PREDICTING SEPSIS AMONG ICU ADMISSIONS USING MIMIC-III DATA

Patryk Szpetnar, Lloyd Linton Jones, Ihona María Correa de Cabo - 2024

***Abstract* —** Sepsis, a life-threatening reaction to infection, poses a formidable challenge in contemporary healthcare due to its common occurrence and high mortality rates. This research focuses on developing a predictive model using the Medical Information Mart for Intensive Care (MIMIC-III) dataset, comprising de-identified health-related data from over forty thousand ICU patients. Leveraging this rich dataset, the project employs machine learning to predict sepsis among ICU admissions, aiming to enhance diagnostic accuracy and timeliness. The model not only contributes to improving patient outcomes but also provides valuable insights into sepsis risk factors and patterns. Extraction of the dataset was performed using SQL, and subsequent data analysis and modeling utilized Python for its simplicity.

The exploratory data analysis revealed key variables indicative of sepsis, including vital signs and blood measures amongst others. Statistical summaries, gender-based sepsis distributions and numerical variable distributions were explored to understand the dataset's characteristics. The project underscores the potential of data-driven approaches in healthcare analytics, showcasing its significance in addressing critical medical challenges.

***Non-scientific abstract* —** This project tackles sepsis, a life-threatening infection reaction, using data from over forty thousand ICU patients. We're using cool tech like machine learning to predict sepsis early. It's like a high-tech health detective helping doctors save lives faster. Think of it as giving superpowers to medical data!

Keywords: *Sepsis, hospital mortality, MIMIC, machine learning, predictive medicine.*

I. INTRODUCTION

Sepsis is a life-threatening reaction to an infection and one of the most critical challenges in contemporary healthcare. It happens when the immune system overreacts to an infection and starts damaging the body's own tissues and organs. Bacterial infections cause most cases of sepsis, although it can also be a result of other infections, including viral and fungal infections. Most people who develop sepsis have at least one underlying medical condition and a weakened immune system. [1]

Unfortunately, this condition is both common and deadly, affecting millions of individuals worldwide each year and accounting for a significant proportion of hospital mortality rates. [2]
The complexity of sepsis, characterized by its rapid progression and varied clinical presentation, makes early detection and intervention crucial for improving patient outcomes. [3]

The significance of this project lies in the development of a predictive model using the Medical Information Mart for Intensive Care (MIMIC-III) dataset. MIMIC-III is a comprehensive and publicly available database that includes a vast array of de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center. By leveraging such a rich dataset, this project aims to develop a Machine Learning model to predict whether patients admitted to the ICU have sepsis or not.
The model's objective is to increase the accuracy and timeliness of sepsis diagnosis as well as to provide insights into the risk factors and patterns associated with the condition.

Moreover, the data fed to the model was extracted from the database using SQL, whereas the data analysis and the modeling was performed using Python due to its simplicity.

This project consists of two essential parts: the first involves data exploration, while the second focuses on training machine learning models. The initial phase is dedicated to understanding and analyzing information, and the latter utilizes advanced machine learning techniques to develop predictive models. The project contributes to the broader field of healthcare analytics, demonstrating the potential of data-driven approaches in tackling critical medical challenges.

II. METHODOLOGY

1. *Querying with SQL and python*

The data was obtained from the MIMIC-III database provided by the URV. This database was queried using SQL with python to perform intermediate steps in building the dataset. The entire query process was divided into two stages. The first stage is dedicated to finding subjects who have sepsis and have been admitted to the ICU, while the second stage finds subjects that have been admitted to the ICU but have been diagnosed with diseases other than sepsis. Both stages are broken down into smaller parts in order to remain within the query time limit of 10 minutes. Each of these parts are similar for both stages with some minor differences. The overall process is described below.

