Early Prediction of Sepsis from Clinical Data

Shashwat Nayak, Priya Singh

The University of Texas at Dallas
{Priya.Singh2, Shashwat.Nayak} @utdallas.edu

Abstract

Sepsis is a medical condition that arises due to dysregulated response of the body's immune system to an infection that can cause widespread inflammation, multi organ failure and eventually, lead to death. Prompt recognition and treatment of sepsis are very crucial for improving the chances of recovery and reducing the risk of complications. Machine learning (ML) has shown great potential in medical diagnosis due to its ability to learn from large datasets and identify patterns that may not be easily detectable by humans. Our main goal is to use machine learning algorithms for early prediction of sepsis and conduct a performance analysis of these algorithms to accurately predict the classification. In this study, the prediction of sepsis has been conducted using supervised learning algorithms, including Decision Trees and Logistic Regression. The analysis involved testing both self-implemented and scikit-learn classifiers. The performance of each algorithm was analyzed to determine which model achieved the highest accuracy in predicting sepsis.

Introduction

Severe infection that sets up a powerful immunological reaction in the body is what leads to sepsis. Various pathogens, such as bacterial, viral, or fungal ones, may be the cause of the infection. Pneumonia, urinary tract infections, skin infections, and digestive tract infections are frequently the causes of sepsis. The immune system releases some chemicals to combat the pathogen when it recognizes an infection. In sepsis, these chemicals set off a chain of events that result in widespread inflammation, organ deterioration, and reduced blood flow in sepsis. In some extreme circumstances, it can result in the failure of several organs and ultimately lead to septic shock, a disease that poses a threat to life.

Sepsis management and avoiding its sequelae depend heavily on early identification and treatment. Diagnosing sepsis can be a challenging task as the symptoms can differ greatly from one individual to another. Additionally, the signs of sepsis can be vague and non-specific, making it challenging to identify the condition accurately. Additionally, the signs of sepsis can be vague and non-specific, like an increase in heartbeat rate, high/low body temperature, rapid breathing/shortness of breath etc., making it challenging to identify the condition accurately.

The National Confidential Enquiry into Patient Outcome and Death Report of 2015 exposed a concerning finding about sepsis. The report highlighted that there were significant delays in identifying sepsis in over one-third of cases (36%). This paper presents the use of machine learning algorithms in medical diagnosis that has rapidly gained attention in recent years, and it has shown promising results in improving the accuracy and efficiency of disease detection.

The dataset consists of various vital, laboratory captured and demographic attributes hourly. The laboratory captured attributes are based on advice of medical practitioner and are centric towards actual actual/potential patients. As a result, the dataset is extremely unbalanced and has multiple instances of missing attributes.

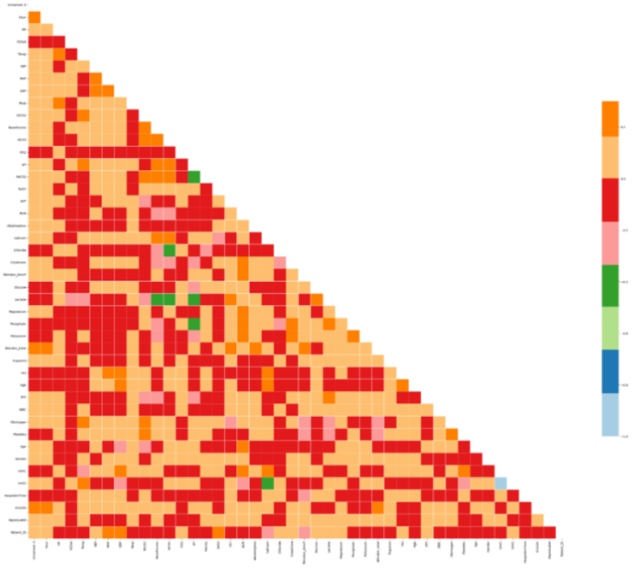
Objective

The purpose of this study is to examine and analyze the performance of some supervised ML models in predicting sepsis. By analyzing the accuracy and reliability of different models, it can provide insights into the effectiveness of these models in sepsis prediction. By assessing the performance of these models, this study seeks to contribute to the ongoing research on the use of ML in medical diagnosis and improve the early detection of sepsis.

