Studying the most relevant risk factors for heart disease

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

January 15, 2022

Abstract

Heart failure represent a huge problem for the world population, considering that 23 million people on the whole planet surface are affected from it. This phenomenon cause high hospitalization costs and most important, a large lives loss. The purpose of this research is to study the possibility of developing a machine learning model that is able to early predict an heart failure case. Moreover, the study shows which are the most relevant risks for heart disease. The UCI medical data-set which has been used in this research, gathers physical, psychological and medical data about patients in four different parts of the world. By using it, we developed a reliable ML model with high performances. In addition the results showed that the most important factors in the predictions are the one related to the blood vessels' size.

Contents

1 Introduction **4**
1.1 Theoretical framework/literature study 4
1.2 Research questions, hypotheses, goals 4

2 Method(s) **5**
2.1 Research Methodology 5
2.2 The dataset 5
2.3 Data cleaning 5
2.4 Exploratory Data Analysis 6
2.5 Model building 6
2.6 Model evaluation 7
2.7 Features importance 7

3 Results and Analysis **7**

4 Conclusion **12**
4.1 Discussion 12
4.2 Future work 12

A Correlation matrix for Cleveland and Long Beach **15**

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].

Classification Algorithm Input classified in categories identified by a numerical code [1]

Supervised learning Predicted labels are known from the training phase [1]

Cardiovascular disease Heart disease [2]

1 Introduction

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 [3] and this leads to high number of hospitalizations with an increasing in national health costs [4]. 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 [6]. Furthermore, several cardiovascular diseases are considered the main cause of death, especially in the most developed countries, where the population is older. [7]

1.1 Theoretical framework/literature study

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 [8]), technological solutions are required in order to reduce the economical weight of this field. [8] Indeed, Big Data analysis and Machine Learning practices 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. [9, 10, 11, 12, 13] 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. [14] Among the various medical issues addressed with the support of ML, there is the category of cardiovascular diseases as well. [15] Indeed, ML has proved to be effective in helping physicians in diagnosing heart diseases before a 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). [16] 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. [15], that involves the usage of our same data-sets. 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 a hybrid random forest model with an accuracy score of 88.7%. [15] This has led to our choice of focusing on 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. [17] We decided to use the study of Mohan et al. because we consider it quite commensurable to our work. However, as we are using the all four data collections, there might be the risk that our research is not comparable with it.

1.2 Research questions, hypotheses, goals

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 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 an ML model for predicting if a patient will experience a 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 risk 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) [18]. 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 four different countries.

2 Method(s)

2.1 Research Methodology

In this study, the analytical method [19] will mainly be involved. However, empirical techniques [19] 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 [20].

2.2 The dataset

Pre-existing data collections coming from the UCI organization [21] 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 practices, 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 analyzed 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.

2.3 Data cleaning

The first activity that we perform before building the ML model was cleaning the data collections stored in the data-set. Indeed, *dirty* data can lead to incorrect predictions and unreliable results. Different techniques have been involved in data cleaning, in order to solve different information quality problems, such as duplicates, missing values, integrity constraints violations, and outliers [22]. A combination of *human-guided* and *automated* techniques [23] was employed. The firsts were adopted for duplicated records removal, empty attributes deletion and data formatting [24] while the seconds were used for missing values retrieval [25]. In particular, for missing values imputation, we utilized the k-Nearest Neighbor (KNN) algorithm that replace the missing data by looking to the k closest neighbors with a known value [26]. This methodology has been chosen because it has been shown to be generally efficient with numerical values [27]. After missing values imputation, we proceeded in analyzing the of positive case (people who suffer from a heart condition) and negative (people without heart condition).

2.4 Exploratory Data Analysis

Once data is acquired, after ensuring its quality is optimal for data analysis, a deep exploration must be performed to better understand the data and be able to perform a proper pre-processing and modelling. This is what is called Exploratory Data Analysis (EDA). The different tasks involved in the ETA are:

- Understanding the dataset issues.
- Get information about data types, shape of the dataset and descriptive metrics.
- Extract information about the relevancy of some features over others.
- Identify outliers, missing values or human error.
- Understand the relationship, or lack of, between variables.
- Maximize the insight into the dataset and minimize the potential error that may occur during the next steps in the analysis process.

Hence, the main purpose of EDA is to explore the structure of the data and find patterns in behavior and distribution of the data. Descriptive analysis is the first approach in EDA to summarize the main characteristics of the dataset. It makes use of descriptive statistics and visual methods to get a deeper insight into the data.
Furthermore, we also computed the correlation matrix for all features (Fig: **??**). Most of them are uncorrelated, but in some cases it is possible to observe a correlation value of 0.8 or smaller.

