Analysis and Early Prediction of Sepsis in Clinical Data

Project Report

CPSC 597

Submitted By:

Anushree Ankola

(893259101)



Supervised by:

Dr. Anand Panangadan

Department of Computer Science

College of Engineering and Computer Science

California State University Fullerton

**Table of Contents**

[1. Abstract 2](#_Toc27379511)

[2. Introduction 3](#_Toc27379512)

[2.1 Literature Review 3](#_Toc27379513)

[2.2 Purpose 3](#_Toc27379514)

[3. Objective 4](#_Toc27379515)

[4. Data 4](#_Toc27379516)

[5. Procedure and Analysis 8](#_Toc27379517)

[5.1 Data Cleaning – Preprocessing & handling missingness 8](#_Toc27379518)

[5.2 Feature Selection & Feature Engineering 9](#_Toc27379519)

[5.3 Handling Data Imbalance 11](#_Toc27379520)

[5.4 Choosing Evaluation Metric 12](#_Toc27379521)

[5.5 Training with ML algorithm – Gradient Tree Boosting 14](#_Toc27379522)

[6. Conclusion 15](#_Toc27379523)

[7. Future Scope 15](#_Toc27379524)

[References 15](#_Toc27379525)

# 1. Abstract

Sepsis is a potentially life-threatening condition that occurs as a response to body's response to infection. When the response for the body's response goes wrong, it might cause tissue damage, organ failure or can even cause death. According to a recent study in the US, nearly 1.8 million people are prone to sepsis and more than 280,000 people die from the condition each year. A famous nation's health protecting agency, CDC or Center for Disease Control & Prevention states that over one third of the people in U.S. hospitals die from Sepsis each year. Internationally, an estimated 30 million people have developed Sepsis and more than 6 million people die from it each year. Among them more than 4.2 million infants and children are prone to the condition. Hospitals shell out $24 billion (which is approximately 13% of U.S healthcare expenses) to treat Sepsis. Early detection of sepsis is very important for improving sepsis outcomes. Each hour of delayed treatment can increase mortality rate by 4-8%. Answering questions such as what the real cause of the condition is, is there a way to predict the condition early, way before it has been conceived during the clinical trials are of utmost importance in this challenge. Hence coming up with a good model that might be able to solve this issue becomes imminent and important. We aim to analyze, evaluate and develop an algorithm that hopes to answer some questions we face while handling to attempt the challenge.

# 2. Introduction

## 2.1 Literature Review

Currently the reliable and early detection of sepsis is quite complicated by nature and can lead to delays in treatment. Nonetheless there has been wide interest in the area and is particularly highlighted in two recent papers. The studies conducted shoes that an increase in the adjusted mortality of people contracted Sepsis who experienced delays in antibiotic therapy. The effect is more seen in-patient suffering from an advance version on sepsis called septic shock. In this case particularly, hourly delays in treatment lead to a 3.6 - 9.9% increase in mortality per hour.

Some professional care societies have proposed clinical criteria for identifying and treating sepsis. However, the core need for early and reliable recognition of Sepsis are yet to be met. Computational approached promise to improve the early detection of Sepsis. It involves applying machine learning techniques to clinical data with the end goal of making time specific predictions at least up to a day before the clinical recognition of Sepsis.

## 2.2 Purpose

The purpose of this project is to analyze, explore and identify the relationship between the features that correspond to a person having sepsis condition and attempt to build a classification algorithm in a open source early detection of sepsis dataset.

Algorithms for early detection of sepsis address a variety of problems and generally go om different direction as the data is tested in different patient cohorts with different clinical variables and a label arising from different clinical criteria from Sepsis. The algorithms seem to overfit on the shared databases and underperform on the test database, especially if they encoded data collection behaviors specific to given hospital system.

Another issue that we are solving in the project is the issue of using a correct evaluation metric as the usual accuracy metric might not necessarily reflect the clinical utility of sepsis detection and treatment. Another thing we are having to consider is that the complexity of the algorithm we come up with is nearly impossible to adequate al the reasons of a person conceiving Sepsis and is open to wide area of discussion among various researcher. Hence making the solution a start to the arena and scope of research open source and available as a starting point for anyone who would like to continue the work.

