Application of Machine Learning for the Prediction of Covid19 through Classification Techniques and Supervised Learning

Miguel Molina-Calderón, MSIG[1], Roberto Crespo-Mendoza, MSIG[1],Ericka Reyes-Cun, Msc[2] ,
Freddy Burgos-Robalino, Msc[2], Darwin Patiño-Pérez, Phd[1],Liliana Sarmiento-Barreiro, MSIG[2], Miguel Botto-Tobar, MSc[1]

[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Guayaquil,Ecuador, miguel.molinac@ug.edu.ec, roberto.crespom@ug.edu.ec,darwin.patinop@ug.edu.ec,miguel.bottot@ug.edu.ec

[2]Universidad de Guayaquil, Facultad de Ciencias Médicas, Guayaquil,Ecuador, ericka,reyesc@ug.edu.ec,freddy.burgosr@ug.edu.ec,liliana.sarmientob@ug.edu.ec

Abstract. – In January 2020, in the city of Wuhan in China, a highly dangerous disease for humans appeared, cataloged as COVID-19 and caused by a virus called SARS-CoV-2; This disease spread from Asia to Europe and then spread throughout the American continent, causing a pandemic that to this day continues to cause irreparable damage. Currently the virus has mutated, and its variants continue to congest health systems in all parts of the world. The spread of the virus has generated information that is available on public research-oriented portals, and that is available for scientists and researchers to have the necessary information so that they can develop strategies to face and stop the disease. Using machine learning techniques, a prediction model has been created, which by applying supervised learning performs an analysis of historical data and has learned to identify patterns, managing to identify the disease.

Keywords: Covid-19, Supervised Learning, Algorithms, Machine Learning.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.425
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

Aplicación de Machine Learning para la Predicción de Covid19 Mediante Técnicas de Clasificación y Aprendizaje Supervisado

Miguel Molina-Calderón, MSIG[1], Roberto Crespo-Mendoza, MSIG[1],Ericka Reyes-Cun, Msc[2] ,
Freddy Burgos-Robalino, Msc[2], Darwin Patiño-Pérez, Phd[1],Liliana Sarmiento-Barreiro, MSIG[2], Miguel Botto-Tobar, MSc[1]
[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Guayaquil,Ecuador, miguel.molinac@ug.edu.ec,
roberto.crespom@ug.edu.ec,darwin.patinop@ug.edu.ec,miguel.bottot@ug.edu.ec
[2]Universidad de Guayaquil, Facultad de Ciencias Médicas, Guayaquil,Ecuador,
ericka,reyesc@ug.edu.ec,freddy.burgosr@ug.edu.ec,liliana.sarmientob@ug.edu.ec

Resumen. – En enero de 2020, en la ciudad de Wuhan en China, apareció una enfermedad altamente peligrosa para los humanos, catalogada como COVID-19 y causada por un virus llamado SARS-CoV-2; está enfermedad se propagó desde Asia hacia Europa y luego se extendió por todo el continente americano, causando una pandemia que hasta la presente sigue causando daños irreparables. Actualmente el virus ha mutado y sus variantes siguen congestionando a los sistemas de salud en todas partes del mundo. La propagación del virus ha generado información que se encuentra disponible en los portales públicos orientados a la investigación, y que están disponibles para que científicos e investigadores cuenten con la información necesaria para que puedan desarrollar estrategias para enfrentar y frenar la enfermedad. Mediante el uso de técnicas de machine learning, se ha creado un modelo de predicción, que aplicando aprendizaje supervisado realiza un análisis a los datos históricos ha aprendido a identificar patrones logrando identificar la enfermedad.

Palabras Claves: Covid-19, Aprendizaje Supervisado, Algoritmos, Machine Learning.

Abstract. – In January 2020, in the city of Wuhan in China, a highly dangerous disease for humans appeared, cataloged as COVID-19 and caused by a virus called SARS-CoV-2; This disease spread from Asia to Europe and then spread throughout the American continent, causing a pandemic that to this day continues to cause irreparable damage. Currently the virus has mutated, and its variants continue to congest health systems in all parts of the world. The spread of the virus has generated information that is available on public research-oriented portals, and that is available for scientists and researchers to have the necessary information so that they can develop strategies to face and stop the disease. Using machine learning techniques, a prediction model has been created, which by applying supervised learning performs an analysis of historical data and has learned to identify patterns, managing to identify the disease.

Keywords: Covid-19, Supervised Learning, Algorithms, Machine Learning.

I. Introduccion

Desde la aparición del SARS-CoV-2 o coronavirus ver Fig. 1, responsable de la enfermedad denominada covid-19 que se originó en Wuhan-China a finales del 2019, los especialistas de la salud a nivel mundial han sumado esfuerzos por encontrar los mecanismos que permitan frenar la propagación del virus así como la rápida identificación en las personas[1]. El coronavirus es un tipo de virus común que causa una infección en la nariz, los senos paranasales o la parte superior de la garganta.

La mayoría de los virus corona no son peligrosos ellos se propagan de la misma manera que otros coronavirus, principalmente a través del contacto de persona a persona por lo que las infecciones varían de leves a graves según[2]; actualmente existe mucha data relacionada con la enfermedad que se encuentra registrada en portales públicos, la misma que está disponible para que científicos e investigadores puedan realizar investigaciones[3]. La pronta identificación del Covid-19, induce a que surjan investigaciones orientadas a buscar soluciones que ayuden a los profesionales de la salud a identificar la enfermedad de forma rápida, para poder implementar tratamientos oportunos a los afectados y así poder controlar la propagación[4].

