Aprendizaje estructural de una Red Bayesiana para la predicción de eventos asociados al COVID-19 en México

Cirilo Tino Salgado
Estudios de Posgrado e Investigación
Instituto Tecnológico de Acapulco
Acapulco de Juárez, México
mm20320018@acapulco.tecnm.mx

Miriam Martinez Arrroyo
Estudios de Posgrado e Investigación
Instituto Tecnológico de Acapulco
Acapulco de Juárez, México
miriam.ma@acapulco.tecnm.mx

Mario Hernández Hernández
Estudios de Posgrado e Investigación
Instituto Tecnológico de Chilpancingo
Chilpancingo de los Bravo, México
mario.hh@chilpancingo.tecnm.mx

Eduardo de la Cruz Gámez
Estudios de Posgrado e Investigación
Instituto Tecnológico de Acapulco
Acapulco de Juárez, México
eduardo.dg@acapulco.tecnm.mx

***Resumen*—El COVID-19 es una enfermedad infecciosa surgida a finales del 2019 en Hubei, China. La fácil transmisión de esta enfermedad ha ocasionado una pandemia mundial que afecta de manera indirecta diversos sectores de la sociedad. En este trabajo, se propone modelar a través de una red bayesiana las dependencias probabilísticas entre los factores que inciden en un diagnóstico positivo de COVID-19 en la población mexicana. El modelo propuesto fue construido a partir del algoritmo de aprendizaje estructural PC, que considera la creación automática de Grafos Dirigidos Acíclicos a partir de un conjunto de datos de comorbilidades clínicas y datos personales de pacientes sospechos a COVID-19, proporcionado por la Dirección General de Epidemiología. Se ha identificado que la probabilidad de defunción depende de la probabilidad de ingresar a una unidad de cuidados intensivos y la probabilidad de requerir intubación mecánica. Así mismo, el tipo de paciente (ambulatorio u hospitalizado) depende probabilísticamente de comorbilidades del paciente, tales como: hipertensión, neumonía y enfermedad renal crónica.**

***Index Terms*—Redes bayesianas, aprendizaje estructural, COVID-19.**

I. Introducción

El COVID-19 es una enfermedad infecciosa causada por el Síndrome Respiratorio Agudo Grave (Sars- CoV-2), surgida en diciembre del 2019 en Wuhan, capital de la provincia de Hubei, China. La transmisión de esta enfermedad se da a través de gotículas de saliva que expulsa una persona enferma al toser, estornudar o hablar; por lo que diversos organismos de salud proponen medidas de distanciamiento social como método de prevención.

En México, la Secretaría de Salud ha optado por medidas de distanciamiento social, la disminución de actividades no esenciales y acciones de reconversión hospitalaria. A pesar de los esfuerzos realizados, durante el año 2020 la Dirección General de Epidemiología (DGE) reportó 1,426,094 diagnósticos positivos por COVID-19, en los que 125,807 pacientes desafortunadamente perdieron la vida.

La Secretaría de Salud en conjunto con organismos públicos descentralizados realizan periódicamente un análisis exploratorio de datos, estimaciones puntuales sobre la tasa de positividad y la tasa de mortandad por COVID-19 en México, así como el análisis de datos a través de sistemas de información geográfica y la aplicación del modelo Gompertz para predecir el comportamiento de la pandemia causada por esta enfermedad, [1].

Adicionalmente, la DGE lleva a cabo un registro de las comorbilidades asociadas a pacientes sospechosos de portar el virus Sars-Cov-2, por lo que se han identificado patologías clínicas que incrementan el riesgo de un desenlace fatal. El análisis y estudio de este conjunto de datos puede ser útil para determinar con precisión la probabilidad de eventos asociados al COVID-19 en México.

En este sentido, se propone modelar las relaciones de dependencia probabilística entre las distintas variables del conjunto de datos proporcionado por la DGE a partir de una red bayesiana. Este tipo de modelo gráfico probabilístico ha sido utilizado con éxito en diversas investigaciones. En [2] se aborda la necesidad de utilizar una red bayesiana para determinar las variables que inciden en la tasa de positividad, la tasa de negatividad y la tasa de mortandad por COVID-19 en Reino Unido. Los autores lograron determinar que la tasa de mortandad por Covid-19 depende de factores demográficos y ambientales de la población, en adición con la calidad de la atención médica brindada.

