UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

PROYECTO FINAL

INTELIGENCIA ARTIFICIAL
HORARIO MARTES-JUEVES 10-12 PM.

NOMBRES:
ADRIANA ISABEL RIOS
MARIA CAMILA LOPERA
SARA PÉREZ HIGUITA

DOCUMENTOS:
1.044.509.774
1.044.507.600
1.152.471.199

DOCENTE: RAÚL RAMOS POLLÁN

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERIA, PROGRAMA DE BIOINGENIERIA
NOBIEMBRE DEL 2022

1. 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)	en UCI (horas desde el ingreso en UCI)
HospAdmTime Horas entre ingresos hospitalarios e ingreso	Resultado (columna 41)
en UCI ICULOS Duración de la estadía	SepsisLabel Para pacientes con sepsis, SepsisLabel es 1 si t≥tsepsis−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.

2. 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 pruebas 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)

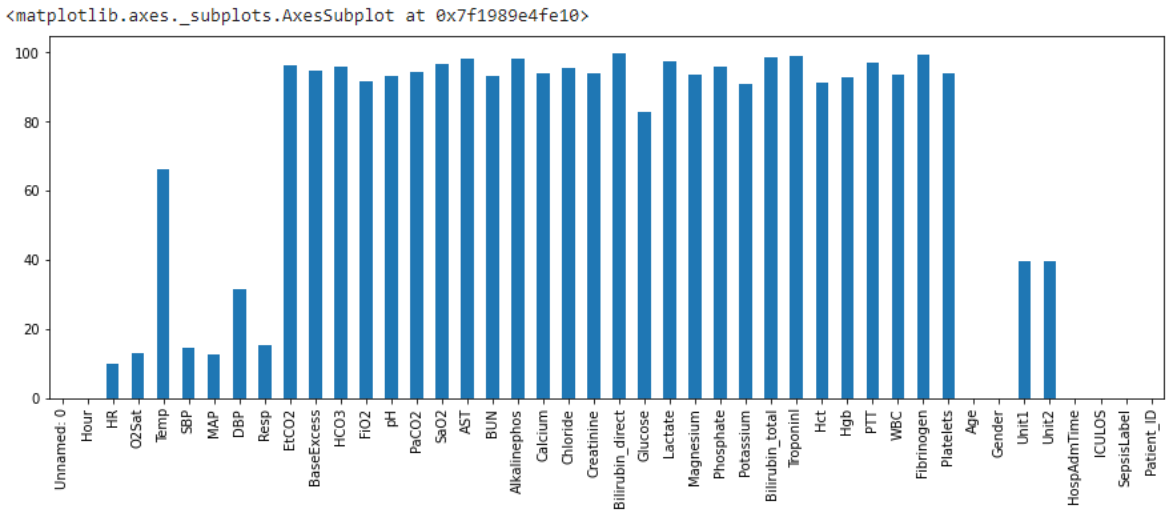

Figura 1*. Inspección visual en grafica de los datos*

3. 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.

Procesamiento de los datos

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.