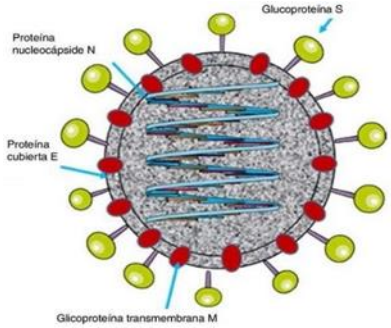


Fig. 1 SARS-CoV2

A. *Variantes del COVID-19*

Los virus cambian constantemente a través de la mutación, cuando un virus tiene una o más mutaciones nuevas, se le llama variante del virus original las mismas que causan mucha preocupación [5]. Identificar a las variantes se ha vuelto un tema de estudio, puesto que todo inicia por los síntomas que generalmente aparecen de 2 a 14 días después de la exposición al virus; de forma general los síntomas son: fiebre, escalofríos, tos, falta de aire o dificultad para respirar, fatiga, dolor corporal o muscular, dolor de cabeza, pérdida del sentido del gusto o del olfato, dolor de garganta, congestión o secreción nasal, náuseas o vómitos y diarrea[6].

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

-Delta (B.1.617.2). Variante catalogada como el doble de contagiosa que las variantes anteriores que podría causar una enfermedad más grave[7]. El mayor riesgo de transmisión es entre las personas no vacunadas. Las personas que están completamente vacunadas pueden contraer infecciones revolucionarias de la vacuna y transmitir el virus a otras personas. Sin embargo, parece que las personas vacunadas transmiten el COVID-19 durante un período más corto que las personas no vacunadas. Si bien las investigaciones sugieren que las vacunas contra el COVID-19 son un poco menos efectivas contra la variante delta, las vacunas contra el COVID-19 de Pfizer-BioNTech, Moderna y Janssen/Johnson & Johnson aún parecen brindar protección contra el virus cuyos síntomas característicos son: dolor de cabeza, dolor de garganta, secreción nasal, perdida de gusto y olfato.

-Omicron (B.1.1.529). Esta variante se propaga más fácilmente que las otras variantes, incluida la delta. Pero aún no está claro si omicron causa una enfermedad más grave. Sin embargo, se espera que las vacunas COVID-19 sean efectivas para la prevención de esta variante que reduce la eficacia de algunos tratamientos con anticuerpos monoclonales[8]. Los síntomas son más leves y similares a los de un resfriado, además causa con menos frecuencia la pérdida del gusto y del olfato. Las personas por lo general sienten dolor de garganta, congestión nasal, dolor de cabeza y dolores musculares.

Las variantes alfa, gamma y beta continúan siendo monitoreadas, pero se están propagando a niveles mucho más bajos en los EE. UU. El variante mu también está siendo monitoreada[9].

B. *Tratamientos*

Actualmente no existe un medicamento antiviral aprobado que cure el COVID-19, por lo que no hay una cura eficaz; los antibióticos no tienen éxito para las infecciones virales como la COVID-19[10]. Se continúa trabajando en la investigación de diferentes fármacos que hasta el momento ha dado como resultado una píldora fabricada por Merck y Ridgeback Biotherapeutics, llamada molnupiravir la misma que reduce en un 30% los riesgos de hospitalización y muerte por COVID-19 si se toma durante los cinco días siguientes a la aparición de los síntomas [11]. Existen otros tratamientos que están destinados a aliviar los síntomas y pueden incluir: analgésicos (ibuprofeno o paracetamol), jarabe para la tos, medicamentos para descansar, así como la ingesta de líquidos[12].

C. *Detección de Covid-19 mediante IA*

El aprendizaje automático o *Machine Learning* es una subcategoría de la inteligencia artificial que automatiza eficazmente el proceso de creación de modelos analíticos[13] y permite que las máquinas se adapten a nuevos escenarios de forma independiente para resolver cualquier tipo de problema que demande de un análisis algorítmico[14]. El funcionamiento varía según la tarea y el algoritmo utilizado para lograrlo, sin embargo, en esencia, un modelo de aprendizaje automático es una computadora que analiza datos e identifica patrones, y luego usa esos conocimientos para completar mejor la tarea asignada[15].

Cuando las tareas dependen de un conjunto de datos o reglas se pueden automatizar mediante el aprendizaje automático y dependiendo de la situación, los algoritmos de aprendizaje automático funcionan con más o menos intervención/refuerzo humano[16]; cuando la computadora recibe un conjunto de datos etiquetados que le permite aprender cómo realizar una tarea humana se dice que el algoritmo aplica aprendizaje supervisado. Aquí la intervención del ser humano es de mucha importancia a la hora de etiquetar adecuadamente e conjunto de datos y los algoritmos intentan replicar el aprendizaje humano[17].