2.5 Model building

In the model building phase, different classification techniques have been involved in order to select the one that was providing the best results. The different ML algorithms used for the model construction are:

- **Logistic Regression**: this algorithm allows obtaining the binary value of a class by using a logistic function to model the prediction [28];
- **KNN Classifier**: this algorithm based each prediction by considering the k-nearest points with a known label [29].
- **Decision Tree Classifier**: this algorithm is based on the usage of decision tree, where leaves represent class labels and branches represent the connection among the different features [30].
- **Neural Network Classifier**: this algorithm builds on the idea of having a network of functions, where each one produces an output based on its parameter (one of the predictors) [31].
- **Random Forest Classifier**: this algorithm uses different decision trees and then select the one that produce the most reliable result [17].

Before building the final model, we also perform the task of features selection, in order to use predictors that provide the best results and to exclude those that might corrupt the model output. We decided to focus on filter methods because they are portable, and they have low computational costs [32]. Two filter techniques for features selection and extraction were involved in this task in order to ensure the results' reliability:

- **F-score based**: this is a simple methodology for features selection that works only with two classes that computes the discriminating power of each feature [33].
- **Mutual information based**: this methodology compares pairs of features by evaluating information they can provide about the final label [34].

2.6 Model evaluation

After completing the building phase, we proceeded with its evaluation. Five metrics were involved in the evaluation process, in order to validate the model performances from different perspectives:

- **Accuracy**: the amount of correct predictions [20].
- **Recall**: the quantity of positive cases correctly identified [20].
- **Precision**: the proportion of correct positive cases identified with respect to the real amount of positive cases [20].
- **F1-score**: the harmonic mean between precision and recall for a classification problem [20].
- **Confusion matrix**: this is not a real metric, but a more visual technique to represent the proportion of correct or incorrect prediction [20].

2.7 Features importance

The feature importance was computed in order to get insights about the predictors and to understand which one influence the prediction the most. In this phase, each feature received a value representing its importance in the prediction. It was possible to obtain the features' importance only for the Random Forest based model and for the Decision Tree based model. This is due to the fact that features relevance is based on the probability of reaching a specific node in the tree [35].

3 Results and Analysis

The first results that we obtained were represented by the number of positive and negative instances for each data collection that are reported in the following histograms (Fig 6). Indeed, one of the most important data features to analyze is class balance. Considering variable "num" as the predictable variable, we observe four possible labels taking integers from 0 to 4, where label 0 refers to cases with a low likelihood of having heart disease, and high otherwise. From figure 6 it is possible to see that the two data collections with the most unbalanced data are on the right column (Switzerland and Hungary respectively). Due to corrupted data, we were able to use only two of the four data collections available within the data-set. Hence, the results reported below refer to Cleveland and Long Beach.

