Nicholas Dragun

Predicting Sepsis Proposal

Sepsis is a fatal disease caused by an infection in the body that poisons the blood and creates organ failure. For clinicians managing this disease has always been a challenge and from a larger perspective hospitals struggle with the high costs involved in successfully treating patients whose outcome has a great impact on the community and hospital system as a whole (Shankar, et al., 2021).

As electronic health records became more prominent over the past decade, the result was an accumulation of healthcare data throughout the health system. Along with the vast amounts of data came the creation of the subject of Data Science. A means of managing, developing, governing, analyzing, and predicting using large data sets.

The most common technique to quantify sepsis is the Sequential Organ Failure Assessment (SOFA) score. The score consists of six parameters covering the respiratory system, nervous system, cardiovascular system, liver, coagulation, and kidneys. Each parameter is assigned a score from 0 to 4, ranging from lowest to highest respectively (Shankar, et al., 2021).

The proposed project aim is to use machine learning techniques in Data Science to help predict the presence of sepsis among a patient population. This in turn will allow for early detection and faster treatment, thus helping cure the patient before the condition reaches severe sepsis or septic shock. The proposed methods will utilize logistic regression, k nearest neighbor, and support vector machine models to assess a large dataset and help predict those with sepsis. The project will go a step further and compare each of these models to determine if there is a statistically significant difference between the outcomes of these models. The ROC curve and AUC score along with the F-1 score, accuracy, and precision will be used to assess each model’s effectiveness.

The sepsis dataset was obtained through the open-source platform of Kaggle.com. Data cleansing, wrangling, and analysis will be carried out to best keep the integrity of the dataset and limit the number of null values especially in the fields that the models will draw their predictions from. In all, the proposed project would not only help clinicians detect sepsis early but also help to eliminate the financial burden of health systems.

Capstone II

1. Sepsis Prediction – Kaggle dataset
   1. What variables are best to use for sepsis prediction
   2. Compare traditional algorithm vs neural networks in prediction
   3. Determine the best method
2. Cardiovascular Disease Prediction
   1. Risk factors that contribute
   2. Which factors contribute the most to developing cardiovascular disease
   3. Compare two algorithms and determine which is more reliable
3. Length of Stay
   1. Determine the factors that contribute to increased LOS in the hospital
   2. Which factors contribute more
   3. Is there any bias in the algorithms used?