**Predicting Sepsis: An algorithm comparison**

Nicholas Dragun

UNF- Data Science Bootcamp

February 20, 2024

**Introduction**

Sepsis is a life-threatening condition that results in organ failure and ultimately death if not treated in a timely fashion. It most likely affects those with weakened immune systems but can affect people of all ages and backgrounds. Early detection of sepsis is key to saving lives and reducing healthcare costs.

Rapid treatment allows healthcare providers to initiate treatment promptly by administering antibiotics and IV fluids to help stabilize the patient. Preventing complications is key as sepsis can rapidly deteriorate into severe sepsis and septic shock. Thus, early detection and treatment increase one’s survival rate. Early detection not only increases survival rate but from a hospital system perspective, reduces the patient’s length of stay and costs savings.

Efforts to improve the early detection of sepsis include the development of clinical decision-support tools, screening protocols, and educational initiatives for healthcare providers. By raising awareness, implementing best practices, and leveraging technology, healthcare systems can enhance their ability to identify sepsis early, ultimately saving lives and reducing healthcare costs.

**Methods of Predicting Sepsis**

A dataset was utilized from Kaggle.com which is composed of over 1 million records and whether the patient had sepsis or not. This dataset was cleansed of null values and permutation was applied for missing values of the four main categories of measure for sepsis. After duplicate patient encounters were deleted and the null values were replaced with the mean for that measure for each patient, exploratory data analysis was then developed. The goal of this project was to apply the models of K-Nearest Neighbor (KNN), Logistic Regression, and Support Vector Machine (SVM) algorithm. These algorithms were then analyzed on their performance using the ROC curve and precision and accuracy measures. This allowed us to determine whether one model was superior to another.

**Model Results**

KNN was conducted first on the newly cleansed dataset. Figure 1 expresses the results of this model. One can determine the accuracy of 0.98 and carries an Area Under the Curve (AUC) score of 0.54.

**Figure 1**

*K-Nearest Neighbor*

A graph with a line

Description automatically generated

Figure 2 expresses the results of the ROC curve for the SVM model. In contrast to the KNN, the SVM model shows a line that mimics that of the random dotted line. As a result, the SVM model at first glance is not as refined as the KNN model. Though the appearance of the ROC curve may not be as suitable as the KNN, the AUC score is 0.52, just 0.2 points under that of the KNN.

**Figure 2**

*Support Vector Machine*

*A graph of a curve

Description automatically generated with medium confidence*

The last model that was tested was the logistic regression model. In theory, this model would hold the most relevant values and be the most accurate given that the dataset is binary which is what logistic regression is best suited for. In all, it was determined that a comparison of these models did not show a statistically significant difference in the prediction of sepsis. Further research could be performed such as XGBoost and other deep neural network models that may result in a larger AUC score and more accurate results.

**Exploratory Data Analysis (EDA)**

Exploratory data analysis is an approach to analyzing datasets to summarize their main characteristics. This is an essential step in any data analysis project and is key to understanding the data using visuals. For this sepsis prediction project, data was collected and cleansed, and summary statistics were performed on the main predictive variables for sepsis. Figure 3 expresses the distributions of values for the six main predictive variables for sepsis in histogram form.

**Figure 3**

*Six main predictive variables for sepsis distribution*

*A screenshot of a graph

Description automatically generated*

Figure 4 expresses a correlation heat map between the variables described above. Some of the greatest correlations are systolic blood pressure and mean arterial pressure. As well as, diastolic blood pressure and mean arterial pressure.

**Figure 4**

*Correlation matrix*

*A screenshot of a computer screen

Description automatically generated*

**Conclusion**

The hypothesis that there is no difference between predictive models was proven true in that there was no statistically significant difference between predictive models. The exploratory data analysis expresses that there is some common correlations between MAP, DBP, and SBP. Further research should be conducted using more detailed models using deep neural networks such as XGBoost. These expanded models would be even more accurate and help clinicians predict sepsis at a greater rate given a patient’s blood test results among vital signs. Thus, helping to reduce deaths from this disease and save lives. In addition, the hospital systems experience less strain surrounding the treatment of this condition.