II. MATERIALES Y METODOS

A. *Materiales*

Para poder resolver la problemática, mediante el análisis de datos se contará con un *dataset* de 1400 registros con información de pacientes afectados con Covid-19, donde se registran 900 pacientes graves y 500 pacientes no grave según la Fig. 2 mediante aprendizaje supervisado, se contará con computador con características para ejecutar programas de inteligencia artificial, de forma local se trabajará sobre Anaconda. Además, se necesita una conexión a internet con buen ancho de banda, puesto que la herramienta de programación para el desarrollo de los modelos es Python[18] ejecutándose sobre una máquina virtual en Colab[19]

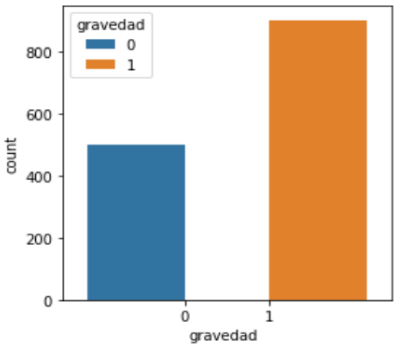


Fig. 2 Estado de Gravedad

B. *Métodos de Predicción*

La metodología de predicción mediante clasificación , estará basada en arboles de decisión o *Decision Tree* [17] como en la Fig. 3 y una mejora de ellos aplicando el impulso del gradiente o *Gradient Boosting* donde se evaluará el rendimiento del modelo, puesto que se ha determinado que con modelos basados en arboles como lo es el algoritmo *Random Forests* que según [20] se puede obtener un buen rendimiento.

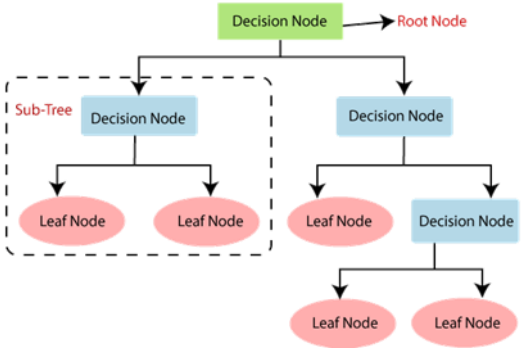


Fig. 3 Árbol de Decisión

El árbol de decisión comúnmente es usado en la investigación de operaciones, específicamente en el análisis de decisiones, para ayudar a identificar una estrategia con mayor probabilidad de alcanzar una meta, pero también es una herramienta popular en el aprendizaje automático según [21]; puesto que cada nodo del árbol actúa como un caso de prueba para algún atributo y cada arista que desciende de ese nodo corresponde a una de las posibles respuestas al caso de prueba[22]. Con *Decision Trees*[23] podemos resolver un problema mediante regresión o clasificación, sin embargo puesto que tenemos un *dataset* cuya variable objetivo maneja 2 tipos de afectación/categoría se debe usar árboles para clasificación, dado que estos se utilizan para variables objetivo-categóricas según [24]. Es de suma importancia establecer alguna de las medidas de frecuencia usada en la configuración de árbol:

-Gini: La impureza de Gini es una medida de la frecuencia con la que un elemento elegido al azar del conjunto se etiquetaría incorrectamente si se etiquetara al azar de acuerdo con la distribución de etiquetas en el subconjunto.

$$I_G(n) = 1 - \sum_{i=1}^{J} (\rho_i)^2 \quad (1)$$

-Entropía: Para la realizar una división de nodos de forma aleatoria.

$$\text{Entropy} = \sum_{i=1}^{C} -\rho_i * \log_2(\rho_i) \quad (2)$$

-Varianza: Funcionan bien para el escenario de clasificación, combina en el caso de la regresión la medida dividida más común utilizada es solo la varianza ponderada de los nodos[25].

$$\text{Variance} = \frac{\sum(X-\bar{X})^2}{n} \quad (3)$$

-Ganancia de información: la ganancia de información o IG es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con su clasificación objetivo.

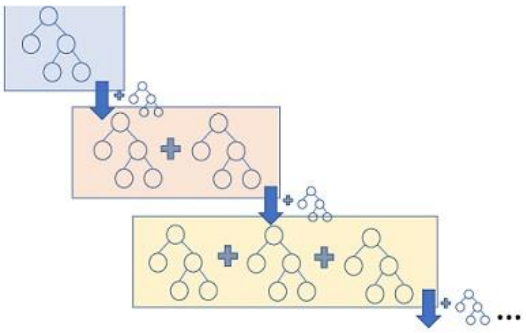


Fig. 4 *Gradient Boosting*

El *Boosting* y su mejora el *Gradient Boosting* como según la Fig. 4 , es un enfoque de *Machine Learning* basado en la idea de crear una regla de predicción altamente precisa mediante la combinación de muchas reglas relativamente débiles e imprecisas[29]. Es una técnica secuencial que funciona según el principio de un conjunto[31]. Combina un conjunto de aprendizaje débil y ofrece una precisión de predicción mejorada. En cualquier instante t, los resultados del modelo se ponderan en función de los resultados del instante anterior t-1.

Los resultados pronosticados correctamente reciben una ponderación más baja y los que no se clasifican correctamente reciben una ponderación más alta. Cabe señalar que un aprendizaje débil es uno que es ligeramente mejor que adivinar al azar[32]. Para un proceso de clasificación el modelo tiene características que están acorde con las marcades en el dataset, pero para los problemas que se resuelven mediante regresión usando el *gradient boosting* el modelo matemático a seguirse estría acorde con el que se visualiza en la Fig 5.

