

Minería de Datos para el Descubrimiento de Factores de Mortalidad en Pacientes Hospitalizados por COVID-19

C. Montserrat Alejandra Cuevas Muñoz¹, M. en C. C. Guillermo García Lambert²,
Dr. en Med. Hugo Mendieta Zerón³, Dra. en Hum. María Luisa Pimentel Ramírez⁴, Dr. en C. I. Ángel Gabriel
Estévez Pedraza⁵, Dra. en C.S. Laura Soraya Gaona Valle⁶

Resumen— En esta investigación se utilizó la metodología de Descubrimiento de Conocimiento en Bases de datos (KDD) y árboles de decisión, técnica de minería de datos para la obtención de los factores de mortalidad con mayor ganancia de información de una base de datos de un establecimiento médico del Estado de México que cuenta con la información de 1,014 pacientes hospitalizados positivos a SARS-CoV-2 (instancias) y 256 datos clínicos (atributos).

El objetivo fue descubrir los principales factores de mortalidad en pacientes hospitalizados por SARS-CoV-2 que permitan la identificación de reglas comprensibles, aplicando minería de datos, con el propósito de apoyar a la toma de decisiones.

Con los resultados obtenidos, fue posible obtener 24 reglas que inciden en el motivo de egreso del paciente ya sea por mejoría o por muerte.

Palabras clave—COVID-19, minería de datos, árboles de decisión, KDD.

Introducción

La Organización Mundial de la Salud (OMS) dio a conocer a la epidemia de enfermedad respiratoria por coronavirus (COVID-19) causada por el virus SARS-CoV-2, como una emergencia de salud pública de preocupación internacional el 30 de enero de 2020, sin embargo, para el día 11 de marzo de 2020 después de reportarse 118,000 casos y 4,291 decesos en 114 países, se caracterizó como pandemia (Organización Mundial de la Salud, 2020).

Conforme incrementaron los contagios y los pacientes empeoraban clínicamente, algunos de ellos presentaron mayores complicaciones que otros, incluso hubo pacientes que se mantuvieron asintomáticos durante todo su proceso de enfermedad. Debido a esto, se realizaron estudios, en los que factores como la edad y sexo resultaron ser de riesgo para los casos de pacientes positivos a SARS-CoV-2. No obstante, otras condiciones relacionadas con enfermedades crónicas también se han asociado con un mayor riesgo (Farzan et al., 2021); cardiovasculares, hepatopatías, hipertensión, diabetes, las enfermedades respiratorias, neumonía, cáncer, obesidad y los pacientes inmunodeprimidos se asociaron con un mayor riesgo de complicaciones, hospitalización y en el peor panorama la muerte (Manta et al., 2022).

Ante la emergencia sanitaria que se vivió en todo el mundo, fue necesario brindar atención médica inmediata, en la que instituciones como el Centro Médico Lic. Adolfo López Mateos (CMALM), organismo de Salud Pública dependiente del Instituto de Salud del Estado de México, prestó su servicio. Los registros hospitalarios de los pacientes contagiados habilitaron la conformación de una base de datos que se integró en tres etapas: al ingreso, durante su estancia y al egreso.

Al tener disponible esta información, es conveniente analizar la existencia de patrones, clasificaciones, agrupaciones, predicciones, etc. Por medio de herramientas tecnológicas como la Inteligencia artificial, la cual se puede comprender, como cualquier tipo de sistema computacional artificial que aparenta ser capaz de comportarse inteligentemente, (Abellán-García Sánchez et al., 2021).

La inteligencia artificial está orientada primordialmente hacia el aprendizaje automático o *machine learning*, el cual consiste en buscar algún patrón de datos para crear algún algoritmo capaz de predecir o tomar

¹ C. Montserrat Alejandra Cuevas Muñoz, egresada de la Licenciatura en Bioingeniería Médica, Facultad de Medicina, Universidad Autónoma del Estado de México. mcuevasm001@alumno.uaemex.mx

² M. en C. C. Guillermo García Lambert, Profesor de Medio Tiempo de la Licenciatura en Bioingeniería Médica, Facultad de Medicina, Universidad Autónoma del Estado de México. ggarcial@uaemex.mx

³ Dr. en Med. Hugo Mendieta Zerón, Director de la Facultad de Medicina, Universidad Autónoma del Estado de México. hmendietaz@uaemex.mx

⁴ Dra. en Hum. María Luisa Pimentel Ramírez, Profesora de Tiempo Completo, Facultad de Medicina, Universidad Autónoma del Estado de México. mlpimentelr@uaemex.mx

⁵ Dr. en C. I. Ángel Gabriel Estévez Pedraza, Profesor de Tiempo Completo, Unidad Académica Profesional Tianguistenco, Universidad Autónoma del Estado de México. aestevezp@uaemex.mx

⁶ Dra. en C. S. Laura Soraya Gaona Valle, Jefa del Departamento de Investigación, Centro Médico "Lic. Adolfo López Mateos". gaonav_81@yahoo.com.mx

decisiones teniendo como objetivo principal optimizar y predecir (De et al., 2021).

Una subcategoría del aprendizaje automático es el aprendizaje supervisado, que requiere de un conjunto de datos etiquetados para generar predicciones en la toma de decisiones. Algunas técnicas de esta herramienta son: árboles de decisión, *K-Nearest Neighbors*, *KNN*; *Support-Vector Machines*, *SVM*, regresión, clasificadores Bayesianos, entre otros (Beatriz Beltrán Martínez, n.d.).

En este sentido, la minería de datos se especializa en el análisis de métodos y algoritmos que sean capaces de automatizar la extracción de información sintetizada, novedosa, válida, comprensible y útil localizando patrones con el fin de predecir, obtener una optimización de información en grandes bases de datos (García Lambert, 2012). Por otra parte, el árbol de decisión es una técnica utilizada para obtener modelos a partir de una muestra de datos, a fin de generar reglas dependientes. Además, existen algoritmos capaces de construir árboles de decisión para facilitar la toma de decisiones a partir de las reglas generadas (Witten & Frank, 2005), dentro de los cuales destacan los algoritmos C4.5, C5.0 y J48, debido a que manejan variables dependientes continuas y categóricas, capaces de determinar un conjunto de reglas lógicas, conducidas por una partición no binaria. Para la construcción de árboles de decisión se requiere de una metodología que dirija al árbol de decisión, como la de KDD (*Knowledge Discovery and Data Mining*), que permite llevar de manera sistemática los pasos para la obtención de patrones (Hernández Orallo et al., 2005). La Figura 1 presenta un diagrama del KDD, proceso utilizado para extraer conocimiento partiendo de grandes cantidades de datos.

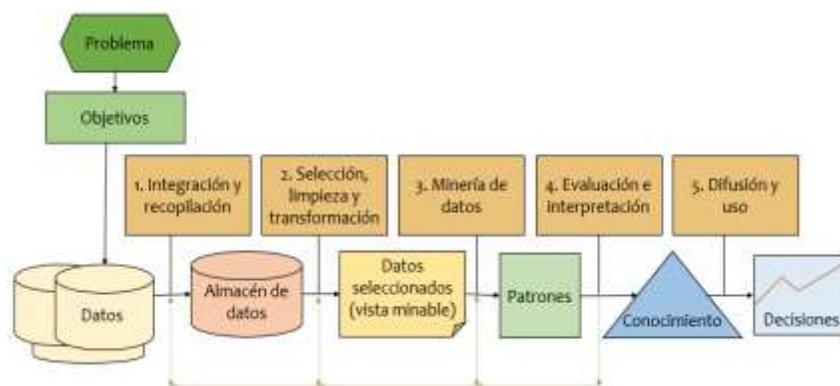


Figura 1. Metodología de KDD.

Metodología

Problema

No existen análisis de la información de los históricos clínicos de ingreso, estancia y egreso del ingreso de los pacientes positivos a SARS-CoV-2 en la región del Valle de Toluca, la cual es usualmente almacenada en bases de datos y repositorios internos. Las técnicas de IA pueden clasificar, agrupar y optimizar la gestión de los datos, proveyendo de información significativa que evidencie las áreas de oportunidad en la atención futura de pacientes contagiados por SARS-CoV-2.

Objetivo

Analizar la información de una base de datos con registros de pacientes positivos a SARS-CoV2 que fueron ingresados y egresados en el CMALM para la obtención de los principales factores de mortalidad mediante reglas comprensibles, aplicando minería de datos.

Integración y recopilación

Se obtuvo una base de datos del CMALM la cual consta de 256 datos clínicos (atributos) y un total de 1,014 pacientes hospitalizados positivos a SARS-CoV-2 (instancias) cuya información se integra desde el ingreso, permanencia y egreso de los pacientes de marzo a diciembre del 2020.

Selección, Limpieza y transformación

Se eliminaron 48 de los atributos debido a que la información en las instancias era faltante o redundante por lo que posterior a esta etapa, resultaron 208 atributos.

Minería de datos

Se consideró la información de todos los pacientes hospitalizados positivos a SARS-CoV-2, es decir 1,014 instancias y 208 atributos, utilizando la validación cruzada con 10 iteraciones como técnica empleada para la evaluación e interpretación de las reglas generadas.

Para realizar la clasificación de los datos se utilizó la herramienta computacional Weka, de aplicación de minería de datos destinada al aprendizaje automático, considerando como etiqueta de clasificación el *motivo de egreso del hospital* (uno de los atributos de la base de datos), la cual tiene como posibles valores la mejoría del paciente y la muerte.

Resultados

Para la realización del árbol de decisión se utilizó el algoritmo J48 debido a que maneja variables dependientes continuas y categóricas capaces de determinar un conjunto de reglas lógicas, que se rigen por una partición no binaria, es decir, puede tener más de una rama por nodo. Para elegir el atributo que tomará el lugar de nodo raíz, se evaluarán la entropía y la ganancia de información de todos los atributos, y de este modo se continua con las próximas ramificaciones, obteniendo así el árbol de decisión de la Figura 2.

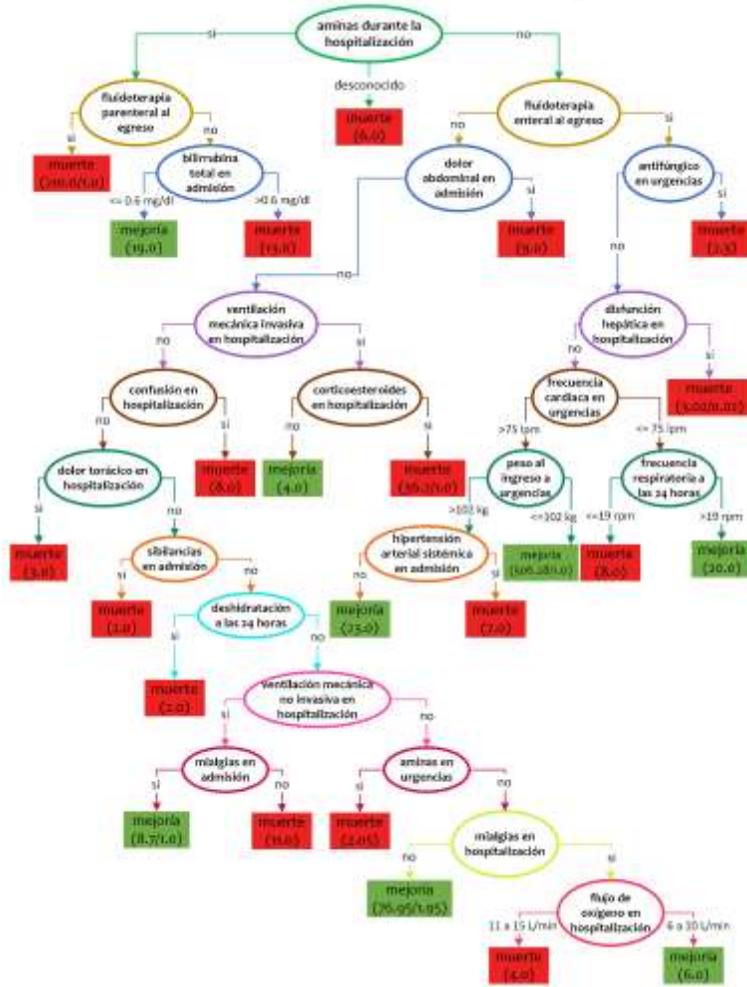


Figura 2. Árbol de decisión.

La administración de aminas a los pacientes durante su hospitalización fue el factor con mayor ganancia de información, destacándose como una vertiente en el pronóstico de los pacientes. De igual manera, los valores presentes en las hojas del árbol de decisión indican el total de pacientes correctamente clasificados (verdaderos positivos), así como aquellos incorrectamente clasificados (falsos positivos) cuyos datos, en algunos casos, fueron racionales. Para una interpretación correcta, se optó por redondeo, ya que la cantidad de pacientes corresponde a los números enteros, de igual manera de distinguieron las hojas del árbol (clase) en color rojo para la clase de muerte como motivo de egreso, así como el color verde indicando la mejoría al egreso de los pacientes.

Los factores con mayor ganancia de información son los atributos que conforman los nodos del árbol de decisión, quedando el 8.6% del total atributos originales, es decir, de 256 atributos originales, solamente 22 tienen mayor ganancia de información. A continuación, se describen las 24 reglas obtenidas, así como las instancias que cumplieron con la clasificación. Quedado que se empleó un algoritmo que se rige por una partición no binaria, fue posible tener más de una rama por nodo, lo cual fue conveniente para atributos con más de una solución.

Reglas obtenidas

1. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias, no presentó disfunción hepática durante su hospitalización, presentó una frecuencia cardíaca mayor a 75 latidos por minuto en urgencias y presentó un peso menor o igual a 102 kg al ingreso a urgencias, entonces el motivo de egreso fue por mejoría (506 instancias)
2. Si al paciente le administraron aminas durante su hospitalización y se le administró fluidoterapia parenteral al egreso entonces el motivo de egreso fue por muerte (210 instancias)
3. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, no requirió de ventilación mecánica no invasiva durante hospitalización, no se le administraron aminas en urgencias y no presentó mialgias en hospitalización entonces el motivo de egreso fue por mejoría (77 instancias)
4. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, requirió de ventilación mecánica invasiva durante su hospitalización y se le administraron corticoesteroides durante su hospitalización entonces el motivo de egreso fue por muerte (37 instancias)
5. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias, no presentó disfunción hepática durante su hospitalización, presentó una frecuencia cardíaca mayor a 75 latidos por minuto en urgencias, presentó un peso mayor a 102 kg al ingreso a urgencias y no presentó hipertensión arterial sistémica en admisión entonces el motivo de egreso fue por mejoría (23 instancias)
6. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias, no presentó disfunción hepática durante su hospitalización, presentó una frecuencia cardíaca menor o igual a 75 latidos por minuto en urgencias y presentó una frecuencia respiratoria mayor a 19 respiraciones por minuto a las 24 horas entonces el motivo de egreso fue por mejoría (20 instancias)
7. Si al paciente le administraron aminas durante su hospitalización, no se le administró fluidoterapia parenteral al egreso y presentó un nivel total de bilirrubina menor o igual a 0.6 mg/dl en admisión entonces el motivo de egreso fue por mejoría (19 instancias)
8. Si al paciente le administraron aminas durante su hospitalización, no se le administró fluidoterapia parenteral al egreso y presentó un nivel total de bilirrubina mayor a 0.6 mg/dl en admisión entonces el motivo de egreso fue por muerte (13 instancias)
9. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, requirió de ventilación mecánica no invasiva durante hospitalización y no presentó mialgias en admisión entonces el motivo de egreso fue por muerte (11 instancias)
10. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso y presentó dolor abdominal en admisión entonces el motivo de egreso fue por muerte (9 instancias)
11. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, requirió de ventilación mecánica no invasiva durante hospitalización y presentó mialgias en admisión entonces el motivo de egreso fue por mejoría (9 instancias)

12. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias, no presentó disfunción hepática durante su hospitalización, presentó una frecuencia cardíaca menor o igual a 75 latidos por minuto en urgencias y presentó una frecuencia respiratoria menor o igual a 19 respiraciones por minuto a las 24 horas entonces el motivo de egreso fue por muerte (8 instancias)
13. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización y presentó confusión durante su hospitalización entonces el motivo de egreso fue por muerte (8 instancias)
14. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias, no presentó disfunción hepática durante su hospitalización, presentó una frecuencia cardíaca mayor a 75 latidos por minuto en urgencias, presentó un peso mayor a 102 kg al ingreso a urgencias y presentó hipertensión arterial sistémica en admisión entonces el motivo de egreso fue por muerte (7 instancias)
15. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, no requirió de ventilación mecánica no invasiva durante hospitalización, no le administraron aminas en urgencias, presentó mialgias durante su hospitalización y requirió de un flujo de oxígeno de 6 a 10 L/min durante su hospitalización entonces el motivo de egreso fue por mejoría (6 instancias)
16. Si al paciente se desconoce si le administraron aminas durante su hospitalización entonces el motivo de egreso fue por muerte (6 instancias)
17. Si al paciente no se le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, no requirió de ventilación mecánica no invasiva durante hospitalización, no se le administraron aminas en urgencias, presentó mialgias durante su hospitalización y requirió de un flujo de oxígeno de 11 a 15 L/min durante su hospitalización entonces el motivo de egreso fue por muerte (4 instancias)
18. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, requirió de ventilación mecánica invasiva durante su hospitalización y no le administraron corticoesteroides durante su hospitalización entonces el motivo de egreso fue por mejoría (4 instancias)
19. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso, no se le administró un antifúngico en urgencias y presentó disfunción hepática durante su hospitalización entonces el motivo de egreso fue por muerte (3 instancias)
20. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización y presentó dolor torácico en hospitalización entonces el motivo de egreso fue por muerte (3 instancias)
21. Si al paciente no le administraron aminas durante su hospitalización, se le administró fluidoterapia enteral al egreso y se le administró un antifúngico en urgencias entonces el motivo de egreso fue por muerte (3 instancias)
22. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión, no presentó deshidratación a las 24 horas, no requirió de ventilación mecánica no invasiva durante hospitalización y se le administraron aminas en urgencias entonces el motivo de egreso fue por muerte (2 instancias)
23. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización y presentó sibilancias en admisión entonces el motivo de egreso fue por muerte (2 instancias)
24. Si al paciente no le administraron aminas durante su hospitalización, no se le administró fluidoterapia enteral al egreso, no presentó dolor abdominal en admisión, no requirió de ventilación mecánica invasiva durante su

hospitalización, no presentó confusión durante su hospitalización, no presentó dolor torácico en hospitalización, no presentó sibilancias en admisión y presentó deshidratación a las 24 horas entonces el motivo de egreso fue por muerte (2 instancias)

De acuerdo con la clase de interés *motivo de egreso del hospital*, las instancias correctamente clasificadas fueron 963 con lo que se obtuvo un 97.17% de precisión de la clasificación, mientras que las instancias positivas incorrectamente clasificadas fueron 14, así como las instancias negativas incorrectamente clasificadas también fueron 14. La Tabla 1 presenta la tabla de contingencia resultante.

		Predicción de la clase	
		muerte	mejoría
Clase real	muerte	315	14
	mejoría	14	648

Tabla 1 Matriz de confusión/Tabla de contingencia

Conclusiones

Las técnicas de inteligencia artificial como la minería de datos resultan ser de gran apoyo para el diagnóstico y tratamiento de pacientes si son utilizadas para fortalecer la toma de decisiones por parte de los profesionales de la salud. Algoritmos como el J48 generan reglas comprensibles para realizar pronósticos certeros.

De los 208 atributos, solamente se utilizaron 22 de ellos para la generación del árbol de decisión, lo que significa que solamente el 10.58% de ellos contiene ganancia de información, asimismo, los modelos encontrados en la presente investigación pueden ser utilizados para posteriores trabajos en los que se pretenda predecir el egreso de los pacientes y así dar prioridad a los pacientes que presenten los factores clínicos con mayor ganancia de información para evitar el egreso por muerte.

Los resultados demuestran la necesidad de aplicar minería de datos para generar conocimiento y así tomar decisiones de una manera estratégica, es necesaria la implementación de la inteligencia artificial en los sistemas datos que recogen en la historia clínica de un paciente, pues pueden conllevar a obtener conocimiento útil para el apoyo al diagnóstico y tratamiento del paciente y así aprovechar la información que se genera en los hospitales.

Referencias

- Abellán-García Sánchez, F., Barón López, F. J., del Barrio Seoane, J., Fernández, O., García Bustelo, X. R., & Íñiguez Romo, A. (2021). *Inteligencia artificial en el campo de la salud. Un nuevo paradigma: aspectos clínicos, éticos y legales*.
 Beatriz Beltrán Martínez, M. (n.d.). *Minería de datos*.
 De, T., De Grau, F. I., Armando, J., & Miralles, V. (2021). *Minería de datos aplicada a datos biológicos*.
 García Lambert, Guillermo; García-Hernández, René Amulfo; Ledeneva, Y. (2012). *Descubrimiento de factores que inciden en la eficiencia terminal de estudiantes de educación superior con árboles de decisión*. Universidad Autónoma del Estado de México.
 Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2005). *Introducción a la Minería de Datos* (P. P. Hall (ed.); Issue 677).
 Manta, B., Sarkisian, A. G., García-Fontana, B., & Pereira-Prado, V. (2022). *Fisiopatología de la enfermedad COVID-19*. *Odontostomatología*, 24. <https://doi.org/10.22592/ode2022n39e312>
 Organización Mundial de la Salud. (2020, March). *La OMS caracteriza a COVID-19 como una pandemia. La OMS Caracteriza a COVID-19 Como Una Pandemia*.
 Russel, S. (2004). *Inteligencia artificial: Un enfoque práctico* (Pearson Prentice Hall (ed.)).
 Witten, I. H., & Frank, E. (2005). *Practical machine learning tools and techniques. in Data Mining: Practical machine learning tools and techniques* (2nd ed.). Elsevier Inc.
 Zhong, N., Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., ... Zhu, S. (2020). *Clinical Characteristics of Coronavirus Disease 2019 in China*. *New England Journal of Medicine*, 382(18), 1708–1720. <https://doi.org/10.1056/nejmoa2002032>