DOI: https://doi.org/10.46480/esj.4.2.41

Predicción de Covid19 con el uso del Algoritmo *Random Forest* y Redes Neuronales Artificiales

Prediction of Covid19 with the use of Random Forests Algorithm and Artificial Neural Networks

Darwin Patiño Pérez[1], Ricardo Silva Bustillos[2], Celia Munive Mora[3], y Miguel Botto Tobar[4]

RESUMEN

En la actualidad el SARS-CoV-2 o Covid19 como se lo conoce, presenta variantes o mutaciones que se propagan rápidamente afectando a las personas, sin que los profesionales de la salud pueden detectarlo oportunamente para dar un tratamiento adecuado y así poder controlar su propagación. En este manuscrito se describe la implementación de un modelo de análisis y predicción de la propagación del Covid19, que mediante técnicas de inteligencia artificial relacionadas con Machine Learning, permitirán aplicar estrategias de aprendizaje supervisado a los programas desarrollados en el lenguaje de programación Python, para que al procesar los grandes volúmenes de datos puedan aprender de las experiencias pasadas y permitan procesar nuevas entradas generando la información de predicción de forma rápida y confiable. El enfoque de hacer un análisis sobre un conjunto de datos extraído de una fuente abierta, servirá para posteriormente realizar un análisis exploratorio de lo procesado. Se realizaron tres predicciones que son: Si el paciente tiene SARS-CoV-2, días transcurridos hasta la mortalidad y la mortalidad por covid, utilizando algoritmos de clasificación y regresión que de acuerdo a estudios previos permitieron seleccionar y aplicar el modelo algorítmico *Random Forest* y Redes Neuronales Artificiales cuyas métricas de confiabilidad nos permiten aceptar las predicciones esperada para una adecuada toma de decisión.

Palabras clave: Covid19, *Dataset*, Redes Neuronales, *Random Forest,* Clasificación, Regresión, Predicción.

ABSTRACT

Currently SARS-CoV-2 or Covid19 as it is known, has variants or mutations that spread rapidly affecting people, without the health professionals being able to detect it in a timely manner to give an adequate treatment and thus be able to control its spread. This manuscript describes the implementation of an analysis and prediction model of the spread of Covid19, which through artificial intelligence techniques related to Machine Learning, will allow the application of supervised learning strategies to programs developed in the Python programming language. so that when processing large volumes of data they can learn from past experiences and allow new inputs to be processed, generating prediction information quickly and reliably. The approach of making an analysis on a data set extracted from an open source, will serve to later carry out an exploratory analysis of the processed. Three predictions were made, which are: If the patient has SARS-CoV-2, days elapsed until mortality and mortality from covid, using classification and regression algorithms that, according to previous studies, allowed the selection and application of the Random algorithmic model Forest and Artificial Neural Networks whose reliability metrics allow us to accept the expected predictions for an adequate decision making.

Keywords: Covid19, Dataset, Neural Networks, Random Forest, Classification, Regression, Prediction.

Fecha de recepción: Marzo 30, 2020.
Fecha de aceptación: Septiembre 15, 2020.

Introducción

Existe mucha controversia a medida que se propaga un virus, aún más cuando el virus es nuevo ya que la afectación se deriva a todos los organismos sociales, pero sin duda la más afectada es el área de la salud. Actualmente hay probabilidades que hacen que se quiera saber hasta donde se llegara a propagar el virus y para eso se eligió el tema de aplicar un modelo que prediga hasta dónde puede llegar un virus, con base, los datos que se van recopilando y aplicando un algoritmo que ayude a deducir este problema

Mundialmente hemos sido afectados por un tipo de coronavirus, estos corresponden a microorganismos que surgen de forma periódica en distintos partes del mundo y que son causantes de la denominada infección Respiratoria Aguda con distinto nivel de calibre.

El SARS-COV-2 (Covid-19) tiene decenas de millones de afectados en personas a nivel mundial(Jhu, n.d.) desde que fue declarada la pandemia por la Organización mundial de la salud(OMS) (WHO, n.d.).

[1] Ph.D. Universidad de Guayaquil, Ecuador. E-mail: darwin.patinop@ug.edu.ec
[2] Ph.D. Villanova University, Pensilvania, EEUU. E-mail: ricardo@alluriam.com
[3] BS, DeSales University, Pensilvania, EEUU. E-mail: cm3877@desales.edu
[4] Msc, Universidad de Guayaquil, Ecuador. E-mail: miguel.botto@ug.edu.ec

Como citar: Patiño Pérez, D., Silva Bustillos, R., Munive Mora, C., & Botto-Tobar, M. (2020). Predicción de Covid19 con el uso del Algoritmo Random Forest y Redes Neuronales Artificiales. *Ecuadorian Science Journal*. 4(2), 101-110. DOI: https://doi.org/10.46480/esj.4.2.41

El virus tiene una afectación tanto física como psicológica (Pfefferbaum & North, 2020) los síntomas son tos, dolor de garganta y/o fiebre, hasta un cuadro de neumonía, es una emergencia de salud pública grave, particularmente mortal en poblaciones y comunidades vulnerables en la que los proveedores de atención no están suficientemente preparados. La OMS recomienda mantener una distancia de al menos un metro con los demás y que el uso de la mascarilla sea una parte normal de su interacción con otras personas (Xie et al., 2020).

La problemática que se plantea es la propagación rápida del covid19, pero tomando en cuenta escenarios e indicadores que se registran en un Dataset(Shakouri et al., 2021), así como un conjunto de datos(Tworowski et al., 2021) que estando disponibles se necesitarían técnicas de *machine learning* para crear un modelo de análisis y predicción favorable para la problemática descrita.

En la actualidad existen muchos métodos para llegar a encontrar un análisis estructurado de lo que se desea investigar, de la misma manera hacer predicciones acerca de un tema, se debe hacer un levantamiento de información acerca de lo que se va a investigar, para poder aplicarlo al modelo. Para poder hacer el levantamiento de información se debe recurrir varias fuentes oficiales para extraerla, es dificultoso la recolección de información en ciertas fuentes oficiales ya que se extraerá cantidades grandes de información en donde se probará el modelo y ocupará una gran cantidad de recursos.