1. From the D_ICD_DIAGNOSES table, diseases are selected where the title includes 'sepsis' for the sepsis cohort, while for the non-sepsis cohort the title excludes 'sepsis' and 'septicemia'.
2. Admission diagnoses that match the diseases found in step 1 are paired with the relevant admission IDs collected from the DIAGNOSES_ICD table. In the case of the sepsis cohort these results are limited by the diagnosis sequence number (SEQ_NUM <= 10) and in the case of the non-sepsis cohort these were limited to only primary diagnoses (SEQ_NUM = 1). Both of these limitations were necessary to remain with query times during later steps.
3. Only the admissions from step 2 that have corresponding ICU stay/s are kept. This is achieved by checking if the admission has any matching ICUSTAY_IDs in the ICUSTAYs table. This is done for both cohorts.
4. For each data entry, the relevant patient information from the PATIENTS table is collected. This includes the gender and age of the patient. The age is calculated as the difference in time between the date of birth and the recorded ICU INTIME. This calculation accounts for the adjusted patient dates of birth related to patients over 89 years-old as specified in the MIMIC documentation.
5. The corresponding CHARTEVENTS data was joined onto the data entry of each unique ICU stay. The data fetched from the CHARTEVENTS table was limited to routine vital signs measurements which included heart rate, blood pressure, temperature, and respiratory rate.
6. The corresponding LABEVENTS data was joined onto the data entry of each unique ICU stay. The data fetched from the LABEVENTS table was limited to measurements related to blood counts, lactate, glucose, creatinine, bilirubin, just to name a few.
7. The data entries for each cohort were grouped by admission, using HADM_ID, and averaged per measurement. This is an essential step in ensuring that the amount of data is not too large and that rows in the dataset would not repeat similar information.

A further limitation was placed on step 3, where a set of distinct patients were sudo-randomly sampled from the population of patients that fulfilled the criteria in previous steps, however this was only implemented for the non-sepsis patients. The number of admissions in this cohort were far too many to handle within the query time limit. For example, there would be over 12 million applicable entries in just the LABEVENTS table that would be called if no sampling was utilized. This number would be even larger for the CHARTEVENTS table since it contains significantly more rows than LABEVENTS.

2. *Exploratory Data Analysis*

Exploratory Data Analysis can help interpreting data in order to extract meaningful insights, identify patterns, and make informed decisions.

The dataset contains a total of 4420 rows (different HADM_ID) and 68 columns. Since there are a wide number of columns, it has been decided to visualize and inspect the ones that were thought to be more representative for sepsis from a medical perspective. These variables are:

- Vital signs (Blood pressure, Heart rate, Respiratory rate and Temperature): Sepsis often presents abnormalities in vital signs, such as increased heart rate, low blood pressure, increased respiratory rate, and fever or hypothermia.
- White Blood Cells count (WBC): Both very high and very low WBC counts can be indicative of sepsis, as it is a response to infection.
- Markers of organ function (Creatinine, Bilirubin): Elevated levels might indicate organ dysfunction, which can be a consequence of severe sepsis or septic shock.
- Glucose: Hyperglycemia can occur in sepsis due to stress response.
- Lactate: High levels of lactate can indicate tissue hypoxia and is often associated with worse outcomes in sepsis.
- Blood gas measures (pCO2): Abnormalities in arterial blood gasses can occur in sepsis.
- Other blood and urine tests: Blood culture tests aid in identifying the presence of bacterial infections. Other blood tests can show liver or kidney dysfunction, blood clotting issues and lowered oxygen levels. Urine samples can indicate urinary tract infections and additionally a reduction in urine volume is often symptomatic of sepsis.

