algorithms: PC and Hill Climb Search. The results made it possible to determine that diabetes, hypertension and obesity are the main factors that affect aggravating events of COVID-19. Likewise, the probability of death depends on the patient's age group and whether or not he was intubated. The Bayesian network as a classifier obtains at least 94% precision and accuracy when classifying aggravating events of COVID-19.

Keywords: Bayesian networks, Structural learning, Probabilistic inference, COVID-19.

1. Introducción

El COVID-19 es una enfermedad transmitida a través del virus SARS-CoV-2, surgida a finales del año 2019 en Wuhan, China, (OMS, 2020). Hasta este momento, se ha identificado que la transmisión del virus se da a través de gotículas que expulsa una persona enferma al toser, estornudar o hablar; saludar de mano a una persona enferma y al tocar superficies contaminadas y luego llevarse las manos a la boca, ojos y

nariz. Diversos organismos de salud proponen medidas de distanciamiento social como método de prevención, lavado constante de manos y uso de mascarillas, (OPS, 2022).

Hasta finales de 2021, la Organización Mundial de la Salud (OMS) reportó 287,088,164 diagnósticos acumulados de COVID-19, con una tasa de defunción del 1.9% a nivel global, (OMS, 2022). En México, la Secretaría de Salud (SSA) a través de la Dirección General de Epidemiología (DGE) reportó la importación del primer caso de COVID-19 el 28 de

*Autor para la correspondencia: mm20320018@acapulco.tecnm.mx

Correo electrónico: mm20320018@acapulco.tecnm.mx (Cirilo Tino Salgado), miriam.ma@acapulco.tecnm.mx (Miriam Martínez Arroyo), mario.hh@chilpancingo.tecnm.mx (Mario Hernández Hernández), eduardo.dg@acapulco.tecnm.mx (Eduardo de la Cruz Gómez), jsnb101184@hotmail.com (José Samuel Noguera Bautista)

febrero del 2020, desde entonces la cifra de casos confirmados por el virus ha incrementado considerablemente, (Secretaría de Salud, 2020).

México terminó el año 2021 con una tasa de incidencia positiva de COVID-19 de 3,085.7 por cada 100,000 habitantes. Además, reportó 3,979,723 casos positivos acumulados de COVID-19 y 299,428 defunciones, (Gobierno de México, 2021).

La sintomatología del COVID-19 es similar a la de otras enfermedades respiratorias como: influenza y resfriado común. Sin embargo, el COVID-19 se caracteriza por la presencia de fiebre, tos y cefalea. A menudo, también es acompañado por al menos uno de los siguientes signos: disnea, artralgias, mialgias, odinofagia, rinorrea, conjuntivitis o dolor torácico, (Secretaría de salud, 2022). Adicionalmente, existen comorbilidades que agravan la condición de pacientes positivos a SARS-CoV-2, tales como: diabetes, hipertensión, EPOC, enfermedades cardiovasculares, obesidad, enfermedad renal crónica y cáncer, (Gobierno de México, 2021).

La OMS ha autorizado el uso de al menos 8 biológicos contra el COVID-19. No obstante, no es posible asegurar la efectividad de estas vacunas ante el surgimiento de nuevas variantes de SARS-CoV-2, (OPS, 2022). Considerando que cada variante de este virus propicia una sintomatología diferente, es necesario identificar a través del tiempo los factores de riesgo que agravan la condición de los pacientes.

En México, la identificación de factores de riesgo asociados a pacientes positivos a COVID-19 se ha realizado a través de análisis exploratorio de datos, (CONACYT, 2022). No obstante, diversos autores proponen el uso de redes bayesianas para identificar y modelar las relaciones de dependencia probabilísticas entre los factores asociados a esta enfermedad. En Reino Unido, se ha abordado la necesidad de utilizar modelos gráficos probabilísticos para determinar las variables que inciden en las principales estadísticas relacionadas al COVID-19, tales como: la tasa de positividad, la tasa de negatividad y la tasa de mortandad, (Fenton y otros, 2020).

Las redes bayesianas también han sido empleadas para verificar la eficiencia de la clasificación de COVID-19 propuesta por la OMS, (de Terwangne y otros, 2020). Los autores resaltan las propiedades de este algoritmo; señalan que son modelos potentes que representan de forma compacta la distribución de probabilidad conjunta de las variables de estudio. Además, precisan que, una vez definido el modelo es posible calcular la probabilidad posterior de la variable objetivo dado un conjunto de variables observadas.

En China, se propone un sistema de evaluación de riesgos de COVID-19 basado en redes bayesianas, (Wei y otros, 2020). El sistema estima la probabilidad de riesgo epidémico de un paciente considerando un conjunto de síntomas observados del individuo y sus ubicaciones geográficas en los últimos 14 días. En el mencionado estudio, se precisa que, para realizar inferencias más exactas, es indispensable actualizar las probabilidades de las variables de estudio en tiempo cercano al real.

La optimización de una red bayesiana para la clasificación de las distintas fases asociadas al COVID-19: positivo, negativo, defunción y positivo activo puede ser vista en (Ojugo & Otakore, 2021). Los autores consideran una muestra de 4,687 registros con 54 variables de estudio obtenidas de pacientes del Laboratorio de Epidemiología del Centro Médico Federal en Delta, Nigeria. Entre las variables de

estudio se encuentran: información personal, régimen de salud; proximidad entre el individuo y el centro médico; reacciones secundarias y estado del tratamiento. Las relaciones de dependencia fueron creadas a partir del conocimiento empírico de médicos especialistas y algoritmos de aprendizaje estructural.

Considerando las bondades de los modelos gráficos probabilísticos en tareas de clasificación e inferencia, en este documento se propone la creación de una red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México. El desarrollo de esta investigación tiene dos objetivos principales:

1. En primera instancia, identificar las relaciones de dependencia probabilística que permitan detectar grupos vulnerables a eventos agravantes (defunción, ingreso a unidad de cuidados intensivos e intubación mecánica) de COVID-19.
2. Así mismo, determinar la probabilidad de eventos asociados al COVID-19 en México, por ejemplo: determinar la probabilidad de que un paciente ingrese a una unidad de cuidados intensivos dado un conjunto de datos generales del paciente y sus comorbilidades.