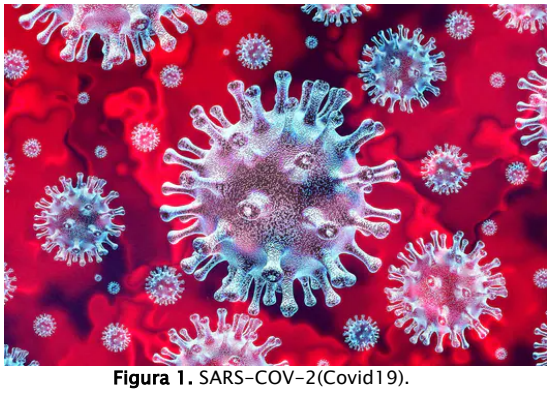

Figura 1. SARS-COV-2(Covid19).
Fuente: Propia

En el presente trabajo de investigación tiene como objetivo Implementar un modelo de predicción basado en técnicas de *machine learning* (Brownlee, 2016) que analice datos relacionados con Covid-19 y pueda predecir niveles de afectación en las personas a nivel mundial. Las pruebas de predicción que se realizarán con los algoritmos de *machine learning* usando redes y arboles estarán basados en:

- El uso de algoritmos de clasificación y regresión en Redes Neuronales que permitan conformar un modelo de predicción.
- Que se ajuste al tipo de predicción que se va a encontrar.
- Obtener predicciones que estén dentro del rango al que pertenecen.
- Evitar sobreajustes del modelo que se entrenó con el conjunto de datos.

Para ellos las Redes Neuronales Artificiales (RNA) han constituido en los últimos tiempos un foco de investigación importante y con una actividad intensa, siendo un paradigma de aprendizaje computacional muy extendido en la resolución de problemas de diversas áreas de la Ingeniería y la Ciencia (De La Hoz, De La Hoz, & Fontalvo, 2019).

Las Redes Neuronales Artificiales son un modelo computacional inspirado en el funcionamiento del cerebro del ser humano, y que poseen una elevada capacidad de generalización y de tratamiento de problemas (Raschka & Mirjalili, 2019) el cual se ha elegido para las predicciones, en el que emplea un modelo que toma en cuenta la afectación en las personas en cuanto a covid-19, ya que ha demostrado realizar predicciones adecuadas en enfermedades (B S, 2021).

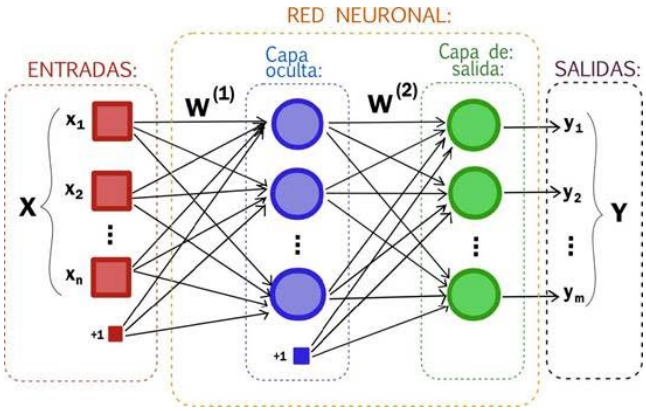

Figura 2. Red Neuronal Artificial
Fuente: Propia

En consecuencia, a los estudios previos e indagaciones del tema de proyecto propuesto, hay que tomar en cuenta de donde se ha tomado todos los datos para desarrollarlo, encontrando diversas fuentes que dan a reconocer cada parte del tema de proyecto, para visualizar la relación que tiene en consecuencia de lo investigado. Las llamadas redes neurales o modelos conexionistas han ido progresivamente utilizándose como herramientas de predicción y clasificación. De forma breve una red neural es un sistema informático reticular (de inspiración neuronal) que aprende de la experiencia mediante la auto modificación de sus conexiones (Lee, Kim, Jeong, & Choi, 2018).

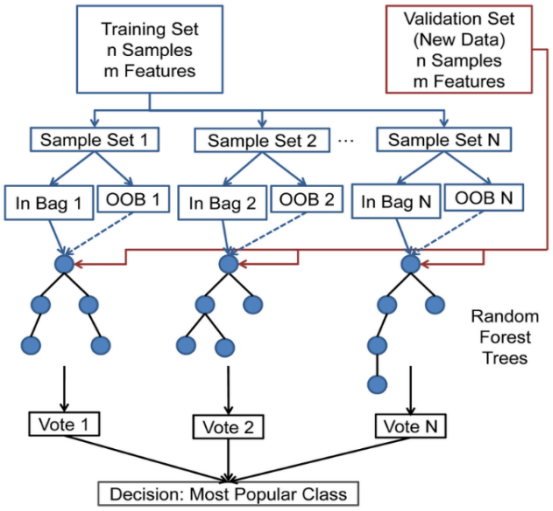

Figura 3. Modelo Random Forest
Fuente: Propia

Random Forest, son bosques de decisión aleatorios formados por un conjunto de árboles de decisión, estos bosques, se forman mediante un algoritmo que introduce una aleatoriedad para reducir la correlación entre los árboles(Breiman, 1999). Este es un algoritmo de muy buenos precedentes ya que pertenece como técnica de aprendizaje supervisado su función es generar diversos arboles de decisión dependientes de un Dataset que este en entrenamiento, el cual arrojara predicciones de acuerdo con lo que se quiera ajustar.

La ventaja de *Random Forest* hacen que se convierta en una técnica ampliamente utilizada en muchos campos, por ejemplo, teledetección(para clasificación de imágenes) (Chowdhury, Chatterjee, & Banerjee, 2019), bancos (para detección de fraudes y clasificación de clientes para otorgamiento de crédito) (Chowdhury et al., 2019), medicina (para analizar historiales clínicos a fin de identificar enfermedades potenciales en los pacientes), finanzas (para pronosticar comportamientos futuros de los mercados financieros) y comercio electrónico (para pronosticar si un cliente comprará, o no, cierto producto), entre otros.

La validación en cuanto a resultados en las 3 predicciones propuestas de la afectación de Covid en las personas, el empleo de modelos predicción para enfermedades infecciosas ha aumentado en grado significativo en los postreros años, debido que suministran información ventajosa para tomar decisiones, y establecer medidas activas en el control o erradicación de una enfermedad infecciosa.

A continuación, se describe las 3 predicciones que se ha realizado con el algoritmo de Redes Neuronales y también se realizara una comparación de resultados de *Random Forest*.

1. Si el paciente tiene SARS-CoV-2 o no.
2. Días transcurridos hasta la mortalidad.
3. Mortalidad.

Y no solo eso, teniendo en cuenta que nuestro modelo tenga buenos resultados después de haber hecho las pruebas correspondientes y el tratamiento adecuando del conjunto de datos, también hemos puesto a prueba el modelo utilizando Random Forest para la realización de las mismas 3 predicciones en las cuales hemos comparado los resultados.