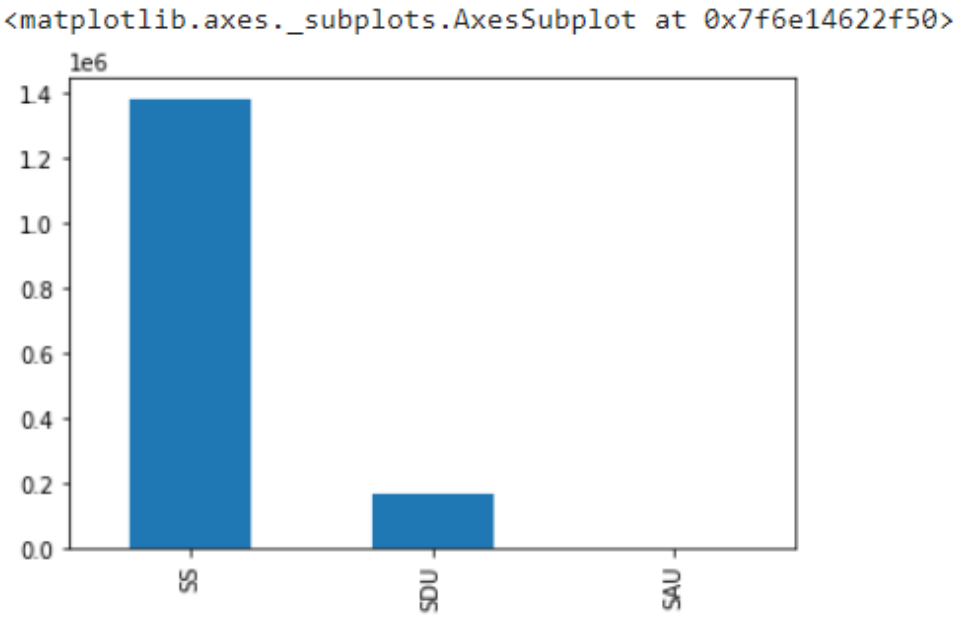

Figura 3: *Segunda inspección visual de los datos.*

Una vez realizada la inspección visual, se procedió a obtener el número total de personas que se encuentran en los diferentes grupos, estos valores se visualizan en la siguiente tabla.

Clasificación	Número de pacientes
Pacientes con sepsis (CS)	2932
Pacientes con sepsis antes de la admisión en la UCI (SAU).	426
Pacientes con la infección después de la admisión de UCI (SDU)	2506
Pacientes que no presentan sepsis (SS)	37404

Tabla 1. Clasificación de grupos de pacientes.

Modelos supervisados

Cálculo de SIRS

El síndrome de respuesta inflamatoria sistémica (SIRS), es una afección grave por la que se inflama todo el cuerpo, esto puede ser causado por una infección bacteriana graves (sepsis), un trauma o pancreatitis. Se caracteriza por presentar un pulso rápido, una presión arterial baja, temperatura del cuerpo alta o baja y un recuento de glóbulos blancos alto o bajo.

con esto podremos evaluar las condiciones de cada paciente para dar una predicción a su posible respuesta inflamatoria.

Para realizar este cálculo, primero se buscó en las columnas de las variables presentes en nuestros datos si los elementos eran NaN o no, para ello se empleó la función Np.isnan(), la cual devuelve el resultado como una matriz booleana y se compararon entre sí para obtener una condición general. Luego, para obtener la cantidad de estos elementos, se utilizó la función Np.where(), la cual según la condición que se obtuvo para cada variable, busca los elementos que la cumplan en las diferentes columnas. Y, por último, se procedió a realizar un gráfico de barras verticales con los resultados obtenidos, usando la función Plot.bar().

Una vez obtenida la información del cálculo de SIRS, se procedió a realizar un análisis con histogramas, gráficos de barras y correlaciones para determinar los efectos del fosfato alcalino, exceso de base y lactato en la incidencia de sepsis.

Histograma

Es la representación gráfica en forma de barras, que nos simboliza la distribución de un conjunto de datos. Sirve para obtener una primera vista general de la distribución de los datos respecto a una característica [4]. Para obtener esta representación gráfica, se utilizó la función histplot, con la cual se trazan histogramas univariados o bivariados para mostrar distribuciones de conjuntos de datos.

Gráficos de barras

Las gráficas de barras presentan barras rectangulares con longitudes proporcionales a los valores que representan, estos se utilizan para comparar dos o más valores [5]. Se empleó la función Plot.bar(), para obtener estos gráficos.

Correlación

La correlación es un tipo de asociación entre dos variables numéricas evaluando la tendencia (creciente o decreciente) de sus datos [6]. Para obtenerla, se utilizó la función paripplot, la cual traza múltiples distribuciones bivariadas por pares en un conjunto de datos.