En el presente documento, inicialmente, se describen los métodos y materiales empleados para generar una red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México. Seguidamente, se realiza una comparación entre los algoritmos de aprendizaje estructural empleados para la generación de estos modelos. Finalmente, se realizan consultas para determinar la probabilidad de los eventos agravantes anteriormente mencionados.

2. Materiales y métodos

En esta sección, se describen los materiales y métodos empleados para el diseño e implementación de la red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México. La presente investigación se caracteriza por ser un estudio retrospectivo, observacional, transversal y analítico sobre el conjunto de datos abiertos de pacientes positivos a SARS-CoV-2, disponible en (DGE, 2022). La *Figura 1* ilustra de manera gráfica la secuencia de procesos empleados para este estudio.



Figura 1. Proceso general para la creación de la red bayesiana para la inferencia de eventos asociados al COVID-19 en México.

2.1. Análisis descriptivo de datos

Se utilizó una muestra de 16,587 pacientes diagnosticados con COVID-19 a través de pruebas PCR durante el periodo 2020 – 2021. Entre los datos proporcionados se encuentra información sociodemográfica, comorbilidades e información general del individuo. Las variables de interés consideradas en esta investigación han sido definidas por médicos especialistas, que han estado en la primera línea de atención al COVID-19. Las características de los pacientes presentes en la muestra de estudio son descritas en la *Tabla 1*. Adicionalmente, se consideró el sector salud que atendió al paciente y su entidad de residencia. A partir del conjunto muestral fue posible obtener el conjunto de entrenamiento y el conjunto de prueba, con el 75% y 25% de registros, respectivamente.

2.2. Redes bayesianas

Una red bayesiana es un modelo gráfico probabilístico de datos multivariados. De acuerdo con (Scutari & Denis, 2021), una red bayesiana está constituida por los siguientes elementos.

- Un conjunto de variables aleatorias $X = \{X_1, X_2, \dots, X_n\}$ que describen eventos o medidas de interés. La distribución de probabilidad multivariada de X se conoce como distribución global de los datos. Por su parte, cada variable aleatoria $X_i \in X$ está asociada a una distribución de probabilidad que comúnmente es llamada distribución local.
- Un grafo acíclico dirigido, denotado por $G(V, A)$. Cada nodo $v \in V$ está asociado con una variable aleatoria X_i . Cada arista $a \in A$ representa una dependencia probabilística directa. Es decir, si un nodo v_j está conectado a un nodo v_k , la probabilidad de v_k depende de la probabilidad de v_j ; en contra parte, v_k es independientemente condicional si no existe una arista que conecte a v_j con v_k .

$$P(X) = P(X_1 \dots X_n) = \prod_{i=1}^n P(X_i \vee Pa(X_i)) \quad (1)$$

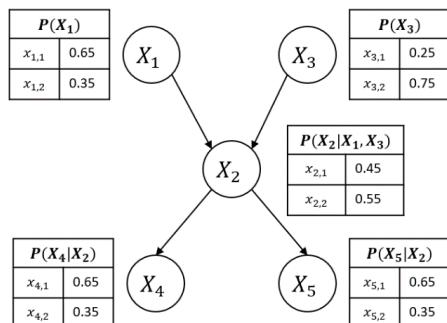


Figura 2. Representación gráfica de una red bayesiana.

La ecuación (1) permite determinar la probabilidad conjunta de X . También es posible observar que, la probabilidad de cualquier nodo está condicionada por la probabilidad de sus nodos padres. De manera interna en una red bayesiana (véase **Error! No se encuentra el origen de la referencia.**), los nodos independientes están asociados a una tabla de

distribución de probabilidad marginal, mientras que los nodos dependientes se asocian a una tabla de distribución de probabilidad condicional.

Tabla 1. Características de los pacientes.

Variable	Valor	Frecuencia	Proporción
Sexo	Femenino	8,350	50.3%
	Masculino	8,237	49.7%
Tipo paciente	Ambulatorio	14,075	84.9%
	Hospitalizado	2,512	15.1%
Defunción	Si	1257	7.6%
	No	15,330	92.4%
Intubado	Si	310	1.9%
	No	2,184	13.2%
	No aplica	14,093	84.9%
Neumonía	Si	1,892	11.4%
	No	14,695	88.6%
Embarazo	Si	133	0.8%
	No	8,187	49.4%
	No Aplica	8,267	49.8%
Indígena	Si	134	0.8%
	No	15,699	94.6%
	No especifica	754	4.5%
Diabetes	Si	1,836	11.1%
	No	14,751	88.9%
EPOC	Si	144	0.9%
	No	16,443	99.1%
Asma	Si	336	2%
	No	16,251	98%
Inmuno-supresión	Si	125	0.8%
	No	16,462	99.2%
Hipertensión	Si	2,430	14.7%
	No	14,157	85.3%
Otra comorbilidad	Si	295	1.8%
	No	16,292	98.2%
Cardiovascular	Si	218	1.3%
	No	16,369	98.7%
Obesidad	Si	2,072	12.5%
	No	14,515	87.5%
Renal crónica	Si	215	1.3%
	No	16,372	98.7%
Tabaquismo	Si	1,145	6.9%
	No	15,442	93.1%
Otro caso	Si	6,378	38.5%
	No	9,531	57.5%
	Se ignora	678	4%
UCI	Si	188	1.1%
	No	2,307	13.9%
	No aplica	14,092	85%
Grupo edad	< 18	966	5.8%
	18 – 40	2,628	15.8%
	40 – 60	5,615	33.9%
	> 60	7,378	44.5%
Fase de atención	Temprana	15,369	92.7%
	Pulmonar	1,102	6.6%

Inflamatoria	116	0.7%
--------------	-----	------

Las relaciones de dependencia probabilísticas presentes en una red bayesiana pueden ser establecidas a partir del conocimiento empírico de expertos o a través de la ejecución de algoritmos de aprendizaje estructural que consideran algún conjunto de datos. Generar la estructura inicial de una red bayesiana puede llegar a ser incluso un problema NP-Completo, (Liu y otros, 2012). El aprendizaje estructural de una red bayesiana puede efectuarse a partir de alguno de los siguientes enfoques: basado en restricciones, basado en puntuación y aprendizaje híbrido.

