Detección de COVID-19 a partir de síntomas mediante modelos de clasificación de machine learning

Carlos Alex Nina Guardapuclla
Data Science Research Perú
Cusco, Perú
124799@unsaac.edu.pe

***Abstract*—**Ante la crisis de salud que atraviesa el mundo, el presente trabajo tiene como objetivo desarrollar modelos de machine learning para la detección del COVID-19 basado en base a los síntomas presentado por los pacientes, para ellos los datos fueron extraídos de Kaggle, también se tomó como referencia la metodología CRIPS-DM para el desarrollo, en donde se procedió a realizar un análisis de los datos, y finalmente el desarrollo de los modelos, obteniéndose los siguientes resultados para el algoritmo árbol de decisión un 95% de accuracy, regresión logística un 94% y Naive Bayes un 83%, concluyendo que el modelo de árbol de decisión tiene un mejor desempeño debido a las características del dataset.

Keywords—COVID-19, Machine learning, CRISP-DM, árbol de decisión, regresión logística y Naive Bayes.

I. INTRODUCCIÓN

Dada la situación actual que atraviesa la sociedad debido a una cepa mutante del coronavirus, el SARS-CoV-2, el cual ha provocado una grave crisis económica, social y sanitaria en todo el mundo.

El Covid-19 en comparación con la gripe o influenza, se propaga con mayor facilidad, sin embargo, ambos son enfermedades respiratorias contagiosas, pero provocados por diferentes tipos de virus[1].

Por lo cual, ante la inminente necesidad por encontrar solución a los diferentes problemas de la sociedad, principalmente en el ámbito de la salud, la inteligencia artificial, el aprendizaje automático y la estadística desempeñan un rol muy importante, ya que ayudan al ser humano a encontrar soluciones para situaciones altamente complejas.

Las investigaciones en su mayoría realizan las tareas de predicción y detección de COVID-19 en base a reconocimiento de imágenes; por tal motivo, este proyecto propone una predicción en base a los síntomas de los pacientes, haciendo uso del algoritmo árbol de decisión, Naive Bayes y regresión logística.

II. DESCRIPCIÓN DEL PROYECTO

A. Marco teórico

El desarrollo del proyecto toma como referencia la metodología CRISP-DM, las etapas desarrolladas son:

1) COVID-19

La enfermedad por coronavirus (COVID-19) es una enfermedad infecciosa provocada por el virus SARS-CoV-2, la mayoría de las personas que padecen COVID-19 sufren síntomas de intensidad leve a moderada y se recuperan sin necesidad de tratamientos especiales. Sin embargo, algunas personas desarrollan casos graves y necesitan atención médica [1].

2) Machine Learning

Es una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. Sin embargo, machine learning no es un proceso sencillo. Conforme el algoritmo ingiere datos de entrenamiento, es posible producir modelos más precisos basados en datos [2].

3) CRISP-DM

Es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software [3].

4) Árbol de decisión

Esta técnica de machine learning toma una serie de decisiones en forma de árbol. Los nodos intermedios (las ramas) representan soluciones. Los nodos finales (las hojas) nos dan la predicción que vamos buscando [4].

5) Regresión logística

Es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula [5].

6) Naive Bayes

Naive Bayes o el Ingenuo Bayes es uno de los algoritmos más simples y poderosos para la clasificación basado en el Teorema de Bayes con una suposición de independencia entre los predictores. Naive Bayes es fácil de construir y particularmente útil para conjuntos de datos muy grandes, el clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características [6].

B. Metodología

El desarrollo del proyecto toma como referencia la metodología CRISP-DM, las etapas desarrolladas son:

A. Comprensión del problema

B. Obtención y compresión de los datos

C. Preprocesamiento de los datos

D. Modelado

C. *Comprensión del problema*

Este proyecto tiene como objetivo detectar si una persona contrajo la enfermedad de la COVID-19, a partir de los síntomas que presenta el paciente, mediante algoritmos de clasificación, los cuales son:

- Árbol de decisión
- Naive Bayes
- Regresión logística

Para así poder diferenciar dichas afectaciones de otras causadas por diferentes tipos de virus o bacterias que puedan afectar de manera similar al del virus SARS-CoV-2.