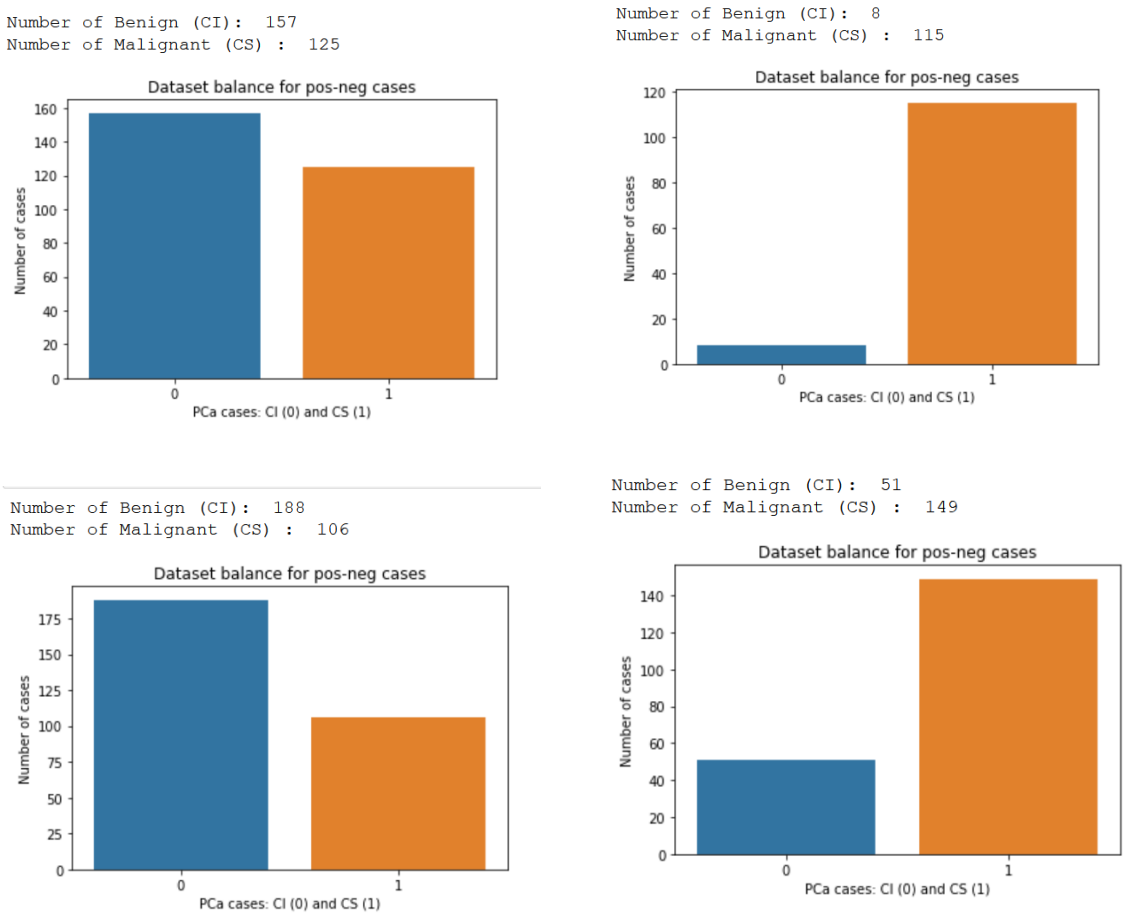

Figure 1: Data distribution histogram

Furthermore, we also computed the correlation matrix for all features (see Appendix A). Most of them are uncorrelated, but in some cases it is possible to observe a correlation value of 0.8 or smaller. The values obtained during the model evaluation and the features' importance computation are reported in this section. The performances of all the models were assessed in order to confirm our hypothesis that Random Forest is the most suitable algorithm for this scenario. The results are reported in two tables per each data collection: the first one refers to models built with F-score based features selection, while the second one to models constructed with mutual information based technique.

	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Logistic Regression	1.000000	1.000000	1.0	1.000000	0.952824	0.898947	1.000000	0.945623	0.985915	0.962963	1.000000	0.981132
KNN Classifier FS	0.947867	0.887755	1.0	0.940541	0.943522	0.887895	0.987500	0.933592	0.985915	0.962963	1.000000	0.981132
KNN Classifier	0.886256	0.755102	1.0	0.860465	0.834330	0.644211	1.000000	0.781708	0.859155	0.629630	1.000000	0.772727
Decision Tree Classifier FS	0.947867	0.887755	1.0	0.940541	0.943411	0.877368	1.000000	0.931967	0.985915	0.962963	1.000000	0.981132
Decision Tree Classifier	1.000000	1.000000	1.0	1.000000	0.924363	0.918421	0.923103	0.918192	0.901408	0.925926	0.833333	0.877193
Neural Network Classifier FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Neural Network Classifier	1.000000	1.000000	1.0	1.000000	0.891362	0.838421	0.924624	0.878018	0.915493	0.925926	0.862069	0.892857
Random Forest Classifier FS	0.947867	0.887755	1.0	0.940541	0.948173	0.887895	1.000000	0.938354	0.985915	0.962963	1.000000	0.981132
Random Forest Classifier	1.000000	1.000000	1.0	1.000000	0.924695	0.879474	0.954474	0.913159	0.971831	0.925926	1.000000	0.961538

Figure 2: Models performances in Cleveland with feature selection based on F-score

	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression MI	0.928910	0.846939	1.0	0.917127	0.928904	0.846316	1.000000	0.915761	0.943662	0.851852	1.000000	0.920000
Logistic Regression	1.000000	1.000000	1.0	1.000000	0.952824	0.898947	1.000000	0.945623	0.985915	0.962963	1.000000	0.981132
KNN Classifier MI	0.928910	0.846939	1.0	0.917127	0.928904	0.846316	1.000000	0.915761	0.943662	0.851852	1.000000	0.920000
KNN Classifier	0.886256	0.755102	1.0	0.860465	0.834330	0.644211	1.000000	0.781708	0.859155	0.629630	1.000000	0.772727
Decision Tree Classifier MI	0.928910	0.846939	1.0	0.917127	0.928904	0.846316	1.000000	0.915761	0.943662	0.851852	1.000000	0.920000
Decision Tree Classifier	1.000000	1.000000	1.0	1.000000	0.924363	0.918421	0.923103	0.918192	0.901408	0.925926	0.833333	0.877193
Neural Network Classifier MI	0.928910	0.846939	1.0	0.917127	0.928904	0.846316	1.000000	0.915761	0.943662	0.851852	1.000000	0.920000
Neural Network Classifier	1.000000	1.000000	1.0	1.000000	0.891362	0.838421	0.924624	0.878018	0.915493	0.925926	0.862069	0.892857
Random Forest Classifier MI	0.928910	0.846939	1.0	0.917127	0.928904	0.846316	1.000000	0.915761	0.943662	0.851852	1.000000	0.920000
Random Forest Classifier	1.000000	1.000000	1.0	1.000000	0.924695	0.879474	0.954474	0.913159	0.971831	0.925926	1.000000	0.961538