Resultados, métricas y curvas de aprendizaje

Para los datos presentados anteriormente, el cálculo del SIRS se muestran por medio de un gráfico de barras que toma 5 variables que se condicionan para la clasificación de los datos en tres variables importantes que darán un acercamiento a la predicción de la sepsis. Las 4 variables son: temperatura (Temp) normal entre 36°- 38°, ritmo cardiaco (HR) normal cuando es menor a 90 latidos por minuto, respiración (Resp) normal a 20 y condicionada por la presión parcial de dióxido de carbono de la sangre arterial (PaCO2) que es normal en niveles superiores a 32 y, por último, la cantidad normal de glóbulos blancos (WBC) entre 4000 y 12000 GB por microlitro.

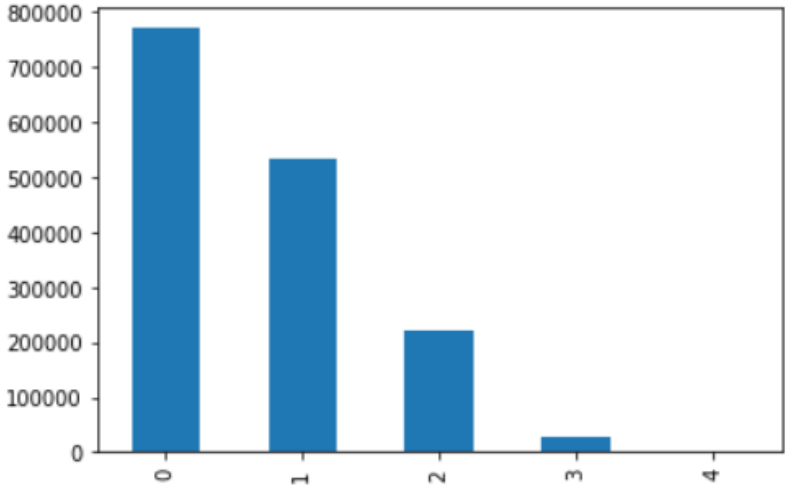

Figura 4: *Grafico de barras para el cálculo del SIRS.*

Se considera que un paciente presenta síndrome de respuesta inflamatoria sistemática al menos 2 de las variables presentadas y la Figura 4 muestra que los cambios más significativos se encuentran el recuentro normal de glóbulos blancos en sangre, este factor se clasifica como fundamental para la predicción de la sepsis y puede obtenerse por pruebas de hemoglobina en sangre.

Para los datos recolectados, se toman tres variables que determinan la incidencia de la sepsis en los pacientes, el fosfato alcalino, exceso de base, lactato en sangre, con esto se hacen los histogramas y la correlación que muestra como la presencia de una o dos de ellas tiene consecuencias en el desarrollo de la sepsis.

La Figura 5 presenta los datos de lactato con relación a pacientes que desarrollaron sepsis, el sesgo a la derecha indica que los valores son atípicos, presenta asimetría positiva (o a la derecha) ya que la cola o media es más larga que la derecha, esto nos dice que la diferencia entre los valores de lactato para los pacientes con sepsis y pacientes sanos es grande.

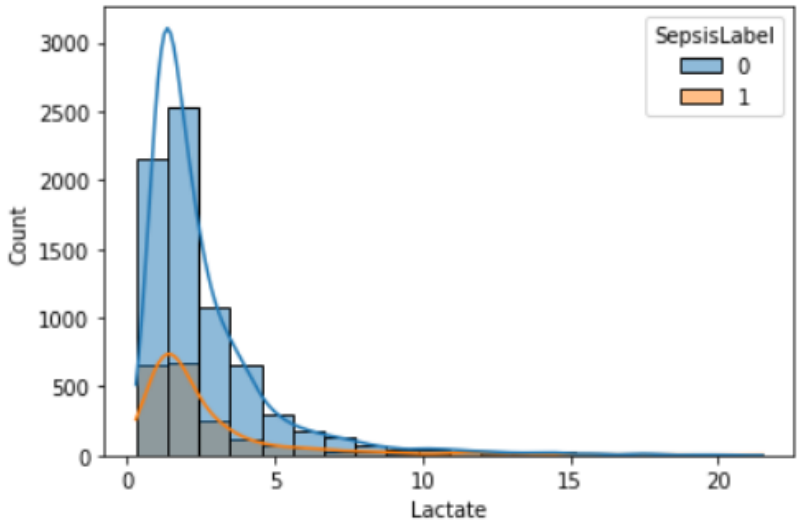

Figura 5: *Histograma de lactato Vs pacientes con sepsis.*

En comparación al resultado anterior, el histograma presentado para el exceso de bases en sangre los datos son simétricos y presenta una distribución normal entre el promedio de los valores de los pacientes con sepsis, de esto se deduce que es uno de los marcadores para el

pronóstico de la sepsis, indica la deficiencia de valores de oxígeno en sangre cumpliendo con el cálculo del SIRS.

