Studying the most relevant risk factors for heart disease

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

October 6, 2021

Contents

List of Acronyms and Abbreviations

Before reading this document, the reader should familiarize with some technical expressions, in order to fully understand how the authors are planning to conduct the research. An overview of the different terms is provided below, following the descriptions included in the cited volumes [1, 2, 3].

Classification Algorithm	Input classified in categories identified by a numerical code [1]
Supervised learning	Predicted labels are known from the training phase [1]
Decision tree	Predictive model used in ML classification practises [3]
Random Forest	Classification algorithm based on the construction of multiple decision trees [2]
Cardiovascular disease	Heart disease [4]

1 Aims, Objectives, Goals, Research questions, hypotheses

The aim of this project is to support medical authorities in fighting heart diseases with the help of Artificial Intelligence. Precisely, we would like to induce a progress in medical diagnosis through a Machine Learning-based algorithm that lead to early disease detection (based on Big Data), supporting the physician to increase the living chances of an individual. Hence, 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. The research questions we want to address are the following:

> Which are the most important features in a ML model for predicting if a patient will experience an heart disease? And by using these features is it possible to build a reliable prediction model?

We hypothesize that early prediction of heart disease will be possible given certain health information about a patient. Moreover, old males have a greater possibility than other individuals in developing hearth diseases [5]. 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 strongly related to these four facts. In the end we expect to obtain more generalized results since data belongs to subjects living in 4 different countries.

2 Background and rationale

This project builds on the idea of using Artificial Intelligence in the healthcare sector, specifically, for heart disease. We will adopt a Machine Learning supervised classifier to detect potential cases of heart failure, as well as ranking the risk factors that will influence the prediction the most. 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.

For what concern the problem, we found that 23 million people worldwide are affected from heart failure [7] and this leads to high number of hospitalizations with an increasing in national health costs [8]. Doctors and medical authorities cannot deal with large amount of 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 [5]. Hence it is possible to state that heart disease is the world's biggest killer [9]. Furthermore, several cardiovascular diseases are considered the main cause of death, especially in the most developed countries, where the population is older. [10]

3 Theory/literature

In the recent years numerous computer science techniques have been involved in the healthcare field in order to support medical authorities in patients treatments. Moreover, as the healthcare sector requires large investments from governments (it is enough to consider the example of the USA where medical expenses represent the 17% of the Gross Domestic Product [11]), technological solutions are required in order to reduce the economical weight of this field. [11] Indeed, Big Data analysis and Machine Learning practises are constantly spreading inside medical organizations and occupying an increasingly central role in helping physicians. For what concerns Big Data, healthcare can strongly benefit from this technology, because it provides the availability of large amounts of information, since it is able to manage structured, semi-structured, and unstructured data in petabytes and more. [12, 13, 14, 15, 16] On the other hand, Machine Learning (ML) techniques are useful in healthcare realm since they can improve the relationship between doctors and patients. Indeed, starting from raw data, machine learning models can provide physicians with diagnosis, medical suggestions and predictions about the manifestation of a new pathology. [17] Among the various medical issues addressed with the support of ML, there is the category of cardiovascular diseases as well. [18] Indeed, ML has proved to be effective in helping physicians in diagnosing heart diseases before an heart failure occurs. In the previous studies, numerous techniques in data mining and neural networks have been adopted in order to assess the severity of a specific cardiovascular condition, such as K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB). [19] Moreover, there are various researches which, by using machine learning techniques, tried to predict the risk of developing heart diseases. One of these is the study conducted by Mohan et al. [18] Here nine ML models were developed with different algorithms and their performances were evaluated. In the end, they concluded that the best results were obtained by an hybrid random forest model with an accuracy score of 88.7%. [18] This has led our choice of adopting a Random Forest algorithm for developing our model. This technique builds various decision trees and aggregate them in order to obtain the best final result. [2]

4 Research Methodology

In this study the analytical method [20] will mainly be involved. However, empirical techniques [20] may be adopted in the analysis of the data set, such as in the exclusion of empty records or no-sense information. On the other hand, for the model and features importance evaluation only the analytical method will be used since we are going to compute several performance metrics such as Accuracy, Precision, Recall, F1-score and Confusion matrix. [21]

5 Participants, Procedures, Data collection and analysis

Pre-existing data collections coming from the UCI organization [22] will be used. The first reason behind choosing this data collection lays down the fact that it includes information gathered from four different parts of the world. Hence, it will be possible to obtain more generalized results at the end of the study. Second, this data set has been selected because suitable for Machine Learning practises as the risk of contracting a cardiovascular disease is expressed in a numerical way. This collection stores 76 attributes per each subject who has been involved in the study. The participants were living in four different parts of the world: Switzerland (CH), Hungary (H), Cleveland (USA) and Long Beach (USA). For each location a different number of individuals was analysed and stored in four different tables: 303 from Cleveland, 294 from Hungary, 123 from Switzerland and 200 from Long Beach. The anonymity is kept by omitting the personal information of each participant. Indeed id code, name and date of birth are not available inside the data collection.