La optimización de una red bayesiana encargada de clasificar las distintas fases del COVID-19: diagnóstico positivo, diagnóstico negativo, defunción por COVID-19 y diagnóstico positivo activo puede ser vista en [3]. Los autores proponen

el uso de algoritmos de aprendizaje estructural basados en medidas para determinar la estructura de la red bayesiana.

Los resultados indican una precisión del 93 % para la clasificación de pacientes en alguna de las distintas fases de COVID-19.

La precisión de la clasificación de COVID-19 propuesta por la Organización Mundial de la Salud (OMS) se verifica a partir de un modelo basado en inferencia de redes bayesiana (COVID EPI-SCORE) en [4]. El estudio considera 127 observaciones de 295 pacientes positivos al virus Sars-Cov-2 tomados de una institución médica al sur de Bélgica. Los autores definen la estructura de la red bayesiana a partir de los resultados obtenidos por un Manto de Markov Aumentado que considera como medida de puntuación la Longitud Mínima de Descripción (MDL). Los autores resaltan las propiedades de las redes bayesianas; señalan que son modelos potentes que representan de forma compacta la distribución de probabilidad conjunta de las variables de estudio. Además, resaltan la oportunidad de calcular la probabilidad posterior de la variable objetivo dado un conjunto de variables observadas. Este modelo fue capaz de realizar clasificaciones con un $91\,\%$ de precisión.

En este documento se describe la generación de un Grafo Dirigido Acíclico (GDA) para una red bayesiana encargada de determinar la probabilidad de riesgo de eventos asociados al COVID-19 en México. Inicialmente, se describen de manera general los fundamentos teóricos de las redes bayesianas. Seguidamente, se presenta la metodología empleada para la construcción del modelo propuesto. Finalmente, se describen los resultados y conclusiones obtenidas de esta investigación.

I-A. Redes bayesianas

Una red bayesiana puede ser vista como un modelo basado en grafos con una estructura probabilística de datos multivariados. En [5] se indica que una red bayesiana esta constituida por los siguientes elementos.

- Un conjunto de variables aleatorias $X = \{X_1, X_2, ..., X_n\}$ 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 GDA, denotado por $G(V, A)$. Cada vértice $v \in V$ está asociado con una variable aleatoria X_i. Cada arista $a \in A$ representa una dependencia probabilística directa entre dos vértices de la red. Es decir, si un vértice v_j esta conectado a un vértice v_k, la probabilidad de v_k depende de la probabilidad de que ocurra v_j; en contra parte, v_k es independiente de v_j si no existe un arista que conecte a v_j con v_k.

De manera interna, los vértices padres de un red bayesiana están asociados con una tabla de distribución de probabilidad marginal; mientras que los vértices hijos de la red se encuentran asociados a una tabla de distribución de probabilidad condicional. Una tabla de distribución de probabilidad condicional contiene la probabilidad de ocurrencia de los eventos posibles del nodo v dados los eventos posibles de sus nodos padres. El tamaño de estas tablas esta definido por el número de eventos posibles del nodo v elevado al número de padres del mismo nodo. Estas probabilidades comúnmente son definidas a través de un conjunto de datos o el conocimiento empírico de un experto.

En concreto, una red bayesiana permite conocer la probabilidad conjunta de X, en otras palabras, el producto de las probabilidades de cada vértice v_i dada la probabilidad sus padres, tal y como se observa en (1).

$$P(v_1, v_2, ..., v_n) = \prod_{i=1}^{n} P(v_i | Padres(v_i)) \tag{1}$$

En la figura 1 se observa la representación de una red bayesiana, es posible visualizar la presencia de relaciones de dependencia secuencial en los vértices X_1, X_2 y X_5; así mismo, se aprecia una relación convergente en el vértice X_2 que proviene de los vértices X_1 y X_3. Además, es visible una relación divergente en el vértice X_2 con dirección a los vértices X_4 y X_5.

En situaciones donde se desconoce la estructura de la red bayesiana, es necesario aplicar algoritmos de aprendizaje estructural que determinen las relaciones de dependencia probabilística entre los vértices de la red. A medida que el número de variables de estudio aumenta, el número de posibles relaciones de dependencia probabilística también aumenta, en consecuencia el número de grafos posibles se incrementa, por lo que resulta complejo elegir de entre tantas opciones un GDA G que represente la distribución global de X.

