Project Milestone 1

Adil Khan

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

Sep 5, 2022

Topic

Heart Failure Predictions Based on Patients' Health Attributes

Research Question/Abstract:

According to estimates, cardiovascular diseases account for 17.9 million annual deaths worldwide, or 31% of all fatalities. Heart failure arises when the heart is unable to pump enough blood to meet the body's needs. Machine learning, particularly when used with medical data, may be a helpful tool for both forecasting the prognosis of each patient displaying heart failure symptoms as well as for identifying the most important clinical features (or risk factors) that may culminate in heart failure. (Latha, 2019). 2019 (Lawler).

Machine learning may help scientists with feature evaluation as well as clinical prediction. A new tool for clinicians to utilise in assessing whether a patient with heart failure will survive or not may be created by using machine learning and data science to clinical practise in the healthcare sector. In example, while attempting to evaluate whether a patient would survive after experiencing heart failure, clinicians frequently place an emphasis on serum creatinine and ejection fraction (Chicco, 2020). In order to build a prediction model that will enable us to forecast a patient's chance of developing heart failure based on the existing health data, the objective of this research is to examine and grasp the data.

Dataset:

The dataset is publicly available on Kaggle website and can be access through the link provided, https://www.kaggle.com/andrewmvd/heart-failure-clinical-data. There are 13 characteristics in the dataset that provide light on clinical, physical, and lifestyle data related to patients. The medical

records of 299 patients with heart failure were gathered from April to December 2015 at the Allied Hospital and Faisalabad Institute of Cardiology in Punjab, Pakistan (Ahmad, 2017). Between the ages of 40 and 95, the sample's 194 men and 105 women represented a wide age range. The dataset includes categorical factors including anaemia, hypertension, diabetes, sex, smoking, and mortality events. In data that has been boolean-typed, these category characteristics are represented. If a patient's hematocrit level was below 36%, they were deemed anaemic (Chicco, 2020). Age, creatinine phosphokinase (CPK), ejection fraction, platelets, serum creatinine, serum sodium, and time make up the remaining characteristics. Continuous variables are used to represent these characteristics. Creatinine phosphokinase (CPK) is released into the circulation when muscle tissue is injured. Consequently, heart failure may be a sign of elevated CPK levels in the blood (Chicco, 2020). High levels of serum creatinine may be caused by renal failure; serum creatinine is a byproduct of creatinine produced during muscle catalysis (Stephens, 2019). In the dataset, the death event variable identifies whether a patient passed away or lived before the conclusion of the follow-up period, which on average lasted 130 days. (Ahmad, 2017).

Methods

The strategy was divided into three stages in order to create a predictive model. The activities that must be completed before moving on to the next step are included in each phase.

• **Phase 1** – exploratory data analysis. The initial stage in every data science analysis task is this phase. Since there are 13 variables total in the dataset, several of them may or may not be correlated, particularly with the feature in question, the death event variable. We must thus picture and comprehend how they are distributed. We must also make careful to look for outliers and missing numbers.

• **Phase 2** – this is the feature selection phase. Once we are aware of how each variable relates to and correlates with our main variable, death event. Then, to base the prediction model on, we may choose the attributes that have the most effect and connection with the primary variable we have chosen. To determine which machine learning model is best for our prediction model, we will test a variety of them. Decision trees, logistic regression, support vector machines, k-nearest neighbours, random forests, and other machine learning models will all be employed in this model.

• **Phase 3** – The characteristics will be applied to the construction of the prediction models in this step after being chosen. The dataset's existing data will be used to execute and train the model.

Ethical Considerations

The leading cause of mortality worldwide, cardiovascular diseases (CVDs), claim 17.9 million lives annually, or 31% of all fatalities worldwide. The majority of cardiovascular illnesses may be avoided by employing population-wide measures to target behavioural risk factors such cigarette use, poor eating and obesity, inactivity, and problematic alcohol consumption.

Early detection and management of people with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors like hypertension, diabetes, hyperlipidemia, or already established disease) are essential, and a machine learning model can be very helpful in this regard.

References

1. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS ONE. 2017; 12(7):0181001.

2. Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

3. Dalen, J. E., Alpert, J. S., Goldberg, R. J., & Weinstein, R. S. (2014). The Epidemic of the 20th Century: Coronary Heart Disease. The American Journal of Medicine, 127(9), 807–812. https://doi.org/10.1016/j.amjmed.2014.04.015

4. Faggella, D. (2020, March 4). 7 Applications of Machine Learning in Pharma and Medicine. Emerj. https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/

5. HealthITAnalytics. (2018, September 18). Using Big Data, Machine Learning to Reduce Chronic Disease Spending. https://healthitanalytics.com/news/using-big-data-machine-learning-to-reducechronic-disease-spending

6. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203.

7. Lawler, R. (2019, July 15). How doctors are using machine learning to improve health outcomes. Samsung NEXT. https://samsungnext.com/whats-next/how-doctors-are-using-machine-learningto-improve-health-outcomes/

8. Stephens C. What is a creatinine blood test? https://www.healthline.com/health/creatinine-blood. Accessed 25 Jan 2019.

9. Stephens, W. (2019, June 19). Machine Learning Can Predict Heart Attack or Death More Accurately Than Humans. AJMC. https://www.ajmc.com/view/machine-learning-can-predict-heart-attackor-death-more-accurately-than-humans

10. Stewart, J., Addy, K., Campbell, S., & Wilkinson, P. (2020). Primary prevention of cardiovascular disease: Updated review of contemporary guidance and literature. JRSM Cardiovascular Disease, 9, 204800402094932. https://doi.org/10.1177/2048004020949326.