D. *Obtención y compresión de los datos*

Los datos fueron obtenidos de Kaggle, el cual cuenta con 1001034 registros y 15 columnas.

Diccionario del dataset:

Target: flag-sospechoso, valores [0,1].

Features: tos, cefalea, congestión-nasal, dificultad-respiratoria, dolor-garganta, fiebre, diarrea, náuseas, anosmia-hiposmia, dolor-abdominal, dolor-articulaciones, dolor-muscular, dolor-pecho y otros-síntomas, valores [0,1].

E. *Preprocesamiento de los datos*

Esta etapa comprende las etapas de limpieza y estandarización de los datos, lo cual permite depurar valores que puedan alterar los resultados a obtener por el modelo, para ello, se aplicaron diferentes técnicas para verificar la calidad de los datos.

Técnicas y herramientas utilizadas:

- Lenguaje de programació Python
- Preprocesamiento de datos con NumPy
- Visualización de datos con Matplotlib, Seaborn.
- Análisis estadístico con NumPy y Pandas.
- Construcción del modelo con ScikitLearn.
- Evaluación del modelo con librerías de ScikitLearn.

F. *Modelado*

1) *Elección de los algoritmos*

Debido a las características que presentan el dataset se ha elegido el algoritmo árbol de decisión, Naive Bayes y regresión logística.

2) *Métricas de evaluación:*

Las métricas a utilizar para evaluar el desempeño de los modelos son la matriz de confusión, accuracy, F1 score y curva ROC.

III. Resultados

A. *Análisis exploratorio de los datos*

En la fig.1 se muestra el porcentaje de pacientes que presentaron los diversos síntomas, en la cual se observa que los síntomas más frecuentes son tos, dolor de garganta, congestión nasal, fiebre, cefalea y dificultad respiratoria.

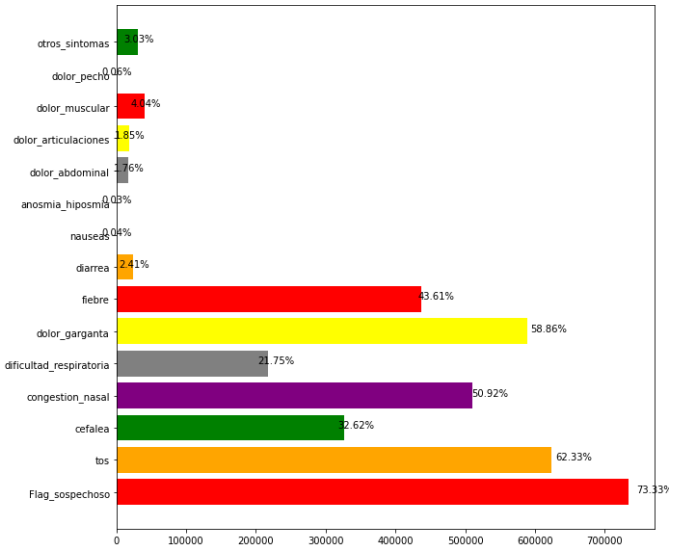

Fig. 1. Porcentaje de pacientes que presentaron los diversos sintomas

En la fig.2 se observa que los pacientes que dieron positivo a COVID-19 en su mayoría presentaron tos, por otro lado, hubo pacientes que dieron positivo sin presentar dicho síntoma.

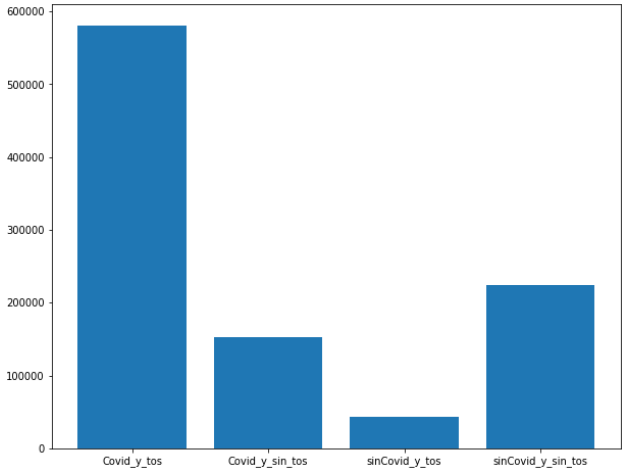

Fig. 2. Comparacion pacientes con y sin COVID-19 vs tos

En la fig.3 se observa que la mayoría de los pacientes que dieron positivo a COVID-19 presentaron dolor de garganta, lo cual indica que es un síntoma a tener en cuenta para acudir al médico.

