UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERIA
INTELIGENCIA ARTIFICIAL

Proyecto 1
Entrega 1

Integrantes
Manuela Arteaga Arango
Rosa Puerta Campo

Documentos
1.152.710.365
1.002.389.631

Docente
Raúl Ramos Pollan

Medellín-Antioquia
2023

Problema predictivo a resolver

La OMS declaró al COVID-19 como una pandemia causada por el SARS-CoV-2, representando una amenaza global para los sistemas de salud debido a la alta demanda que sobrepasa las capacidades de las UCI, y su fácil contagio empeora la situación. Aunque se ha logrado cierto control, siguen apareciendo contagios y disminuye el autocuidado. Para frenar la pandemia, es crucial lograr la inmunidad global a través de la administración de vacunas, una herramienta históricamente efectiva contra enfermedades infecciosas.

En menos de un año, se desarrollaron vacunas efectivas contra el SRAS-CoV-2 (Mathieu, E. et al., 2021). El desafío actual es garantizar su acceso en todo el mundo, no solo en países ricos [5]. Our World in Data recopila datos internacionales sobre vacunación, actualizados diariamente, para monitorear el progreso.

El objetivo de este proyecto consiste en predecir la evolución futura de la pandemia, a partir del desarrollo de un modelo capaz de proporcionar predicciones precisas sobre el número de casos nuevos y posiblemente la tasa de mortalidad del COVID-19 en diferentes países o regiones. El objetivo central sería construir un modelo que pueda prever con precisión el número de casos nuevos de COVID-19 en el futuro. Esto implica utilizar datos históricos y características relevantes para hacer predicciones que se acerquen lo más posible a la realidad. El proyecto permitirá predecir la evolución de la pandemia a medida que avanza en el tiempo. Esto es crucial para la planificación de recursos de salud, la toma de decisiones gubernamentales y la preparación para posibles aumentos en la propagación del virus [5].

Dataset que se va a utilizar

El dataset a utilizar se encuentra en un desafío de Kaggle (https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset?resource=download) el cual presenta información de la enfermedad del COVID-19, recopilada en diferentes regiones del mundo. El dataset está compuesto por un conjunto de datos .csv, el cual se caracteriza por presentar un tamaño de 166326 filas y 67 columnas, en las cuales se tiene información de la situación de la pandemia en una región en específico, como es el caso del total de muertes, total de vacunados, pacientes hospitalizados, entre otros.

En los datos más relevantes se encuentra los siguientes:

- location: Contiene el país a donde está asociado el registro
- date: Contiene la fecha del registro
- total_cases: Contiene el conteo de casos positivos totales hasta el momento

- new_cases: Contiene la información de nuevos casos de Covid reportados.
- total_deaths: Contiene la información de las muertes causadas por el covid hasta ese día.
- new_deaths: Contiene la información de las nuevas muertes reportadas en esa fecha.
- icu_patients: Contiene el conteo de pacientes en UCI hasta ese día.
- hosp_patients: Contiene el conteo de pacientes que han sido hospitalizados hasta el momento.
- people_vaccinated: Contiene el conteo de personas que han sido vacunadas hasta el momento.
- new_vaccinations: Contiene el número de nuevos vacunados ese día.

Métricas de desempeño requeridas

1. Machine Learning

- **Matriz de confusión:** Es una matriz en donde se coteja la información real con la información predicha, en la búsqueda de establecer qué tan fiable o no es el modelo. La cual es de la siguiente forma:

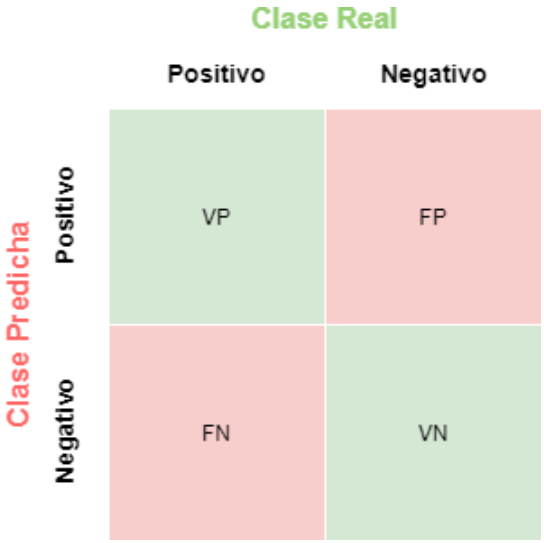

Imagen 1. Matriz de confusión [3].