# 3. Objective

The goal of performing this project was to the development of an algorithm for early prediction of Sepsis using routinely available clinical data. Early prediction particularly is potentially a lifesaving, while late or missed predictions are potentially life-threatening, and false algorithms might consume hospital resources and erode trust un the algorithm itself. The dataset seems a bit complicated, and hence we will first handle missingness and imbalance in the set as well try to come up with newer features before implementing the algorithm.

The algorithm will be designed at various levels and incorporating the clinical data, automatically identifying a patient's risk of conceiving Sepsis and also make a positive or negative prediction of at least by six hours and no more than twelve hours before the onset time of Sepsis. To evaluate the algorithm, we have for the time being selected to plot the precision-recall curve with an F1 score that can assess the algorithm and penalize them for wrongly classified data points.

# 4. Data

The challenge data is obtained from two geographically distinct U.S hospital systems with two different EMR or Electronic Medical Record systems. The two hospitals are Beth Israel Deaconess Medical Center (hospital system A), Emory University Hospital (hospital system B). These data were collected over 10 years with the approval of appropriate Institutional Review Boards. Data were collected for 40,336 patients from both the hospitals, and each patient has their own data files. The data from both the hospitals were posted on the Physio net Challenge 2019 and were available for the public and for download.

The data consisted of a combination of hourly vital sign summaries, laboratory values, and patients’ demographics. To be more specific, the dataset consisted of 40 clinical variables where there were 8 vital sign variables, 26 laboratory values, and 6 static patient descriptions. These data included over 2.5 million hourly time windows and 15 million data points. As mentioned, the data is extracted from the EMR and underwent series of preprocessing steps prior to formal analysis and development. The patient features were condensed into hourly time window for simplified model development and testing. The patient data is labeled in accordance with the Sepsis-3 clinical criteria. The missing and erroneous data has been present in the dataset.

A close up of a piece of paper

Description automatically generated

The following is the Clinical data used for this project

1. HR Heart rate (beats per minute)
2. O2Sat Pulse oximetry (%)
3. Temp Temperature (deg C)
4. SBP Systolic BP (mm Hg)
5. MAP Mean arterial pressure (mm Hg)

6. DBP Diastolic BP (mm Hg)

7. Resp Respiration rate (breaths per minute)

8. EtCO2 End tidal carbon dioxide (mm Hg)

9. BaseExcess Excess bicarbonate (mmol/L)

10. HCO3 Bicarbonate (mmol/L)

11. FiO2 Fraction of inspired oxygen (%)

12. pH pH

13. PaCO2 Partial pressure of carbon dioxide from arterial blood (mm Hg)

14. SaO2 Oxygen saturation from arterial blood (%)

15. AST Aspartate transaminase (IU/L)

16. BUN Blood urea nitrogen (mg/dL)

17. Alkalinephos Alkaline phosphatase (IU/L)

18. Calcium Calcium (mg/dL)

19. Chloride Chloride (mmol/L)

20. Creatinine Creatinine (mg/dL)

21. Bilirubin direct Direct bilirubin (mg/dL)

22. Glucose Serum glucose (mg/dL)

23. Lactate Lactic acid (mg/dL)

24. Magnesium Magnesium (mmol/dL)

25. Phosphate Phosphate (mg/dL)

26. Potassium Potassiam (mmol/L)

27. Bilirubin total Total bilirubin (mg/dL)

28. TroponinI Troponin I (ng/mL)

29. Hct Hematocrit (%)

30. Hgb Hemoglobin (g/dL)

31. PTT Partial thromboplastin time (seconds)

32. WBC Leukocyte count (count/L)

33. Fibrinogen Fibrinogen concentration (mg/dL)

34. Platelets Platelet count (count/mL)

35. Age Age (years)

36. Gender Female (0) or male (1)

37. Unit1 Administrative identifier for ICU unit (MICU); false (0) or true (1)

38. Unit2 Administrative identifier for ICU unit (SICU); false (0) or true (1)

39. HospAdmTime Time between hospital and ICU admission (hours since ICU admission)

40. ICULOS ICU length of stay (hours since ICU admission)

41. SepsisLabel For septic patients, SepsisLabel is 1 if t ≥ tsepsis −6 and 0 if t < tsepsis −6. For non-septic patients, SepsisLabel is 0.