Materiales y Métodos

Materiales

En este proyecto se utilizó un conjunto de herramientas como Anaconda con extensión JupiterLab (Källén & Wrigstad, 2021), Google Colab(Fuat, 2018), Google Drive (Gallaway & Starkey, 2013) se enlaza con Google colab extensión gratuita de Google para probar el programa en un ambiente donde todos los investigadores tienen acceso, todas estas herramientas sirvieron para el desarrollo de la investigación.

Conjunto Datos

Se extrajo un Dataset desde la página oficial de Datos abiertos registrados en Kaggle, que contiene datos que sirven para la experimentación que se hizo en este caso el análisis y predicción de la propagación de un virus y la afectación en las personas, este conjunto de datos contiene campos que ayudan con las variables que se requiere para cada predicción.

Métodos

Análisis Exploratorio

Posteriormente se hizo una limpieza profunda del Dataset Master.csv donde se realizó varios procesos de rutina como eliminación de datos nulos, conversiones de mayúsculas y minúsculas, datos tipo carácter a numéricos, etc.

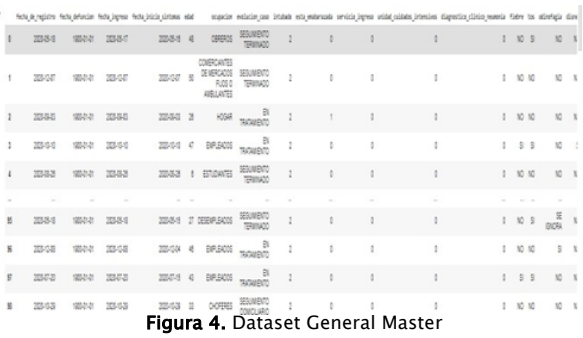

Figura 4. Dataset General Master

Fuente: Propia

El conjunto de datos tiene la cantidad total de 908941 registros después del análisis exploratorio se procedió a escoger las variables de las cuales 49 son de uso exclusivo para las predicciones y posteriormente se procedió hacer la limpieza de la data.

N	NOMBRE DE LA VARIABLE	DESCRIPCION
1	Fecha de registro	Fecha de registro
2	fecha defunción	fecha defunción
3	fecha ingreso	fecha ingreso
4	fecha inicio síntomas	fecha inicio síntomas
5	Sexo	Sexo
6	Edad	Edad
7	nacionalidad -----> No es necesario	nacionalidad -----> No es necesario
8	Ocupación	Ocupación
9	Entidad residencia--------> Dirección	Entidad residencia--------> Dirección
10	municipio residencia -------> Delegaciones	municipio residencia -------> Delegaciones
11	tipo paciente--No es necesario, puede ser reemplaza servicio ingreso	tipo paciente-No es necesario, puede ser reemplaza servicio ingreso
12	Evolución caso	Evolución caso
13	Intubado	Intubado
14	Está embarazada	Está embarazada
15	servicio ingreso	servicio ingreso
16	Unidad cuidados intensivos	Unidad cuidados intensivos
17	diagnóstico clínico neumonía	diagnóstico clínico neumonía
18	Fiebre	Fiebre
19	Tos	Tos
20	Odinofagia	Odinofagia
21	Disnea	Disnea
22	Irritabilidad	Irritabilidad
23	Diarrea	Diarrea
24	Dolor torácico	Dolor torácico
25	Calofríos	Calofríos
26	Cefalea	Cefalea
27	Mialgias	Mialgias
28	Artralgias	Artralgias
29	Ataque al estado general	Ataque al estado general
30	Rinorrea	Rinorrea
31	Polipnea	Polipnea
32	Vomito	Vomito
33	Dolor abdominal	Dolor abdominal
34	Conjuntivitis	Conjuntivitis
35	Cianosis	Cianosis
36	Inicio súbito síntomas	Inicio súbito síntomas
37	Diabetes	Diabetes
38	Época	Época
39	Asma	Asma
40	Inmunosupresivo	Inmunosupresivo
41	Hipertensión	Hipertensión
42	VIH sida	VIH sida
43	Otra condición	Otra condición
44	Enfermedad cardiaca	Enfermedad cardiaca
45	Obesidad	Obesidad
46	insuficiencia renal crónica	insuficiencia renal crónica
47	Tabaquismo	Tabaquismo
48	toma muestra	toma muestra
49	Resultado definitivo	Resultado definitivo

Figura 5. Variables Exclusivas para predicciones

Fuente: Propia

Uso de Algoritmos

A partir de una serie de datos, utilizando algoritmos de aprendizaje automático, es posible interpretar la información para transformarla en conocimiento (Brownlee J., 2019).

Para la presente investigación se utiliza los algoritmos de clasificación y regresión con aprendizaje supervisado (Brownlee, 2016) ya que sirven para las tres predicciones que tendrá el modelo.

Algoritmo de Clasificación

En cuanto algoritmos de clasificación hace la referencia a que encuentra patrones en los datos y los clasifica en grupos en donde señala que La clasificación es una técnica de Minería de Datos que se utiliza para averiguar con qué grupo una instancia de datos está relacionada dentro de un determinado conjunto de datos (Neelamegam & Ramaraj, 2013).

Se debe tener en cuenta en que cuando se aplica algoritmos de clasificación la variable que esta por predecirse en un conjunto de categoría como, por ejemplo: Binarios, múltiples u ordenados.

Algoritmo de Regresión

En las tareas de regresión, el programa de aprendizaje automático debe estimar y comprender las relaciones entre las variables. El análisis de regresión se enfoca en una variable dependiente y una serie de otras variables cambiantes, lo que lo hace particularmente útil para la predicción y el pronóstico (Giraldo Mejía & Vargas Agudelo, 2019).

Dependiente del machine learning supervisado, que establece la relación entre varias características y una variable continua, que se clasifica en los siguiente:

-Regresión lineal. La correlación y la regresión lineales simple son métodos estadísticos que estudian la relación lineal existente entre dos variables. Antes de profundizar en cada uno de ellos, conviene destacar algunas diferencias: La correlación cuantifica como de relacionadas están dos variables, mientras que la regresión lineal consiste en generar una ecuación (modelo) que, basándose en la relación existente entre ambas variables, permita predecir el valor de una a partir de la otra (Pardo. A, & Ruiz, 2005).

-Regresión múltiple La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3…) (Juan, 2003).

