INTELIGENCIA ARTIFICIAL
PROYECTO ENTREGA #2

ADRIANA ISABEL RIOS
MARIA CAMILA LOPERA
SARA PÉREZ HIGUITA

DOCENTE: RAÚL RAMOS POLLÁN
HORARIO MARTES-JUEVES 10-12 PM.

UNIVERSIDAD DE ANTIOQUIA
AGOSTO DEL 2022

Introducción

- **Problema predictivo por resolver.**

El cuerpo humano tiene mecanismos de defensa que generan diferentes respuestas en contacto con las distintas afecciones que puedan presentarse. La sepsis es una respuesta extrema del cuerpo ante una infección de alto nivel. Es considerada una emergencia médica ya que, sin un tratamiento oportuno, puede provocar rápidamente daños en los tejidos, insuficiencia orgánica e incluso la muerte. La detección temprana de la sepsis es esencial para mejorar el pronóstico, tiene el potencial de salvar vidas y limitar los recursos hospitalarios requeridos para atender la emergencia.
Por esto, con los valores de la sintomatología de las sepsis expresadas en el dataset a continuación, se tendrá como propósito desarrollar un modelo predictivo para predecir la sepsis tempranamente (6 horas antes de la predicción clínica) y falsas alarmas en dado caso.

- **Dataset a utilizar.**

El dataset fue seleccionado de la plataforma Kaggle [https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis], que tiene 1552210 filas y 44 columnas.

- **Parámetros registrados.**

- Signos vitales (columnas 1-8)

FC Frecuencia cardíaca (latidos por minuto)
O2Sat Pulsioximetría (%)
Temp Temperatura (Grados C)
PAS PA sistólica (mm Hg)
PAM Presión arterial media (mm Hg)
PAD PA diastólica (mm Hg)
Resp Tasa de respiración (respiraciones por minuto)
EtCO2 Dióxido de carbono corriente final (mm Hg)

- Valores de laboratorio

(columnas 9-34)
BaseExcess Medida de exceso de bicarbonato (mmol/L)
HCO3 Bicarbonato (mmol/L)
FiO2 Fracción de oxígeno inspirado (%)
pH N/A
PaCO2 Presión parcial de dióxido de carbono de la sangre arterial (mm Hg)
SaO2 Saturación de oxígeno de la sangre arterial (%)
AST Aspartato transaminasa (UI/L)
BUN Nitrógeno ureico en sangre (mg/dL)
Fosfatasa alcalina Fosfatasa alcalina (UI/L)
Calcio (mg/dL)
Cloruro (mmol/L)
Creatinina (mg/ dL)

Bilirrubina directa Bilirrubina directa (mg/dL) Glucosa Glucosa sérica (mg/dL) Lactato Ácido láctico (mg/dL) Magnesio (mmol/dL) Fosfato (mg/dL) Potasio (mmol/L) Bilirrubinatotal Bilirrubina total (mg/dL)
Troponina I Troponina I (ng/mL)
Hct Hematocrito (%)
Hgb Hemoglobina (g/dL)
PTT Tiempo de tromboplastina parcial (segundos) Recuento de
leucocitos WBC (recuento 10^3/µL) Fibrinógeno (mg/ dL) Plaquetas (recuento 10^3/µL)

- Datos demográficos (columnas 35-40)

Edad Años (100 para pacientes de 90 años o más)
Sexo Mujer (0) o Hombre (1)
Unidad 1 Identificador administrativo de la unidad de UCI (MICU) Unidad 2 Identificador administrativo de la unidad de UCI (UCI)
HospAdmTime Horas entre ingresos hospitalarios e ingreso
en UCI ICULOS Duración de la estadía en UCI (horas desde el ingreso en UCI)

Resultado (columna 41)
SepsisLabel Para pacientes con sepsis, SepsisLabel es 1 si $t \geq t_{sepsis} - 6$ y 0 si t

- **Métricas de desempeño requeridas.**

El rendimiento del modelo se evaluará utilizando datos de discriminación y calibración independientes.
Para la evaluación del desempeño del modelo se utilizará el puntaje F1, que incluye las métricas de precisión, y sensibilidad; lo cual resulta conveniente para el contexto clínico trabajado, ya que en este caso la distribución de las clases es desigual.