Figure 3: Models performances in Cleveland with feature selection based on mutual information

In the two tables above, models performances on Cleveland data collection are illustrated. It can be noticed that Random Forest and Logistic Regression are the models with the best performances. They present high values of accuracy, recall, precision and F1-score, especially when these are computed on the test sets. For both, the average value of the four metric is greater than 0.90. An interesting and unexpected result is that all the models have similar performances when built by using feature selection, both F-score and Mutual information. This may be due to the restricted number of features involved in the prediction. However, it can be stated that with features selection we are able to see a performance increase, especially in those algorithms with low scores such as KNN, which has a recall of 0.63.

	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression FS	0.926667	0.899083	1.000000	0.946860	0.926667	0.899134	1.000000	0.945842	0.92	0.900	1.000000	0.947368
Logistic Regression	1.000000	1.000000	1.000000	1.000000	0.893333	0.870996	0.980952	0.920748	0.92	0.925	0.973684	0.948718
KNN Classifier FS	0.933333	0.926606	0.980583	0.952830	0.920000	0.908225	0.981304	0.942350	0.90	0.875	1.000000	0.933333
KNN Classifier	0.873333	0.880734	0.941176	0.909953	0.820000	0.853680	0.899948	0.871434	0.76	0.750	0.937500	0.833333
Decision Tree Classifier FS	0.966667	0.963303	0.990566	0.976744	0.913333	0.917749	0.963123	0.938396	0.86	0.900	0.923077	0.911392
Decision Tree Classifier	1.000000	1.000000	1.000000	1.000000	0.986667	0.981385	1.000000	0.990471	0.92	0.950	0.950000	0.950000
Neural Network Classifier FS	0.933333	0.908257	1.000000	0.951923	0.926667	0.908658	0.991304	0.946520	0.90	0.900	0.972973	0.935065
Neural Network Classifier	1.000000	1.000000	1.000000	1.000000	0.853333	0.880519	0.918426	0.897372	0.88	0.900	0.947368	0.923077
Random Forest Classifier FS	0.966667	0.963303	0.990566	0.976744	0.920000	0.926840	0.963123	0.943269	0.88	0.925	0.925000	0.925000
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000	0.900000	0.972294	0.901839	0.934431	0.92	0.925	0.973684	0.948718

Figure 4: Models performances in Long Beach with feature selection based on F-score

	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Test Accuracy	Test Recall	Test Precision	Test F1
Logistic Regression MI	0.846667	0.908257	0.883929	0.895928	0.806667	0.861905	0.876419	0.863455	0.86	0.875	0.945946	0.909091
Logistic Regression	1.000000	1.000000	1.000000	1.000000	0.893333	0.870996	0.980952	0.920748	0.92	0.925	0.973684	0.948718
KNN Classifier MI	0.880000	0.899083	0.933333	0.915888	0.826667	0.862338	0.899025	0.877336	0.88	0.900	0.947368	0.923077
KNN Classifier	0.873333	0.880734	0.941176	0.909953	0.820000	0.853680	0.899948	0.871434	0.76	0.750	0.937500	0.833333
Decision Tree Classifier MI	0.993333	0.990826	1.000000	0.995392	0.813333	0.870563	0.871014	0.869858	0.82	0.850	0.918919	0.883117
Decision Tree Classifier	1.000000	1.000000	1.000000	1.000000	0.986667	0.981385	1.000000	0.990471	0.92	0.950	0.950000	0.950000
Neural Network Classifier MI	0.860000	0.899083	0.907407	0.903226	0.800000	0.870996	0.863014	0.861603	0.90	0.900	0.972973	0.935065
Neural Network Classifier	1.000000	1.000000	1.000000	1.000000	0.853333	0.880519	0.918426	0.897372	0.88	0.900	0.947368	0.923077
Random Forest Classifier MI	0.993333	0.990826	1.000000	0.995392	0.806667	0.843723	0.885588	0.863362	0.84	0.850	0.944444	0.894737
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000	0.900000	0.972294	0.901839	0.934431	0.92	0.925	0.973684	0.948718

Figure 5: Models performances in Long Beach with feature selection based on mutual information

In the two tables above models performances on Long Beach data collection are illustrated. It can be noticed that in this case also the decision tree, together with Random Forest and Logistic Regression, provides good results. They present high scores of accuracy, recall, precision and F1-score especially when this is computed on the test sets, with an average value greater than 0.92. An interesting and unexpected result is that using features selection in this case reduces models performances, hence we kept all the attributes for performing predictions on this data collection.