Uno de los algoritmos de aprendizaje estructural más utilizados dentro del enfoque basado en restricciones es el algoritmo PC, (Spirtes y otros, 2000). El algoritmo parte de un grafo no dirigido, en donde todos sus nodos se conectan entre sí. Para asignar una direccionalidad a las aristas es necesario realizar pruebas de independencia condicional e identificar las immoralidades presentes en el grafo. Es decir, para cada par de variables X_i y X_j se determina su dependencia o independencia dado un conjunto de eventos Z . En conjuntos de datos numéricos se recomienda hacer uso del coeficiente de correlación de Pearson, mientras que para conjuntos de datos categóricos es posible determinar la independencia a partir de las pruebas de independencia χ^2 . Por su parte, si en un camino $X_i - Z - X_j$ no existe una arista que conecte a X_i con X_j y Z no fue contemplada en la prueba de independencia de X_i y X_j existe una immoralidad y por tanto es posible definir una direccionalidad $X_i \rightarrow Z$ y $X_j \rightarrow Z$.

Otro algoritmo de aprendizaje estructural ampliamente utilizado es Hill Climb Search, este algoritmo de aprendizaje híbrido se basa en el concepto de búsqueda local, si bien en la mayoría de las ocasiones no retorna la solución óptima global, puede retornar soluciones aceptables.

$$G^* = \underset{G \in G^n}{\operatorname{argmax}} f(G; D) \quad (2)$$

La ecuación (2) indica el objetivo de Hill Climb Search; busca identificar de entre un conjunto de estructuras posibles, la estructura G^* que maximice una función de puntuación respecto al conjunto de datos de entrenamiento D . Algunas de las medidas de puntuación más utilizadas son: Bayesian Criterion Information (BIC), Bayesian Dirichlet equivalent uniform (BDeu) y K2 Score.

Hill Climb Search parte de una estructura inicial sin aristas. Seguidamente, mientras el algoritmo presente mejores resultados en cuanto a la función de puntuación, se realizan los siguientes procesos:

1. Adición de nodos vecinos: Para cada nodo X_i y cada nodo $X_j \notin Pa_G(X_i)$ se establece $X_i \rightarrow X_j$ sí y solo si no se introduce un ciclo en la estructura G , se calcula la diferencia entre la estructura generada y la actual, es decir: $dif = f(G + \{X_i \rightarrow X_j\}; D) - f(G; D)$.
2. Eliminación de nodos vecinos: Para cada nodo X_i y cada nodo $X_j \in Pa_G(X_i)$ se calcula la diferencia entre la estructura generada y la estructura actual, es decir, $dif = f(G - \{X_j \rightarrow X_i\}; D) - f(G; D)$.
3. Intercambiar la dirección de nodos vecinos: Para cada nodo X_i y cada nodo $X_j \in Pa_G(X_i)$, tal que al

invertir $X_i \rightarrow X_j$ no se genera un ciclo en la estructura G , se calcula $dif = d_1 - d_2$ donde $d_1 = f(G - \{X_j \rightarrow X_i\}; D) - f(G; D)$ y $d_2 = f(G + \{X_j \rightarrow X_i\}; D) - f(G; D)$.

4. Posteriormente, se verifica que exista una mejora en la estructura G , para ello es necesario obtener la máxima diferencia obtenida en los tres pasos anteriores, para después aplicar los cambios a la estructura general y continuar nuevamente con este proceso hasta que no exista mejora o se llegue al número máximo de iteraciones.

Es importante mencionar que cada una de las diferencias generadas en los procesos de adición, eliminación e intercambio de direccionalidad de los nodos son almacenadas en memoria, por lo que el tiempo de ejecución del algoritmo Hill Climb Search se ve reducido, (Gámez y otros, 2011).

2.3. Criterios de evaluación

Existen diversos criterios de evaluación para los modelos gráficos probabilísticos. Como se mencionó anteriormente, las redes bayesianas pueden ser utilizadas para tareas de inferencia y clasificación. Considerando los objetivos de esta investigación, en primera instancia es necesario evaluar que tanto se ajusta el modelo a la realidad y posteriormente, evaluar el modelo como clasificador para las variables: intubado, UCI y defunción. En este sentido, para evaluar el ajuste del modelo al conjunto de prueba se propone hacer uso de métricas de puntuación para algoritmos de aprendizaje estructural, tales como: BIC y BDeu.

$$BIC = -2\ln L(\hat{\theta}|y) + k\ln(n) \quad (3)$$

BIC es un criterio basado en la log-verosimilitud para determinar el mejor modelo de entre un conjunto finito de modelos probabilísticos a comparar. En (3) es posible observar la definición de BIC, donde $\hat{\theta}$ es el modelo estimado, y el conjunto de datos, k indica el número de parámetros de $\hat{\theta}$ y n el número de variables de estudio.

$$BDeu(G, D, \alpha) = \prod_{i=1}^N \prod_{j=1}^{q_i} \left[\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ik} + n_{ijk})}{\Gamma(\alpha_{ijk})} \right] \quad (4)$$

La ecuación (4) es la expresión algebraica de BDeu, una función que recibe como parámetros una estructura inicial G , un conjunto de datos D y α un parámetro de la distribución Dirichlet, N indica el número de nodos de la estructura G , r_i indica el número de estados de X_i , q_i es el número de configuraciones de $\prod X_i$ y $n_{ij} = \sum_k n_{ijk}$. La interpretación de BIC y BDeu no es sencilla, son útiles para comparar dos o más modelos probabilísticos, a medida que su valor se incrementa se obtiene un mejor ajuste. Una revisión detallada acerca de BIC y BDeu pueden ser vistas en (Neath & Cavanaugh, 2012) y (Scutari, 2016), respectivamente.