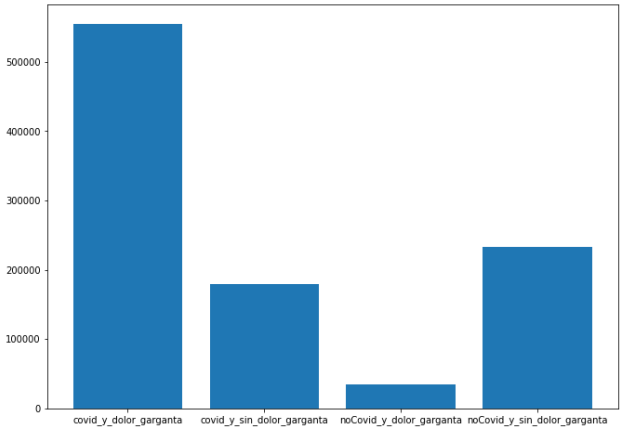

Fig. 3. Comparacion pacientes con y sin COVID-19 vs dolor de garganta

En la fig.4 se observa que los pacientes positivos a COVID-19 en su mayoría no presentaron dificultad respiratoria, esto puede deberse a diversos factores, sin embargo, este síntoma ha de tenerse en cuenta, ya que esta enfermedad podría confundirse con una simple gripe.

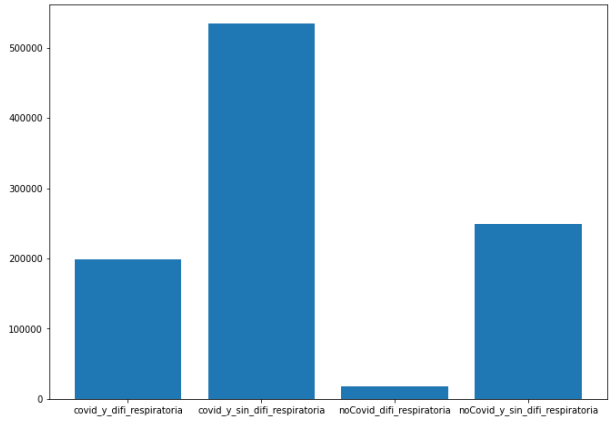

Fig. 4. Comparacion pacientes con y sin COVID-19 vs dificultad respiratoria

En la fig.5 se observa que los pacientes positivos a COVID-19 un numero razonable presento cefalea, sin embargo, en la mayoría de pacientes este síntoma no lo se presentó, lo que da a entender que esta enfermedad puede pasar desapercibida sino se realiza un seguimiento adecuado.

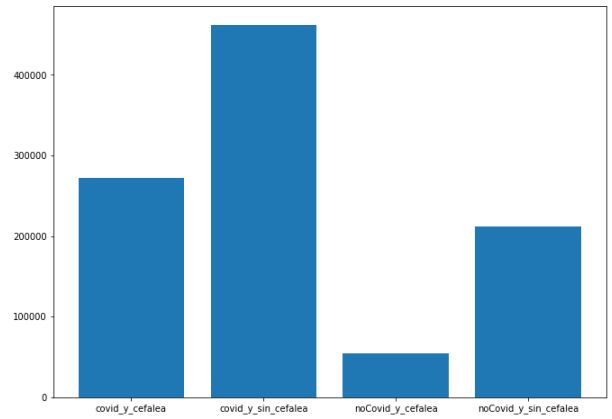

Fig. 5. Comparacion pacientes con y sin COVID-19 vs cefalea

En la fig.6 se observa que los pacientes positivos a COVID-19 presentaron fiebre, sin embargo, el otro grupo de pacientes que no presentaron este síntoma dan a entender que esta enfermedad puede causar síntomas variables y dificultar su diagnóstico llegando a confundirse con la gripe o enfermedades similares.