En la literatura actual existen diversos enfoques para determinar la estructura de una red bayesiana: aprendizaje estructural basado en medidas, aprendizaje estructural basado en restricciones y aprendizaje estructural híbrido. El aprendizaje estructural basado en medidas busca maximizar una valor que indique la calidad de un GDA G, tal y como se observa en (2), donde D es un conjunto de datos.

$$arg_{max}\ score(G, D) \tag{2}$$

El paradigma de aprendizaje estructural basado en restricciones se enfoca en identificar las dependencias probabilísticas presentes en las variables de estudio a partir de pruebas estadísticas como la prueba de significancia estadística χ^2 de Pearson, la razón de verosimilitud G^2, así como pruebas de independencia condicional. Por su parte, el aprendizaje estructural híbrido busca combinar técnicas basadas en restricciones y medidas; incluyendo algunas técnicas basadas en meta-heurísticas.

Entre los métodos de aprendizaje estructural más utilizados se encuentra el algoritmo PC propuesto en [6]. En este algoritmo se considera un conjunto de datos con n variables de estudio para generar de forma inicial un Grafo No Dirigido (GND) denotado por $G'(V', A')$. Cada vértice $v' \in V'$ está asociado a una variable aleatoria del conjunto de datos; mientras que el conjunto de aristas A' esta dado por $A' = \{(v'_i, v'_j)\}$ para $i, j = 1, 2, ..., n$ e $i \neq j$. La figura 2 ilustra la estructura

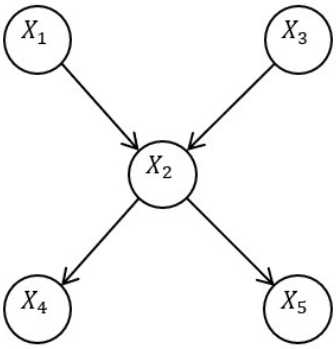

Figura 1. Representación de una red bayesiana.

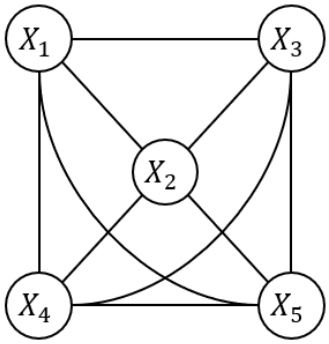

Figura 2. Generación del GND en el algoritmo PC.

inicial de un GND G' en el algoritmo PC para la generación de la red bayesiana vista en la figura 1.

Seguidamente, se procede a generar el esqueleto del GND G' a partir de pruebas de independencia. Inicialmente, el método propuesto para identificar relaciones de independencia era la D-Separación a través de pruebas de independencia condicional. Sin embargo, es posible emplear otras técnicas como pruebas de significancia χ^2 de Pearson para conjuntos de datos categóricos y pruebas sobre el coeficiente de correlación de Pearson para conjuntos de datos continuos. Estas técnicas verifican la independencia que puede existir en conjuntos de variables A y B dado un conjunto de variables C de forma análoga al método de D-Separación.

Una definición del concepto de D-Separación puede ser vista a continuación. Sean A, B y C tres subconjuntos disjuntos de vértices en un grafo G, se dice que C separa a A de B si y solo si no existe un camino no dirigido U entre A y B, de modo que cada colisionador en U tiene un descendiente en C y ningún otro vértice de U está en C. El algoritmo PC realiza el proceso de D-Separación a través de pruebas de independencia condicional para los pares de vértices v'_i y v'_j dado un conjunto C, cuya cardinalidad es determinada por el orden de las pruebas de independencia condicional. En consecuencia, si existe evidencia suficiente que respalde la independencia entre los vértices $v'_i, v'_j|C$ se procede a eliminar el arista correspondiente.