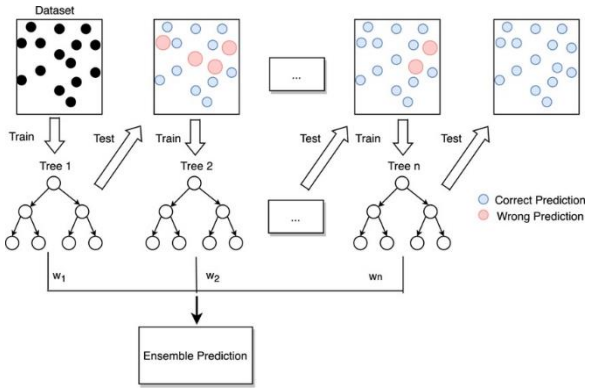


Fig. 5 Proceso Funcional del *Gradient Boosting*

C. Métricas de Evaluación

Las métricas se utilizan para monitorear y medir el rendimiento de un modelo[21] (durante el entrenamiento y la prueba), y no es necesario que sean diferenciables según [33] por lo que es muy importante elegir la métrica adecuada para evaluar los modelos de clasificación que se usaran para resolver el problema.

-Matriz de Confusión.

Uno de los conceptos clave en el rendimiento de la clasificación es la matriz de confusión (también conocida como matriz de error), que, aunque no es una métrica expone una visualización tabular de las predicciones del modelo frente a las etiquetas de verdad en el terreno[34]. Cada fila de la matriz de confusión representa las instancias de una clase real o actual y cada columna representa las instancias de una clase predicha.

TABLA I
MATRIZ DE CONFUSION

		Predicción	
		NO	**SI**
Realidad Actual	**NO**	TN	FP
	SI	FN	TP

Cuando el modelo de ML no acierta

FP:(False Positive) Falso Positivo, es cuando el modelo ha predicho que SI y en realidad es un NO.

FN:(False Negative) Negativos Falsos, es cuando el modelo ha predicho que NO pero en realidad es SI.

Cuando el modelo de ML si acierta

TP(True Positive), Verdaderos Positivos, cuando el modelo ha predicho que SI y en realidad SI.

TN:(True Negative), Verdaderos Negativos, aquí el modelo ha predicho que NO y en realidad es un NO.

2)La Exactitud (Accuracy)
La exactitud es una de las métricas más simple y usada de clasificación, que en muchas situaciones puede inducir a que un modelo malo parezca que es mucho mejor de lo que es[20].

$$\text{Accuracy} = \frac{(TN+TP)}{(FP+TP)+(TN+FN)} \quad (5)$$

3)Precisión(Precision)
Hay muchos casos en los que la precisión de la clasificación no es un buen indicador del rendimiento del modelo. Uno de estos escenarios es cuando la distribución de clases está desequilibrada (una clase es más frecuente que otras). En este caso, incluso si predice todas las muestras como la clase más frecuente, obtendrá una alta tasa de precisión, lo que no tiene ningún sentido (porque su modelo no está aprendiendo nada y solo está prediciendo todo como la clase superior).

$$\text{Presicion} = \frac{TP}{(TP+FP)} \quad (6)$$

4- Exhaustividad (Recall)
La exhaustividad es otra métrica importante, que se define como la fracción de muestras de una clase que el modelo predice correctamente[34]. Esta métrica informa sobre la cantidad que el modelo de ML es capaz de identificar. De manera formal está definida en (7):

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (7)$$

5- Valor F(F1_Score)
F1_score según (4) es la media armónica de precisión y exhaustividad(recall) , se utiliza para combinar esas dos medidas en un sólo valor y facilita la comparación del rendimiento combinado de la precisión y la exhaustividad entre varias soluciones[35].

$$F1\_score = 2 * \frac{presicion*recall}{(presicion+recall)} \quad (8)$$

La versión generalizada de F1-score se define en (9), como un caso especial de F_β cuando $\beta = 1$.

$$F_\beta = (1+\beta^2) * \frac{presicion*recall}{\beta^2*(presicion+recall)} \quad (9)$$

6- Sensibilidad y Especificidad(Sensitivity , Specificity)
La sensibilidad y la especificidad son otras dos métricas populares que se utilizan principalmente en campos relacionados con la medicina y la biología, y se definen como:

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{(TP+FN)} \quad (10)$$

$$\text{Specificity} = True\_Negative\_Rate = \frac{TN}{(TN+FP)} \quad (11)$$

7- Curva ROC
Se denomina Curva característica Operativa del Receptor[37] a un gráfico como en la Fig. 6, que muestra el rendimiento de un clasificador binario en función de su umbral de corte[38]. La curva muestra la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR) [39] para varios valores de umbral. Muchos de los modelos de clasificación son probabilísticos, ellos predicen la probabilidad de que una muestra sea de un infectado.

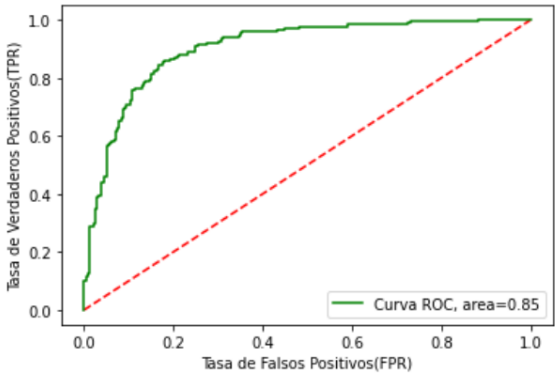


Fig.6 Ejemplo de Curva ROC

III.METODOLOGIA

La metodología utilizada está basada en técnicas de *machine learning* donde los pasos son:

1)Tratamiento de los Datos
2)Creación del Modelo
3)Fase de entrenamiento
4)Fase de Prueba
5)Evaluación del modelo con sus métricas
6)Aplicación del modelo

Tratamiento de los Datos:
La base de datos de registros actualizados constituida por los casos confirmados de COVID-19 a nivel mundial, la selección de características, y el entrenamiento de los datos. La base de datos contiene los registros que serán utilizados para el entrenamiento de los algoritmos de aprendizaje supervisado por clasificación usando árboles de decisión y Gradiente Boosting.