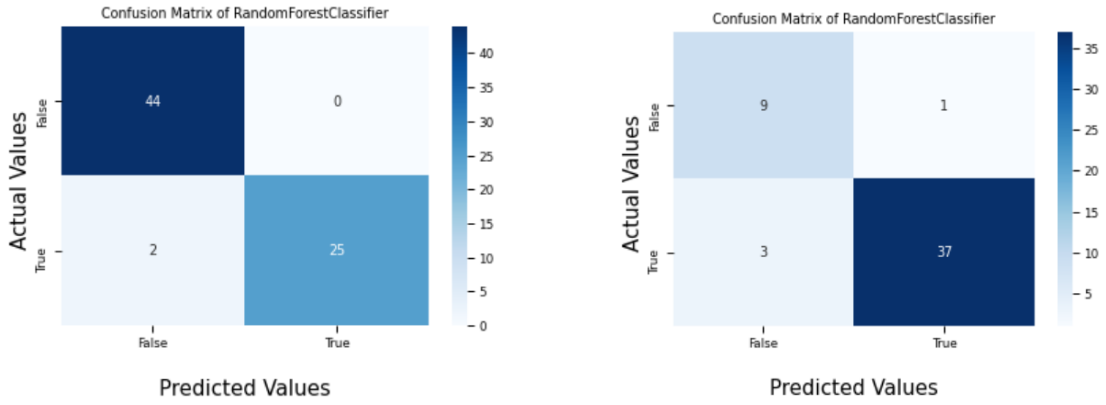

Figure 6: Confusion matrix for Cleveland and Long Beach

The confusion matrix reported above is just related to the Random Forest classifier, that was the one offering the best results. It can be seen that the number of incorrect predictions slightly increases in Long Beach, where we have a number of false negative of 3 and false positive of 1.