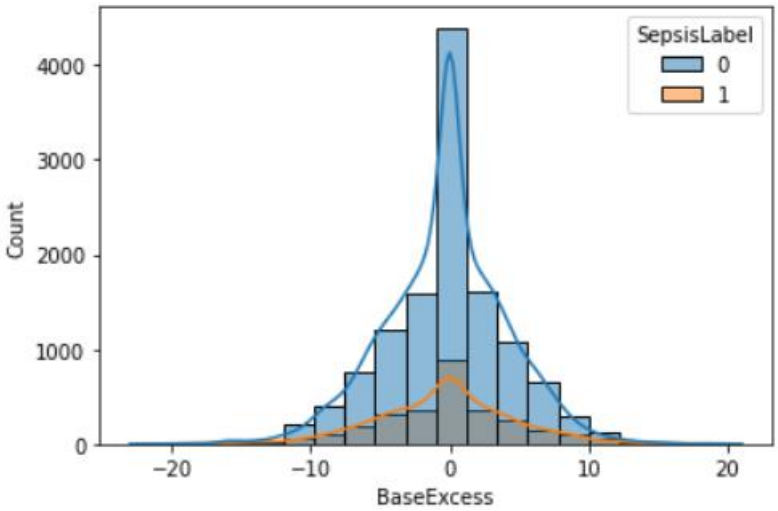

Figura 6: *Histograma de exceso de bases en sangre Vs pacientes con sepsis.*

Para la Figura 7 se presenta el tercer marcador que implica los fosfatos alcalinos en sangre, el histograma presenta sesgo a la derecha con media significativa, esto implica daño en el higado, lo que desencadena falla sistemica y disminucion del recuento de globulos blancos en sangre, un factor fundamental para el diagnostico de sepsis.

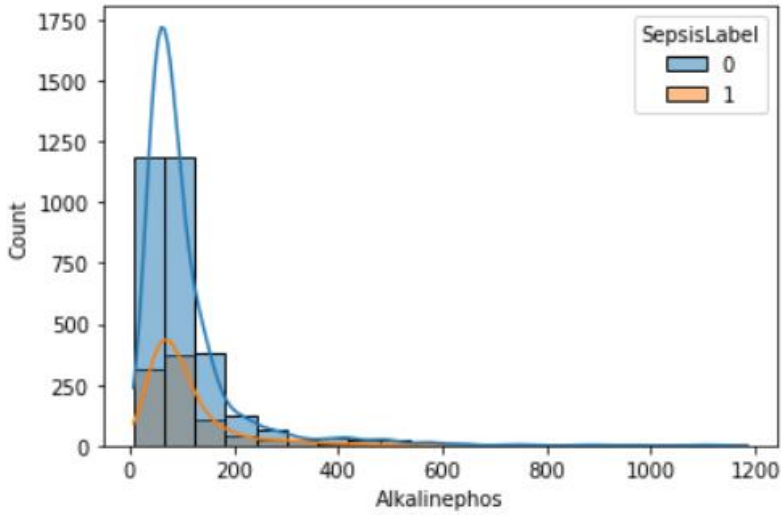

Figura 7: *Histograma de fosfato alcalinos Vs pacientes con sepsis.*

En el cálculo de la correlación presentamos los gráficos de la Figura 8 y Figura 9 tienen variaciones entre 3.5 y -0.5, presentan correlación entre los tres marcadores estudiados para el cálculo del SIRS y por ende diagnóstico temprano de la sepsis.