El levantamiento de información en base a expedientes y registros médicos, se analizaron diferentes bases de datos públicas (locales y extranjeras) que se adaptaron a las métricas preestablecidas fundamentadas por la sintomatología de un paciente infectado por el virus.

La selección de características se filtraron registros de un patrón sintomatológico, los cuales sirvieron para crear una nueva base de datos, esta estaría conformada por las siguientes columnas; edad, dificultad para respirar, saturación, dolor de cabeza, dolor abdominal, dolor muscular dolor de garganta, tos, fiebre, diarrea, fatiga, pérdida de olfato, pérdida de apetito y gravedad. De los campos mencionados previamente, la gravedad de los pacientes fue

fundamentada a través de las características mencionadas en la siguiente tabla:

```
df = pd.read_csv('sintomas.csv', sep=';', encoding='UTF-8')

print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1400 entries, 0 to 1399
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   edad            1400 non-null   int64
 1   dif_resp        1400 non-null   int64
 2   saturacion      1400 non-null   int64
 3   dolor_cabeza    1400 non-null   int64
 4   dolor_abdom     1400 non-null   int64
 5   dolor_muscular  1400 non-null   int64
 6   dolor_garganta  1400 non-null   int64
 7   tos             1400 non-null   int64
 8   temperatura     1400 non-null   int64
 9   diarrea         1400 non-null   int64
 10  fatiga          1400 non-null   int64
 11  perdida_olf     1400 non-null   int64
 12  perdida_apt     1400 non-null   int64
 13  gravedad        1400 non-null   int64
```

Fig.7 Dataset Sintomas

El dataset está conformado por 1400 registros de los cuales 500 están con aislamiento en casa y 900 hospitalizados con 12 características relativas a la enfermedad, la variable de salida es binaria donde 0 corresponde a que el paciente no está infectado y 1 si está infectado según Fig 12:

1.Dificultad para respirar (representada por 0 = no; 1 = sí)
2.Saturación: Cantidad de oxígeno disponible en la sangre.
3.Dolor de cabeza (representada por 0 = no; 1 = sí)
4.Dolor abdominal (representada por 0 = no; 1 = sí)
5.Dolor muscular (representada por 0 = no; 1 = sí)
6.Dolor de garganta (representada por 0 = no; 1 = sí)
7.Tos (representada por 0 = no; 1 = sí)
8.Temperatura: Valor en C°.
9.Diarrea (representada por 0 = no; 1 = sí)
10Fatiga (representada por 0 = no; 1 = sí)
11.Pérdida de olfato (representada por 0 = no; 1 = sí)
12.Pérdida de apetito (representada por 0 = no; 1 = sí)

Creación de los Modelos

Puesto que se utiliza un dataset que contiene 1400 registros de pacientes atendidos en el hospital público, para los modelos de ML se tomó el 80% de los registros, determinados en la base de datos para su entrenamiento y el valor de porcentaje faltante fue orientado a pruebas. La técnica de clasificación empleó la extracción de características, para el entrenamiento y para las respectivas pruebas.

Para determinar la gravedad de las personas contagiadas de COVID-19 se aplicaron los algoritmos de aprendizaje supervisado como: árboles de decisión *Decisión Tree* y una de las variantes denominado *Gradient Boosting* tomados de la biblioteca de Sklearn de Python según Fig. 8

```
# Decision Tree - Modelo de Árbol de Decisión para Clasificación
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

algoritmo = DecisionTreeClassifier(criterion="entropy", max_depth=6)

algoritmo=algoritmo.fit(X_train, y_train)
y_prds = algoritmo.predict(X_test)
accuracyDT=  accuracy_score(y_test, y_prds)
print("Decision Tree con {:0.3f} % de precisión".format(accuracyDT))
accMOD.append(accuracyDT)

#GradientBoosting
from sklearn.ensemble import GradientBoostingClassifier
gbk = GradientBoostingClassifier(random_state=100,
                                 n_estimators=150,min_samples_split=100, max_depth=6)
gbk.fit(X_train, y_train)
y_prds = gbk.predict(X_test)

accuracyGB =  accuracy_score(y_test, y_prds)
print("Gradient Boosting con {:0.3f} % de precisión".format(accuracyGB))
accMOD.append(accuracyGB)
```

En la técnica de árboles de decisión se definieron los parámetros “criterion=gini”; para especificar la función de impureza, la cual, valida el desempeño en la división de los datos, y “max_depth=3” para especificar la profundidad máxima del árbol.

Fig.8 Modelos de Clasificación

La creación de la matriz de confusión fue muy importante para poder probar las métricas de clasificación, la matriz resultante de las pruebas desarrolladas por el modelo según la Fig.9.