The following table (*Table 1*) represents a summary of the dataset statistics:

	count	mean	std	min	25%	50%	75%	max
AGE	4420.0	65.471267	16.712993	15.000000	54.000000	67.000000	79.000000	100.000000
Heart Rate	4414.0	86.526021	14.535802	35.739130	76.251953	85.866900	96.355535	147.380952
Respiratory Rate	4413.0	19.704616	3.908950	7.000000	17.000000	19.310345	22.070175	61.849315
Arterial Blood Pressure mean	2174.0	79.261471	21.059357	-16.000000	71.981308	77.650411	84.672195	353.666667
Temperature Celsius	492.0	38.197893	8.910640	-2.000000	36.449823	36.894444	37.535714	99.000000
WBC Count	58.0	8.870402	5.870440	2.000000	5.275000	7.250000	10.587500	37.000000
Creatinine	4418.0	1.522348	3.392895	0.100000	0.763636	1.016667	1.600000	208.550000
Bilirubin, Total	3438.0	1.904282	4.290460	0.100000	0.400000	0.650000	1.385625	49.754545
Glucose	4419.0	134.015806	44.489027	60.000000	108.000000	122.833333	145.275862	773.333333
Lactate	3765.0	2.242519	1.595794	0.300000	1.350000	1.837500	2.587500	19.050000
pCO2	3140.0	41.416932	9.624603	12.666667	35.875179	40.049043	45.000000	104.000000

Table 1.- Summary statistics (own source)

A valuable categorical variable is the gender (Male as M and Female as F). In the following bar plot (*Figure 1*) it is possible to observe the target variable to predict (sepsis) with respect to the different genders. A sepsis value of 0 means that there is no sepsis and a value of 1 means that there is sepsis. It is possible to see that there are more males with sepsis in the dataset.

Figure 1.- Sepsis divided by gender (own source)

In *Figure 2* it is seen that the dataset is a bit unbalanced (there are a bit more patients without sepsis than with sepsis).

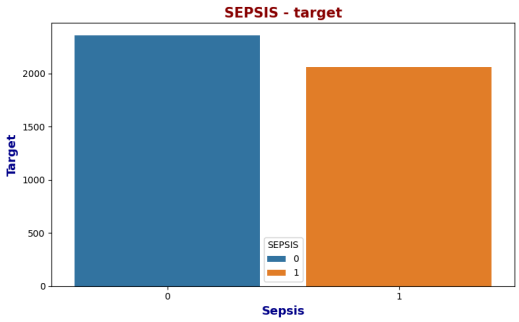


Figure 2.- Sepsis division (own source)

Plotting the distribution of numerical variables is also important to identify outliers and understand the spread and variability of the data. Moreover, different machine learning algorithms may make different assumptions about the data distribution. Therefore, knowing the actual distribution of the variables can guide the modeling techniques.

Following *Figure 3, Figure 4* and *Figure 5*, it is possible to see that the heart rate, the respiratory rate and the pCO2 (partial pressure of carbon dioxide) follow a normal distribution, although the respiratory rate and the pCO2 are a bit skewed.

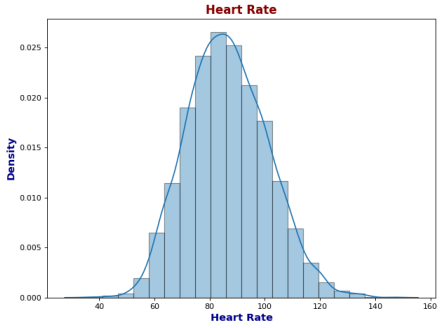


Figure 3.- Heart rate distribution (own source)

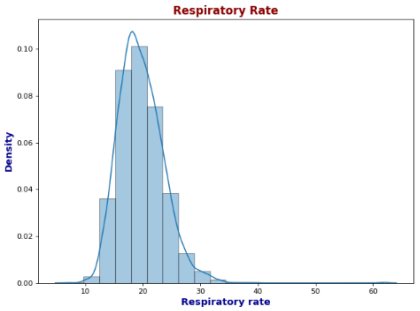


Figure 4.- Respiratory rate distribution (own source)