# 5. Procedure and Analysis

The above figure shows the systematic approach taken to solve the challenge of predicting sepsis in patients.

The dataset the project has is an hourly time sequence record for each patient. But each of the patient records do not have any time component associated with them (timestamp). So, there can be two ways of approaching the problem. First, we can impute the time component (as well as the innate Patient ID) as a part of the dataset. Let us call this time series approach. Another approach is to ignore the time component and treat the record as independent and identically distributed. This approach can predict Sepsis at each hour without any past patient record. This approach will be called non-time series approach. The project we are doing will be the first sting of approaching the problem of early sepsis prediction, hence we find it appropriate in taking the non-time series approach and get a baseline prediction.

In this approach, as mentioned, we will be ignoring the time component associated with each hourly patient record and treat them as an individual and identically distributed. The dataset is split into two parts - Training set and testing set. The training set is in-turn split so we can perform some validation before testing it on the test set.

## 5.1 Data Cleaning – Preprocessing & handling missingness

While finding the data distribution, we found that many variables had a lot of missing values. More than ten laboratory values had missing values in more than 92% of the feature. Hence data imputation on them for reasonably tricky as we could not agree upon the correct mean of the dataset. It would be better to come up with newer features instead of trying to clean up with newer ones and then impute the missing ones as a new category.

A picture containing object, antenna

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## Feature Selection & Feature Engineering

Feature Selection:

Two Approaches employed for Feature Selection:

1. Checked correlation of features contributing to the presence of Sepsis

A close up of a map

Description automatically generated

1. Read health magazines and Research journals such as

* [US National Library of Medicine](https://www.nlm.nih.gov/), [National Institutes of Health](https://www.nih.gov/)
* [Centers for Disease Control and Prevention](https://www.cdc.gov/)
* [Sepsis - The American Journal of Medicine](https://www.amjmed.com/article/S0002-9343(07)00556-6/abstract)

and filtered out the most named indicator of Sepsis.

Outcome: Heart rate, Pulse Oximetry, Body temperature, Blood Pressure (SBP, DBP), Mean Arterial Pressure, Respiration rate, Frac of inspired oxygen, Age, Gender, Hospital Admission Time and ICU length of stay.

Feature Engineering & label encoding

Developed 8 new features and are described:

1. **new\_age** : has 3 categorical values – old, young and adult
2. **new\_hr, new\_temp, new\_o2sat, new\_bp, new\_resp, new\_map, new\_fio2**: has 3 categorical values – normal, abnormal and missing

A screenshot of a cell phone

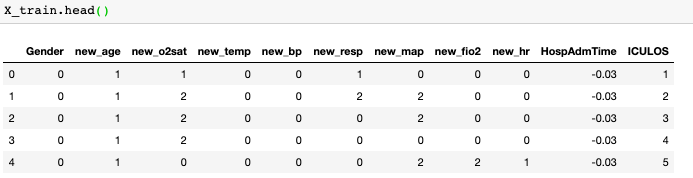
Description automatically generated

Next, **performed feature section again** on them and selected all above features, plus **Gender, Hospital Admission Time and ICU length of Stay** for further processing as a training set.

A screenshot of a cell phone

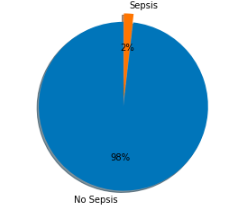
Description automatically generated

All these are categorically values. They are **encoded** so that it is easier to run a ML algorithm.



## Handling Data Imbalance

We can see from the data distribution that around 98% of patients does not have sepsis and just 2% have Sepsis.



**A screenshot of a cell phone

Description automatically generated**

There are many ways of handling imbalance in a dataset. In Machine Learning there are 3 prominent ways to handle them and they have been debated over and again in many scenarios.