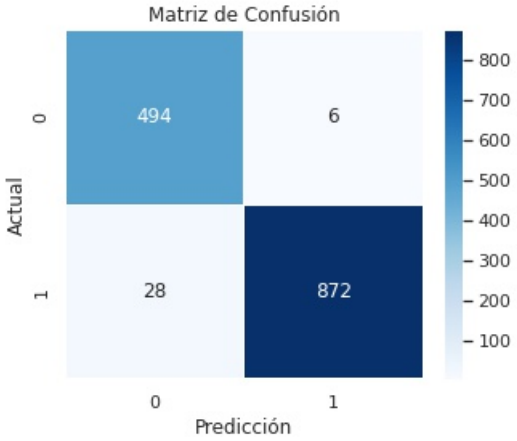


Fig.9 Matriz de Confusión

La matriz de confusión también conocida como matriz de error o tabla de contingencia y cuya descripción se refleja en la Tabla I también sirve para determinar las métricas expuestas desde (5) hasta (11).

-De 900 registros de personas graves, el modelo ha predicho correctamente 872 como graves y ha clasificado incorrectamente 28 de ellos como no graves. Si nos referimos a la clase "graves" como positiva y a la clase que no están graves como clase negativa, entonces 1400 muestras pronosticadas como infectado se consideran positivas verdaderas 872, y las 28 muestras predichas como no infectado son negativas falsas.
-De 500 registros de no infectados, el modelo clasificó 494 de ellos correctamente y clasificó incorrectamente 6 de ellos. Las 494 muestras clasificadas correctamente se denominan verdadero negativo y las 6 se denominan falso positivo.

En Fig.10 se refleja el nivel de exactitud que se ha alcanzado con el Set Datos usando *Decision Tree* y se refleja de manera general se ha obtenido una exactitud del 97%.

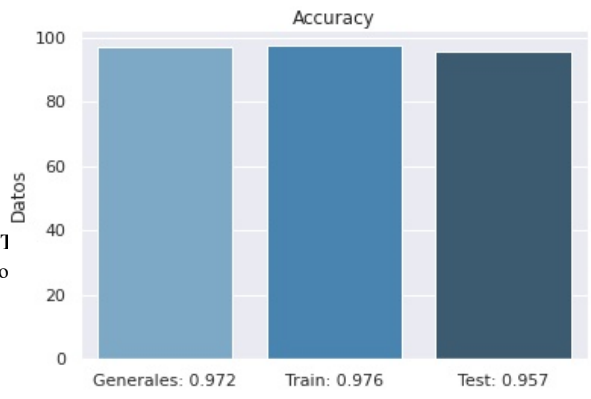


Fig.10 Exactitud alcanzada con *Decisión Tree*

El estudio de curva ROC (receiver operating characteristic curve) Fig. 11, representa una técnica estadística para definir la exactitud diagnóstica de pruebas, siendo empleadas con los siguientes objetivos específicos: precisar la posición de corte de una progresión continua en la que se obtiene la sensibilidad y especificidad mayor, y valorar el desempeño diferenciado de la prueba diagnóstico, de pacientes con covid-19 que necesitan ser hospitalizados versus pacientes que pueden llevar el tratamiento desde su casa de manera aislada y controlada.

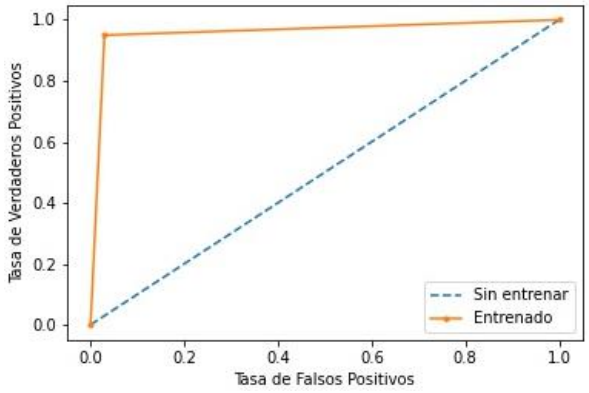


Fig.11 Curva ROC

A través de la librería sklearn se obtuvo la curva de precisión-sensibilidad Fig. 12 y sus valores asociados. Luego fue empleada la función precision_recall_curve(), que elige como indicadores las salidas reales y las posibilidades alcanzadas para la clase positiva. Esta función retorna vectores de precisión, sensibilidad y las entradas para los valores mencionados.

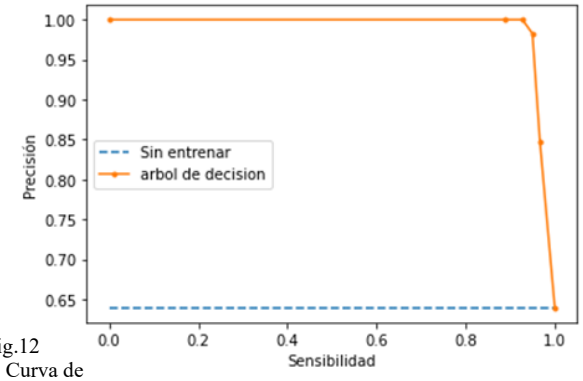
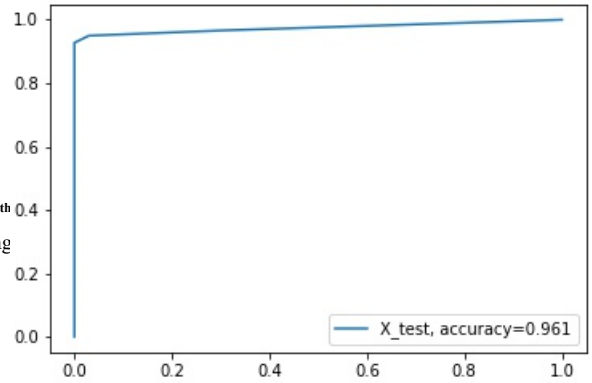