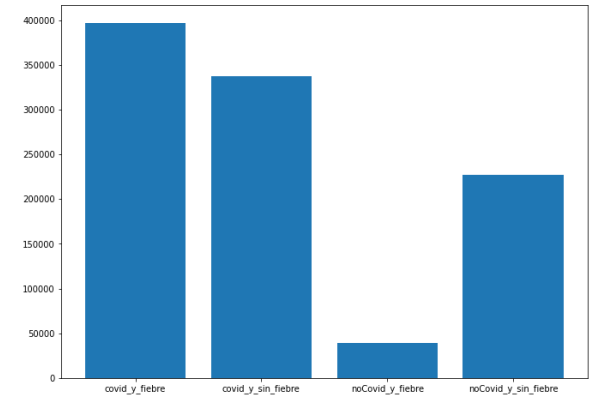

Fig. 6. Comparacion pacientes con y sin COVID-19 vs fiebre

A continuación de la fig.7 se muestra el grado de correlación de los diversos síntomas, de donde se puede abstraer que los síntomas que podrían determinar si un paciente tiene COVID-19 son la tos, dolor de garganta, fiebre, congestión nasal, dificultad respiratoria, cefalea y dolor de pecho, los demás síntomas se presentaron en con poca frecuencia, esto podría deberse a las condiciones u otras enfermedades que el paciente puede tener.

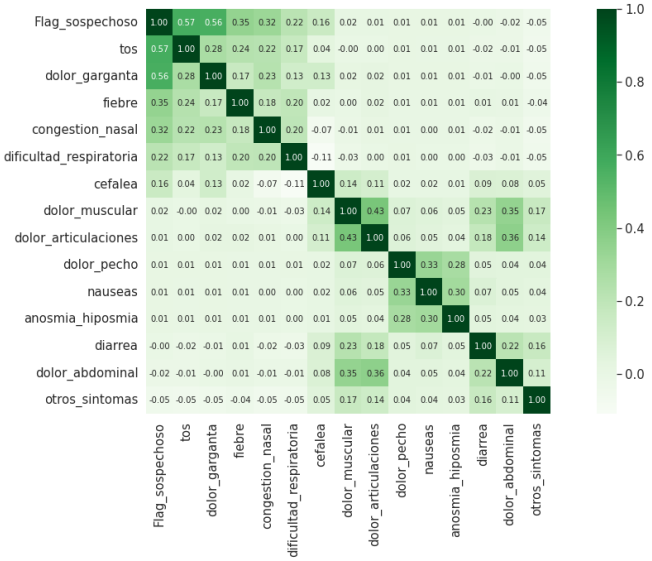

Fig. 7. Matriz de correlación

B. *Modelo*

A continuación, se muestran las tablas con los resultados obtenidos por los modelos de clasificación de machine learning elegidos.

TABLE I. MATRIZ DE CONFUSION ARBOL DE DECISION

	NEGATIVO	POSITIVO
FALSO	92942	13449
VERDADERO	4687	289336

TABLE II. MATRIZ DE CONFUSION NAÏVE BAYES

	NEGATIVO	POSITIVO
FALSO	94842	11549
VERDADERO	55488	238535

TABLE III. MATRIZ DE CONFUSION REGRESION LOGISTICA

	NEGATIVO	POSITIVO
FALSO	92726	13665
VERDADERO	8957	285066

La fig.8 muestra la gráfica de la matriz de confusión, de donde se observa que el algoritmo árbol de decisión y regresión logística tuvieron como resultado mayor acierto de verdaderos positivos.

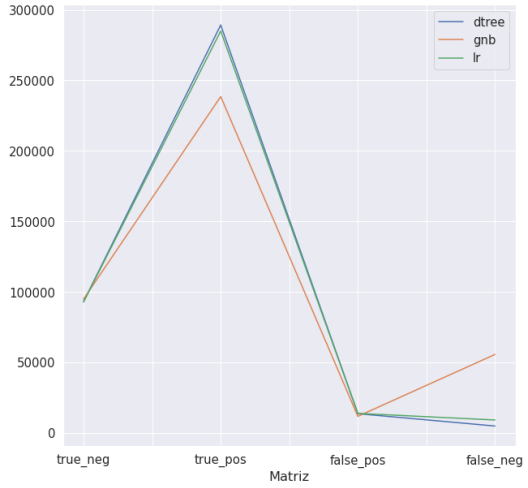

Fig. 8. Gráfica de la matriz de confusión

En las figuras 9,10, 11 y 12 se observa el desempeño obtenido por los diferentes modelos aplicando diferentes métricas de evaluación.