- **Crítica sobre desempeño deseable en producción.**

Se utilizará una función de utilidad. El modelo de predicción de sepsis en pacientes debería de tener un porcentaje de acierto >90%, y también un falso positivo <10%, ya que es una patología grave en la que no conviene una predicción tardía que agravará lacondición clínica ni un falso pronóstico que conllevará al desaprovechamiento de recursos hospitalarios.

- **Exploración descriptiva del dataset**

Los datos presentados en el dataset dado contienen valores de diferentes pruebas para diferentes pacientes, las medidas tomadas en cada prueba se presentan en las columnas, y las diferentes tomas se presenta en las filas.
En la inspección inicial de los datos se cargan y se visualizan los primeros datos para confirmar la información anterior, luego, se verifica el tipo de variable para garantizar que se tengan todos los valores. Se grafican y se observan valores desequilibrados de una media

proporcional, lo que indica que faltan muchos datos en algunas pruebas. (Figura 1)

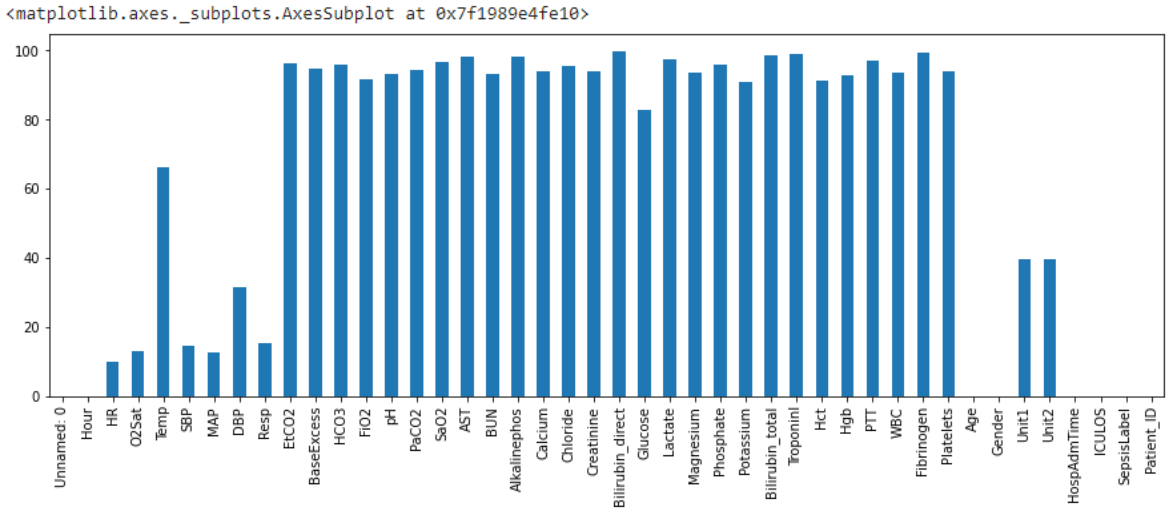

Figura 1. *Inspección visual en grafica de los datos*

- **Iteraciones de desarrollo**

Se busca inicialmente el equilibrio de los datos, para ello existen varias formas como por ejemplo el sobre muestreo que garantice la mínima perdida de datos. Sin embargo, en este caso, para facilitar el procesamiento y disminuir el gasto computacional se opta por separar los datos de los pacientes que contrajeron sepsis en dos grupos: los que contrajeron sepsis antes de entrar a UCI y los que adquieren la infección después de la admisión.