Posteriormente, el algoritmo PC agrega direccionalidad para cada tripleta de vértices en el grafo G', es decir, dados tres vértices v'_i, v'_j y $v'_k \in G'$, donde v'_i y v'_j son adyacentes; v'_j y v'_k también son adyacentes y v'_i y v'_k no son adyacentes, el camino $v'_i - v'_j - v'_k$ tendrá relaciones de dependencia $v'_i \rightarrow v'_j \leftarrow v'_k$ si y solo si v'_j no esta presente en el conjunto de separación (v'_i, v'_k). Finalmente, el algoritmo PC establece que si existe una relación $v'_i \rightarrow v'_j$, v'_j y v'_k son adyacentes, mientras que v'_i y v'_k no son adyacentes y en el camino $v'_j - v'_k$ no existe una direccionalidad, entonces se creará una relación $v'_j \rightarrow v'_k$. Por su parte, si existe un camino entre $v'_i - v'_j$ y existe un arista que conecta v'_i con v'_j, el camino $v'_i - v'_j$ puede establecerse a partir de un arista dirigida de $v'_i \rightarrow v'_j$.

II. Metodología

En esta investigación se considero el conjunto de datos abiertos de COVID-19 en México, proporcionado por la DGE en [7]. El conjunto de datos original contempla 40 variables de estudio entre información socioeconómica, datos personales y comorbilidades que padece cada paciente sospechoso de padecer COVID-19. El conjunto de datos fue procesado para reducir y categorizar algunas variables de estudio. Particularmente, se eliminaron variables de información socioeconómica. A partir de la variable fecha de defunción se creo la variable defunción que indica el desenlace fatal del paciente asociado a una instancia del conjunto de datos; por su parte, la variable edad fue categorizada en los siguientes intervalos: infante = [0,18), joven = [18, 40), adulto = [40, 60) y mayor = $Edad \geq 60$.

Finalmente, el conjunto de datos a analizar quedo constituido por 20 variables: sexo, tipo paciente, defunción, intubado, neumonía, embarazo, diabetes, EPOC, asma, inmunosupresión, hipertensión, enfermedad cardiovascular, obesidad, enfermedad renal crónica, tabaquismo, ingreso a una unidad de cuidados intensivos, edad, otras comorbilidades, contacto directo con un paciente positivo a COVID-19 y clasificación final. Una descripción más detallada de las variables de estudio puede ser vista en la tabla I.

Inicialmente, se contemplo la ejecución de diversos algoritmos de aprendizaje estructural a fin de comparar las estructuras generadas mediante medidas de puntuación. Entre estos algoritmos se encuentran: Chow and Liu, Búsqueda Exhaustiva y PC. Considerando el número de variables aleatorias presentes en el conjunto de datos a analizar, no resulta viable generar y comparar estructuras a partir del algoritmo de Búsqueda Exhaustiva. El algoritmo de Chow and Liu retorna un conjunto ordenado de forma descendente de $n - tupas$ con la cantidad información mutua que proporcionan, sin embargo, este algoritmo no proporciona direccionalidad, por lo que se opto por descartarlo.

Se ejecuto el algoritmo PC con muestras estratificadas por edad con conjuntos muestrales del $10\,\%$, $5\,\%$ y $1\,\%$. Sin embargo, los GDA generados no presentaron cambios significativos en las relaciones de dependencia probabilística, por lo que se opto por utilizar el conjunto muestral del $1\,\%$,

Tabla I
VARIABLES DE ESTUDIO

Nombre	***Descripción***
Clasificación final	Diagnóstico del paciente a COVID-19 [a]
Edad	Grupo de edad del paciente
Diabetes	Presencia o ausencia de diabetes
Tipo paciente	Indica si el paciente es ambulatorio u hospitalizado
Cardiovascular	Presencia o ausencia de enfermedad cardiovascular
EPOC	Paciente con enfermedad pulmonar obstructiva crónica
Renal crónica	Indica si el paciente padece una enfermedad renal crónica
Otro caso	Indica el contacto directo del paciente con un caso positivo a COVID-19
Obesidad	Señala la presencia de obesidad en el paciente
Intubado	Indica si el paciente requirió intubación mecánica
Neumonía	Presencia o ausencia de neumonía en el paciente
Defunción	Señala el desenlace fatal de un paciente
Tabaquismo	Indica si el paciente consume tabaco
Hipertensión	Indica si el paciente es hipertenso
UCI	Señala si el paciente ingreso a una unidad de cuidados intensivos
Otra comorbilidad	Indica si el paciente padece otra comorbilidad
Embarazo	Señala si el paciente está en estado de embarazo
Inmunosupresión	Indica si el paciente es inmunodeficiente
Sexo	Género del paciente
Asma	Indica si el paciente es asmático

es decir 73,378 instancias del conjunto de datos original; cada estructura generada identifico las relaciones de dependencia a partir de pruebas de significancia estadística χ^2 de Pearson debido a que el conjunto de datos está integrado en su totalidad por variables categóricas.