	precision	recall	f1-score	support
0	0.95	0.87	0.91	106391
1	0.96	0.98	0.97	294023
accuracy			0.95	400414
macro avg	0.95	0.93	0.94	400414
weighted avg	0.95	0.95	0.95	400414

Fig. 9. Metricas para árbol de decisión

	precision	recall	f1-score	support
0	0.63	0.89	0.74	106391
1	0.95	0.81	0.88	294023
accuracy			0.83	400414
macro avg	0.79	0.85	0.81	400414
weighted avg	0.87	0.83	0.84	400414

Fig. 10. Metricas para Naive Bayes

	precision	recall	f1-score	support
0	0.91	0.87	0.89	106391
1	0.95	0.97	0.96	294023
accuracy			0.94	400414
macro avg	0.93	0.92	0.93	400414
weighted avg	0.94	0.94	0.94	400414

Fig. 11. Metricas para regresion logística

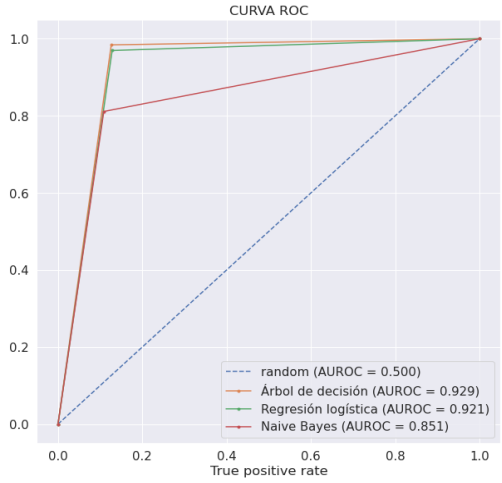

Fig. 12. Curva ROC

En base a los resultados obtenidos se observa que el algoritmo árbol de decisión presenta un mejor desempeño con un accuracy de 0.95 y AUROC de 0.929, seguido por el algoritmo de regresión logística con un accuracy de 0.94 y AUROC de 0.921 y por último Naive Bayes con un accuracy de 0.83 y AUROC de 0.851.

IV. Conclusiones y trabajos futuros

Se concluye que el algoritmo árbol de decisión se adapta mejor a datos de tipo binarios al igual que la regresión logística, sin embargo, Naive Bayes presento un menor desempeño debido a que la lógica lo realiza en base a probabilidades.

En relación a los resultados del análisis de los síntomas que causa la enfermedad COVID-19 se concluye que ante la presencia de alguno o en conjunto de estos síntomas (tos, dolor de garganta, fiebre, dificultad respiratoria, cefalea y congestión nasal) acudir al centro de salud para su detección y tratamiento adecuado, ya que es una enfermedad que en su mayoría podría llegarse a confundir con enfermedades similares como la gripe o influenza.

Se espera que este trabajo ayude a futuras investigaciones y en el desarrollo de un aplicativo, lo cual permitirá a los centros de salud agilizar la detección, realizando un diagnóstico temprano de esta enfermedad.

V. Referencias bibliograficas

[1] OMS.(23 de diciembre de 2021).Coronavirus disease covid19.Organizacion mundial de la salud. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted

[2] IBM.(2022). ¿Qué es Machine Learning?. https://www.ibm.com/pe-es/analytics/machine-learning

[3] Sngular.(2022).CRISP-DM: La metodología para poner orden en los proyectos.https://www.sngular.com/es/data-science-crisp-dm-metodologia/

[4] IArtificial.(19 de Septiembre de 2020). Árboles de decisión. https://www.iartificial.net/arboles-de-decision-con-ejemplos-en-python/#more-1498

[5] Amat R.(agosto, 2016).Regresion logistica simple.Ciencia de datos. https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

[6] Gonzalez Ligdi. (20 septiembre de 2019). AprendeIA.https://aprendeia.com/algoritmo-naive-bayes-machine-learning/