Modelo de Red Neuronal

El concepto de redes neuronales es un paradigma informático de procesamiento de información inspirado en los sistemas neuronales biológicos(Agasi et al., 2018) donde se considera a la neurona biológica como un elemento de almacenamiento de información y consisten en un conjunto de elementos simples de procesamiento llamados nodos o neuronas conectadas entre sí (Haglin, Jimenez, & Eltorai, 2019).

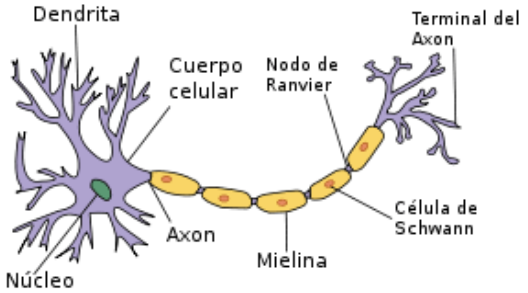

Figura 6. Neurona Biológica.
Fuente: Propia

Existen numerosas formas de definir a las redes neuronales; desde las definiciones cortas y genéricas hasta las que intentan explicar más detalladamente qué son las redes neuronales.

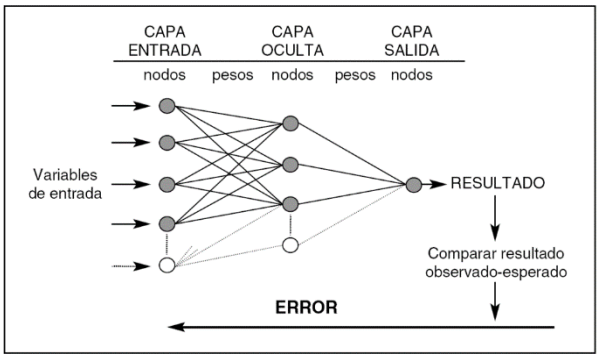

Figura 7. Arquitectura de una red Neuronal
Fuente: Propia

A continuación, se especifica los tipos de activaciones que existen en un algoritmo de Redes neuronales, especificando que en nuestras predicciones se utilizaron dos activaciones esenciales que fueron Relu y Sigmoid, como se observa en la siguiente figura.

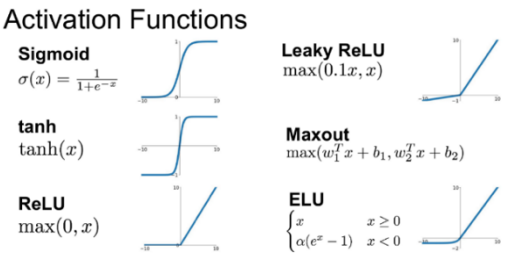

Figura 8. Tipos de activación para RNA[31].
Fuente: Propia

La idea base de este modelo es el de imitar el sistema más complejo que se conoce hasta ahora, el cerebro. Éste está formado por millones de células llamadas neuronas. Estas neuronas son unos procesadores de información sencillos con un canal de entrada de información, un órgano de cómputo y un canal de salida de información (Amato et al., 2013) es un algoritmo de cálculo que se basa en una analogía del sistema nervioso. La idea general consiste en emular la capacidad de aprendizaje del sistema nervioso, de manera que la RN aprenda a identificar un patrón de asociación entre los valores de un conjunto de variables predictoras (entradas) y los estados que se consideran dependientes de dichos valores (salidas)(Haglin et al., 2019).

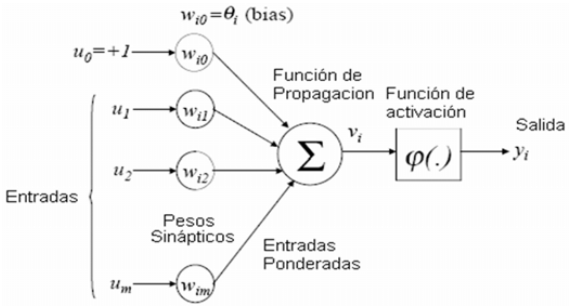

Figura 9. Neurona Artificial
Fuente: Propia

Modelo Random Forest

Random Forest o Bosques Aleatorios a un tipo de clasificador que agrupa un conjunto de árboles estructurados distribuidos de manera idéntica que arrojan un voto unitario para la clase más popular su uso es tanto para tareas de clasificación como regresión usando voto mayoritario y ponderación respectivamente(Yiu, 2019). La combinación de dichos arboles claro está que bajo ciertas condiciones proporciona un mejor resultado dando como resultado un método más preciso, estable, dinámico que busca el equilibrio entre el sesgo y la varianza del bosque("A Survey on Decision Tree Algorithms of Classification in Data Mining," 2016).

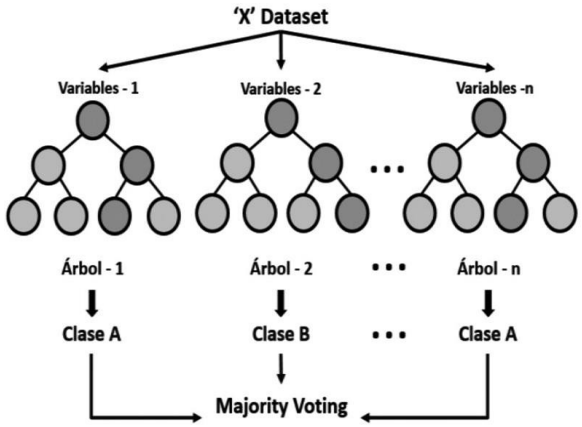

Figura 10. Arquitectura de Random Forest
Fuente: Propia

Aplicación de los Métodos para la Predicción

Redes Neuronales

A. Si el paciente tiene SARS-CoV-2 o no.

1) Arquitectura de predicción

En la siguiente imagen se identifica la arquitectura de redes neuronales por clasificación para la predicción: si el paciente SARS-CoV-2 o no, utilizando una capa de entrada, dos capas ocultas, capa de salida con activación sigmoide.

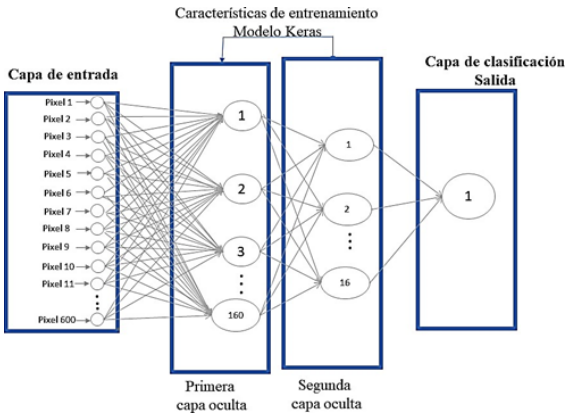

Figura 11. Arquitectura de Predicción
Fuente: Propia

2) Entrenamiento del Dataset.

