UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE BIOINGENIERÍA
Inteligencia Artificial

Proyecto

Entrega 1

Nombres
Nataly Rodriguez Ateourtua
Juan Camilo Cataño Zuleta
Daniel Enrique López Yepes

Docente
Raul Ramos Pollan

Documentos
1.001.505.918
1.001.738.289
1.000.407.016

Medellín - Antioquia
2022

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

   La Organización Mundial de la Salud (OMS) debido al aumento exponencial del número de casos, caracterizó el 11 de marzo al COVID-19, causado por el SARS-CoV-2, como pandemia, siendo una amenaza en los sistemas de salud de todo el mundo, debido a la alta demanda que se necesitaba abastecer, ya que se encontraban al límite de sus capacidades de UCI, presentando un sistema de salud abrumado con posibles limitaciones en la realización de pruebas para la detección de SARS-CoV-2, donde probar cada caso sería poco práctico y los resultados de las pruebas podrían retrasarse, incluso si se estudia una pequeña población. Es por esto que se desea desarrollar un modelo que permita predecir por medio del dataset utilizado los casos confirmados de COVID-19 entre los casos sospechosos durante una visita a la sala de urgencias, teniendo en cuenta que un caso se considera sospechoso basándose en los resultados de las pruebas de laboratorio comúnmente recolectadas.

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

   El dataset a utilizar se encuentra en un desafío de Kaggle, donde se proporciona datos anónimos de pacientes atendidos en el Hospital Israelita Albert Einstein, en São Paulo, Brasil, a quienes se les recolectaron muestras para realizar el SARS-CoV-2 RT-PCR y pruebas de laboratorio adicionales durante su estadía en el hospital. Es de resaltar que los datos clínicos se estandarizaron para tener una media de cero y una desviación estándar unitaria.

   El dataset está compuesto por un conjunto de datos .xlsx, el cual es nombrado como dataset.xlsx, se caracteriza por presentar un tamaño de 5644 filas y 111 columnas, en las cuales se tiene información del paciente, como su identificación y la edad y resultados de distintas pruebas de laboratorio realizadas, entre los datos suministrados se encuentra:

   - 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. Analizaremos si los datos que se encuentran tienen o no sentido para la aplicación que necesitamos, además de 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: 1). Optimización: Los modelos trabajan de forma más rápida cuando la información a procesar es más pequeña. y 2). Facilitar el manejo de pruebas, pues obtener grandes volúmenes de datos de un paciente para determinar si tiene o no COVID-19 puede no ser muy óptimo. La exactitud se representa como un porcentaje o un valor entre 0 y 1.
- 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 me brindará información sobre qué tan fiable o no es este algoritmo y qué tan viable es ponerlo en producción.
- Sensibilidad: indicará la proporción de resultados positivos que están siendo predichos correctamente por el modelo entre todos los positivos reales. La 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.
- 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:

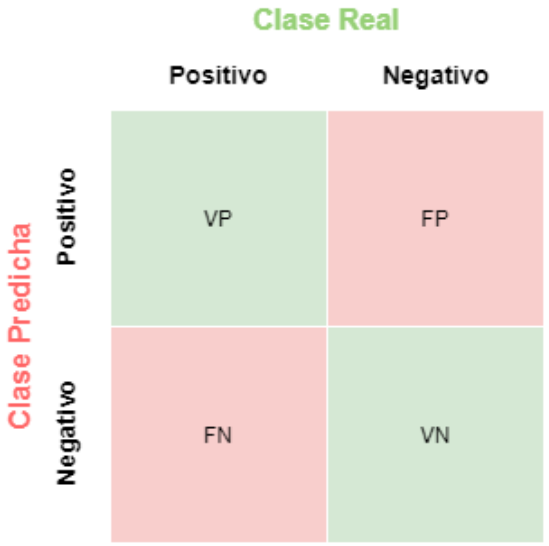

Imagen 1. Matriz de confusión.

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:

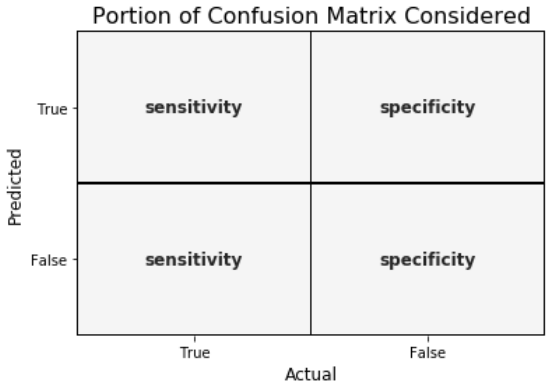

Imagen 2. Especificidad y sensibilidad en la matriz de confusión.

- 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/