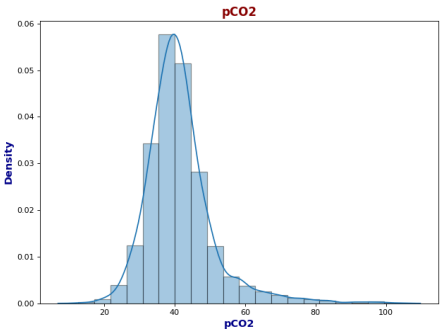


Figure 5.- pCO2 (own source)

The highest correlations of the features with respect to the variable to predict (*Figure 6*) can be interesting in order to determine which variables are more important and contribute the most to sepsis. As we can see there are always small correlations:

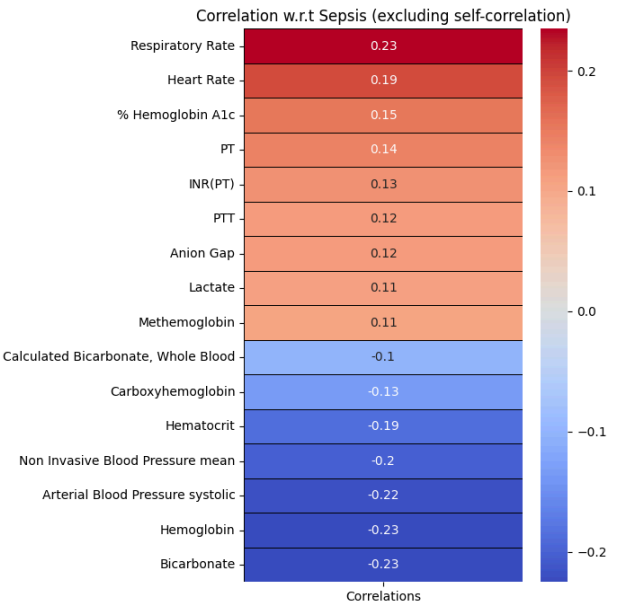


Figure 6.- Correlation with respect to sepsis (own source)

Finally, *Figure 7* represents the mean values of numeric variables from sepsis and non-sepsis patients, respectively. These values can help to compare differences between patients with and without sepsis.

It is possible to see that the mean of WBC Count, Creatinine, Bilirubin, Heart Rate, Respiratory Rate,

Lactate, and Glucose is a bit higher in Sepsis patients. Increased WBC count can reflect the immune body response to sepsis, whereas increased creatinine and bilirubin may indicate kidney and liver dysfunction, respectively. An increased heart rate can be a compensatory mechanism in response to systemic infection and inflammation, as the body tries to maintain adequate blood flow and oxygen delivery to tissues. On the other hand, an elevated respiratory rate can be a response to increased metabolic demands during sepsis. As mentioned before, high lactate levels are indicative of tissue hypoxia and impaired cellular metabolism, often seen in severe sepsis. Finally, hyperglycemia in sepsis can result from stress-induced insulin resistance and increased hepatic glucose production.

However, it is possible to see that the Arterial Blood Pressure mean and the pCO2 levels are a bit higher in non-sepsis patients.
Higher arterial blood pressure in non-sepsis patients might reflect the absence of sepsis-induced vasodilation. In sepsis, blood vessels can become dilated and permeable, leading to a drop in blood pressure. Additionally, higher levels of pCO2 in non-sepsis patients could indicate better maintained respiratory function compared to sepsis patients.

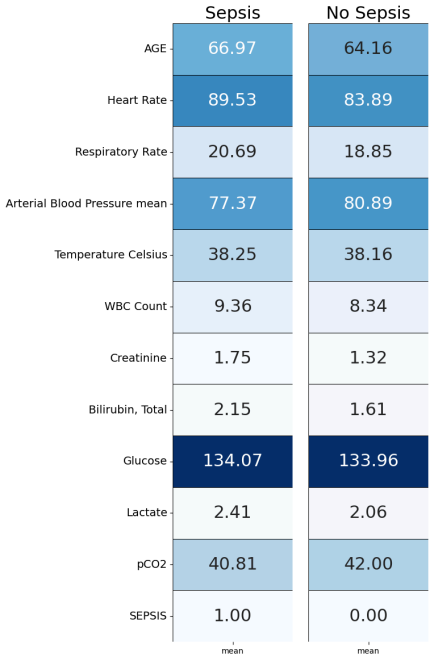


Figure 7.- Mean values for sepsis and no sepsis (own source)

3. *Machine learning modeling and prediction*

After exploring and visualizing the dataset, the data was processed and prepared before the modeling.