Una vez realizado las épocas para el entrenamiento del modelo se obtuvo algunos datos de referencia para demostrar la efectividad del entrenamiento y deducir que es el adecuado para el modelo predictivo. Se encontró el valor mínimo de perdida y precisión en este entrenamiento que se visualizara a continuación:

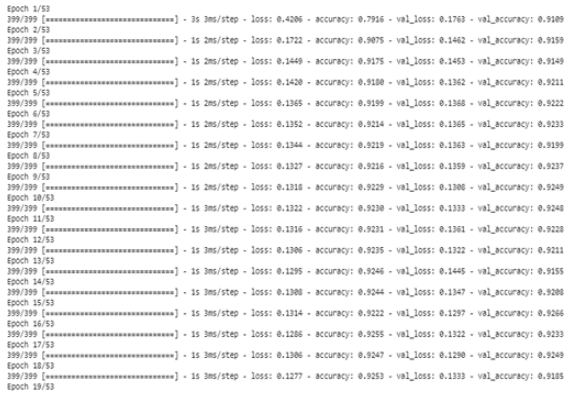

Figura 12. Entrenamiento con los Datos
Fuente: Propia

3) Predicción

En la Figura 13 se observa la comparación entre la curva de datos training y validación en base a precisión donde el valor de y es accuracy y x tiene el valor de épocas de entrenamiento. Como resultado obtenido para la primera predicción que es si el paciente tiene SARS-CoV-2 o no.

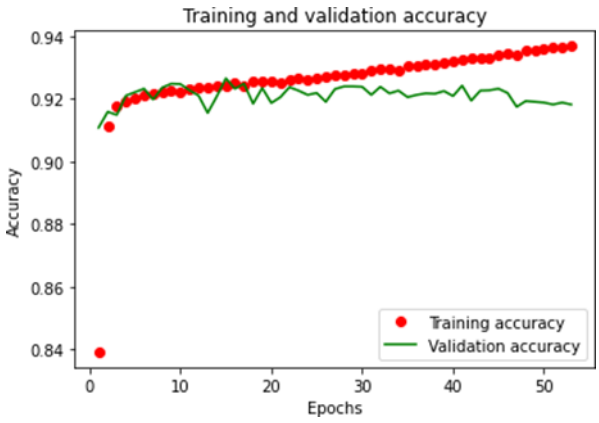

Figura 13. Comparación entre los datos de prueba y predicción
Fuente: Propia

4) Precisión del Modelo

Como resultado se muestra la precisión del modelo en base a la primera predicción que se realizó que es Si tiene SARS- CoV-2, utilizando el modelo de redes neuronales con algoritmo de clasificación que es 93.53341288782816%

```
print("La precisión es de {}%".format(((cm[0][0] + cm[1][1])/83800)*100))
La precisión es de 93.53341288782816%
```

Figura 14. Precisión del Score
Fuente: Propia

B. Días transcurridos hasta su Mortalidad.

1)Arquitectura de predicción

En la siguiente imagen se identifica la arquitectura de redes neuronales por regresión para la predicción: días transcurridos, utilizando una capa de entrada, tres capas ocultas, y una capa de salida:

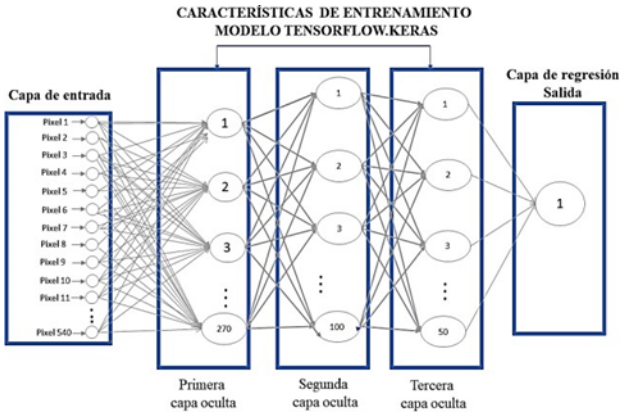

Figura 15. Arquitectura de para la Predicción

Fuente: Propia

2)Entrenamiento con el Dataset.

Una vez realizado las épocas que mejor se ajustaron al entrenamiento del modelo se obtuvo algunos datos de referencia para demostrar la efectividad del entrenamiento y deducir que es el adecuado para el modelo predictivo. Se encontró el valor mínimo de perdida y precisión en este entrenamiento que se visualizara a continuación:

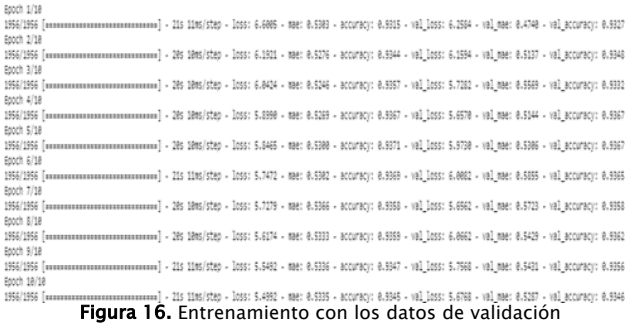

Figura 16. Entrenamiento con los datos de validación

Fuente: Propia

3)Predicción

En esta grafica se muestra los valores en y de número de días que han sido evaluados y en el eje de x total de registros que se encuentra en el conjunto de datos evaluado.

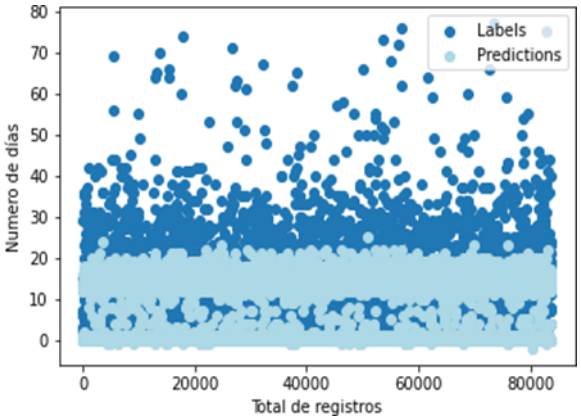

Figura 17. Predicción de Días transcurridos hasta su mortalidad

Fuente: Propia

4)Precisión del modelo

Como resultado se muestra la precisión del modelo en base a la primera predicción que se realizó que es Si tiene covid o no, utilizando el modelo de redes neuronales con algoritmo de clasificación que es 93.7%

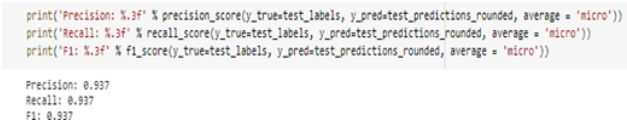

Figura 18. Precisión del Score

Fuente: Propia

C. Mortalidad

1)Arquitectura de predicción

En la siguiente imagen se identifica la arquitectura de redes neuronales por regresión para la predicción: Mortalidad, utilizando una capa de entrada, una capa oculta y la capa de salida.

