Prediction of heart disease for doctor's decision support

Fandio Njylla Esdras E.
BS in Computer Science at National Advanced School of Engineering Yaounde, CM,
fandioemma@gmail.com

Mukam Augusta Priscille
BS in Computer Science at National Advanced School of Engineering Yaounde, CM,
mukamaugusta5@gmail.com

I. Introduction

This document produced within the frame of the **Machine Learning Pipo Competition** presents a decision-making aid system useful for cardiologists to determine patients suffering from cardiac diseases.

II. Problem presentation and Solution

A. Problem presentation

Cardiovascular disease is the number one killer worldwide: more people die from cardiovascular disease each year than from any other cause. An estimated 17.7 million deaths are attributable to cardiovascular disease, that 31 percent of total global mortality. Among these deaths, an estimated 7.4 million are due to coronary heart disease and 6.7 million to a stroke (2015 estimation). Over three-quarters of deaths from cardiovascular disease occur in low- and middle-income countries, mainly in Africa.[1]
On the other side in Cameroon in particular, there are only 40 cardiologists working (2014 figures) for a population of more than 22 million inhabitants, population whose rate of prevalence with hypertension (one of the causes of heart disease) is 30 percent.[2].
From these two analyzes, it can be concluded that in Cameroon, there is a small number of cardiologists and a very large number of patients and potential patients. We also deduce an overload of work from cardiologists preventing them from doing their job properly and assisting all patients. Hence the question: **How can the doctor's work be optimized to allow him to manage the large number of patients effectively?**

B. Proposed Solution

To overcome this problem, we proposed a decision support system named **CardioHelp**, that will allow the doctor to quickly know the critical cases on which he will have to focus. In practice, this is a questionnaire which, when completed, makes it possible to know whether the patient is probably suffering from an illness and whether his case should be monitored.

C. Motivation

Cardiovascular disease has serious socio-economic repercussions in in terms of cost, health care, absenteeism and national productivity on individuals, families and communities (each treatment costs 3 Millions for the patient and 3 Millions for the state).
Furthermore, according to statistics, 2 out of 3 patients do not know that they have the disease until it enters the critical phase. This is due to the fact that the consultations necessary to treat it are generally expensive and many people prefer to avoid them or don't even know them.
Our motivation to develop this project lies in reducing consultation costs so that more people can access them and know their situation much sooner.

III. Methodology

To be able to build a reliable model allowing to determine the presence or not of a heart disease, we carried out several stages:

1) The research of the important factors entering into the decision
2) The Research/development of a dataset
3) The creation and validation of a model
4) The development of an interface for the doctor

A. Research of the important factors entering into the decision

In this phase, it was a question for us of determining all the factors determining the condition of a patient.
After various researches carried out on sites with medical aim and in medical articles, it was observed that the principal mean used by the cardiologist to know the state of the heart of a patient is to examine the results of his electrocardiogram (ECG). It is a quick examination taking only a few minutes, painless and non-invasive, devoid of any danger. It can be done in a doctor's office, hospital, or even at home. However, its interpretation remains complex and requires methodical analysis and some experience from the clinician. It makes it possible to highlight various cardiac anomalies and has an important place in diagnostic examinations in cardiology.
[3] presents the 76 attributes that can be analyzed from an electrocardiogram. [4] presents a study made by doctors and engineers to determine between those attributes, the different ones that has a big importance to predict the heart disease. From this document, 14 attributes were the most important for the prediction of heart disease:

- age
- gender
- chest pain type
- resting blood pressure in mm/Hg
- serum cholesterol in mg/dl
- fasting blood pressure¿120mg/dl ? (yes/no)
- resting electrocardiographic results
- maximum heart rate achieved
- exercise induced angina
- ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels colored by flourosopy
- thalassemia

B. The Research/development of a dataset

After having the different features that are used to predict the health disease, we looked for data from the electrocardiograms of Cameroonian patients. But after several searches, we realized that structured data (in digital form) does not exist. We therefore conducted studies to find out whether the results of the EKG depend on factors such as region, race, country of origin ... and it emerged that the results of the ECG depend little on these factors. We therefore continued our search to find databases from elsewhere that could help us. The database found comes from the Cleverand hospital in the United States ([3]).
After having analysed this dataset we did some operations such as : clean, transform (by one hot encoding...), ... After those operations the size of the dataset was 303. We add the following observations from the dataset:

- There are 138 results corresponding to women and 165 corresponding to men
- 45.54 percent of the patients did'nt have heart disease.

C. The creation and validation of a model

As we did'nt have a lot of data, we pre-selected 4 principal models:

- K-Nearest Neighbour Classification (k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.) which had an accuracy of 77.05 percent
- Logistic Regression (Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable) which had an accuracy of 86.89 percent
- XGBBoost (It is an implementation of gradient boosting machines created by Tianqi Chen, now with contributions from many developers) which had an accuracy of 86.88 percent
- Random Forest Classifier which had an accuracy of 88.52 percent.

The different models were implemented in **Keras**.

1) Presentation of the Random Forest Classifier: To understand that concept, we need first of all to understand the concept of decision trees:
A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. They are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It correct the overfitting habit of the decision trees on the training set.

2) Results: After having trained those models, the best model we had was based on a Random Forest Classifier with 1000 decision trees inside. His result was 88.52 percent on accuracy.

D. The development of an interface for the doctor

We have developed a web application to allow the doctor, after an electrocardiogram, to enter the various results necessary to predict the client's condition. The result returned after validation of the form is the probability that the patient is sick. We developed this interface using Angular and Flask as a Restful server

IV. Conclusion and Future work

The results obtained from the model a little bit satisfying but not enough to be used in hospitals. To correct that, we need to collect more representative data (coming from Cameroon if possible) and to equalize the ratio of illness patients. After that to be sure the model works, we need to validate it with the help of a doctor.

References

[1] https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] http://africapress.over-blog.com/article-cameroun-sante-seulement-40-cardiologues-en-service-cameroon-122951713.html

[3] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] S.Chellammal, R. Sharmila. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019. https://www.ijrte.org/wp-content/uploads/papers/v8i2S3/B11630782S319.pdf