Dataset and Features

We have obtained this data from PhysioNet/Computing in Cardiology Challenge 2019 wherein the data of ICU patients from two hospital systems has been available publicly. The dataset has 40 (+1 including label) features consisting of 8 vital signs, 26 laboratory values, 6 demographic data and finally, the outcome Sepsis label.

Vital Signs	HR, O2Sat, Temp, SBP, MAP, DBP, Resp, ETCO2
Lab values	BaseExcess, HCO3, FiO2, pH, PaCo2, SaO2, AST , BUN Alkanethiols, CalciumChloride, Creatinine, Bilirubin, Glucose, Lactate, Magnesium, Phosphate, Potassium , Bilirubin_total, Troponin ,Hct ,Hgb ,PTT ,WBC ,Fibrinogen, Platelets
Demographics	Age,Gender,Unit1,Unit2, HospAdmTime , ICULOS
Target	Sepsis label



We split the above dataset into 2 parts for 2 hospitals each. Further we split it into training and test sets using 80:20 split. While performing the split, we observed that the ratio of sepsis to non-sepsis is extremely skewed. So, we have performed under sampling in ratio of 1:2 to get better results.
Over 41 attributes, we analyzed several distributions (standard, mean, gaussian etc..) and percentages of NAN attributes. We removed all the attributes with more 25% of missing values to reduce noise. For the rest, we also performed imputation to handle rest of the missing values to reduce the skewness of dataset.

Out of 41 attributes only 19 attributes are well defined (post processing) for the model building. We split this refined dataset as mentioned above.

Methodology

A medical expert's view of solving above problem involves rule-based and scoring method to predict the onset of sepsis. For our convenience, we have implemented ID3 Decision tree classifier and Logistic Regression (with Regularization). We have also used Scikit-Learn to validate our implemented algorithm and, we used this library to figure which algorithm is possibly best to categorize and predict the correct label. We have used Decision Tree, Logistic Regression, Naïve Bayes and Xtreme Gradient Boosting for our analysis.

Performance and Analysis

We have implemented our own version of Decision Tree (ID3) and Logistic Regression (with L2 Regularization).

The table below gives a brief about how much our algorithm stands w.r.t to various metrics.

Algorithm	Accuracy	Precision	Recall	F1	Auc-Roc
ID3 (dpt-5)	0.66	0.59	0.01	0.02	0.503
ID3 (dpt-10)	0.67	0.62	0.06	0.11	0.52
LR (without L2)	0.5	0.38	0.85	0.53	0.59
LR (with L2)	0.68	1	0.036	0.07	0.5

For comparison w.r.t scikit, the table below gives brief about how much our algorithm fairs against scikit implemented algorithm.

Algorithm	Accuracy	Precision	Recall	F1	Auc-Roc

Decision tree	0.89	0.8	0.9	0.85	0.89
LR	0.74	0.73	0.37	0.49	0.65
Neural Network	0.77	0.7	0.57	0.63	0.72
Naïve Bayes	0.74	0.7	0.42	0.53	0.66
XGB	0.85	0.82	0.72	0.77	0.82

Analysis

On comparing our implementation with scikit algorithm, we can observe marginal difference between the two.

For decision tree, the accuracy differs around 20%. This could be possibly due to scikit's different algorithm for tree building (CART in scikit whereas we have used ID3).

Also, we have observed an increase in accuracy when we increased the depth of Decision tree in our own implementation, so, it might be possible that it might perform better with a greater depth, but we need to be careful since it can also overfit.

For Logistic Regression, without the Regularization, we observe the poorest accuracy of 50%. However, on application of L2 regularization, we observe the jump to 68%which is comparable to scikit's LR implementation which we have tested it with L2 settings. (Only 6% difference).

For scikit based algorithms, we also observe accuracy in range of 70-80%. LR, Neural Network, and Naive Bayes models did not perform as well. The LR and Neural Network models had relatively low recall scores, indicating that they had a higher false negative rate. The Naive Bayes model had a relatively low F1 score, suggesting that it had a lower balance between precision and recall. This could be due to nature of our dataset(imbalanced) and the fact that relationship between features is non-linear.

However, tree-based algorithms XGB (weighted regression trees) and decision tree showed similar and better accuracy. This could be due the fact that they can capture non-linear relationships and interactions much better than other algorithms.