III. RESULTADOS

La figura 3 ilustra el GDA generado por el algoritmo PC. Está estructura fue utilizada para determinar las tablas de probabilidad conjunta asociadas a cada vértice de la red bayesiana. El modelo es validado verificando que las tablas de probabilidad condicional cumplan los axiomas de probabilidad y que las relaciones de dependencia identificadas sean plausibles con la realidad.

La red bayesiana para la predicción de riesgo de COVID-19 en México proporciona información sobre diversos escenarios asociados a la enfermedad. Por ejemplo, la tabla II detalla las probabilidades de cada evento de la variable aleatoria INTUBADO dadas las probabilidades de los eventos de la variable UCI. Se tiene que, la probabilidad de ser intubado dado que se ha ingresado a una unidad de cuidados intensivos es de 0.484(48.4 %).

Tabla II
TABLA DE DISTRIBUCIÓN DE PROBABILIDAD DE LA VARIABLE INTUBADO

UCI	***Si***	***No***	***No aplica***	***Desconocido***
Si	0.484	0.078	0.0	0.0
No	0.515	0.921	0.0	0.0
No aplica	0.0	0.0	1.0	0.0
Desconocido	0.0	0.0	0.0	1.0

Además, se observa una relación de convergencia en el vértice DEFUNCIÓN que proviene de los vértices UCI e INTUBADO, en otras palabras, la probabilidad de que un paciente presente un desenlace fatal está condicionada por la probabilidad de que el paciente haya ingresado a una unidad de cuidados intensivos y haya requerido intubación mecánica. Lo anterior se verifica a partir de las tablas de distribución de probabilidad, la tabla III presenta las probabilidades de defunción dados los posibles eventos de las variables UCI e INTUBADO, se observa que la probabilidad de un desenlace fatal dado que el paciente fue ingresado a una unidad de cuidados intensivos y fue intubado es de $0{,}627(62{,}7\,\%)$.

También es posible observar que el tipo de paciente es definido de acuerdo a la presencia o ausencia de comorbilidades (hipertensión, neumonía, enfermedad renal crónica), factores triviales (ingreso a una unidad de cuidados intensivos y la necesidad de intubación mecánica) y si el paciente tuvo un desenlace fatal.

El modelo propuesto también fue capaz de identificar relaciones de dependencia esperadas, por ejemplo: la probabilidad de embarazo depende del sexo del paciente; y, la presencia o ausencia de enfermedades como hipertensión, diabetes, enfermedades renales crónicas e inmunosupresión permiten inferir el rango de edad del paciente.

IV. CONCLUSIONES Y TRABAJOS FUTUROS

Una red bayesiana aprendida estructuralmente por el algoritmo PC (consistente de pruebas de significancia χ^2 de Pearson en cada tripleta de vértices) permitió identificar las relaciones de dependencia entre los factores que inciden en eventos asociados al COVID-19 en la población mexicana, tales como: la probabilidad de requerir atención hospitalaria dada un conjunto de comorbilidades del paciente, la probabilidad de ser intubado dado que el paciente ha sido ingresado a una unidad de cuidados intensivos.

La estructura de la red bayesiana no presentó cambios significativos en las relaciones de dependencia probabilística con conjuntos muestrales del 10 %, 5 % y 1 %, por lo que se opto por utilizar el conjunto muestral del 1 %. El modelo propuesto