The research will be divided in different phases and all the coding tasks will be realized in Python:

- data pre-processing: in this phase we will perform data-formatting, anomaly detection, missing values imputation and data cleaning;
- feature selection: in this phase we will select a subset of the 76 attributes that will be used as input of the ML model;
- model construction: a model based on the Random Forest algorithm will be realized. During the model training and testing, records from all the four tables will be involved.
- performances evaluation: in this phase several metrics will be computed in order to understand the performances of the model. These are Accuracy, Precision, Recall, F1-score and Confusion matrix. [21]
- features importance: in this phase we will determine, by using the method of partial dependence [23], the features that influence the predicted value the most.

6 Expected outcomes

As mentioned in the hypothesis, we expected the predicted value to be strongly related to age, sex and presence of hypertension (HT) or valvular disease (VHD). [5, 6] Moreover, after analysing the paper of

Mohan et al. [18], we expect our model accuracy to be not lower than 80%, so that it can be considered acceptable. In the end, we plan to find some similar patterns in the development of cardiovascular diseases among subjects living in 4 different parts of the world.

7 Milestones/schedule, budget

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)

8 Risks

There are some risks that we need to take into account. First, it should be considered that the authors of the research do not possess any medical background, hence the features selection might present some errors due to lack of specific knowledge. Moreover, the restricted size of the data-set might cause the model to be under-fit [1], which means that it might be not able to fully understand the relationship between the input set X and the outcomes set Y.

9 Outline

The final report of this research will have the following structure:

- Title and Authors
- Abstract
- Introduction
  - Literature study
  - Research questions and hypothesis
  - Goals
- Method

  - Research methodology
  - Theoretical framework
  - Data pre-processing
  - Feature selection
  - Model construction
  - Model evaluation
  - Feature importance

- Results
- Discussion and Future work
- Conclusion

References

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[2] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[3] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[4] P. Libby and P. Theroux, "Pathophysiology of coronary artery disease," *Circulation*, vol. 111, no. 25, pp. 3481–3488, 2005.

[5] 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.

[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] 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.

[8] 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.

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

[10] "Country comparisons median age." [Online]. Available: https://www.cia.gov/the-world-factbook/field/median-age/country-comparison

[11] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A study of machine learning in healthcare," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2017, pp. 236–241.

[12] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, "A look at challenges and opportunities of big data analytics in healthcare," in *2013 IEEE international conference on Big Data*. IEEE, 2013, pp. 17–22.

[13] B. Kayyali, D. Knott, and S. Van Kuiken, “The big-data revolution in us health care: Accelerating value and innovation,” *Mc Kinsey & Company*, vol. 2, no. 8, pp. 1–13, 2013.

[14] A. Tattersall and M. J. Grant, “Big data–what is it and why it matters,” 2016.

[15] R. Bhardwaj, A. Sethi, and R. Nambiar, “Big data in genomics: An overview,” in *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 45–49.

[16] T. Daveport, “Industrial-strength analytics with machine learning,” *The Wall Street Journal*, vol. 240, 2013.

[17] D. Maddux, “The human condition in structured and unstructured data,” *Acumen Physician Solutions*, 2014.

[18] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE access*, vol. 7, pp. 81 542–81 554, 2019.

[19] M. Durairaj and V. Revathi, “Prediction of heart disease using back propagation mlp algorithm,” *International Journal of Scientific & Technology Research*, vol. 4, no. 8, pp. 235–239, 2015.

[20] P. Bock, *Getting it right: R&D methods for science and engineering*. Academic Press, 2001.

[21] 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.

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

[23] C. Molnar, T. Freiesleben, G. König, G. Casalicchio, M. N. Wright, and B. Bischl, “Relating the partial dependence plot and permutation feature importance to the data generating process,” *arXiv preprint arXiv:2109.01433*, 2021.