En este caso, los marcadores presentados agrupan variables fisiológicas importantes que presentaba el set de datos estudiados. El lactato y la base estándar medidos al ingreso a la UCI son de utilidad pronóstica en los pacientes críticamente enfermos, pues sus niveles séricos predicen mortalidad a través de la puntuación en sí misma [7]. Los resultados obtenidos muestran una mayor correlación entre el aumento de lactato y el exceso de bases en sangre, esto debido a que el exceso de bases en sangre afecta de manera directa el metabolismo y el aumento de lactato produce acidosis láctica en los tejidos. El exceso de

bases en sangre supera el déficit que ocasiona los otros dos marcadores, pero ayuda al progreso de ambos.

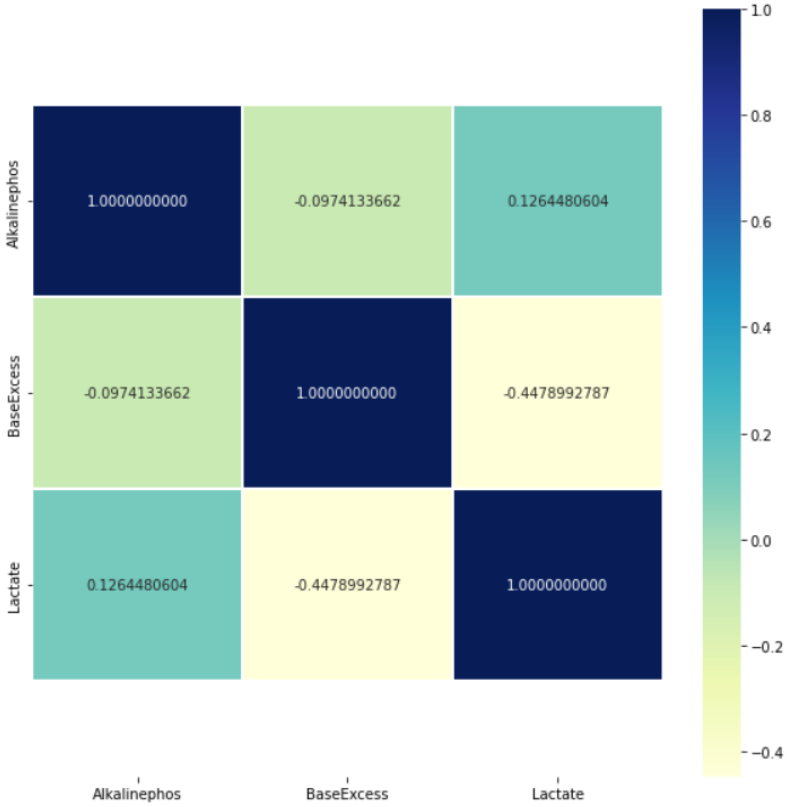

Figura 8: *Correlación de los marcadores 1.*

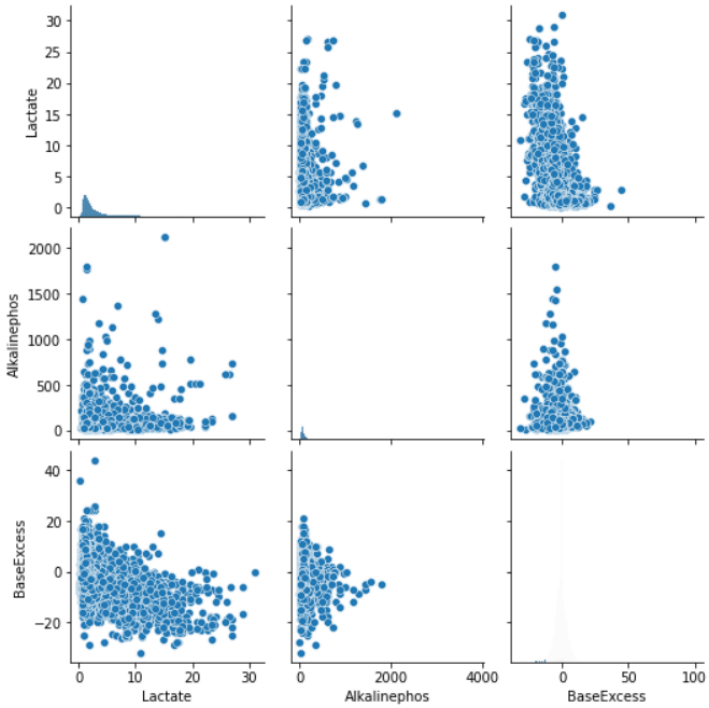

Figura 9: *Correlación de los marcadores 2.*

4. Retos y consideraciones de despliegue

Sabemos que la inteligencia artificial es un conglomerado de conceptos y tecnologías, que presentan un gran potencial en el desarrollo de la cuarta revolución industrial, pero esta se encuentra limitada por ciertos factores tanto técnicos como humanos.

En el desarrollo de nuestro proyecto, se presentaron limitantes como la falta de recursos necesarios para ejecutar una buena predicción al momento de procesar grandes cantidades de datos, pues esto va de la mano con las capacidades del hardware.

La clasificación de las variables importantes se vuelve un proceso fundamental en el procesamiento de los datos, ya que al tener una cantidad de datos grande la perdida de información se vuelve inevitable y esto interfiere con el resultado de la predicción o cualquier otra función que pretenda realizar el algoritmo.

No es conveniente recargar al algoritmo y generar mayor gasto computacional por tener en cuenta variables que no son tan significativas a la hora de diagnosticar una enfermedad; es por ello que, además de tener destrezas para el desarrollo de algoritmos, es conveniente conocer las características de la enfermedad para obtener resultados óptimos.

5. Conclusiones