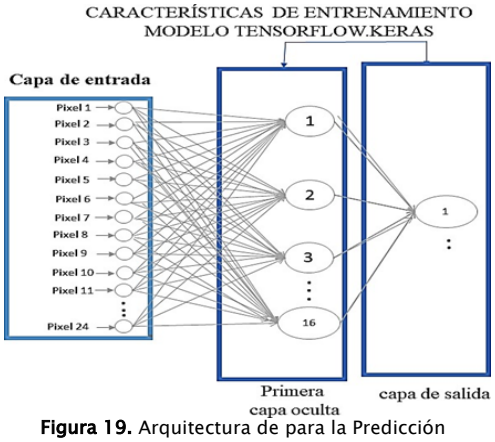

Figura 19. Arquitectura de para la Predicción

Fuente: Propia

2)Entrenamiento con el Dataset.

En este apartado podemos ver el entrenamiento del modelo, en base a las épocas, de la segunda predicción que se realizó que es Mortalidad, utilizando el modelo de redes neuronales con algoritmo de regresión.

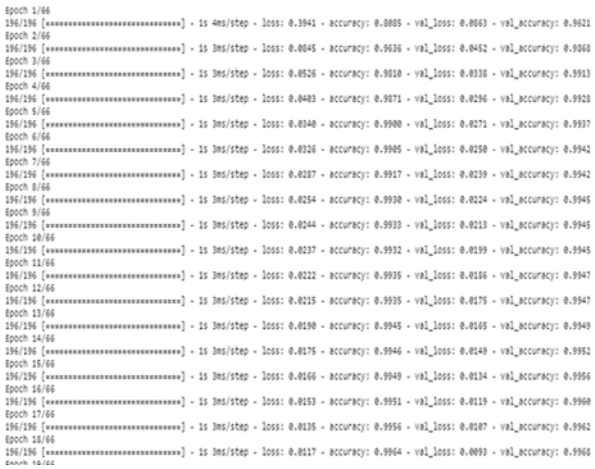

Figura 20. Predicción de Días transcurridos hasta su mortalidad
Fuente: Propia

3)Predicción

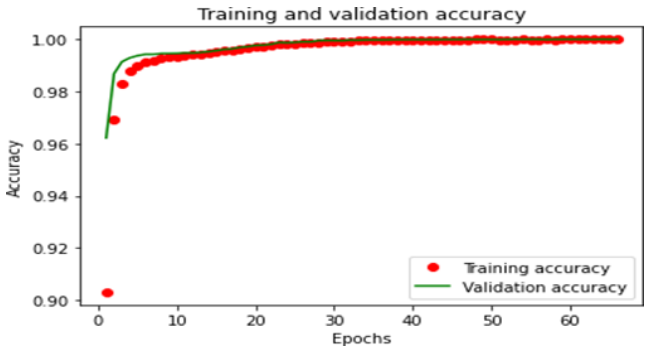

Figura 21. Predicción de Días transcurridos hasta su mortalidad
Fuente: Propia

En la Figura.21 se observa la comparación entre la curva de datos training y validación en base a precisión donde el valor de y es *accuracy* y el eje de x tiene el valor de épocas de entrenamiento. Como resultado obtenido para la primera predicción que es si el paciente tiene covid o no. Se denota que es una buena predicción.

4)Precisión del modelo

Como resultado se muestra la precisión del modelo en base a la primera predicción que se realizó que es Si tiene covid o no, utilizando el modelo de redes neuronales con algoritmo de clasificación que es 91.88066825775655%

```
[136] print("La precisión es de {}%".format(((cm[0][0] + cm[1][1])/83800)*100))

      La precisión es de 91.88066825775655%
```

Figura 22. Precisión del Score
Fuente: Propia

Random Forest

A. Si el paciente tiene SARS-CoV-2 o no.

1.Predicción

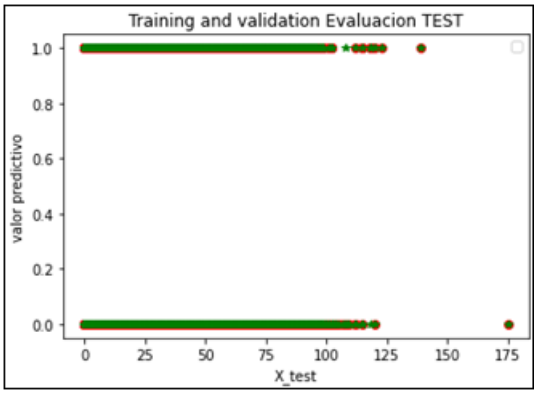

Figura 23. Precisión del Score
Fuente: Propia

2.Precisión de entrenamiento del conjunto de datos

```
RF_Model.fit(X_rf, y_rf)
print(RF_Model.score(X_rf, y_rf))

0.951629958287946
```

Figura 24. Precisión del Score
Fuente: Propia

3.Tabla de resultados obtenidos de la primera predicción, está compuesta de score, error, y valor de predicción.

Tabla 1: Propiedades de Composites

Accuracy	Error	Predicción
0.95136	0.219932	95.2%

Fuente: Propia

En la descripción de la tabla 1 se especifica los valores encontrados en la predicción- Si el paciente tiene SARS-Cov- 2 o no. Donde los valores de score en accuracy es de 0.95136, el error es 0.219932, y resultado de la predicción en porcentaje es de 95.2%.