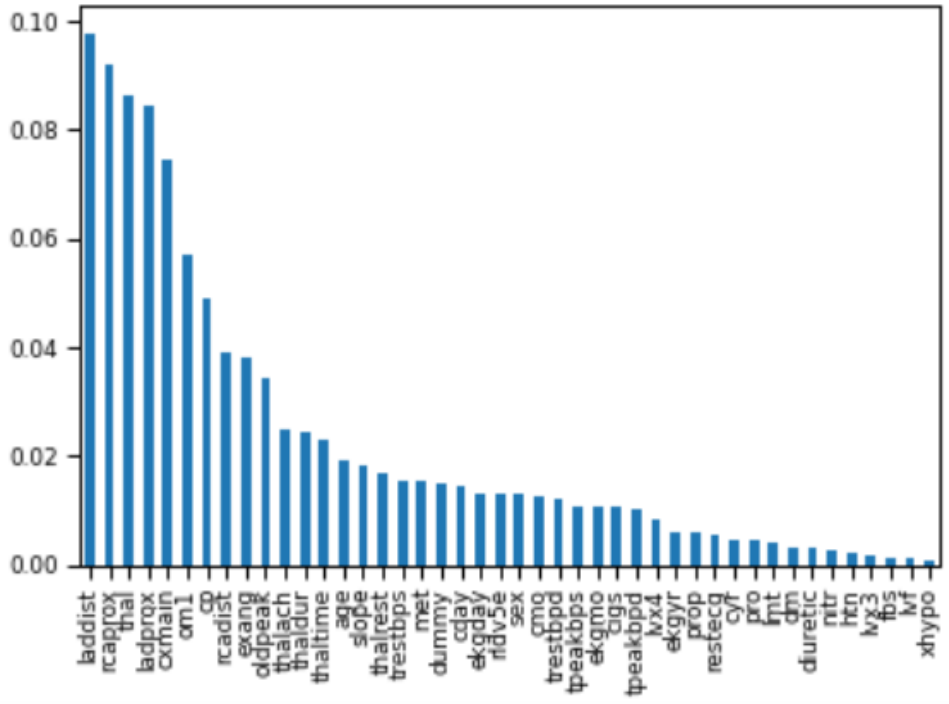

Figure 7: Features importance in Cleveland

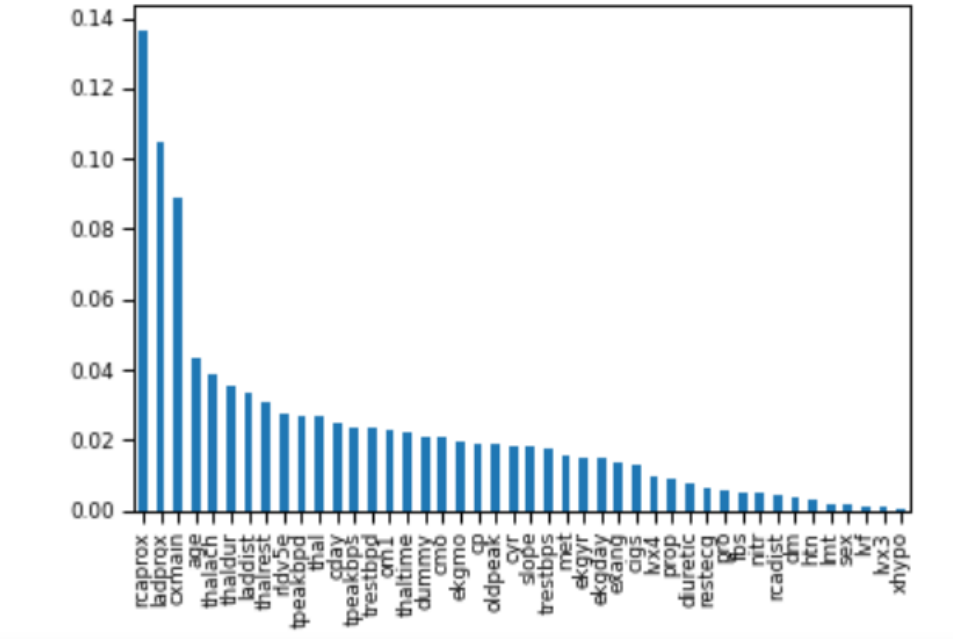

Figure 8: Features importance in Long Beach

The histograms above reports the features' importance for the Cleveland (Fig 7) and Long Beach (Fig 8) data collection. It is possible to observe that the features' relevance changes based on the geographical location. In Cleveland, the two most important attributes in the prediction are *laddist* and *rcaprox*, while in Long Beach the two most important one are *rcaprox* and *ladprox*. However, it is possible to notice a similarity: the attribute *rcaprox*, which is a measure referring to volume of the right coronary artery, is relevant in both the predictions.

4 Conclusion

4.1 Discussion

By looking at the results, it can be affirmed that the research questions have been partially fulfilled. Indeed, we discovered that the most important features in an ML model for predicting if a patient will experience a heart disease are the one related to medical parameters that measure the size of blood vessels (*rcaprox* or *laddist*). However, the answer to the second research question is that it is not possible to build a reliable ML model by using only these two attributes. Indeed, medical, physiological, and psychological information is necessary for having a medical picture about the heart condition of a patient: without this data the prediction model has low performances and does not provide trustful results. This was the main reason behind the choice of including in the model building also the attributes that appeared to be strong correlated. For what concerns the hypothesis, they were partially confirmed. First, we showed that it is possible to early predict heart disease it is possible, but for ensuring the reliability no data needs to be excluded. Second, we saw that are heart disease are strongly related to the blood vessels shape. Hence, they do not depend on the subject's age, by they are connected to the presence of hypertension (HT) or vascular disease (VHD) that might reduce the vessels' size. Finally, we are able to generalize the results obtained for the features' importance. Indeed, we saw that the volume of arteries is important in two different world parts, Long Beach and Cleveland. However, the impossibility of using half of the available data collections (Switzerland and Hungary) represented a limitation for our research. It is necessary to mention that in the work of Moan et al. [15] they used a hybrid algorithm for building the prediction model, that we on the contrary exclude from our analysis. This was done because by adopting the pure version of Random Forest, it is possible to have a clear view of the feature importance within the model. Moreover, by involving all the data collections' valid attributes in the model construction, we managed to obtain better performances than the ones reported in the study of Moan et al. [15]. The comparison among the two studies is observable on Table 1. However, it is possible only to compare the model's performances on the Cleveland data collection, as the others were excluded in the correlated study.

Table 1: Comparison between RF performances on Cleveland dataset in this study and the one of Moan et al.

Metric	THIS STUDY	MOAN ET AL.
Accuracy	0.97	0.86
Precision	1.00	0.87
F-score	0.96	0.92

4.2 Future work

There are different ways to improve and extend the work realized in this study. For example, it will be interesting to produce partial dependencies graphs in order to analyze the model behavior in relation to the predictors' trends. Another interesting extension, might be to gather the same information collected in UCI data-set, from people living in other parts of the world (maybe repeating the measure in Switzerland and Hungary considering that these data collection were corrupted). In addition, future research might focus on discovering different patients subgroups [36], which means that certain individuals are more likely to develop heart failure with some factors than others. In this way, it will be possible to offer people a personalized treatment based on the medical category they belong to.