Firstly, the gender column was numerically encoded using label encoding to ensure that the characters representing male and female (M and F) were compatible with the machine learning algorithms.

Next, the data was divided into features (X) and labels (y) and subsequently split into training and test sets. The split followed an 80-20 ratio, allocating 80% of the data for training the models and reserving the remaining 20% for assessing their performance. This approach ensures that the models are trained on a substantial portion of the data while providing an independent set for evaluation, helping gauge their generalization capabilities and allowing fair evaluation of model performance.

Following the splitting of data were two feature engineering stages. The first was imputing the NANs by implementing the strategy of replacing them with the mean of each feature. Both the strategies of using the mean and the median were attempted, however the mean yielded better results. The second was normalizing each feature, which was done using z-score normalization. With the feature engineering stages completed, the dataset is prepared.

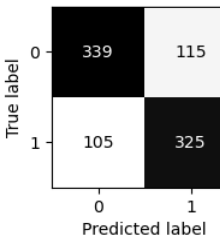
In the machine learning investigation, three distinct algorithms, namely XGBClassifier, neural networks and k-nearest neighbors, were employed to tackle the classification task focused on predicting the presence of sepsis. The choice of XGBClassifier, a gradient boosting algorithm, was motivated by its proven effectiveness in various domains, robustness to overfitting, and capability to capture complex relationships within the data [4]. The selection of a neural network was driven by its capacity to learn intricate patterns and relationships within the medical dataset. The adaptability and capability of neural networks to handle complex data structures influenced this choice [5]. Finally, K-nearest neighbors (KNN) is a simpler and intuitive algorithm suitable for classification tasks where instances with similar features tend to belong to the same class [6]. The combination of these algorithms provides a diverse set of approaches, potentially capturing different aspects of the underlying patterns in the data.

Grid search was utilized to fine-tune the hyperparameters of the XGBoost and KNN algorithms. Hyperparameters are critical settings that influence the learning process of the algorithm. The grid search involved systematically testing different combinations of hyperparameters to identify the configuration that maximizes the model's performance. For example, in the XGBoost algorithm these parameters included booster type, learning rate (ETA), gamma, and tree method. This automated exploration of hyperparameter space is crucial for optimizing the algorithm and enhancing its predictive capabilities.

The performance of the machine learning algorithms was assessed using several metrics. The main metric and the one used to optimize parameters was the F1-score. Other key metrics included accuracy, precision, and recall and area under the ROC curve (AUC-ROC).

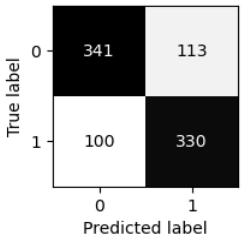
III. RESULTS

XGBClassifier
without parameter optimization



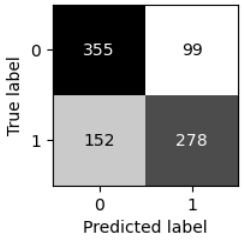
Accuracy	0.7590
Precision	0.7449
Recall	0.7674
F1 Score	0.7560
AUC-ROC	0.7593

XGBClassifier
including parameter optimization
using GridSearchCV



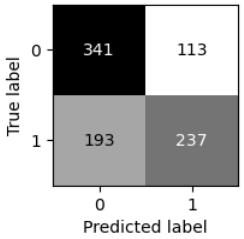
Accuracy	0.7511
Precision	0.7386
Recall	0.7558
F1 Score	0.7471
AUC-ROC	0.7513