B. Días Transcurridos hasta la Mortalidad

1.Predicción

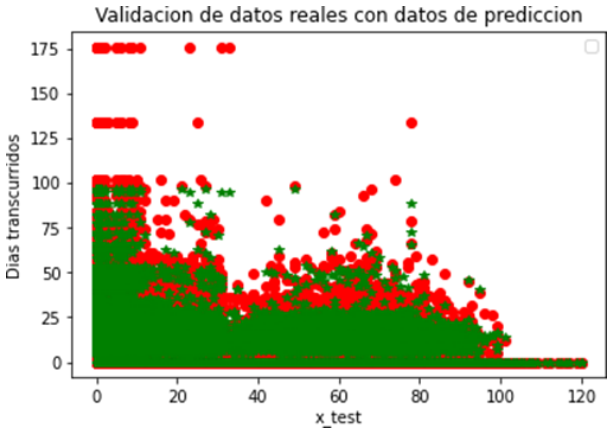

Figura 25. Precisión del Score
Fuente: Propia

2.Precisión del entrenamiento del conjunto de datos

```
RF_r.fit(X_train, y_train)
print('Presicion del Modelo')
print(RF_r.score(X_train, y_train))
```

```
Presicion del Modelo
0.8545640371433613
```

Figura. 26. Score del entrenamiento.
Fuente: Propia

3.Tabla de resultados obtenidos de la segunda predicción que se conforma de score, error, y resultado de la predicción.

Tabla 2: Descripción de Resultados

Accuracy	Error	Predicción
0.8352540724233488	0.13358468891157824	97.4%

Fuente: Propia

En la descripción de la tabla 2 se especifica los valores encontrados en la predicción- Días Transcurridos hasta la Mortalidad. Donde los valores de score en *accuracy* es de 0.835 , el error es0.134 , y resultado de la predicción en porcentaje es de 97.4%.

C)Días Transcurridos hasta la Mortalidad

1.Predicción

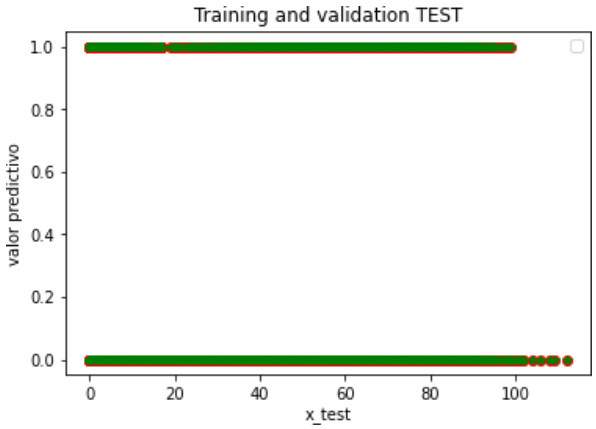

Figura 27. Predicción – Mortalidad
Fuente: Propia

2.Precisión del entrenamiento del conjunto de datos

```
[50] RF_Model.fit(X_train, y_train)
     print(RF_Model.score(X_train, y_train))
```

```
0.9999795429904057
```

Figura 28. Score de Entrenamiento
Fuente: Propia

3.Tabla de resultados obtenidos de Mortalidad, está compuesta de score, error, y valor de predicción.

En la descripción de la tabla 3 se especifica los valores encontrados en la predicción – Mortalidad. Donde los valores de score en accuracy es de 0.99998, de error 0.004523 y resultado de la predicción en porcentaje es de 100%.

Tabla 3. Descripción de resultados

Accuracy	Error	Predicción
0.99998	0.004523	100 %

Fuente: Propia

Análisis de las Predicciones

•En redes neuronales para realizar las predicciones de Si el paciente tiene SARS-CoV-2 o no y Mortalidad se aplicó método de clasificación con la estructura en base al modelo Keras, que está conformada por capas de entrada, capas ocultas y capa de salida, posteriormente se procedió hacer el entrenamiento del conjunto de datos, luego se definió las épocas y métricas, después de varias pruebas se encontró la época con mejor score en error, *accuracy* y validación, se graficó todos estos parámetros para describir la predicción junto a la precisión del modelo.

•En la predicción de Días transcurridos hasta la mortalidad se aplicó el método de regresión lineal con la estructura en base al modelo keras, al mismo se le definió neuronas en la capa de entrada, capas ocultas, y de salida. Posteriormente se entrenó el conjunto de datos, se definió las épocas y métricas de regresión, luego de varias pruebas se encontró la época con mejor score en base a mae, error, accuracy y validación, se graficó todos estos parámetros para definir la predicción junto precisión del modelo.

•El modelo de Random Forest es una estructura diferente a la de Redes Neuronales, donde las predicciones de Si el paciente tiene SARS-CoV-2 o no y Mortalidad se aplicó método de clasificación, (randomforestclassifier), en la estructuración de la predicción se definen cuantos n_estimators se usaran, procesadores, random_state, se procedió a entrenar el conjunto de datos como resultado dio un buen score así que no se procedió hacer iteraciones, luego se obtiene el margen de error, el accuracy posteriormente la predicción en porcentaje.

•En la predicción de Días transcurridos hasta la mortalidad se aplicó el método de regresión (randomforestregression), en la estructuración de la predicción se definen cuantos n_estimators se usarán, procesadores, random_state, se procedió a entrenar el conjunto de datos luego se obtiene el margen de error, el accuracy posteriormente la predicción en porcentaje todo esto es representado por medio de graficas.

Resultados

En este trabajo de investigación se logró analizar y predecir la propagación de un virus en este caso SARS-CoV-2, la afectación que causa en las personas por medio de tres predicciones que se realizaron con el uso de técnicas Machine Learning. Se evaluó al conjunto de datos, se aplicó el preprocesamiento, análisis exploratorio, posteriormente se eligió las variables que se necesitaban para cada predicción y después realizar el entrenamiento y aplicar los algoritmos de Redes Neuronales y Random Forest, en las diversas predicciones. Aplicando métodos de regresión y clasificación dependientes de cada predicción, en la siguiente tabla de comparación de resultados se muestra los scores resultantes de cada predicción en base a los algoritmos.

Tabla 4: Descripción de Resultados

		Redes Neuronales	Random Forest
Si el paciente tiene SARS-CoV-2 o No	Error	0.0319	0.21993
	Accuracy	0.9863	0.95163
	Porcentaje de Predicción	91.83%	95.20%
	Tipo de algoritmo	Clasificación	
Días transcurridos hasta la Mortalidad	Error	5.4992	0.133584689
	Accuracy	0.9345	0.835254072
	Porcentaje de Predicción	93.70%	97.40%
	Tipo de Algoritmo	Regresión	
Mortalidad	Error	0.0012	0.00452
	Accuracy	0.9998	0.99998
	Porcentaje de Predicción	93.53%	100%
	Tipo de Algoritmo	Clasificación	

Fuente: Propia

En la tabla 4 se observa la descripción de los resultados obtenidos por cada predicción en base a los dos algoritmos que se utilizaron para el modelo, haciendo comparación del *accuracy* error, el porcentaje de la predicción y el tipo de algoritmo que se empleó.