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

[3] 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.

[4] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] A. Tattersall and M. J. Grant, "Big data–what is it and why it matters," 2016.

[12] 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.

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

[14] D. Maddux, "The human condition in structured and unstructured data," *Acumen Physician Solutions*, 2014.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," in *Proceedings of the 2016 international conference on management of data*, 2016, pp. 2201–2206.

[23] X. Chu, I. F. Ilyas, and P. Papotti, "Discovering denial constraints," *Proceedings of the VLDB Endowment*, vol. 6, no. 13, pp. 1498–1509, 2013.

[24] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *arXiv preprint arXiv:1208.1927*, 2012.

[25] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi, "A cost-based model and effective heuristic for repairing constraints by value modification," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005, pp. 143–154.

[26] J. Kaiser, "Dealing with missing values in data," *Journal of systems integration*, vol. 5, no. 1, pp. 42–51, 2014.

[27] S. Zhang, "Nearest neighbor selection for iteratively knn imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012.

[28] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[29] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.

[30] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[31] R. Féraud and F. Clérot, "A methodology to explain neural network classification," *Neural networks*, vol. 15, no. 2, pp. 237–246, 2002.

[32] Q. Song, H. Jiang, and J. Liu, "Feature selection based on fda and f-score for multi-class classification," *Expert Systems with Applications*, vol. 81, pp. 22–27, 2017.

[33] S. Ding, "Feature selection based f-score and aco algorithm in support vector machine," in *2009 Second International Symposium on Knowledge Acquisition and Modeling*, vol. 1. IEEE, 2009, pp. 19–23.

[34] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[35] S. Ronaghan, "The mathematics of decision trees, random forest and feature importance in scikit-learn and spark," Nov 2019. [Online]. Available: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark

[36] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and information systems*, vol. 29, no. 3, pp. 495–525, 2011.