Fig.12 Curva de Sensibilidad

La función auc() selecciona como entradas la sensibilidad y la precisión regresando el valor del área bajo la curva, el cual puede ser tomado como síntesis del desempeño del modelo. Para conseguir el valor de AUC, se emplea la función roc_auc_score() (dato de entrada conocido previamente). En esta ocasión retorna la estimación de AUC, contenida entre 0.5 (clasificador al azar) y 1.0 (clasificador óptimo).



VI. Resultado, Discusion y Conclusion

Se determino que los dos modelos Fig. 13 se comportan adecuadamente, pero se refleja además que la mejor exactitud obtenida es con el *Gradient Boosting* 96% que por lo visto se puede decir que es un buen resultado para conseguir la clasificación y predicción con todos los dataset con lo que se puede determinar si un paciente necesita ser hospitalizado o mandarlo con aislamiento domiciliario.

Fig.13 *Gradient Boosting* - Curva ROC

Para determinar los síntomas característicos del covid-19 se efectuó un análisis de diferentes estudios, investigaciones, y revistas científicas encontradas en "Google Scholar" y en Mendeley así como historiales clínicos de acceso público reflejados en una serie de dataset disponibles en Kaggle y en el Johns Jopkins University.

Para desarrollar los modelos se seleccionó la técnica de aprendizaje supervisado, árboles de decisión y *gradient boosting* tal como se refleja en Fig 13, debido a que no existía un algoritmo previo enfocado al reconocimiento e identificación del nivel de gravedad de pacientes infectados con covid-19 en donde se exploren diversos arboles con distintas profundidades para valor la mejor de las configuraciones.

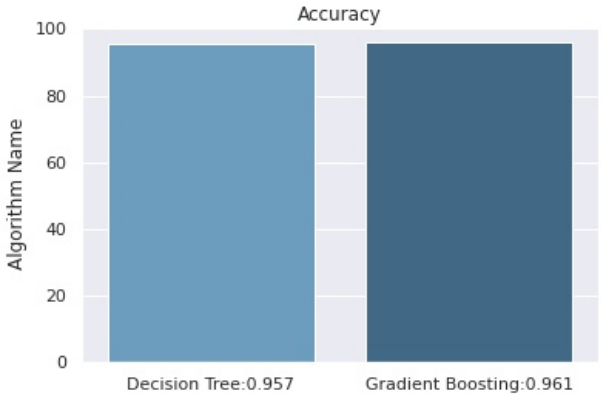


Fig.14 Cuadro Comparativo

Se puede concluir que los modelos de aprendizaje supervisado basados en arboles de decisión o *Decisión Tree* y *Gradient Boosting* han aprendido a realizar adecuadamente la predicción mediante clasificación, en el caso del árbol de decisión se obtuvo como resultado que la precisión de la predicción es del 95% frente al *Gradient Boosting* que alcanzó el 96% de precisión según se aprecia en la Fig. 14.

Referencias

[1] P. C. M, S. E, V. C. MA, and L. J. M, "COVID-19, a worldwide public health emergency," *Rev. Clin. Esp.*, vol. 221, no. 1, pp. 55–61, Jan. 2020.

[2] johns hopkins Hospital, "Covid19 DataCenter." [Online]. Available: https://coronavirus.jhu.edu/map.html.