Por su parte para evaluar el modelo como clasificador en las variables: intubado, UCI y defunción, se propone hacer uso de los siguientes criterios de puntuación: precisión, sensibilidad, exactitud y el valor F. Estas métricas de puntuación son

expresadas en términos de los posibles eventos que pueden darse al efectuar la clasificación:

- Verdaderos positivos (vp): pacientes clasificados a través de la red bayesiana como intubados y que de acuerdo con el conjunto de prueba fueron intubados.
- Falso positivo (fp): pacientes clasificados a través de la red bayesiana como intubados y que de acuerdo con el conjunto de prueba no fueron intubados.
- Falso negativo (fn): pacientes clasificados a través de la red bayesiana como no intubados y que de acuerdo con el conjunto de prueba fueron intubados.
- Verdadero negativo (vn): pacientes clasificados a través de la red bayesiana como no intubados y que de acuerdo con el conjunto de prueba no fueron intubados.

Considere que la definición de verdadero positivo, falso positivo, verdadero negativo y falso negativo es análoga para las variables UCI y defunción. Un resumen de los posibles eventos que pueden darse en una tarea de clasificación puede ser visto en la *Figura 3*.

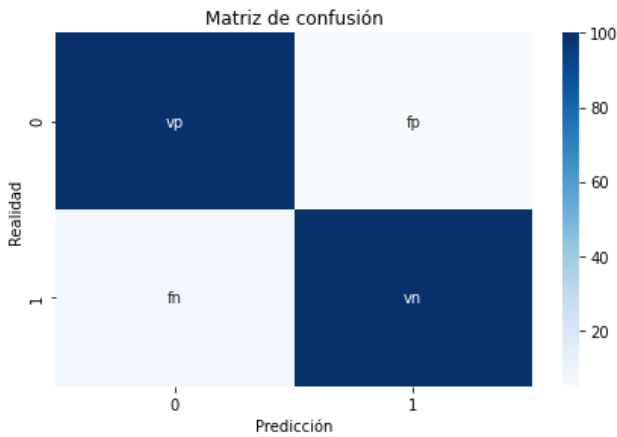


Figura 3. Matriz de confusión.

En (5) se observa que la precisión indica la proporción de verdaderos positivos entre el total de positivos clasificados por el modelo. Así mismo en (6) se presenta la sensibilidad, esta medida expresa la proporción de verdaderos positivos entre los pacientes que realmente presentaron el evento agravante por COVID-19. Por su parte, en (7) se define la exactitud como la proporción de clasificaciones realizadas correctamente por el modelo entre el tamaño del conjunto de prueba.

$$precision = \frac{vp}{vp+fp} \quad (5)$$

$$sensibilidad = \frac{vp}{vp+fn} \quad (6)$$

$$exactitud = \frac{vp+vn}{vp+fp+vn+fn} \quad (7)$$

$$F = 2 \cdot \frac{precision \cdot sensibilidad}{precision + sensibilidad} \quad (8)$$

La ecuación (8) expresa el valor F que resume la precisión y la sensibilidad de forma ponderada, a medida que el valor F se acerca a 1 indica una mejor precisión y sensibilidad. Esta métrica es útil para evaluar algoritmos de clasificación que fueron entrenados a partir de conjuntos de datos no balanceados.

3. Resultados

En esta sección se describen los resultados obtenidos en esta investigación. Inicialmente, se describe la generación y evaluación de estructuras de redes bayesianas a través de los algoritmos PC y Hill Climb Search con las funciones de maximización BIC y BDeu. Seguidamente, se evalúan los modelos como clasificador para las variables intubado, UCI y defunción. Finalmente, se presentan algunas de las consultas que pueden realizarse a través de la red bayesiana.

3.1. Generación y evaluación de Redes Bayesianas

A partir de la librería pgmpy, disponible para su consulta en (Ankan & Panda, 2015); fue posible generar estructuras de redes bayesianas con el algoritmo PC y el algoritmo Hill Climb Search. Se generaron 14 modelos a través del algoritmo PC, utilizando la prueba de independencia χ^2 de Pearson con un nivel de significancia $\alpha = 0.01$. Por su parte, a partir del algoritmo Hill Climb Search se generaron 28 modelos de redes bayesianas, la mitad de ellos considerando a BIC como función de maximización y el resto considerando a BDeu. Cada uno de los modelos generados por Hill Climb Search consideró un $\epsilon = 1e^{-4}$ y un número de iteraciones máximas de $1e^6$.

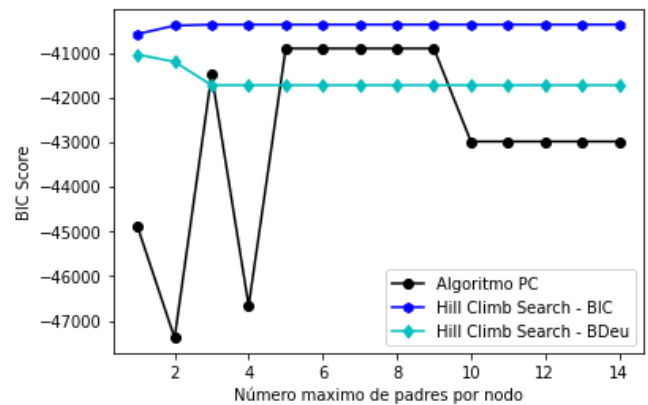


Figura 4. Comparación de aprendizaje estructural de redes bayesianas a través de BIC.

En la *Figura 4* se observan las distintas configuraciones y resultados del aprendizaje estructural de redes bayesianas, generadas a partir de los algoritmos PC y Hill Climb Search; todas las redes bayesianas fueron evaluadas utilizando BIC. Se aprecia que los modelos generados a partir del algoritmo Hill Climb Search que consideraron a BIC como función de maximización y máximo tres padres por nodo son los que presentan un mejor ajuste al conjunto de prueba. Así mismo, es posible visualizar que el algoritmo PC oscila en puntos máximos y mínimos, presentando un comportamiento inestable. Por su parte, el algoritmo Hill Climb Search con BDeu como función de maximización presenta un buen ajuste con un número pequeño de padres por nodo, sin embargo, no

supera el ajuste de Hill Climb Search con BIC como función de maximización.