A Correlation matrix for Cleveland and Long Beach

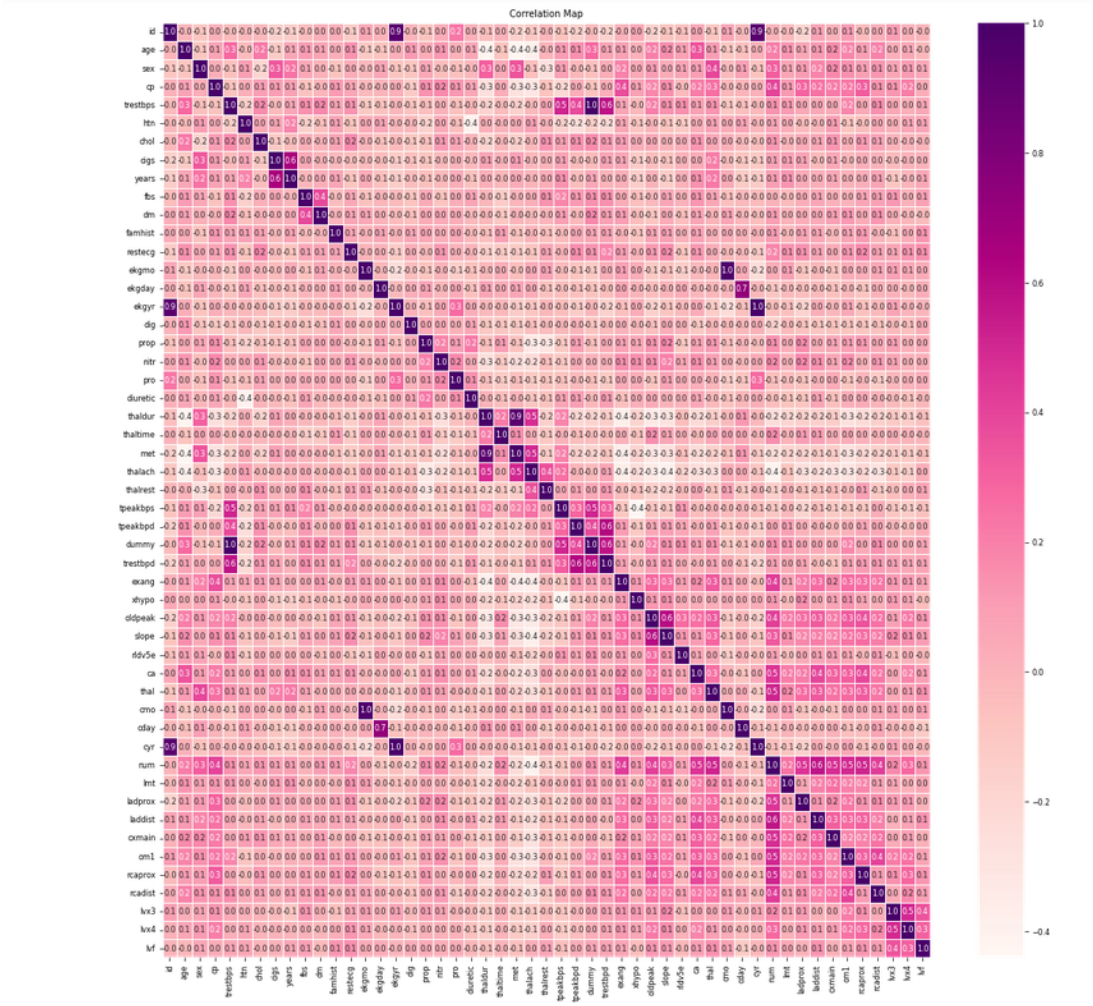

Figure 9: Correlation matrix for Cleveland

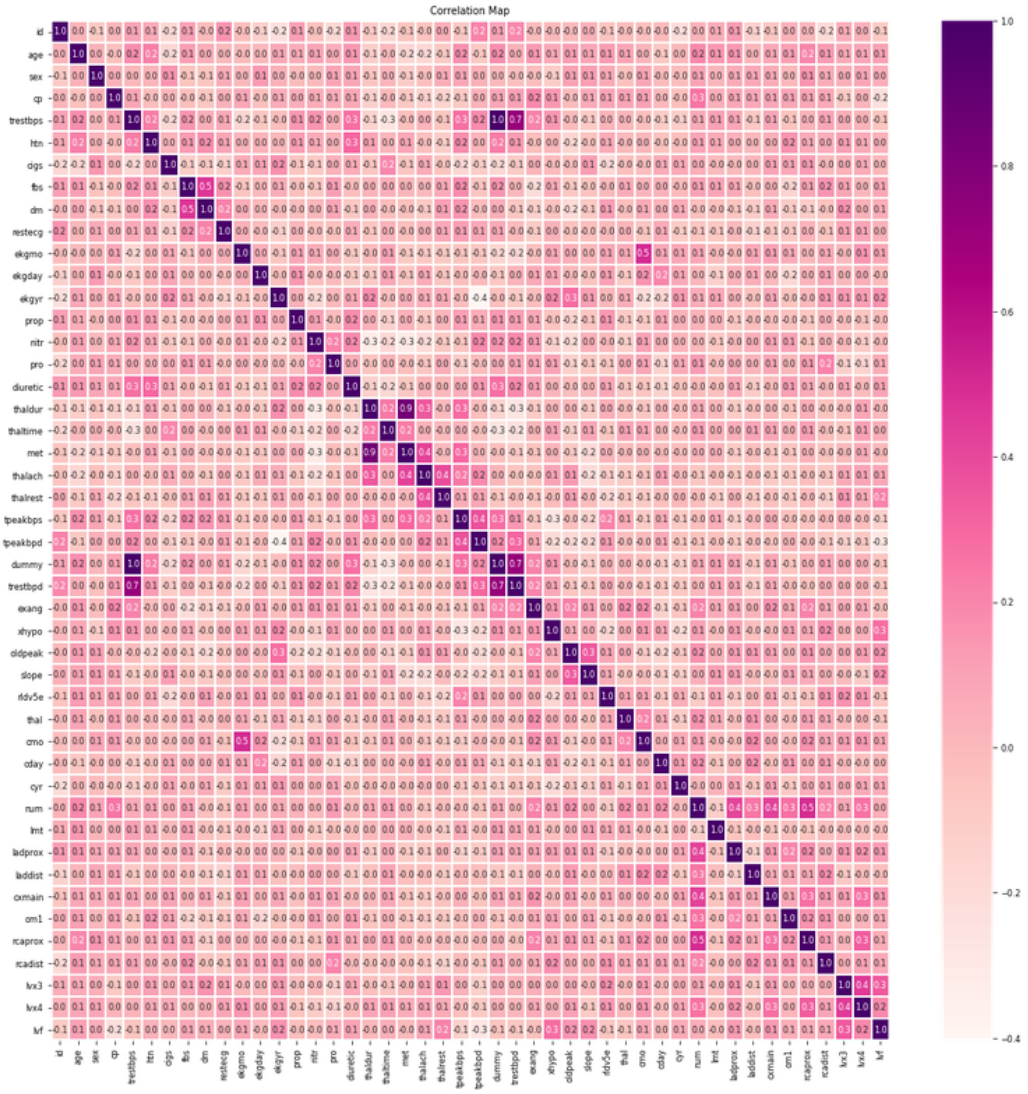

Figure 10: Correlation matrix for Long Beach