Neural network



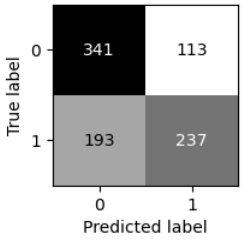
Accuracy	0.7161
Precision	0.7374
Recall	0.6465
F1 Score	0.6890
AUC-ROC	0.7142

KNN Classifier
without parameter optimization



Accuracy	0.6538
Precision	0.6771
Recall	0.5512
F1 Score	0.6077
AUC-ROC	0.6511

KNN Classifier
including parameter optimization
using GridSearchCV



Accuracy	0.6550
Precision	0.6781
Recall	0.5535
F1 Score	0.6095
AUC-ROC	0.6523

IV. DISCUSSION

Regarding the results obtained from the data analysis, it has been observed that patients with sepsis exhibit significantly higher heart rate, respiratory rate, creatinine, bilirubin, and lactate levels compared to the group of patients without sepsis. This is consistent with various studies that report the similar findings that note this data is important in knowing the severity of the patient's sepsis. [7,8,9]

Looking at the correlation table, we can see how the respiratory rate, heart rate, bicarbonate, hemoglobin and arterial blood pressure are more related in our study to sepsis. Although the correlation is very small.

The results obtained from classifying patients into sepsis and non-sepsis categories were not very promising. The best result achieved was an F1 score of 0.756 with the XGBClassifier algorithm with optimized parameters. The neural network had an F1 score of 0.689, and finally, KNN had only an F1 score of 0.6095.

It's essential to consider that individuals without sepsis may have various illnesses that led to them being admitted to the ICU. Many diseases could cause similar symptoms and produce laboratory test results quite comparable to those of sepsis.

Considering the serious and even life-threatening nature of having sepsis, the prediction results that were obtained are not sufficient. In order to enhance the results, it would be beneficial to explore training models based on database query modifications for example experiment with different sample sizes, add new relevant columns and remove potentially unhelpful ones. These modifications would enable new datasets to be created that might yield better results. Also, further optimization of model parameters and different machine learning algorithms could be considered. Special attention should be given to the neural network, which wasn't optimized in this study due to its high computational cost.

V. REFERENCES

[1] Gyawali, Bishal, Karan Ramakrishna, and Amit S. Dhamoon. "Sepsis: The evolution in definition, pathophysiology, and management." SAGE open medicine 7 (2019): 2050312119835043.

[2] Wang, Henry E., et al. "National variation in United States sepsis mortality: a descriptive study." International journal of health geographics 9 (2010): 1-9.

[3] Beckmann, Nadine, et al. "Staging and personalized intervention for infection and sepsis." Surgical infections 21.9 (2020): 732-744.

[4] Kumar, S. Jagadish, et al. "Melanoma Classification using XGB Classifier and EfficientNet." 2021 International Conference on Intelligent Technologies (CONIT). IEEE, 2021.

[5] Tsagkaris, Kostas, Apostolos Katidiotis, and Panagiotis Demestichas. "Neural network-based learning schemes for cognitive radio systems." Computer communications 31.14 (2008): 3394-3404.

[6] Jivani, Anjali Ganesh, et al. "The adept K-nearest neighbour algorithm-an optimization to the conventional K-nearest neighbour algorithm." Transactions on Machine Learning and Artificial Intelligence 4.1 (2016): 52.

[7] Patel, Jayshil J., et al. "The association of serum bilirubin levels on the outcomes of severe sepsis." Journal of intensive care medicine 30.1 (2015): 23-29.

[8] Xiao, Yufei, et al. "Evaluation of qSOFA score, and conjugated bilirubin and creatinine levels for predicting 28-day mortality in patients with sepsis." Experimental and Therapeutic Medicine 24.1 (2022): 1-9.

[9] Mikkelsen, Mark E., et al. "Serum lactate is associated with mortality in severe sepsis independent of organ failure and shock." Critical care medicine 37.5 (2009): 1670-1677.