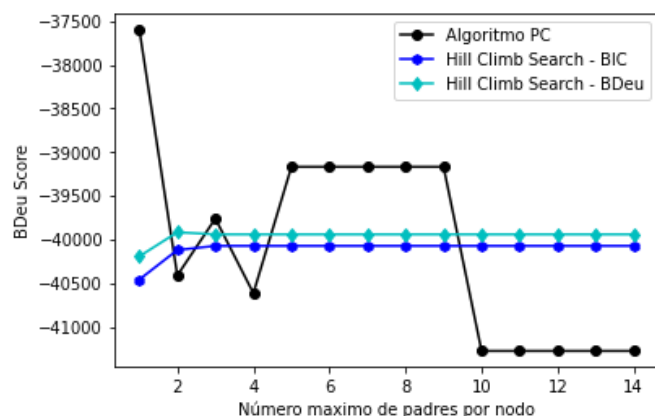


Figura 5. Comparación de aprendizaje estructural de redes bayesianas a través de BDeu.

En la Figura 5, es posible observar el ajuste de los modelos de redes bayesianas considerando la puntuación BDeu. El algoritmo PC presenta un comportamiento inestable, sin embargo, el modelo generado a partir de este algoritmo con máximo un padre por nodo indica un buen ajuste al conjunto de prueba. Por su parte, el algoritmo Hill Climb Search con BDeu como función de maximización indica un buen ajuste a partir de modelos con mínimo dos padres por nodo, lo mismo ocurre con el algoritmo Hill Climb Search con BIC como función de maximización.

3.2. Evaluación de la red bayesiana como clasificador

Cada una de las redes bayesianas creadas a partir del algoritmo PC y Hill Climb Search fueron evaluadas como clasificador para las variables: intubado, UCI y defunción. Considerando los registros de pacientes en el conjunto de datos de prueba se realizaron consultas a través del algoritmo de eliminación de variables, obteniendo la probabilidad de cada uno de los posibles valores que puede tomar la variable en cuestión, el resultado de la clasificación es el valor que presenta una mayor probabilidad.

Tabla 2. Clasificación de la variable intubado a través de la red bayesiana creada por el algoritmo Hill Climb Search con BDeu como función de maximización y solo un padre por nodo.

	Precisión	Sensibilidad	F	Soporte
Si	0.79	0.22	0.35	85
No	0.89	0.99	0.94	547
No aplica	1.0	1.0	1.0	3515
Precisión	0.9826			4147
Exactitud	0.9826			4147

Tal y como se observa en la Tabla 2, la red bayesiana creada a partir del algoritmo Hill Climb Search con BDeu como función de maximización y solo un padre por nodo fue la que mejor clasificó a la variable intubado. Particularmente, presenta una buena precisión para clasificar a aquellos pacientes que no requieren ser intubados, pero han sido hospitalizados; así mismo, la red bayesiana clasifica de forma genuina a los pacientes que no son aptos para recibir intubación mecánica asistida, es decir, aquellos pacientes que son positivos a COVID-19 pero que su atención médica es ambulatoria. Los resultados obtenidos al clasificar a los

pacientes que requieren intubación presentan una precisión de 0.79.

En concreto, la red bayesiana creada a partir del algoritmo Hill Climb Search con BDeu como función de maximización y solo un padre por nodo permite clasificar a la variable UCI con una precisión y exactitud del 98%.

Tabla 3. Clasificación de la variable UCI a través de la red bayesiana creada por el algoritmo Hill Climb Search con BDeu como función de maximización y máximo tres padres por nodo.

	Precisión	Sensibilidad	F	Soporte
Si	0.70	0.30	0.42	53
No	0.94	0.98	0.96	580
No aplica	1.0	1.0	1.0	3514
Precisión	0.9886			4147
Exactitud	0.9886			4147

La red bayesiana generada a partir del algoritmo Hill Climb Search con la función de maximización BDeu y máximo tres padres por nodo clasificó de mejor manera la variable UCI, que indica si un paciente ingresó a la unidad de cuidados intensivos. En la Tabla 3, se aprecia que clasificó de forma inequívoca a los pacientes que fueron diagnosticados con COVID-19, pero su atención médica fue ambulatoria. Así mismo, se obtiene una buena precisión y sensibilidad en los pacientes positivos a COVID-19, que fueron hospitalizados, pero que afortunadamente no ingresaron a la unidad de cuidados intensivos. En la clasificación de pacientes que requirieron ingresar a la unidad de cuidados intensivos, la red bayesiana solo presenta un 70% de precisión, producto del desbalance en las categorías de esta variable.

Tabla 4. Clasificación de la variable defunción a través de la red bayesiana creada por el algoritmo Hill Climb Search con BIC como función de maximización y máximo dos padres por nodo.

	Precisión	Sensibilidad	F	Soporte
Si	0.68	0.54	0.60	297
No	0.96	0.98	0.97	3850
Precisión	0.9491			4147
Exactitud	0.9491			4147

Finalmente, la red bayesiana creada por el algoritmo Hill Climb Search con BIC como función de maximización y dos padres por nodo tuvo un mejor desempeño para la clasificación de la variable defunción. La Tabla 4 ilustra una buena precisión para clasificar a los pacientes que superaron el COVID-19. Sin embargo, la precisión para clasificar a pacientes que fallecieron por COVID-19 es del 68%. En términos generales, la red bayesiana clasifica la variable defunción con una precisión y exactitud superior al 94%.

Es posible afirmar que, las redes bayesianas generadas a partir de Hill Climb Search con las funciones de maximización BIC y BDeu proporcionan mejores resultados para la clasificación de eventos agravantes por COVID-19 en comparación con el algoritmo PC.

Considerando que uno de los objetivos es modelar las relaciones de dependencia probabilísticas entre los distintos eventos asociados al COVID-19 y realizar tareas de clasificación sobre más de una variable, se ha optado por seleccionar una red bayesiana que presente una buena estabilidad en los indicadores BIC y BDeu; así como una buena precisión y exactitud al clasificar las variables: intubado, UCI y defunción.