Donde:

VP: Verdadero positivo. Indicará el número de casos positivos donde el modelo prediga como positivos.

FP: Falsos positivos. Indicará el número de casos negativos donde el modelo prediga como positivos.

FN: Falsos negativos. Indicará el número de casos positivos donde el modelo prediga como negativos.

VN: Verdaderos negativos. Indicará el número de casos negativos donde el modelo prediga como negativos.

- **Sensibilidad:** indicará la proporción de resultados positivos que están siendo predichos correctamente por el modelo entre todos los positivos reales. El objetivo es que el algoritmo tenga una sensibilidad máxima para que pueda esclarecer los casos positivos y negativos, sin embargo, va de la mano con las otras métricas, pues por sí sola no aporta mucha información [3].
- **Especificidad:** Es la verdadera tasa negativa o la proporción de verdaderos negativos a todo lo que debería haber sido clasificado como negativo [3].
- **Precisión:** Se medirá la precisión de nuestro modelo al momento de predecir los casos positivos. Una métrica de suma importancia es determinar si el modelo está clasificando correctamente a los pacientes con la patología, pues es esta la que brindará información sobre qué tan fiable o no es este algoritmo y qué tan viable es ponerlo en producción [3].
- **Exactitud:** Con qué frecuencia nuestros datos son correctos. Analizaremos si los datos que se encuentran tienen o no sentido para la aplicación que necesitamos, además si toda esta información realmente aporta al correcto diagnóstico de un paciente, pues se busca que el modelo sea capaz de establecer una patología con la menor cantidad posible de información por temas de Optimización: Los modelos trabajan de forma más rápida cuando la información a procesar es más pequeña. La exactitud se representa como un porcentaje o un valor entre 0 y 1 [3].

2. Negocio

Nuestro modelo debe lograr una exactitud del 95% o superior y una precisión del 90% o superior. Esta alta precisión es fundamental debido a la gravedad de la patología en cuestión. Es esencial evitar errores en la predicción de casos positivos, ya que, en una aplicación práctica, tales errores pueden tener graves consecuencias para la planificación de recursos de salud, la toma de decisiones gubernamentales y la preparación para posibles aumentos en la propagación del virus [2].

Primer criterio sobre el cual seria el desempeño deseable en producción.

Si el modelo no alcanza una exactitud del 95% y una precisión del 90%, no sería justificable su implementación en producción debido a su falta de fiabilidad. Esto es especialmente crítico dado que su principal beneficio radicaría en respaldar la toma precisa de decisiones. Para poder establecer un modelo capaz de determinar nuevo casos de infecciones en base a unos datos recolectados, es necesario determinar que los datos con los que se está construyendo son útiles y aportan al sistema, es por ello que es de suma importancia analizar el set con el que se cuenta; ahora, cuando el modelo sea generado y entrenado, se esperaque sea capaz de predecir con la

mayor precisión y exactitud posible el número de casos nuevos de COVID-19, haciendo uso exclusivo de información que se le suministre.

Bibliografía

[1] COVID-19 dataset. (s.f.). Kaggle: Your Machine Learning and Data Science Community. Tomado de: https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset

[2] ¿Cómo sé si mi modelo de predicción es realmente bueno?. Tomado de: https://datos.gob.es/es/blog/como-se-si-mi-modelo-de-prediccion-es-realmente-bueno

[3] Métricas De Evaluación De Modelos En El Aprendizaje Automático. Tomado de: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

[4] Interpretabilidad de los modelos de Machine Learning. Tomado de: https://quanam.com/interpretabilidad-de-los-modelos-de-machine-learning-primera-parte/

[5] Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). Tomado de: https://ourworldindata.org/covid-vaccinations