Studying the most relevant risk factors for heart disease

FRANCESCO DI FLUMERI PABLO LASO
frdf@kth.se | plaso @kth.se

September 16, 2021

1 Allocation of responsibilities

In principle, Francesco Di Flumeri is responsible for the data set cleaning, analysis, model evaluation, and features importance computation. Pablo Laso will focus on feature analysis, feature selection (also ranking), and model building.

Reports and presentations are intended to be conducted together. Mainly, the Introduction will be made by Pablo. Methods is a two-person work, so that each writes about what he did by coding. Finally, the Discussion section will be written by Francesco.

2 Organization

The project will be conducted by two KTH computer science students, who are attending the II2202 course. Pre-existing data collections coming from the UCI organization [1] are used. In addition, previous studies about hearth diseases will be analyzed in order to select the most suitable data for the construction of the model.

3 Background

This project builds on the idea of using Artificial Intelligence in the healthcare sector, specifically, for heart disease. We will adopt a Machine Learning classifier to detect potential cases of heart diseases, as well as ranking the most important features (i.e. risk factors) leading to these conditions. Since the amount of data is too large to be analyzed from humans, these algorithms offer a big support in case of working in Big Data sector.

4 Problem statement

23 million people worldwide are affected from hearth failure [2] and this leads to high number of hospitalizations with an increasing in national health costs [3]. Doctors and medical authorities cannot deal with Big Data without the assistance of computers which might help them to intervene before a heart stroke happens, avoiding pain for the patients and reducing the number of surgical interventions. Moreover, there are many risk factors that can lead to hearth disease, and even vary between different regions and ethnicities [4].

5 Problem

Heart disease is the world's biggest killer [5]. Furthermore, several cardiovascular diseases are considered the main cause of death, especially in the most developed countries, where the population is older.

6 Hypothesis

Early prediction of heart disease is possible given certain health information about a patient. Moreover, old males have a greater possibility than other individuals in developing hearth diseases [4]. Additionally, one of the major causes of hearth failure are the presence of hypertension (HT) and valvular disease (VHD) [6]. Therefore, outcomes of prediction are expected to be based especially on these four facts. Overall, we hypothesize that a healthier lifestyle, and avoiding or controlling certain risk factors, can lead to a lower likelihood of suffering from heart disease.

7 Purpose

The purpose of this project is to assist people in choosing better health habits that lower their chances of suffering from a heart disease later in life. A complementary purpose is to progress in medical diagnosis through Artificial Intelligence-based algorithms that lead to early diagnosis (based on Big Data), as well as serving as a decision-support system for physicians.

8 Goal(s)

The goal is to identify the most relevant risk factors, or even finding new ones, as well as building a reliable algorithm that can predict early cases of heart diseases, hopefully allowing physicians to treat them as soon as possible.

9 Tasks

A Machine Learning model shall be built with the aim of early detecting potential patients suffering from heart disease. The data-set presents myriad features related to the patient lifestyle and health status that shall be analyzed and ranked to study their correlation to the pathology.

10 Method

We will use an analytical method to make sense out of the data collection we have, provided by the ICU organization. After data analysis, data cleaning will delete misleading information. We will adopt a supervised-learning, binary classification method. Everything will be implemented in Python. Evaluation will be performed adopting metric such as Accuracy, Precision, Recall, F1-score and Confusion matrix [7].

11 Milestone chart (time schedule)

The project will start on 13.Sep.2021 and end in 17.Jan.2022, at 16:59. There will be the following milestones and deliverables:

September, 17 Project proposal submission.

September, 20 Presentation of your proposed research: Ethics & Sustainability.

September, 27 clean data-set with which to work. Bibliographic research already conducted on similar topics.

October, 8 Research plan: First draft of your research plan, presentation, and peer reviewing

October, 27 Feature selection performed. Correlation and statistical studies performed. Study on most important features and its underlying biological meaning.

November, 29 model completed based upon using the statistical package python. Study the most accurate models, and compare to other studies. Final report: First draft and Presentation with peer review of draft report and presentation

December, 8 report draft for: Written opposition: before final seminar - with peer review

January, 10 evaluation of models. Slides for: Assignment Final seminar - with oral opposition

January, 17 submit final report (the report will have been written in parallel with each of the above steps)

References

[1] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[2] J. McMurray, M. Petrie, D. Murdoch, and A. Davie, “Clinical epidemiology of heart failure: public and private health burden.” *European heart journal*, vol. 19, pp. P9–16, 1998.

[3] M. Gheorghiade and R. O. Bonow, “Chronic heart failure in the united states: a manifestation of coronary artery disease,” *Circulation*, vol. 97, no. 3, pp. 282–289, 1998.

[4] S. Khatibzadeh, F. Farzadfar, J. Oliver, M. Ezzati, and A. Moran, “Worldwide risk factors for heart failure: a systematic review and pooled analysis,” *International journal of cardiology*, vol. 168, no. 2, pp. 1186–1194, 2013.

[5] W. H. Organization. Cardiovascular diseases. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[6] K. Fox, M. Cowie, D. Wood, A. Coats, J. Gibbs, S. Underwood, R. Turner, P. Poole-Wilson, S. Davies, and G. Sutton, “Coronary artery disease as the cause of incident heart failure in the population,” *European heart journal*, vol. 22, no. 3, pp. 228–236, 2001.

[7] Y. Liu, Y. Zhou, S. Wen, and C. Tang, “A strategy on selecting performance metrics for classifier evaluation,” *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, vol. 6, no. 4, pp. 20–35, 2014.