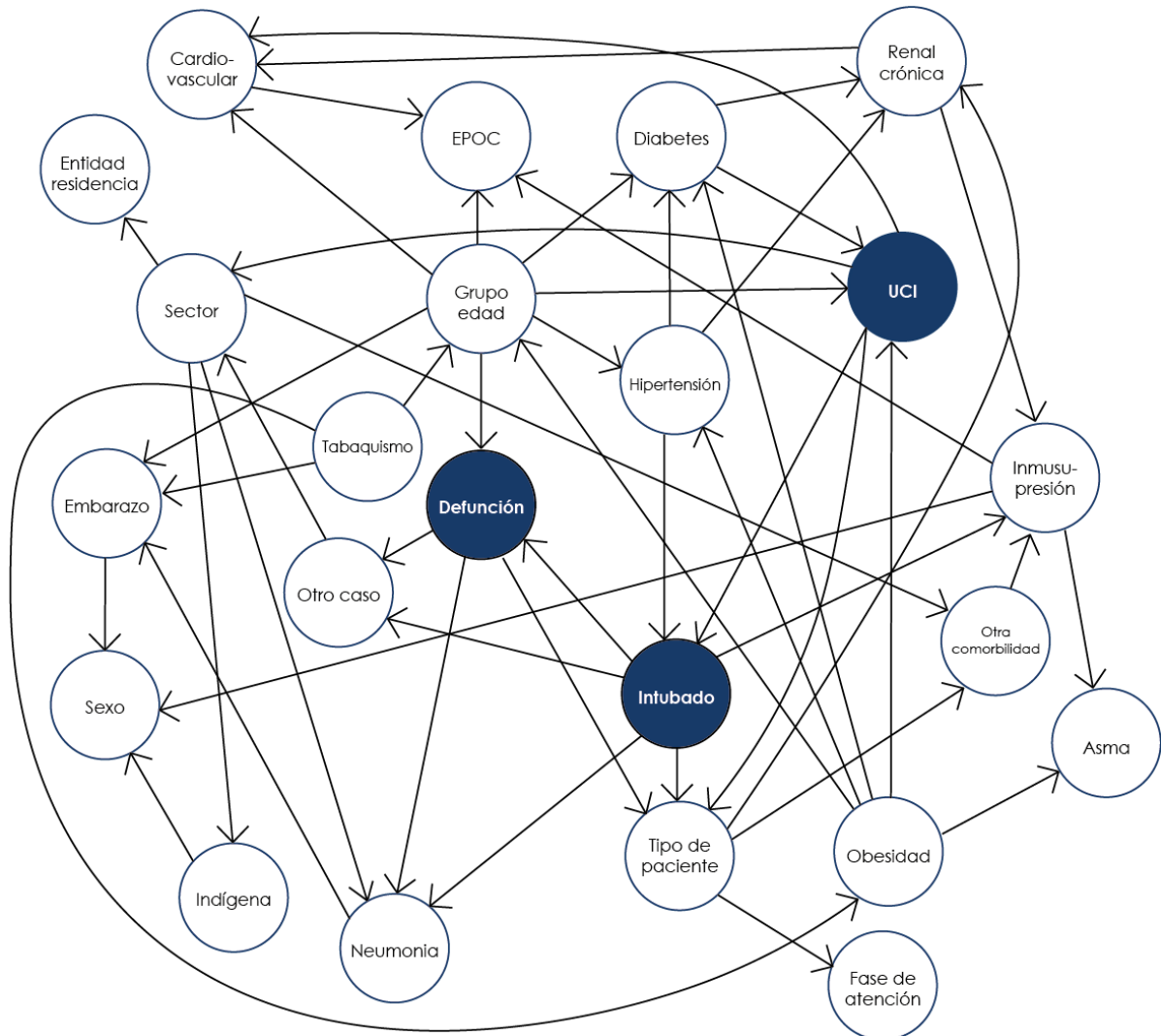


Figura 6. Red bayesiana para la predicción de eventos agravantes de COVID-19 en México.

La Figura 6 ilustra las relaciones de dependencia probabilísticas identificadas por el algoritmo Hill Climb Search con BDeu como función de maximización y máximo tres padres por nodo. La estructura de la presente red bayesiana fue evaluada de forma positiva por BIC y BDeu; además, presenta buenos resultados como clasificador, puesto que logró clasificar de mejor manera la variable UCI. Se observa también que la probabilidad de que un paciente positivo a COVID-19 sea ingresado a la unidad de cuidados intensivos depende probabilísticamente del grupo de edad del paciente y si presenta o no diabetes. Por su parte, la probabilidad de que un paciente positivo a COVID-19 sea intubado depende de si el paciente fue ingresado o no a la unidad de cuidados intensivos y padezca o no hipertensión. Finalmente, la probabilidad de defunción de un paciente con COVID-19 depende probabilísticamente de su grupo de edad y si es o no intubado. En concreto, los factores de riesgo que inciden en eventos agravantes de COVID-19 son: diabetes, obesidad, hipertensión y el grupo de edad.

Es importante recalcar que, está red bayesiana ya ha sido evaluada como clasificador para la variable UCI (véase la Tabla 3), por lo que resulta conveniente presentar su matriz de confusión en la Figura 7; podemos observar que en efecto, clasifica de forma correcta a los 3,514 pacientes del conjunto

de datos que recibieron atención médica ambulatoria y que por tanto no pueden ingresar a la unidad de cuidados intensivos. También podemos observar que, de los 580 pacientes que de acuerdo con la DGE no ingresaron a la unidad de cuidados intensivos, solo 10 fueron clasificados de forma errónea por la red bayesiana. El modelo clasifica con menor precisión a los pacientes que realmente ingresaron a la unidad de cuidados intensivos.

Matriz de confusión - UCI

	Predicción		
Diagnostico PCR	Si	No	No aplica
Si	16	37	0
No	7	570	3
No aplica	0	0	3514

Figura 7. Matriz de confusión de la variable UCI, clasificada por la red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México.

