UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE BIOINGENIERÍA
Inteligencia Artificial

Proyecto
Entrega 1

Nombres
Barbara Paulina Chavarria Rua
Lina Natalia Angulo

Docente
Raul Ramos Pollan

Documentos
1001555256
1006874552

Medellín - Antioquia
2022

1. ***Problema predictivo a resolver:***

   Debido al creciente número de casos, la Organización Mundial de la Salud (OMS) ha calificado el COVID-19, causado por el virus SARS-CoV-2, como una pandemia el 11 de marzo. Esto representa una amenaza para los sistemas de salud en todo el mundo, ya que la alta demanda está abrumado las capacidades de las Unidades de Cuidados Intensivos (UCI) y es posible que se presenten limitaciones en la realización de pruebas para detectar el SARS-CoV-2. Probar cada caso puede resultar poco práctico y los resultados de las pruebas podrían retrasarse, incluso en el caso de una pequeña población. Por lo tanto, se busca desarrollar un modelo que pueda predecir los casos confirmados de COVID-19 entre los casos sospechosos que visitan la sala de urgencias, teniendo en cuenta que se considera un caso sospechoso basándose en los resultados de las pruebas de laboratorio comúnmente recolectadas.

2. ***Dataset que se va a utilizar:***

   El conjunto de datos que se utilizará proviene de un desafío en Kaggle y contiene información anónima de pacientes atendidos en el Hospital Israelita Albert Einstein en São Paulo, Brasil. Durante su estancia en el hospital, se les tomaron muestras para realizar pruebas de laboratorio adicionales, además del SARS-CoV-2 RT-PCR. Es importante destacar que los datos clínicos se han normalizado para tener una media de cero y una desviación estándar unitaria.

   El conjunto de datos se compone de un archivo .xlsx llamado "dataset.xlsx", que consta de 5644 filas y 111 columnas. Proporciona información del paciente, como su identificación y edad, así como los resultados de varias pruebas de laboratorio realizadas. Entre los datos suministrados se incluyen:

   - patient_id, patient_age**- Datos del paciente**
   - sars-cov-2_exam_result- **Resultado del examen sars-cov-2**
   - patient_addmited_to_regular_ward_(1=yes,_0=no) patient_addmited_to_semi-intensive_unit_(1=yes,_0=no) patient_addmited_to_intensive_care_unit_(1=yes,_0=no)- **Estado de ingreso del paciente**
   - hematocrit , serum_glucose, respiratory_syncytial_virus, mycoplasma_pneumoniae, neutrophils, urea, proteina_c_reativa_mg/dl, potassium- **Resultado de pruebas realizadas al paciente**
   - influenza_b, alanine_transaminase , gamma-glutamyltransferase, total_bilirubin, ionized_calcium, strepto_a, magnesium, pco2_(venous_blood_gas_analysis) , fio2_(venous_blood_gas_analysis)- **Resultado de pruebas realizadas al paciente**

- urine_-_esterase, urine_-_aspect, urine_-_ph, urine_-_hemoglobin, urine_-_ketone_bodies, urine_-_nitrite, urine_-_sugar, urine_-_leukocytes, urine_-_crystals- **Resultado de pruebas de orina realizadas al paciente**
- partial_thromboplastin_time (ptt), vitamin_b12, creatine_phosphokinase (cpk), ferritin, arterial_lactic_acid, lipase_dosage, d-dimer, albumin, arteiral_fio2, phosphor, entre otras- **Resultado de pruebas realizadas al paciente**

El dataset presenta como variable target (variable objetivo) :sars_covid_exam_results y entre los tipos de datos cuenta con variables de tipo floats (flotante), objects (objeto) e integer (entero)