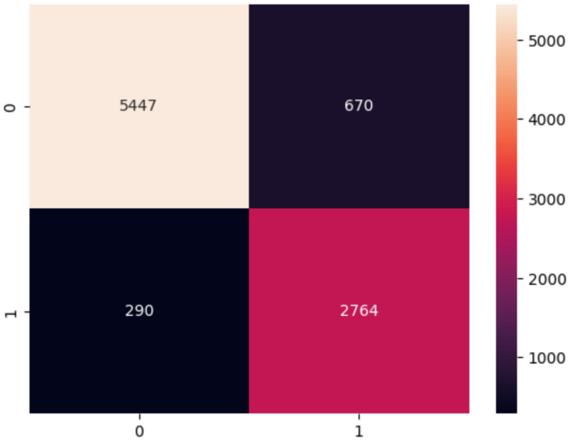
Conclusion

We have presented our own implementation of machine learning algorithms along with scikit algorithms to validate our result and provided brief analysis of where our algorithm fails to classify labels. We have extensively performed data engineering to shift the balance of data to provide results as accurate as possible. Based on what we have done, we would also like to understand the performance on application of machine learning algorithm on unseen datasets.

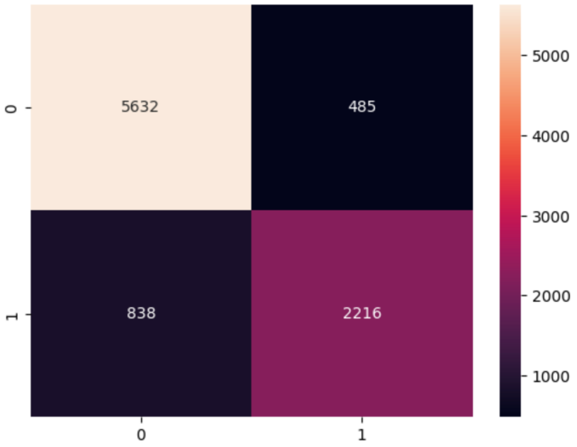
The diagrams below are confusion matrices of respected implementation and scikit algorithms.

