Prediction of heart disease for doctor's decision support

FANDIO Njylla Esdras E., MUKAM Augusta Priscille
fandioemma@gmail.com, mukamaugusta5@gmail.com

BS in Computer Science at National Advanced School of Engineering Yaounde, CM

Problem presentation

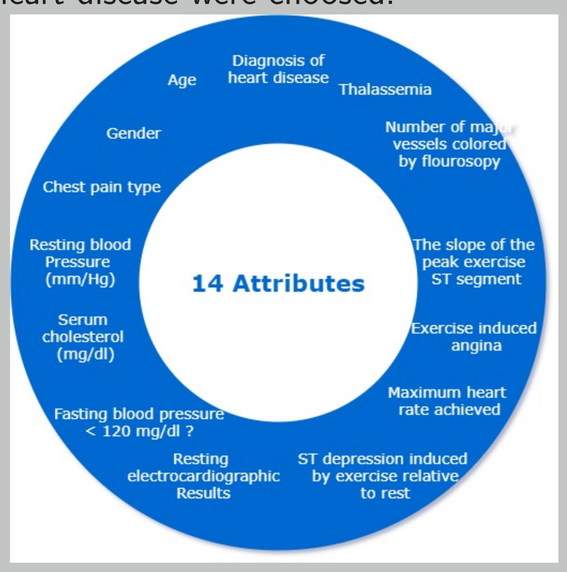
- **Cardiovascular disease** is the number one killer worldwide (**31 percent of total global mortality**). Over three-quarters of deaths from cardiovascular disease occur in low- and middle-income countries, mainly in Africa.
- In Cameroon in particular, there are only **40 cardiologists** working for 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**. 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?**
- To overcome this problem, we proposed a decision support system named **CardioHelp**, that will help 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.

Motivation

- Cardiovascular disease has serious socioeconomic repercussions in in terms of **cost**, **health care**, **absenteeism** and **national productivity** on individuals, families and communities.
- 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 high cost of consultations and the lack of information.
- Our motivation to develop this project lies in reducing consultation costs so that more people can access them and know their situation much sooner.

Methodology

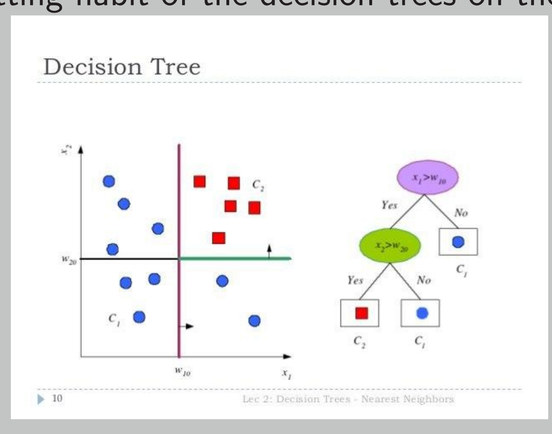
- **The research of the important factors entering into the decision**
  - The principal mean used by the cardiologist to know the state of the heart of a patient is to examine the results of his electrocardiogram (ECG) which highlights many attributes, but 14 attributes the most important for the prediction of heart disease were choosed.



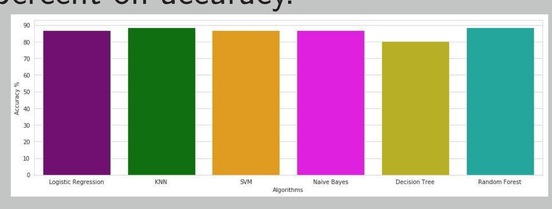
- **The research/development of a dataset**
  - The next step was to look for data from the electrocardiograms of Cameroonian patients. Unfortunately we didn't find any structured data. From then on, we continued our researches along axes such as region, race, country of origin ... to find databases from elsewhere that could help us. So the database found comes from **the Cleverand hospital** in the United States.
  - After having analysed this dataset we did some operations such as : **clean**, **transform**, ... 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.

- **The creation and validation of a model**
  - **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** (it 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.



- **The development of an interface for the doctor** We developed a web application to helps 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.

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.

Machine Learning Pipo Competition, 2020