[3] Google, "Kaggle." [Online]. Available: https://www.kaggle.com.
[4] Y. C. Wu, C. S. Chen, and Y. J. Chan, "The outbreak of COVID-19: An overview," *J. Chinese Med. Assoc.*, vol. 83, no. 3, pp. 217–220, 2020.
[5] A. Latini *et al.*, "COVID-19 and genetic variants of protein involved in the SARS-CoV-2 entry into the host cells," *Genes (Basel).*, vol. 11, no. 9, 2020.
[6] J. S. Cobb and M. A. Seale, "Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (United States) using statistical analyses and a random forest machine learning model," *Public Health*, vol. 185, p. 27, Aug. 2020.
[7] C. España, "Información importante acerca de las variantes | CDC." [Online]. Available: https://espanol.cdc.gov/coronavirus/2019-ncov/variants/about-variants.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fvariants%2Fdelta-variant.html. [Accessed: 18-Feb-2022].
[8] C. EE.UU, "Variante Omicron: lo que necesita saber | Centros para el Control y la Prevención de Enfermedades." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/omicron-variant.html. [Accessed: 18-Feb-2022].
[9] Jhu, "COVID-19 Map - Johns Hopkins Coronavirus Resource Center." [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 07-Jul-2021].
[10] U. of W. Hospitals, "Tratamientos y profilaxis (prevención) para COVID-19 - UW Health COVID-19 Information." [Online]. Available: https://coronavirus.uwhealth.org/es/sintomas-y-cuidados/tratamientos-y-profilaxis-prevencion-para-covid-19/. [Accessed: 18-Feb-2022].
[11] F. EE.UU, "Coronavirus (COVID-19) Update: FDA Authorizes Additional Oral Antiviral for Treatment of COVID-19 in Certain Adults | FDA." [Online]. Available: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-additional-oral-antiviral-treatment-covid-19-certain. [Accessed: 18-Feb-2022].
[12] R. M. Elshazli *et al.*, "Diagnostic and prognostic value of hematological and immunological markers in COVID-19 infection: A meta-analysis of 6320 patients," *PLoS One*, vol. 15, no. 8 August, 2020.
[13] S. Amin, M. I. Uddin, H. H. Al-Baity, M. A. Zeb, and M. A. Khan, "Machine learning approach for COVID-19 detection on twitter," *Comput. Mater. Contin.*, vol. 68, no. 2, 2021.
[14] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 731–739, Sep. 2020.
[15] D. N. Vinod and S. R. S. Prabaharan, "Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19," *Chaos, Solitons and Fractals*, vol. 140, 2020.
[16] T. C. T. Chen, C. L. Liu, and H. D. Lin, "Advanced artificial neural networks," *Algorithms*, vol. 11, no. 7. 2018.
[17] R. Arias Montoya, J. J. Santa Chávez, and J. de J. Veloza Mora, "Aplicación del aprendizaje automático con árboles dedecisión en el diagnóstico médico TT - Application of machine learning with decision trees in medical diagnosis," *Cult. Cuid. enferm*, vol. 10, no. 1, 2013.
[18] P. Mathur, *Machine Learning Applications Using Python*. 2019.
[19] Fuat, "Google Colab Free GPU Tutorial," *DEEP LEARNING TURKEY*, 2018. .
[20] D. P. Pérez, R. S. Bustillos, C. M. Mora, and M. Botto-Tobar, "Prediction of Covid19 with the use of Random Forests Algorithm and Artificial Neural Networks," *Ecuadorian Sci. J.*, vol. 4, no. 2, pp. 101–110, Sep. 2020.
[21] S. Ballı, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," *Chaos, Solitons and Fractals*, vol. 142, 2021.
[22] A. Morillo Alujas, G. Yagüe Utrilla, D. Villalba Mata, and G. Cano López, "Árboles de clasificación y regresión," *Suis*, no. 157, 2019.
[23] G. Prashant, "Decision Trees in Machine Learning – Towards Data Science," *Towar. Data Sci.*, 2017.
[24] I. Menes Camejo, G. A. Medina, P. Moreno Beltrán, and K. G. Carrillo, "Performance of data mining algorithms in academic indicators: Decision Tree and Logistic Regression," *Rev. Cuba. Ciencias Informáticas*, vol. 9, no. 4, 2015.
[25] F. A. Fabris João Pedro de Magalhães Alex Freitas, "A review of supervised machine learning applied to ageing research," *Biogerontology*, vol. 18.
[26] M. A. F. Azlah, L. S. Chua, F. R. Rahmad, F. I. Abdullah, and S. R. W. Alwi, "Review on techniques for plant leaf classification and recognition," *Computers*, vol. 8, no. 4. 2019.
[27] O. Özdemir, M. Batar, and A. H. Işık, "Churn Analysis with Machine Learning Classification Algorithms in Python," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 43, 2020.
[28] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making," *Smart Heal.*, vol. 20, 2021.
[29] V. A. Andrade Saltos and P. Flores M., "Comparativa Entre Classification Trees, Random Forest Y Gradient Boosting; En La Predicción De La Satisfacción Laboral En Ecuador.," *Cienc. Digit.*, vol. 2, no. 4.1., pp. 43–56, 2018.
[30] InteractiveChaos, "Gradient Boosting," *2021*, 2021. .
[31] J. I. Glaser, A. S. Benjamin, R. H. Chowdhury, M. G. Perich, L. E. Miller, and K. P. Kording, "Machine learning for neural decoding," *eNeuro*, vol. 7, no. 4, 2020.
[32] D. P. Perez, R. S. Bustillos, M. Botto-Tobar, and C. M. Mora, "X-Ray Images Analysis by Medium Artificial Neural Network," *Ecuadorian Sci. J.*, vol. 5, no. 1, pp. 55–60, Mar. 2021.
[33] R. Srivatsan, P. N. Indi, S. Agrahari, S. Menon, and S. D. Ashok, "Machine learning based prognostic model and mobile application software platform for predicting infection susceptibility of COVID-19 using healthcare data," *Res. Biomed. Eng.*, 2020.
[34] A. A. Osi *et al.*, "A classification approach for predicting COVID-19 Patient's survival outcome with machine learning techniques," *medRxiv*, 2020.
[35] J. Runge and R. Zmeureanu, "Forecasting energy use in buildings using artificial neural networks: A review," *Energies*, vol. 12, no. 17. 2019.
[36] J. H. Lee, D. H. Kim, S. N. Jeong, and S. H. Choi, "Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm," *J. Periodontal Implant Sci.*, vol. 48, no. 2, 2018.
[37] R. Delgado, "Introducción a la Validación Cruzada (k-fold Cross Validation) en R," *R Pubs*, 2018.
[38] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic

review," *Cancers (Basel).*, vol. 11, no. 9, 2019.
[39] S. Fiorini, F. Hajati, A. Barla, and F. Girosi, "Predicting diabetes second-line therapy initiation in the Australian population via timespan-guided neural attention network," *bioRxiv*, 2019.