Scikit-Learn Classifiers

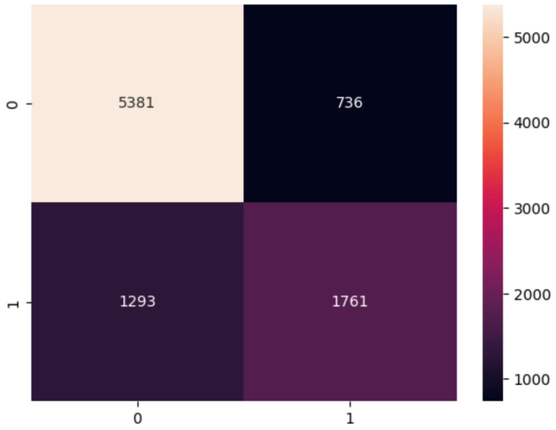
Confusion Matrix : Decision Matrix



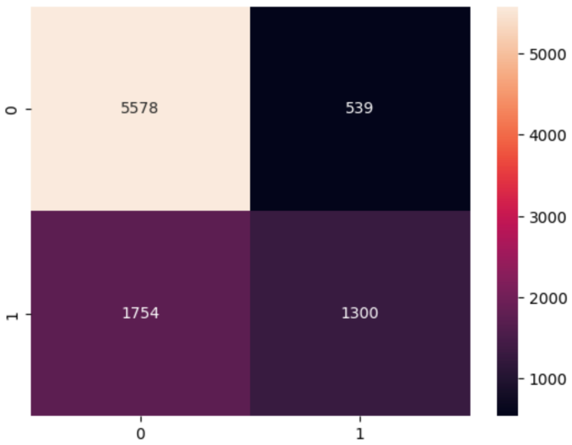
Confusion Matrix : XGBoost



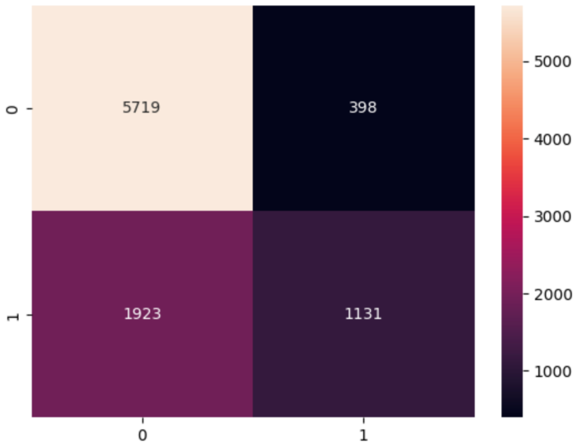
Confusion Matrix : Neural Network



Confusion Matrix : Naïve Bayes

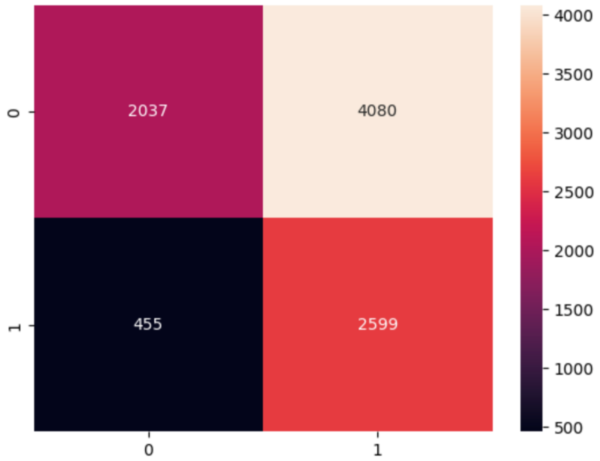


Confusion Matrix : Logistic Regression

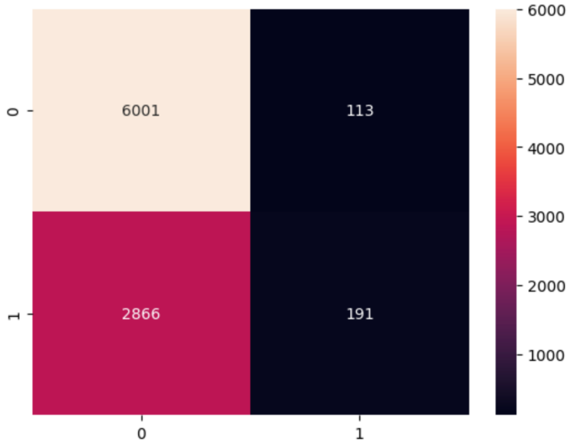


Our Implementation

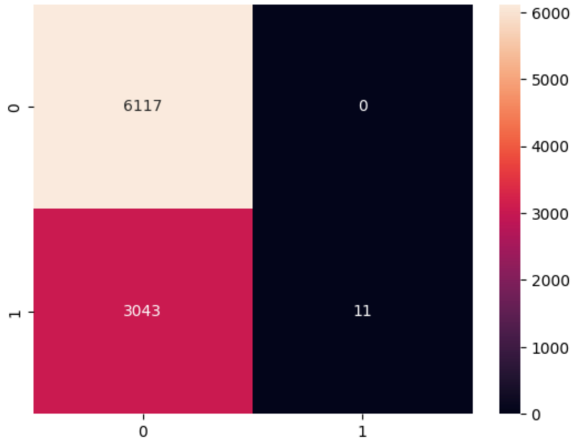
Confusion Matrix : Decision Tree (Depth 5)



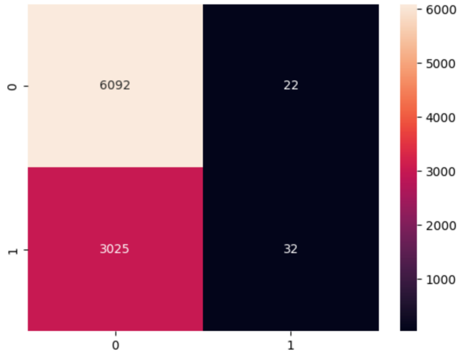
Confusion Matrix : Decision Tree (Depth 10)



Confusion Matrix : Logistics (L2 Regularization)



Confusion Matrix : Logistics (without L2 Regularization)



References

Rober et al.(2015) National Confidential Enquiry into Patient Outcome and Death. Just Say Sepsis! A review of the process of care received by patients with sepsis, p.7

Reyna, Matthew A. PhD1; Josef, Christopher S. MD1; Jeter, Russell PhD1; Shashikumar, Supreeth P. B.Tech2,3; Westover, M. Brandon MD, PhD4; Nemati, Shamim PhD1,3; Clifford, Gari D. DPhil1,2; Sharma, Ashish PhD1. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine 48(2):p 210-217, February 2020. | DOI: 10.1097/CCM.0000000000004145