La red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México clasificó a la variable intubado con una precisión del 98%, tal y como se observa en la *Tabla 5*. Así mismo, en la *Figura 8* se aprecia que la red bayesiana fue capaz de clasificar casi de forma perfecta a las categorías No y No aplica, puesto que de los 547 pacientes que fueron hospitalizados y no requirieron intubación mecánica asistida solo 8 pacientes fueron clasificados de manera incorrecta; mientras que de los 3,515 pacientes para los que no aplicaba la intubación mecánica solo uno fue clasificado de forma errónea. Por su parte, la categoría con una clasificación más deficiente fue Si, se observa que la red bayesiana solo clasificó a 19 pacientes como los pacientes que requieren ingresar a la unidad de cuidados intensivos, siendo que el total de paciente que requieren de este nivel de atención fueron 85.

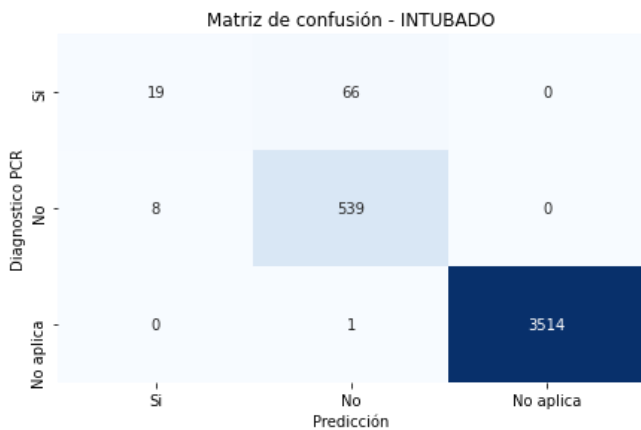


Figura 8. Matriz de confusión de la variable intubado, clasificada por la red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México.

Tabla 5. Clasificación de la variable intubado a través de la red bayesiana para la predicción de eventos asociados al COVID-19 en México.

	Precisión	Sensibilidad	F	Soporte
Si	0.70	0.22	0.34	85
No	0.89	0.99	0.93	547
No aplica	1.0	1.0	1.0	3515
Precisión	0.9819			4147
Exactitud	0.9819			4147

Finalmente, para la clasificación de la variable defunción, la red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México presenta una precisión y exactitud superior al 94%, tal y como se observa en la *Tabla 6*.

Tabla 6. Clasificación de la variable defunción a través de la red bayesiana para la predicción de eventos asociados al COVID-19 en México.

	Precisión	Sensibilidad	F	Soporte
Si	0.58	0.49	0.5	297
No	0.96	0.98	0.9	3850
Precisión	0.9467			4147
Exactitud	0.9467			4147

La red bayesiana tiene una sensibilidad media del 58% para los pacientes que desafortunadamente fallecieron, lo anterior también puede ser visto en la matriz de confusión de la *Figura*

9, donde el número de verdaderos negativos es de 3,780; clasificando solo de forma errónea a 70 pacientes, indicando que fueron pacientes fallecidos a causa del COVID-19. Adicionalmente, la red bayesiana clasificó de forma correcta a 146 pacientes de 297 que lastimosamente fallecieron en la lucha contra esta enfermedad.

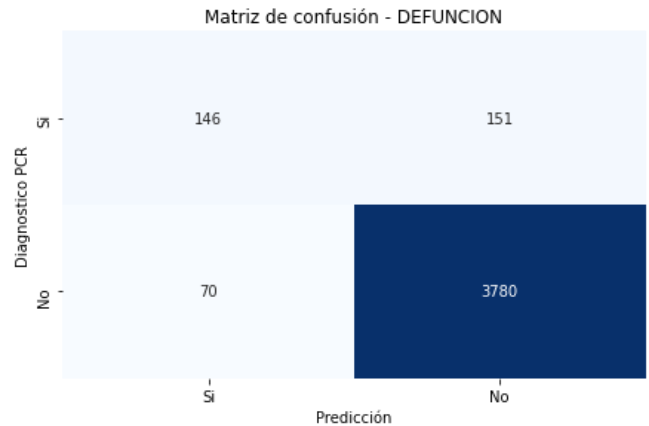


Figura 9. Matriz de confusión de la variable defunción, clasificada por la red bayesiana para la predicción de riesgo de eventos asociados al COVID-19 en México.

3.3. Predicción de riesgo de eventos asociados al COVID-19 en México.

Una de las bondades más importantes de las redes bayesianas es su capacidad de realizar inferencias probabilísticas, es decir, dado un conjunto de información sobre algunas de las variables de la red bayesiana, está es capaz de inferir los valores y la probabilidad de los nodos que no recibieron información. En este sentido y considerando los objetivos de la presente investigación, es posible determinar la probabilidad que toman las variables: intubado, UCI y defunción dado un conjunto de datos generales y comorbilidades del paciente.

$$P(\text{Intubado} = \text{Si} \mid \text{Sector} = \text{SEMAR}, \text{Entidad} = \text{Guanajuato}, \text{Sexo} = \text{Masculino}, \text{Grupo edad} = 40 - 60, \text{Fase atención} = \text{Inflamatoria}, \text{Tipo paciente} = \text{Hospitalizado}, \text{Hipertensión} = \text{Si}, \text{Neumonía} = \text{Si}, \text{UCI} = \text{Si}) = 0.6427 \tag{9}$$

La ecuación (9) determina la probabilidad de ser intubado dado que el paciente fue atendido por la SEMAR en el estado de Guanajuato, es masculino, tiene entre 40 y 60 años, se atendió durante la fase inflamatoria de COVID-19 (después del día 14), fue hospitalizado y además padece hipertensión, neumonía e ingresó a la unidad de cuidados intensivos.

$$P(\text{UCI} = \text{Si} \mid \text{Sector} = \text{ISSSTE}, \text{Entidad} = \text{Guerrero}, \text{Sexo} = \text{Femenino}, \text{Grupo edad} = 40 - 60, \text{Fase atención} = \text{Infección temprana}, \text{Tipo paciente} = \text{Hospitalizado}, \text{Diabetes} = \text{Si}, \text{Cardiovascular} = \text{Si}, \text{Obesidad} = \text{Si}, \text{Intubado} = \text{Si}) = 0.7126 \tag{10}$$

