Heart Disease Prediction

Xavier Santos
Departamento de Eletrónica, Telecomunicações e Informática
Universidade de Aveiro
Aveiro, Portugal
xavier@ua.pt

***Abstract*—This project aims to find the most efficient model to predict the presence of a heart disease on a patient based on 14 preconditions and exam results from a single hospital and then test the trained models against data from hospitals in different hospitals and countries.**

***Index Terms*—machine learning, dataset, heart disease, prediction, logistic regression, naive Bayes, k nearest neighbors, decision tree, random forest**

I. Introduction

In this study, it was analyzed the data form the UCI Machine Learning Repository [5] regarding patient data used to ascertain the presence of a heart disease. In order to predict said disease were used five models of prediction: Logistic Regression, Naive Bayes, K Nearest Neighbors, Decision Tree and Random Forest.

II. Dataset

This dataset used for training contains the data of 303 patients from Cleveland [4] described by 14 attributes each; 5 of which are numerical values while the other 9 represent categories. The ”goal” field is represented by a binary value representing the presence or absence of heart disease in the patient. There were other 3 other datasets from Hungary [1], Switzerland [2], [3] and Long Beach [4] in which the models obtained from the first set were tested after the training. Most of these datasets, however, had many missing values.
The features used are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram, maximum heart rate, exercise induced angina, ST depression, ST slope, number of major vessels and Thallium stress test results. An example of the data is show in Table I.

TABLE I
Data head

	Age	Sex	ChestPainType	RestingBloodPressure	Cholesterol	FastingBloodSugar	RestingECG	MaxHeartRate	ExerciseInducedAngina	ST_depression	ST_slope	NumMajorVessels	ThalliumStressTest	Diagnosis
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	1
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
5	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0
6	62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0	0.0	3.6	3.0	2.0	3.0	1
7	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0
8	63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0	0.0	1.4	2.0	1.0	7.0	1
9	53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0	0.0	7.0	1

III. Data analysis

The contents of Table II show some relevant statistics about the data that is being used for training.

Continuing the analysis, it was verified that some features have more impact on the patient's diagnosis than others. For instance, if a patient has or not exercise induced angina, this is a much stronger indicator of a heart disease than his number of major vessels or fasting blood sugar value as it can be seen in Table III and Figure 1.

TABLE II
Data description

	Age	Sex	ChestPainType	RestingBloodPressure	Cholesterol	FastingBloodSugar	RestingECG	MaxHeartRate	ExerciseInducedAngina	ST_depression	ST_slope	NumMajorVessels	ThalliumStressTest	Diagnosis
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.44	0.68	3.16	131.69	246.69	0.15	0.99	149.61	0.33	1.04	1.60	1.38	5.06	0.46
std	9.04	0.47	0.96	17.60	51.78	0.36	0.99	22.88	0.47	1.16	0.62	6.22	4.47	0.50
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00	3.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	3.00	0.00
50%	56.00	1.00	3.00	130.00	241.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00	3.00	0.00
75%	61.00	1.00	4.00	140.00	275.00	0.00	2.00	166.00	1.00	1.60	2.00	1.00	7.00	1.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	54.44	54.44	1.00

TABLE III
Data correlation

	Diagnosis
Diagnosis	1.000000
ExerciseInducedAngina	0.431894
ST_depression	0.424510
MaxHeartRate	0.417167
ChestPainType	0.414446
ST_slope	0.339213
Sex	0.276816
ThalliumStressTest	0.232627
Age	0.223120
RestingECG	0.169202
RestingBloodPressure	0.150825
Cholesterol	0.085164
FastingBloodSugar	0.025264
NumMajorVessels	0.020634

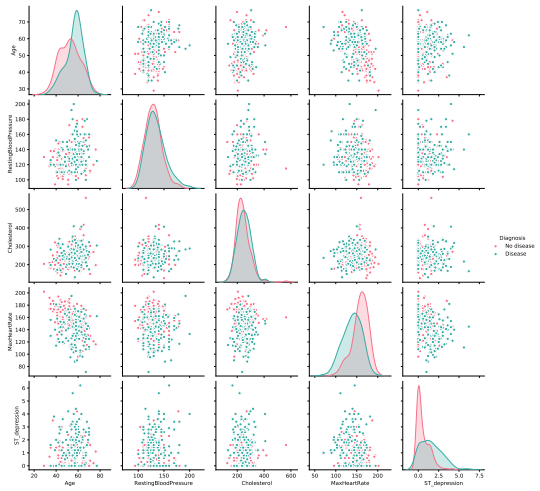

Fig. 1. Scatter plots of the correlation between the features and the diagnosis

From the gathered data it can also be concluded that the percentage of patience without heart problems in this sample is 54.13%.

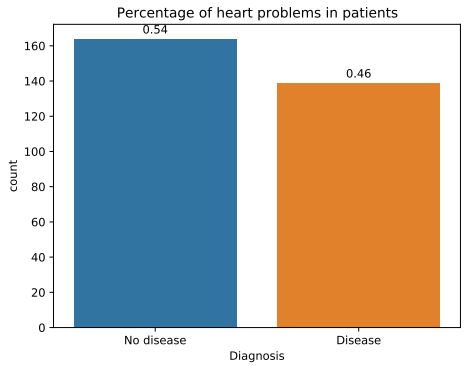

Fig. 2. Distribution of heart disease in patients

IV. Data Preprocessing

Before training the models the data had to be prepared by filling the missing values on some of the features. This was achieved by giving the mean value for that feature if it was a numeric value or the most common value if it was categorical. The data was then shuffled and split in two sets: one for training and another one for testing and cross-validation with the distribution of 80% and 20% respectively.

V. Model training and tuning

A. *Logistic Regression*

B. *Naive Bayes*

C. *K Nearest Neighbors*

D. *Decision Tree*

E. *Random Forest*

VI. Results

VII. Conclusion

References

[1] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
[2] University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
[3] University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
[4] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.
[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.