Primera predicción qué es si el paciente tiene SARS-CoV-2 o no con método de clasificación se puede observar que en RN el error es de 0.0319 y RF 0.219932 en base a la precisión el valor de RN es de 09863 y RF es de 0.95163, en el porcentaje de la predicción de RN es de 91.82 56 801% y de RF es del 95.2%.

Predicción de días transcurrido hasta la mortalidad con método de regresión en la comparación del error en base a RN es decir 5. 4992 y de RF es de 0.13358468891157824 en la comparación de accuracy en RN es de 0.9345 y RF 0.83 5254072423 3488 y en el porcentaje de predicción en base a RN 93.70% y RF 97.4%

Tercera predicción que es Mortalidad con técnica de clasificación se puede que en la comparación del error en RN tiene el valor de 0.0012 y de RF es de 0.004523 en accuracy el valor de RN es de 0.9998y de RF 0.99998 en cuanto a porcentajes de predicción RN es de 93.53% y de RF es 100%

Conclusión

El uso de herramientas de machine learning permitieron desarrollar el modelo de predicción en base a la propagación de un virus en este caso es Covid-19 y la afectación que causa en las personas por medio de tres predicciones.

Utilizando los algoritmos de *Random Forest* y Redes Neuronales Artificiales, se aplicaron las técnicas de clasificación y regresión; donde el algoritmo de *Random Forest* tuvo el mejor score en las predicciones.

El margen de diferencia alcanzado entre la predicción usando Regresión Lineal vs Random Forest es de aproximadamente el 2% en las 2 primeras evaluaciones.

Referencias Bibliográficas

A Survey on Decision Tree Algorithms of Classification in Data Mining. (2016). *International Journal of Science and Research (IJSR)*, *5*(4). https://doi.org/10.21275/v5i4.nov162954

Agasi, O., Anderson, J., Cole, A., Berthold, M., Cox, M., & Dimov, D. (2018). What is an Artificial Neural Network (ANN)? - Definition from Techopedia. *Techopedia*.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, Vol. 11. https://doi.org/10.2478/v10136-012-0031-x

B S, L. (2021). Data Analysis and Data Classification in Machine Learning using Linear Regression and Principal Component Analysis. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2). https://doi.org/10.17762/turcomat.v12i2.1092

Breiman, L. (1999). Randon Forests. *Machinelearning202.Pbworks.Com*.

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms. *Understand Machine Learning Algorithms*.

Brownlee J. (2019). Supervised and Unsupervised Machine Learning Algorithms. *Machine Learning Mastery Pty. Ltd.*

Chowdhury, A. R., Chatterjee, T., & Banerjee, S. (2019). A Random Forest classifier-based approach in the detection of abnormalities in the retina. *Medical and Biological Engineering and Computing*, *57*(1). https://doi.org/10.1007/s11517-018-1878-0

De La Hoz, E. J., De La Hoz, E. J., & Fontalvo, T. J. (2019). Methodology of Machine Learning for the classification and Prediction of users in Virtual Education Environments. *Informacion Tecnologica*, *30*(1), 247–254. https://doi.org/10.4067/S0718-07642019000100247

Fuat. (2018). Google Colab Free GPU Tutorial.

G. Jignesh Chowdary, Suganya. G, P. M. (2020). EFFECTIVE PREDICTION OF CARDIOVASCULAR DISEASE USING CLUSTER OF MACHINE LEARNING ALGORITHMS. *Journal of Critical Reviews*, *7*(18).

Gallaway, T. O., & Starkey, J. (2013). Google Drive. *The Charleston Advisor*, *14*(3). https://doi.org/10.5260/chara.14.3.16

Giraldo Mejía, J. C., & Vargas Agudelo, F. A. (2019). Aplicación de la técnica regresión logística de la minería de datos en el proceso de descubrimiento de conocimiento (KDD) en bases de datos operativas o transaccionales. *Perspectiv@*, *14*(13).

Haglin, J. M., Jimenez, G., & Eltorai, A. E. M. (2019). Artificial neural networks in medicine. *Health and Technology*, Vol. 9. https://doi.org/10.1007/s12553-018-0244-4

Jhu. (n.d.). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Retrieved July 7, 2021, from https://coronavirus.jhu.edu/map.html

Juan, A. M. (2003). Regresión lineal múltiple. *Técnicas Estadísticas Aplicadas Al Análisis de Datos*.

Källén, M., & Wrigstad, T. (2021). Jupyter Notebooks on GitHub: Characteristics and Code Clones. *The Art, Science, and Engineering of Programming*, *5*(3).

https://doi.org/10.22152/programming-journal.org/2021/5/15

Lee, J. H., Kim, D. H., Jeong, S. N., & Choi, S. H. (2018). Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *Journal of Periodontal and Implant Science*, *48*(2). https://doi.org/10.5051/jpis.2018.48.2.114

Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in Data mining : An Overview. *International Journal of P2P Network Trends and Technology (IJPTT)*, *4*(8).

Pardo. A, & Ruiz, M. A. (2005). Análisis de regresión lineal : El procedimiento Regresión lineal. *Guía Para El Análisis de Datos*.

Pfefferbaum, B., & North, C. S. (2020). Mental Health and the Covid-19 Pandemic. *Https://Doi.Org/10.1056/NEJMp2008017*, *383*(6), 510–512. https://doi.org/10.1056/NEJMP2008017

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning & Deep Learning with Python, Scikit-Learn and TensorFlow 2, Third Edition. In *Packt Publishing Ltd*.

Shakouri, S., Bakhshali, M. A., Layegh, P., Kiani, B., Masoumi, F., Ataei Nakhaei, S., & Mostafavi, S. M. (2021). COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis. *BMC Research Notes*, *14*(1). https://doi.org/10.1186/s13104-021-05592-x

Tworowski, D., Gorohovski, A., Mukherjee, S., Carmi, G., Levy, E., Detroja, R., … Frenkel-Morgenstern, M. (2021). COVID19 Drug Repository: Text-mining the literature in search of putative COVID19 therapeutics. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa969

WHO. (n.d.). WHO | World Health Organization. Retrieved July 7, 2021, from https://www.who.int/

Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F., & Liu, J. (2020). Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing. *Radiology*, *296*(2), E41–E45. https://doi.org/10.1148/RADIOL.2020200343

Yiu, T. (2019). Understanding Random Forest - Towards Data Science. *Understanding Random Forest How the Algorithm Works and Why It Is So Effective*.