La ecuación (10) indica que la probabilidad que tiene un paciente diagnosticado con COVID-19 de ser ingresado a la unidad de cuidados intensivos dado que se atendió durante la fase de infección temprana (primeros 7 días después del inicio de síntomas) en una institución de salud del ISSSTE en el

estado de Guerrero, es femenino, de entre 40 y 60 años; presenta diabetes, obesidad, enfermedad cardiovascular y requiere de intubación mecánica asistida es del 71%.

$$P(\text{Defunción} = Si | \text{Sector} = \text{IMSS}, \text{Entidad} = \text{CDMX}, \text{Sexo} = \text{Femenino}, \text{Grupo edad} = \text{Mayor de 60}, \text{Fase atención} = \text{Infección pulmonar}, \text{Tipo paciente} = \text{Hospitalizado}, \text{Neumonía} = Si, \text{Diabetes} = Si, \text{Hipertensión} = Si, \text{Obesidad} = Si, \text{Otro caso} = Si, \text{Intubado} = Si) = 0.8329 \quad (11)$$

La ecuación (11) indica la probabilidad de defunción dado que la paciente es femenina, mayor de 60 años, se atendió durante la fase de infección pulmonar (después del día 7 de la aparición de síntomas) en una institución del IMSS en la Ciudad de México y presenta diabetes, hipertensión, neumonía, fue intubado y previo a su contagio estuvo en contacto directo con un paciente positivo a COVID-19.

4. Conclusiones y discusiones

Los modelos gráficos probabilísticos, en especial, las redes bayesianas permiten modelar las relaciones de dependencia probabilísticas que existen en un conjunto de variables de interés. En esta investigación fue posible identificar los factores asociados a eventos agravantes por COVID-19 en México. La probabilidad de ingresar a la unidad de cuidados intensivos depende de la probabilidad que toman las variables: grupo edad, diabetes y obesidad. Así mismo, la probabilidad de ser intubado depende probabilísticamente de si el paciente presenta hipertensión y si fue ingresado a la unidad de cuidados intensivos. Por su parte, la probabilidad de defunción depende de si el paciente fue intubado o no y el grupo de edad al que pertenece. En concreto, los eventos que inciden en eventos agravantes por COVID-19 son: hipertensión, diabetes, obesidad y el grupo de edad.

Por otra parte, se resalta la importancia de comparar distintos algoritmos de aprendizaje estructural, tales como: el algoritmo PC y Hill Climb Search con dos funciones distintas de maximización: BIC y BDeu. El análisis de resultados permitió observar mayor inestabilidad en el algoritmo PC; Hill Climb Search con BIC y BDeu como funciones de maximización presentan resultados muy similares, sin embargo, para la clasificación de eventos agravantes por COVID-19 el algoritmo Hill Climb Search con BDeu como función de maximización y máximo tres padres por nodo fue el que mejor resultados presentó, puesto que logró clasificar de forma correcta a las variables: intubado, UCI y defunción con una precisión y exactitud de entre el 94% y el 98%.

En trabajos futuros se plantea el diseño e implementación de una aplicación web, capaz de actualizar de forma periódica la red bayesiana que mejor describa las relaciones de dependencia probabilísticas asociadas al COVID-19 en México; y, en consecuencia, conocer la probabilidad de los eventos agravantes surgidos a partir de esta enfermedad.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el financiamiento otorgado para esta investigación. A los departamentos de enseñanza médica y urgencias del Hospital

General del ISSSTE en Acapulco, Guerrero; por las aportaciones realizadas a esta investigación.

Referencias

- Ankan, A., & Panda, A. (01 de 04 de 2015). pgmpy: Probabilistic graphical models using python. Citiseer. pgmpy: <https://pgmpy.org/>
- CONACYT. (01 de 04 de 2022). CONACYT. Proyectos: <https://salud.conacyt.mx/coronavirus/investigacion/proyectos/exploratorios.html>
- de Terwangne, C., Laouni, J., Jouffe, L., Lechien, J. R., Bouillon, V., Place, S., Capulzini, L., Machayekhi, S., Ceccarelli, A., Saussez, S., & Sorgente, A. (2020). Predictive accuracy of covid-19 world health organization (Who) severity classification and comparison with a bayesian-method-based severity score (epi-score). *Pathogens*.
- DGE. (01 de 04 de 2022). *Datos Abiertos Dirección General de Epidemiología*. Datos Abiertos Dirección General de Epidemiología: https://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip
- Fenton, N. E., Neil, M., Osman, M., & McLachlan, S. (2020). COVID-19 infection and death rates: the need to incorporate causal explanations for the data and avoid bias in testing. *Journal of Risk Research*, 862-865.
- Gámez, J. A., Mateo, J. L., & Puerta, J. M. (2011). Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 106-148.
- Gobierno de México. (31 de 12 de 2021). *Covid-19 México*. Covid-19 México: <https://datos.covid-19.conacyt.mx/>
- Gobierno de México. (02 de 08 de 2021). *Guía clínica para el tratamiento de la COVID-19 en México*. Guía clínica para el tratamiento de la COVID-19 en México: https://coronavirus.gob.mx/wp-content/uploads/2021/08/GuiaTx_COVID19_ConsensoInstitucional_2021.08.03.pdf
- Liu, Z., Malone, B., & Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. *BMC bioinformatics*, 1-16.
- Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 199--203.
- Ojugo, A., & Otakore, O. D. (2021). Forging An Optimized Bayesian Network Model With Selected Parameters For Detection of The Coronavirus In Delta State of Nigeria. *Journal of Applied Science, Engineering, Technology, and Education*, 37-45.
- OMS. (10 de 11 de 2020). *Información básica sobre la COVID-19*. Información básica sobre la COVID-19: <https://www.who.int/es/emergencias/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>
- OMS. (01 de 01 de 2022). *WHO Coronavirus (COVID-19) Dashboard*. WHO Coronavirus (COVID-19) Dashboard: <https://covid19.who.int/>
- OPS. (01 de 04 de 2022). *Preguntas frecuentes: Vacunas contra la COVID-19*. Preguntas frecuentes: Vacunas contra la COVID-19: <https://www.paho.org/es/vacunas-contra-covid-19/preguntas-frecuentes-vacunas-contra-covid-19>