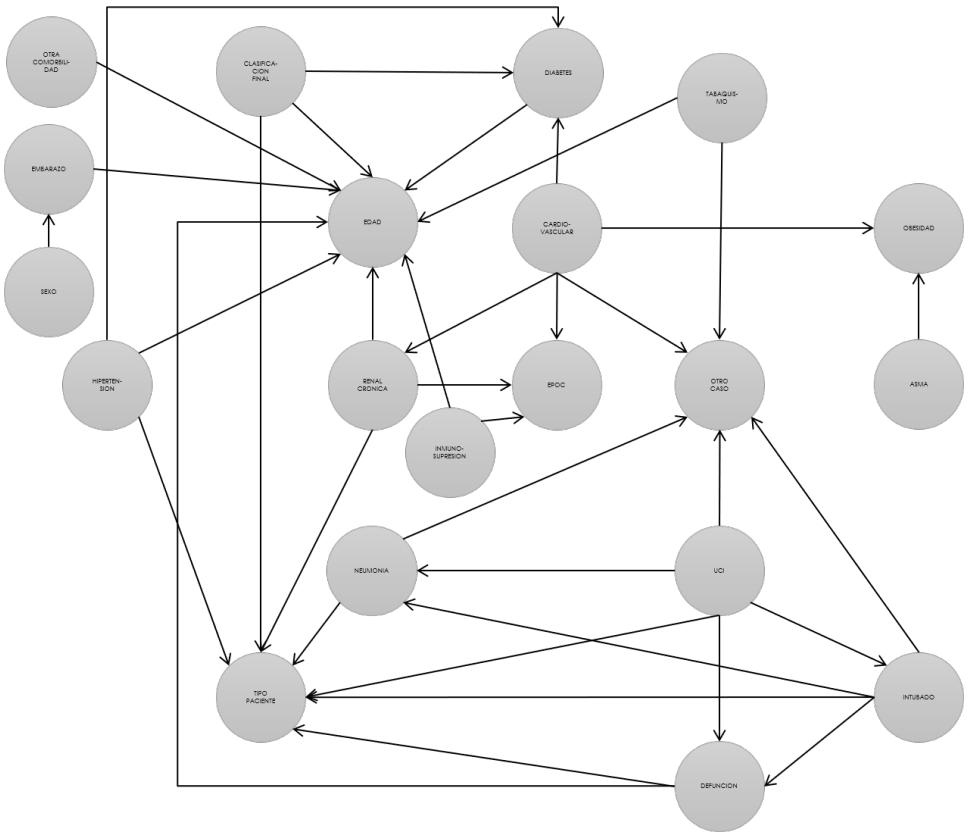

Figura 3. Red bayesiana para la predicción de riesgo de COVID-19 en México.

Tabla III
TABLA DE DISTRIBUCIÓN DE PROBABILIDAD DE LA VARIABLE DEFUNCIÓN

Intubado	***Si***				***No***				***NA***[a]				***D***[b]			
UCI	***Si***	***No***	***NA***	***D***	***Si***	***No***	***NA***	***D***	***Si***	***No***	***NA***	***D***	***Si***	***No***	***NA***	***D***
Si	0.627	0.814	0.5	0.5	0.286	0.294	0.5	0.5	0.5	0.5	0.003	0.5	0.5	0.5	0.5	0.271
No	0.372	0.185	0.5	0.5	0.713	0.705	0.5	0.5	0.5	0.5	0.996	0.5	0.5	0.5	0.5	0.728

[a]No aplica.
[b]Desconocido.

podría ser utilizado para realizar tareas de clasificación e inferencia probabilística sobre variables desconocidas a partir de algoritmos de propagación de probabilidades.

Como trabajos futuros se pretende ejecutar redes bayesianas con conjuntos muestrales de mayor dimensión con ayuda de una supercomputadora. Además de generar un modelo similar para el estado de Guerrero, México.

AGRADECIMIENTOS

Agradezco al Consejo Nacional de Ciencia y Tecnología por el financiamiento otorgado para el desarrollo y difusión de esta investigación. Así como al Instituto Tecnológico de Acapulco por la asesoría brindada.

REFERENCIAS

[1] Coronavirus.conacyt.mx. 2021. PROYECTOS. Recuperado 21 de junio de 2021, de https://coronavirus.conacyt.mx/proyectos/

[2] 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, 23(7-8), 862-865.

[3] 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, 3(1), 37-45.

[4] de Terwangne, C., Laouni, J., Jouffe, L., Lechien, J. R., Bouillon, V., Place, S., ... EPIBASE TEAM. (2020). Predictive accuracy of COVID-19 world health organization (Who) severity classification and comparison with a bayesian-method-based severity score (epi-score). Pathogens, 9(11), 880.

[5] Scutari, M., Denis, J. B. (2021). Bayesian networks: with examples in R. CRC press.

[6] Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D. (2000). Causation, prediction, and search. MIT press.

[7] DGE. (2021, 21 junio). Datos Abiertos Dirección General de Epidemiología. Recuperado 21 de junio de 2021, de http://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip