DESDE 2023
https://revista.ccaitese.com

RICT Revista de Investigación Científica, Tecnológica e Innovación

ISSN: 2992-7315

Publicación Semestral RICT Vol. 2 No. 3 (2024) Páginas 1 – 5

Predicción de la condición de hospitalización para pacientes Covid-19 utilizando modelos de clasificación
Prediction of hospitalization condition for Covid-19 patients using classification models

Alberto Bautista-Loaiza[a], Francisco-Jacob Ávila-Camacho[b]

[a] *Maestría en Ingeniería en Sistemas Computacionales, Tecnológico de Estudios Superiores de Ecatepec, 55210, Ecatepec, Estado de México, México.*
[b] *División de Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México / TES Ecatepec, 55210, Ecatepec, Estado de México, México*

Resumen

Se propone utilizar modelos de inteligencia artificial para predecir si un paciente con COVID-19 requerirá hospitalización basado en síntomas. Con el propósito de apoyar a los servicios de salud que sobrepasan su capacidad de atención en la pandemia. Se utiliza 13,757,682 registros de pacientes de México considerando 12 características que influyen en la evolución de la enfermedad, la recuperación de la información fue a través de la publicación de la dirección de epidemiología del 25 enero del 2022, el entrenamiento de los modelos se lleva a cabo con algoritmos de regresión logística y redes neuronales. Al término de diversos ajustes ambos modelos obtuvieron una precisión cercana al 80%. Se concluye que estos modelos deben considerarse para apoyo de diagnósticos médicos para determinar la necesidad de hospitalización de pacientes con COVID-19 en México. La metodología consistió en la recolección de datos, preprocesamiento, entrenamiento y evaluación del desempeño. Se propone que los modelos entrenados se incorporen en aplicaciones web para facilitar su uso en las áreas de salud.

Palabras clave: Covid-19, regresión logística, clasificación, predicción, machine learning.

Abstract

It is proposed to use artificial intelligence models to predict whether a patient with COVID-19 will require hospitalization based on symptoms. With the purpose of supporting health services that exceed their capacity to care in the pandemic. 13,757,682 patient records from Mexico are used, considering 12 characteristics that influence the evolution of the disease, the recovery of the information was through the publication of the Epidemiology Directorate on January 25, 2022, the training of the models is carried out carried out with logistic regression algorithms and neural networks. After various adjustments, both models obtained an accuracy close to 80%. It is concluded that these models should be considered to support medical diagnoses to determine the need for hospitalization of patients with COVID-19 in Mexico. The methodology consisted of data collection, preprocessing, training and performance evaluation. It is proposed that the trained models be incorporated into web applications to facilitate their use in health areas.

Keywords: Covid-19, logistic regression, classification, prediction, machine learning

1. Introducción

El campo de salud se ha visto beneficiadas a lo largo de la existencia de la inteligencia artificial, la cual ha logrado identificar patrones ocultos en el análisis de estudios de áreas como cardiología, genética, neurología, radiología, oncología, etc. (Ardakani, 2020).

La enfermedad COVID-19 causada por el coronavirus SARS COV 2 ha afectado a millones de personas al rededor del mundo, dado que es una enfermedad de la cual no se tienen mucha información por su corto tiempo de estudio se desconoce de manera específica como reaccionara en diversas personas, ya que no todas las personas tienden a tener los mismos síntomas ni en la misma magnitud lo que ha ocasionado saturación de hospitales con pacientes con

*Autor para la correspondencia: albertoblmsc@gmail.com
Correo electrónico: albertoblmsc@gmail.com (Alberto Bautista-Loaiza), fjacobavila@tese.edu.mx (Francisco-Jacob Ávila-Camacho).

Historial del manuscrito: recibido el 03/11/2023, última versión-revisada recibida el 13/12/2023, aceptado el 15/01/2024, en línea (postprint) desde el 02/02/2024, publicado el 09/04/2024. **DOI:** https://doi.org/10.2992/rict.v2i3.33

síntomas graves (Martínez-Ortega, 2019). Para apoyar en esta situación se considera desarrollar un modelo eficaz de clasificación para realizar predicciones sobre si un paciente con COVID-19 evolucionara de tal forma que requiera ser hospitalizado basado en algoritmos de aprendizaje automático entrenado con casos documentados en la república mexicana. para tal propósito se deberán tener en cuenta los siguientes puntos.

La predicción que llevará a cabo el modelo computacional será solo de apoyo para tomar precauciones, no sustituye la observación médica durante el padecimiento.

Cabe resaltar que el modelo se creara con casos de COVID-19 de la población mexicana por lo tanto un factor será la localidad, ya que el modelo debe considerar las características regionales y con ello una parte de su estilo de vida la cual debe ser lo más similar posible para la etapa de aprendizaje, por lo tanto, el modelo obtenido no podrá aplicarse a personas que no sean residentes de México (Epidemiología, 2022).

La creación de herramientas de este tipo tiene sus orígenes desde la antigüedad, No obstante, fue hasta a mediados del siglo XX que aparecieron herramientas tangibles que se podrían considerar maquinas con aprendizaje. (Ardakani, 2020)

Vaishya, Javaid, Khan, & AbidHaleemb (2020) publicaron un trabajo sobre aplicaciones con inteligencia artificial para la pandemia de COVID-19 puntualizan los principales usos de estas tecnologías en tiempos de pandemia de COVID-19, con estas herramientas se puede evaluar rápidamente síntomas anormales y alertar al servicio médico y con ello tomar decisiones en menor tiempo implementando Support Vector Machine (SVM). (Vaishya, 2020).

Una tarea para estas aplicaciones en una pandemia es apoyar en la creación de medicamentos y vacunas específicas para el virus, ya que en los últimos años la inteligencia artificial ha tomado popularidad en las investigaciones de fármacos mediante análisis de datos con algoritmos de reducción de dimensión junto con el análisis de la información disponible de COVID-19 (Escudero, 2021).

Otra implementación de esta tecnología es la que se emplea en la fase de pruebas en nuevos fármacos donde estas pruebas se realizan en tiempo real, cuando las pruebas tradicionales requerían mucho tiempo y trabajo esto a través de big data y la ciencia de datos (Díaz, 2020). Con la aparición de estas herramientas ayudaron a recortar el tiempo de pruebas de una forma significativa, trabajo que no podría ser realizado de la misma forma por una persona. Con el análisis de datos en tiempo real, los modelos brindan resultados que ayudan a prevenir la propagación de la enfermedad (Chávez Martínez, 2019). La publicación concluye que la inteligencia artificial se convertirá en un aliado indispensable en el futuro contra epidemias y pandemias, así como un gran apoyo en medidas preventivas y correctivas contra muchas otras enfermedades (Vaishya, 2020).

2. Materiales y Método

Para poder crear un modelo de predicción de pacientes COVID-19 se necesita una fuente de información confiable que contenga las principales características que influyen sobre la evolución de los pacientes con esta enfermedad y poder clasificar si el paciente será ambulatorio o requerirá hospitalización. Para llegar a los objetivos se tiene que seguir los siguientes puntos:

- Recabar información acerca de los casos asociados a COVID-19 en México.
- Elaborar una metodología para la exploración de datos con algoritmos de aprendizaje automático.
- Desarrollar y comparar el desempeño de clasificadores desarrollados con diferentes algoritmos de aprendizaje automático.
- Determinar si algún clasificador es adecuado para realizar un pronóstico confiable de que el paciente con Covid-19 requiera ser hospitalizado.

Desde los primeros casos de COVID-19 en México la Dirección General de Epidemiología emitió desde el 12 de abril del 2020 como materia de datos abiertos documentos CVS con casos asociados a personas atendidas por sospechas de COVID-19 con el propósito de facilitar a todos los usuarios que la requieran, el acceso, uso, reutilización y redistribución de la misma (Epidemiología, 2022).

Donde podemos encontrar múltiples características de los pacientes atendidos incluido Tipo Paciente el cual especifica si el paciente fue hospitalizado o fue un caso ambulatorio, Por tanto, podemos usar esta información para entrenar un modelo de predicción sobre qué tipo de paciente será la persona infectada, actualmente ya se tienen vacunas con lo que se concluye que la información de la base de datos contiene personas con y sin vacuna dependiendo la fecha del registro, Por lo tanto se tomarán en cuenta solo los datos a partir de que el sector salud notifico un esquema completo de vacunación en la población adulta mexicana el día 29 de octubre de 2021 y hasta a la fecha 25 de enero del 2022.

Dado que la predicción mantendrá una relación de dependencia entre las características del conjunto de datos es necesario identificarlas como variables dependientes e independientes para determinar que pacientes necesitarán hospitalización y que pacientes serán ambulatorios.

Las variables independientes son los factores de riesgo que posee una persona, es lo que lo hace susceptible a desarrollar síntomas graves, el sector salud ha mencionado cuales son dichos factores que determinan la evolución del COVID-19 (Gutiérrez, 2022), con esta información se identifican los campos del conjunto de datos que se emplean como variables independientes, serían los siguientes.

- ASMA
- CARDIOVASCULAR
- DIABETES
- EDAD
- EMBARAZO
- EPOC
- HIPERTENSION
- INMUSUPR

- OBESIDAD
- RENAL_CRONICA
- SEXO
- TABAQUISMO

Se tomaran para el entrenamiento las 12 variables independientes, para las variables tipo dependientes solo TIPO_PACIENTE ya que si una persona requirió cuidados intensivos así como si presentó un cuadro de neumonía se concluye que dicha persona requirió hospitalización , es por eso que solo la variable tipo paciente es la cual nos indica si la persona contagiada requirió hospitalización o no. Los posibles valores de cada una de las características del conjunto de datos se observan en la tabla 1.

Tabla 1. Diccionario de datos

Característica	Valor	Descripción	Tipo
TIPO_PACIENTE	0	Paciente no requiere ser hospitalizado	Dependiente
	1	Paciente requiere ser hospitalizado	
SEXO	0	Mujer	Independiente
	1	Hombre	
EDAD	0 a 99	Edad numérica del paciente	Independiente
EMBARAZO	0	Si embarazada	Independiente
	1	No embarazada	
DIABETES	0	Padece diabetes	Independiente
	1	No padece diabetes	
EPOC	0	Padece EPOC	Independiente
	1	No padece EPOC	
ASMA	0	Padece asma	Independiente
	1	No padece asma	
INMUSUPR	0	Padece INMUSUPR	Independiente
	1	No padece INMUSUPR	
HIPERTENSION	0	Padece hipertensión	Independiente
	1	No padece hipertensión	
CARDIOVASCULAR	0	Padece enfermedad Cardiovascular	Independiente
	1	No Padece enfermedad Cardiovascular	
OBESIDAD	0	Tiene obesidad	Independiente
	1	No tiene obesidad	
RENAL_CRONICA	0	Padece alguna enfermedad renal	Independiente
	1	No padece enfermedades renales	
TABAQUISMO	0	Es fumador	Independiente
	1	No es fumador	

Al observar la cantidad de datos de entrenamiento podemos notar que nos encontramos con información desbalanceada ya que 860,989 registros son de personas que no requirieron hospitalización y 36,403 de personas que si requirieron hospitalización.

Lo cual equivale a un 95.94 % No hospitalizadas y 4.05 % hospitalizadas del total de los datos.

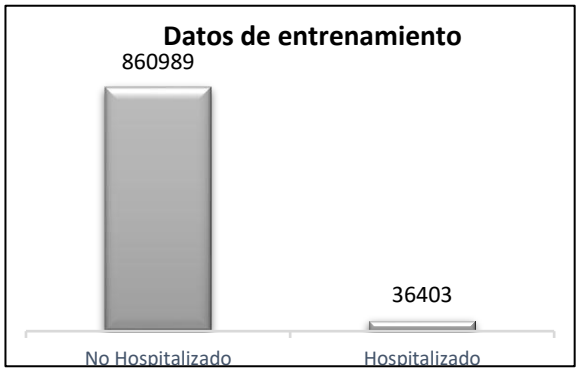

Figura 1. Tipo de pacientes.

Con el IDE Spyder que nos permite utilizar las librerías scikit-learn en Python se crean modelos de clasificación con el conjunto de datos

Se realiza un submuestreo 1 a 1 aleatorio sobre los datos de personas no hospitalizadas que representan el 95.94 % de la información total para obtener un conjunto de datos balanceado posteriormente el 75 % de la información para entrenamiento y el 25 % restante para realizar pruebas.

Para aquellos registros que tienen valores faltantes se les considera la media del campo y se ajustan los datos de entrenamiento como de pruebas a una escala similar para terminar el preprocesamiento de datos.

2.1. Entrenamiento y predicción con regresión logística

Para la correcta elección de algoritmos en la clasificación binaria se evalúan los algoritmos Regresión Lineal, Regresión Logística, Arboles de Decisión, SVR, Clustering, Redes Neuronales y Naive Bayes , que se usan comúnmente para este tipo de aplicaciones, en base a los resultados se observa que los algoritmos regresión logística y redes neuronales obtienen un mejor desempeño por lo cual son estos dos algoritmos a los que se les da continuidad en su configuración.

La Regresión Logística es un Algoritmo Supervisado y se utiliza para la clasificación binaria (González Segoviano, 2023), partiendo de nuestros datos de entrenamiento y prueba procedemos a crear nuestro modelo con la función LogisticRegression que nos proporciona sklearn (AWS, 2023). Para poder visualizar el desempeño del algoritmo de regresión logística obtenemos la matriz de confusión que nos permitirá observar las predicciones correctas e incorrectas de cada clase. En la siguiente tabla se observa la matriz de confusión con la relación entre valores de predicción y valores reales

Tabla 2. Matriz de confusión de predicción aplicando regresión logística

Predicción/Reales	Verdaderos	Falso
Verdadero	7550	1467
Falso	2762	6423

2.2. Entrenamiento y predicción con redes neuronales

Las redes neuronales artificiales son un algoritmo que se puede implementar para la clasificación supervisada binaria, dada su buena fama para aplicaciones médicas implementamos una predicción con este algoritmo (Céspedes, 2021), partiendo de nuestros datos de entrenamiento y con ayuda de Keras la cual es una librería de redes neuronales escrita en de Python (APD, 2019).

Se construye una red neuronal profunda de 3 capas, se declara con el método dense, 12 neuronas en la capa de entrada, 24 neuronas en la capa oculta y una neurona de salida, la función de activación relu para las capas de entrada como oculta y sigmoid para la capa de salida.

Se realiza la compilación de la red neuronal con *binary_crossentropy* como función de perdida que se utiliza para evaluar el grado de error entre salidas calculadas y las salidas deseadas de los datos de entrenamiento. un optimizador Adam y se realiza el entrenamiento con una cantidad de iteraciones de 250. En la siguiente tabla se muestra la matriz de confusión del modelo con redes neuronales una vez realizada la predicción.

Tabla 3. Matriz de confusión aplicando redes neuronales

Predicción/Reales	Verdaderos	Falso
Verdadero	7735	1282
Falso	2801	6384

3. Resultados

En la tabla 4 compararemos los resultados de la matriz de confusión, de cada uno de los algoritmos de regresión logística y redes neuronales.

Tabla 4.Comparación de matriz de confusión de ambos algoritmos

Predicción	Regresión logística	Redes neuronales
Verdaderos positivos	7550	7735
Verdaderos negativos	6423	6384
Falsos positivos	2762	2801
Falsos negativos	1467	1282

Las mediciones obtenidas de ambos modelos se observan en la tabla 5.

Tabla 5. Comparación de las curvas ROC de ambos modelos

Indicador	Regresión logística	Redes neuronales
Precisión	0.8113541534960172	0.8327680667884164
Exactitud	0.7659597846390507	0.7756839907702451
Sensibilidad	0.6986390854654327	0.6950462710941753
Puntaje f1	0.7507897507897509	0.7576998397721203

Se crean las gráficas ROC de ambos modelos que se puede apreciar en la figura 2 las cuales se calculan a partir de los resultados obtenidos y con ayuda de la librería matplotlib.

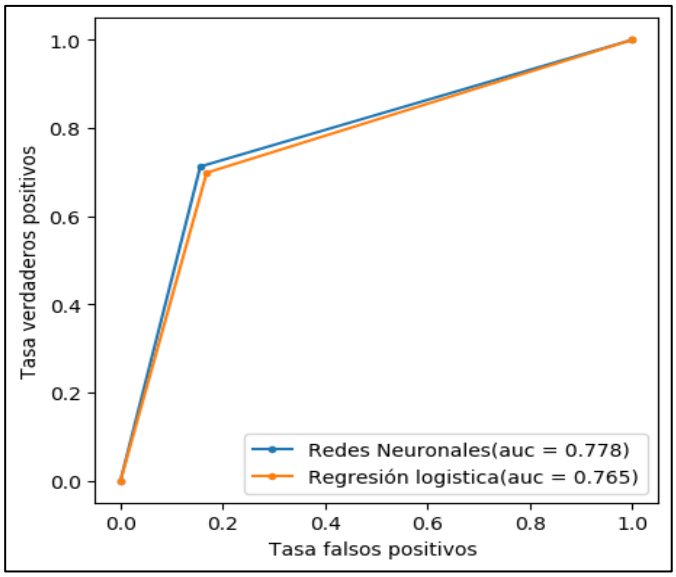

Figura 2. Comparación de las curvas ROC de ambos modelos.

4. Discusión

En el algoritmo de regresión logística no hubo cambios significativos al modificar algún parámetro de entrenamiento. Para el algoritmo de redes neuronales, se realizaron diversas pruebas, la configuración con el mayor desempeño fue con las siguientes características, la capa de entrada en 12, la capa oculta en 24, y la cantidad de interacciones 250, con esta configuración se obtuvo el resultado para redes neuronales mostrado en la tabla 5.

Para esta implementación obtenemos que el algoritmo de clasificación de redes neuronales tiene un mejor desempeño al predecir si un paciente requerirá hospitalización, aunque la ventaja de este algoritmo sobre el de regresión logística no es significativo, siendo ambos óptimos para el objetivo ya que como se aprecia en la Figura 2 la diferencia del área bajo la curva de ambos algoritmos regresión logística y redes neuronales es mínima. Esta razón fue calculada en base al desempeño de cada modelo, a partir del resultado de la razón de verdaderos positivos que se obtiene a partir de la formula:

$$VPR = \frac{VP}{P} = \frac{VP}{VP+FN}$$

Y de la razón de falsos positivos de la formula:

$$FPR = \frac{FP}{N} = \frac{FP}{FP+VN}$$

Con el resultado de las operaciones se puede dibujar la curva ROC y con ello el área bajo la curva de los modelos (Darlington, 2020). Los resultados son los siguientes, para regresión logística el AUC = 765 y para redes neuronales AUC = 778 lo cual es levemente mayor al área bajo la curva de clasificador entrenado con el algoritmo de regresión logística.

5. Conclusiones

Con la generación de clasificadores con ambos algoritmos se determina que se alcanza los objetivos propuestos ya que con la información recolectada se tuvo la capacidad para crear modelos de clasificación de personas infectadas con la enfermedad de COVID-19, con el conjunto de datos creado a partir del preprocesamiento junto con las técnicas de inteligencia artificial lograron generar clasificadores con un 80% de precisión lo cual se considera un modelo que puede ser probado en pacientes futuros al igual como en la literatura expuesta cuando se aplicaron las herramientas con un 70% de efectividad en diversos diagnósticos en la aplicación de algoritmos de regresión y clasificación.

Para incrementar la efectividad de los modelos sería necesario conocer otras características de los pacientes por ejemplo desde cuando padece alguna enfermedad de las mencionadas o días transcurridos desde el primer síntoma de COVID-19, con esta nueva información complementaria se tendrá un incremento en la precisión de los modelos.

Como se aprecia en los resultados todos los clasificadores mostraron un rendimiento similar en las diversas métricas evaluadas, esto por un lado se debe a la naturaleza del conjunto de datos la cual no mostró problemas extremos aun cuando la totalidad de la información presenta un claro desbalanceo, pero sin inexistencia de valores atípicos o alguna dominancia marcada de alguna clase. En base a los resultados obtenidos se determina que la implementación de clasificadores desarrollados con los algoritmos de regresión logística y redes neuronales son confiables para determinar la evolución de pacientes, siempre y cuando la información recolectada sea suficiente ya que para un buen aprendizaje los clasificadores basan su efectividad dependiendo en gran medida de la información con la cual se les alimente y por supuesto a la incorporación de los avances tecnológicos en la inteligencia artificial que se tienen día con día, Esto se reflejara en herramientas de diagnóstico cada vez más innovadoras enfocadas en ayudar al sector salud y enfrentar cualquier tipo de enfermedad.

6. Trabajos futuros

Como continuación de este trabajo de tesis en un futuro próximo a fin de realizar predicciones se plantea la posibilidad de incorporar el modelo clasificación en una aplicación web los cuales son sencillos de utilizar y no requieren conocimientos avanzados de informática, además de ser personalizables a gusto y adaptarse a la mayoría de formas de trabajo actuales donde lo único que tendría que realizar el operador es capturar las características del paciente para obtener el resultado de hospitalizado o ambulatorio sin necesidad de algún conocimiento extra.

Un trabajo que se debe realizar periódicamente es actualizar la información del conjunto de datos con las diversas publicaciones de la dirección general de epidemiología sobre la base de datos publica de COVID-19 ya que recordemos que cada día hay nueva información incluso durante el desarrollo de este trabajo ha surgido información importante de diversas fuentes sobre COVID-19 en México, una continuación al proyecto es buscar mejorar el desempeño de los modelos de clasificación con una retroalimentación sobre el conjunto de datos con más registros o incluso incorporando más características que ayuden a determinar la evolución de la enfermedad en el paciente cada vez con más precisión.

7. Referencias

APD. (04 de 04 de 2019). *APD.* Obtenido de https://www.apd.es/algoritmos-del-machine-learning/

Ardakani, A. A. (30 de 03 de 2020). *National library of medicine.* Obtenido de https://www.sciencedirect.com/science/article/pii/S0010482520301645

AWS. (01 de 01 de 2023). *AWS.* Obtenido de https://aws.amazon.com/es/what-is/logistic-regression/#:~:text=La%20regresi%C3%B3n%20log%C3%ADstica%20es%20una,factores%20bas%C3%A1ndose%20en%20el%20otro.

Céspedes, F. A. (28 de 12 de 2021). *Facultad de Ingeniería de Sistemas e Informática - UNMSM.* Obtenido de https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/download/21862/17543/75813

Chávez Martínez, R. (2019). *Universidad de LIma*. Obtenido de https://repositorio.ulima.edu.pe/handle/20.500.12724/8401

Cid, A. S. (29 de 10 de 2021). *EL PAÍS.* Obtenido de https://elpais.com/mexico/2021-10-29/mexico-concluye-su-plan-de-vacunacion-un-83-de-la-poblacion-mayor-de-edad-tiene-al-menos-la-primera-dosis.html

Darlington, K. (22 de 05 de 2020). *BBVA Open Mind.* Obtenido de https://www.bbvaopenmind.com/tecnologia/inteligencia-artificial/esta-ayudando-la-inteligencia-artificial-contener-la-pandemia-covid-19/

Díaz, J. E. (29 de 07 de 2020). *Universidad de Cundinamarca*. Obtenido de https://revistes.ub.edu/index.php/RBD/article/view/31643

Epidemiología, D. G. (25 de 01 de 2022). *Secretaría de Salud.* Obtenido de https://www.gob.mx/salud/documentos/datos-abiertos-152127

Escudero, X. (24 de 03 de 2021). *Archivo de cardeologia de México.* Obtenido de https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-99402020000500007

Etecé, E. (05 de 08 de 2021). *Editorial Etecé.* Obtenido de https://concepto.de/diagnostico/

González Segoviano, L. J. (30 de 9 de 2023). Regresión logística vs árboles de decisión en el riesgo crediticio. (T. e. RICT Revista de Investigación Científica, Ed.) *RICT Revista de Investigación Científica, Tecnológica e Innovación*, *1*(2), 32-37. Obtenido de https://revista.ccaitese.com/index.php/ridt/article/view/21

Gutiérrez, V. F. (2022). *Ecografía en el manejo del paciente crítico con infección por SARS-CoV-2 (COVID-19): una revisión narrativa.* Obtenido de https://www.sciencedirect.com/science/article/pii/S0210569120301558?via%3Dihub

Martínez-Ortega, A. G. (22 de 12 de 2019). *Pro Sciences.* Obtenido de https://d1wqtxts1xzle7.cloudfront.net/87614928/287162093-libre.pdf?1655404808=&response-content-disposition=inline%3B+filename%3DTecnologias_en_la_inteligencia_artificia.pdf&Expires=1695936710&Signature=HNx9yigjuH3nIQY0ZJx8G5kahLtl7svfwgOCVc72rlUBspJf0K9h

Novás, J. D. (2006). *Revista Cubana de Medicina General Integral.* Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21252006000100007#:~:text=Cuando%20unimos%20los%20s%C3%ADntomas%20y,o%20de%20otra%2C%20cu%C3%A1les%20son

Shibly, K. H. (2020). *sciencedirect.* Obtenido de https://www.sciencedirect.com/science/article/pii/S2352914820305554

Vaishya, R. J. (4 de 08 de 2020). *Facultad de Ingeniería de Sistemas e Informática - UNMSM.* Obtenido de https://www.sciencedirect.com/science/article/pii/S1871402120300771