Para el primer grupo que corresponde a los pacientes con sepsis (CS), se emplea la función unique() [1], la cual permite localizar los valores únicos en los arrays. Esta retorna un array Numpy con estos valores. Luego, se crea un frame con estos datos extraídos, utilizando la función isin() [2] la cual comprobará que cada elemento del data.Patiente_ID si contenga el valor especificado en los elementos CS de entrada, devolviendo una tabla tipo DataFrame de booleanos indicando si cada elemento contiene los valores de la entrada.
Luego, se realiza el mismo procedimiento para el grupo de pacientes correspondiente a los que presentan sepsis antes de la admisión en la UCI (SAU).

Partiendo de los frames anteriores, se pueden construir los frames necesarios para los otros grupos de pacientes que faltan que son lo que presentan la infección después de la admisión de UCI (SDU) y los que no presentan sepsis (SS).

```
SS     1379800
SDU     168764
SAU       3646
Name: sepsisType, dtype: int64
```

Figura 2: *Verificación del procedimiento anterior*

Después de esto se hace una inspección visual de los datos como en el procedimiento inicial para verificar que exista mayor equilibrio y en el siguiente paso poder hacer el cálculo del

síndrome de la respuesta inflamatoria con el que se hará la predicción que se tiene como objetivo.

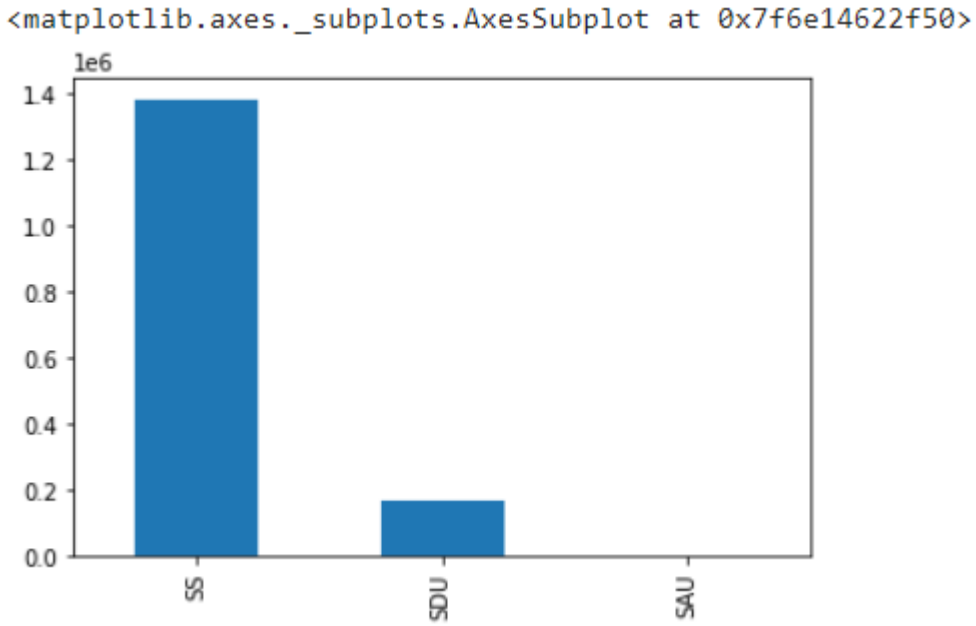

Figura 3: *Segunda inspección visual de los datos.*

Referencias

[1] Joshi S. Función Python numpy.unique() [Internet]. Delft Stack. 2020 [cited 2022 Aug 23]. Available from: https://www.delftstack.com/es/api/numpy/python-numpy-unique/#:~:text=unique()%20M%C3%A9todo

[2] Hu J. Función Pandas DataFrame DataFrame.isin() [Internet]. Delft Stack. 2020 [cited 2022 Aug 23]. Available from: https://www.delftstack.com/es/api/python-pandas/pandas-dataframe-dataframe.isin-function/

[3] Hackathon_1Feb2022 [Internet]. kaggle.com. [cited 2022 Aug 23]. Available from: https://www.kaggle.com/code/binitagiri/hackathon-1feb2022