3. Métricas de desempeño requeridas (de machine learning y de negocio)

- Exactitud: Con que frecuencia nuestros datos son correctos.Evaluaremos si los datos disponibles son apropiados para el propósito deseado, así como si toda la información contribuye realmente a un diagnóstico preciso del paciente. Esto se debe a que queremos que el modelo pueda identificar una patología con la menor cantidad de información posible por dos motivos: 1) para optimizar el rendimiento del modelo, ya que procesa información de manera más eficiente cuando se manejan conjuntos de datos más pequeños; y 2) para facilitar la realización de pruebas, ya que recopilar grandes cantidades de información de un paciente para determinar si tiene o no COVID-19 puede ser ineficiente. La precisión se expresa como un porcentaje o un valor entre 0 y 1.
- Precisión: Se evaluará la precisión de nuestro modelo en la predicción de casos positivos. Es crucial determinar si el modelo está identificando adecuadamente a los pacientes con la patología, ya que esto nos proporcionará información sobre su confiabilidad y su capacidad para ser implementado en la práctica. Esta es una métrica muy importante para nosotros.
- Sensibilidad: La métrica evaluará la fracción de resultados positivos que el modelo predice correctamente entre todos los resultados positivos reales. Es importante que el algoritmo tenga una sensibilidad máxima para poder detectar tanto los casos positivos como negativos. Sin embargo, esta métrica no proporciona información suficiente por sí sola, y debe ser evaluada en conjunto con otras métricas.
- Especificidad: Es la verdadera tasa negativa o la proporción de verdaderos negativos a todo lo que debería haber sido clasificado como negativo.
- 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 matriz se ve de la siguiente manera:

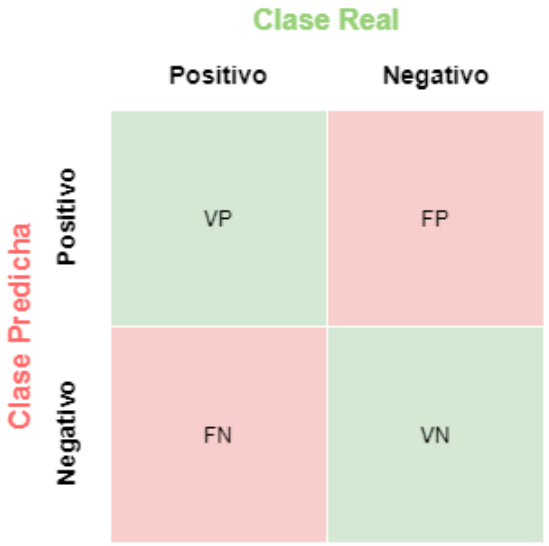

Imagen 1. Matriz de confusión.

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.

La matriz de confusión también puede ser vista de la siguiente manera:

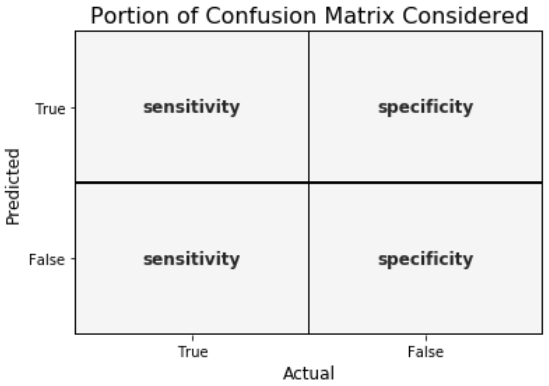

Imagen 2. Especificidad y sensibilidad en la matriz de confusión.

- F1 Score: Combina precisión y sensibilidad. Esta métrica la usamos en caso de que encontremos que los datos no se encuentren balanceados (Que estén sesgados).

En cuanto a la métrica de negocio, nuestro modelo debería de tener una exactitud igual o superior al 95% y una precisión mayor o igual al 90%, esto debido a que es una patología grave y es preferible no fallar una detección de un paciente que verdaderamente tiene la patología, pues en una aplicación real esto puede traer consecuencias como lo es el incremento de casos confirmados de COVID-19 por un mal diagnóstico, poniendo en riesgo a las demás personas.

4. ***Primer criterio sobre cuál sería el desempeño deseable en producción.***

Si el modelo no tiene una exactitud del 95% y una precisión del 90%, no vale la pena ponerlo en producción, debido a la poca fiabilidad de este. Para poder establecer un modelo capaz de determinar si un paciente está infectado o no 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 espera que sea capaz de determinar con la mayor precisión y exactitud posible si un paciente tiene o no covid, haciendo uso exclusivo de información que se le suministre.

Bibliografía

- Diagnóstico de COVID-19 y su espectro clínico| kaggle. Tomado de: https://www.kaggle.com/datasets/einsteindata4u/covid19?resource=download
- ¿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
- 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
- Interpretabilidad de los modelos de Machine Learning. Tomado de: https://quanam.com/interpretabilidad-de-los-modelos-de-machine-learning-primera-parte/