CS598 DLH
Matthew Kennedy
Project Proposal

Citation to the original paper

Hou, N., Li, M., He, L. *et al.* Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 18, 462 (2020). https://doi.org/10.1186/s12967-020-02620-5

General problem

The general problem of L. et al. [1] focuses on the challenge of predicting 30-day mortality in patients diagnosed with sepsis-3 in the ICU setting. Sepsis is a severe condition resulting from the body's response to infection leading to organ dysfunction, and it has a high mortality rate especially among ICU patients. Early and accurate prediction of mortality in these patients is important as it can guide time-critical and appropriate treatment for patients, potentially improving survival outcomes. However, traditional prediction models and clinical scoring systems have limitations in their predictive accuracy and applicability. The study aims to address this problem by developing and validating a machine learning approach using the XGboost algorithm, comparing its performance with traditional models to demonstrate its potential as a more effective tool for predicting mortality in sepsis-3 patients.

Specific approach

The specific approach of L. et al. [1] adopts a machine learning-based approach to improve the prediction of 30-day mortality among sepsis-3 patients in the ICU by using the XGboost algorithm. By analyzing data from the MIMIC-III database, the study selects clinically significant variables through stepwise analysis for model construction. Three predictive models are developed and compared - the three models are a conventional logistic regression model, a SAPS-II score model, and the XGboost model, which is the model of focus for L. et al. [1]. The effectiveness of these models is assessed through ROC curve AUCs and decision curve analysis. The XGboost model's better performance is demonstrated, and its predictive value is further validated using a risk nomogram and clinical impact curve, highlighting its potential to significantly enhance clinical decision-making and patient management for sepsis-3 in critical care settings.

Hypotheses to be tested

The hypothesis tested by L. et al. [1] is that a machine learning model developed using the XGboost algorithm can more accurately predict 30-day mortality among patients diagnosed with sepsis-3 in the ICU setting compared to traditional prediction models and clinical scoring systems. The study aims to demonstrate that the XGboost model using a comprehensive set of clinically significant variables from the MIMIC-III database can outperform conventional logistic regression and SAPS-II score models in predicting the risk of mortality. This ipmroved model can then provide a more effective tool for guiding clinical decisions and interventions in the management of sepsis-3 patients.

Ablations planned
For ablations, I will be looking into removing certain features from the database that were identified as significant in the authors' stepwise analysis. From [1]:

"The detailed process of data extraction is shown in Fig. 1. Following demographic data were extracted: age, gender, ethnicity, weight, height and body mass index (BMI), length of stay in hospital, length of stay in the ICU, hospital expire fag (in-hospital death recorded in the hospital database) at the first ICU admission. Ten, we collected vital signs of the patients from the frst 24 h of ICU stay, including heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), temperature (TEMP), respiratory rate (RR) and oxyhemoglobin saturation (SpO2). Afterwards, laboratory values, such as blood routine examination, liver and kidney function, blood glucose, and arterial blood gas (ABG) were abstracted. Furthermore, advanced cardiac life support (mechanical ventilation, renal replacement therapy, etc.) and accompanied diseases (diabetes, malignant tumour, etc.) were accessed. Because of the high sampling frequency, we use the maximum, minimum and the mean value when incorporating the characteristics of vital signs and related laboratory indicators." (L. et al. p3-4).

Understanding which features are most significant in identifying 30 day mortality with ICU patients diagnosed with sepsis-3 would be helpful in understanding what a model is weighing more heavily in it's predictions.

Description of how I will access the data
The data used in this study is MIMIC III v1.4, which I have accessed through PhysioNet via instructions given by the teaching staff of the course. Additionally, there is a GitHub repository of the data involved in this paper and how to preprocess/prepare this data. The models are not included in the code, but a description of the models are, so I will create my own models using the sklearn and xgboost python libraries.

Discussion of the feasibility of the computation
While the dataset used is >6gb compressed (MIMIC-III), the size of models is largely dependent on me, since the models themselves are not directly available. Thus, I will be able to scale the 3 models mentioned for comparison to whatever size is feasible on my own PC. I currently have a machine with 64gb of ram and a 3950x, as well as a 2080 super, so most models will be feasible.

Will I use the existing code?
For preprocessing the data - yes, as described in the section above "Description of how I will access the data". For creating the models, no - there is no source code available for the models, so model creation, hyperparameters, and model structure features such as hidden layers, will be created to be whatever my PC can handle.