- Para realizar predicciones eficientes y precisas, es de suma importancia tener datos de alta calidad que sean bastante representativos, esto se logra haciendo un preprocesamiento de los datos y estableciendo divisiones de datos de entrenamiento y datos de prueba, con el objetivo de mejorar el desarrollo y la estabilidad de los algoritmos de aprendizaje automático.
- El aprendizaje automático se encarga del desarrollo de algoritmos que pueden aprenderse los datos y gracias a eso hacer predicciones sobre estos mismos, lo cual es de mucha ayuda en diferentes situaciones como es el caso de la predicción temprana de la sepsis.
- La Inteligencia Artificial tiene el potencial de cambiar la medicina y la atención médica y puede proporcionar muchos beneficios para la detección temprana de la sepsis, al considerar cualquier riesgo que pueda tener un individuo de desarrollar la infección.

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

[4] ¿Qué es un histograma? (s/f). Material Didáctico - Superprof. Recuperado el 12 de noviembre de 2022, de https://www.superprof.es/apuntes/escolar/matematicas/estadistica/descriptiva/histograma.html

[5] IBM Documentation. (2021, marzo 5). Ibm.com. https://www.ibm.com/docs/es/odm/8.5.1?topic=charts-defining-bar-chart

[6] Ferrero, R. (s/f). ¿Qué es la correlación estadística y cómo interpretarla? Máxima Formación. Recuperado el 12 de noviembre de 2022, de https://www.maximaformacion.es/blog-dat/que-es-la-correlacion-estadistica-y-como-interpretarla/

[7] García Gómez, G., Salvador, J., Díaz, S., Gabriela, K., Moguel, P., Zepeda, E. M., Antonio Martínez Rodríguez, E., Verónica, M., Sánchez, C., Ganador, D., Premio, A. «., & Shapiro, M. (s/f). Medigraphic.com. Recuperado el 12 de noviembre de 2022, de https://www.medigraphic.com/pdfs/medcri/ti-2019/ti196b.pdf

[8] Cuervo, A., Correa, J., Garcés, D., Ascuntar, J., León, A., & Jaimes, F. A. (2016). Development and validation of a predictive model for bacteremia in patients hospitalized by the emergency department with suspected infection. Revista chilena de infectología: órgano oficial de la Sociedad Chilena de Infectología, 33(2), 150–158. https://doi.org/10.4067/S0716-10182016000200004