1. Under-sampling : Deleting instances from the over represented class is one way of dealing with it. However, this might lead to heavy data loss
2. Over-sampling : Adding copies of instances from the underrepresented class is another way of dealing with it. SMOTE or Syntactic Minority Over - sampling TEchnique can be appied to minority class but it would not be a correct way if working as the heath care records' original data distribution might be lost.
3. Trying different algorithm: It is found that decision tree is said to perform well on a variety of imbalance dataset. The node splitting rule that look at class variables while creating a tree can be forced to address both classes.

In this project, after performing all the preprocessing step, we are going to try one version of decision trees as a classifier.

## Choosing Evaluation Metric

Usually a popular classification metric is basic accuracy score. The ratio of correct prediction to total number of data points in the dataset. However, in our case of imbalance data classes, this metric can be misguiding, as high metrics might not necessarily mean correct classification or correct prediction capacity for minority classes.

While testing this theory with our dataset, the model ran with decision tree gave an accuracy of 99% but it is still lousy prediction capacity on the class that we are truly interested in, which is predicting if a patient has sepsis. To gain more understanding for current model, we are considering Precision and Recall scores as a possibly efficient classification accuracy. The following metric can be used for our imbalance dataset.

1. True Positive Rate (sensitive or recall): It shoes how many relevant samples are selected, i.e how well our model can predict all interested samples in our dataset.

A close up of a logo

Description automatically generated

1. False Positive Rate (also fall-out)

A close up of a logo

Description automatically generated

1. Precision: It shows how many predicted samples are relevant i.e it shows how our mistakes into classifying samples as a correct one if its not true.

A close up of a logo

Description automatically generated

1. True Negative Rate (also specificity)

A picture containing object

Description automatically generated

1. F-1 score: This shows a harmonic average of the precision and recall. It is a good metric for imbalance classification scenario. The range of F1 is in [0,1] where 1 is perfect classification and 0 is a total failure.

A screenshot of a cell phone

Description automatically generated

We ran the algorithm and got the Precision – Recall Curve and the ROC curve as well.

A close up of a map

Description automatically generatedA screenshot of a social media post

Description automatically generated

From Fig, although ROC curve seems promising, we can see that P-R curve is not great at classifying.

## Training with ML algorithm – Gradient Tree Boosting

Although the projection has not tested an efficient algorithm to classify correct data points with sepsis and non-sepsis labels, we try to train using Gradient Tree Boosting.

Here, many models are trained sequentially. It is optimized such that the loss function is y = ax + b + e using Gradient Descent Method. Decision Trees are used as weak learners in Gradient Boosting. The accuracy was found good but need better model building and training.

A screenshot of a cell phone

Description automatically generated

# 6. Conclusion

We have handled the missing ness and imbalance in the large dataset, and we removed missing values greater than 92%. Next we performed feature engineering (8 new features) and selected important features. We aimed to predict the onset of the sepsis by 6 hours and so far, the Machine Learning model employed seem to classify it partially. The project has a scope of continuing with further research on the importance of the features, better model building and under the guidance of a good health science domain expert.

# 7. Future Scope

We need to perform more analysis using techniques, such as Random Forest, Principal Component Analysis etc. in order to establish a high confidence for the current results. We must next use time component Approach; need domain expert. We can think of using SMOTE for handling Imbalance. Work further on Gradient Boosting and in general use a good classifier.

# References

[1] <https://www.physionet.org/content/challenge-2019/1.0.0/>

[2] <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>

[3] <https://towardsdatascience.com/using-bagging-and-boosting-to-improve-classification-tree-accuracy-6d3bb6c95e5b>

[4] <https://towardsdatascience.com/early-detection-of-sepsis-using-physiological-data-78d5f31fab9d>

[5] <https://iopscience.iop.org/article/10.1088/1757-899X/428/1/012004>

[6] <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

[7] <https://www.cdc.gov/>

[8] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6429642/>

[9] <http://www.erogol.com/fighting-class-unbalance-supervised-ml-problem/>

[10]<https://towardsdatascience.com/metrics-for-imbalanced-classification-41c71549bbb5>