Detection of Heart Disease

Mathieu Le Cam

Abstract

In this project, I used the heart disease dataset shared by the University of California Irvine on Kaggle. I tried to answer the following questions: how accurately can we detect a heart disease with given measurements and a trained machine learning algorithm? What are the main features to consider?

I used and compared the accuracy of several classifiers to evaluate if the human subject is healthy or presents a heart disease. The three best performing classifiers are the Linear SVM, Gaussian process and Neural net with an accuracy of 85.2%.

Motivation

World Health Organization reported in 2016 that 17.9 million people died from Cardio Vascular Diseases (CVDs) in 2016, representing 31% of all global deaths.

- CVDs are the number one cause of death globally: more people die annually from CVDs than from any other cause.
- Over three quarters of CVD deaths took place in low- and middle-income countries.
- Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2015, 82% are in low- and middle-income countries, and 37% are caused by CVDs.

It is of **high importance** to **improve** the **methods** for **early detection of CVDs**.

Dataset(s)

This study is based on the heart disease dataset shared by the University of California Irvine on Kaggle.

The database contains 303 instances of measurements taken on human subjects; and 14 attributes describing the health of the human subject (age, sex, heart rate, cholesterol, chest pain, blood pressure at rest...). The presence of heart disease in the patient is represented by a binary value: 0 (no disease) and 1 (presence of disease).

Data Preparation and Cleaning

No data preparation or cleaning was required in this study.

The dataset provided did not present any missing data; the attributes are integers or floats. The dataset could be used in the current state.

```
df.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

```
df.dtypes
```

```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

Research Question(s)

In this study, I focused on the following questions:

- How accurately can we detect a heart disease from given measurements and using different machine learning algorithms?
- What are the main features to consider in the prediction algorithm?

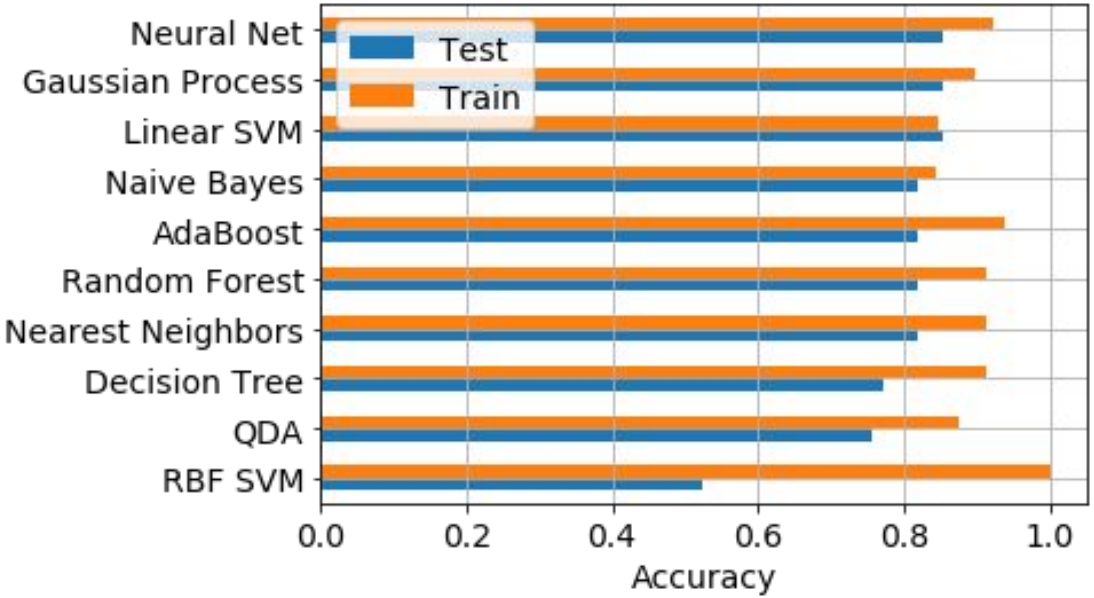
Methods

I used the following approach for the **classification problem** of this study:

1. **Preliminary analysis** of the data: check for missing values and data types (replaced if required)
2. **Inspect** the data: look at the distribution of each attribute (outliers?) and possible intercorrelation between attributes
3. **Split** the data into a training (on which the model is fitted) and testing set (used for model evaluation)
4. **Normalize** the data so that all the attributes are equivalent in magnitude
5. **Train** different classifiers and **compare the accuracies** on the test set
6. Analyse the **main important features** identified by the classifiers

Findings: Classifiers Accuracy Comparison

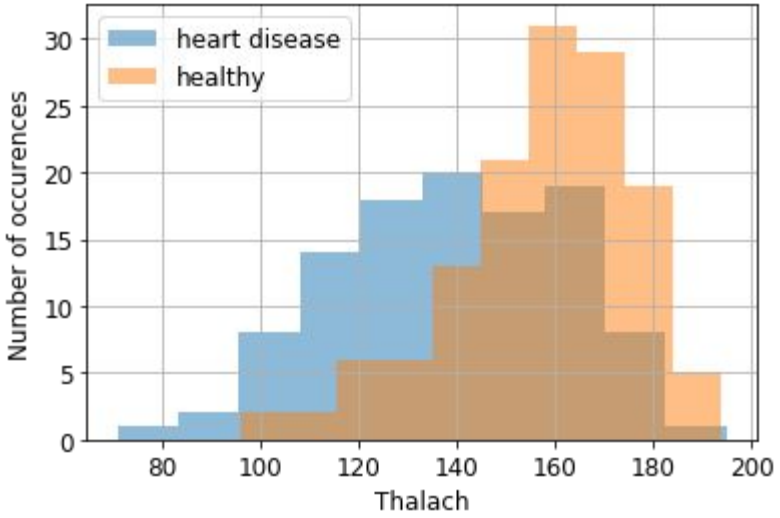
Ten different classifiers were trained and tested on unseen data to evaluate if the patient presents a CVD. The three best performing classifiers are the **Neural Net**, **Gaussian Process** and **Linear SVM** with an accuracy of **85.2%** on the testing set.

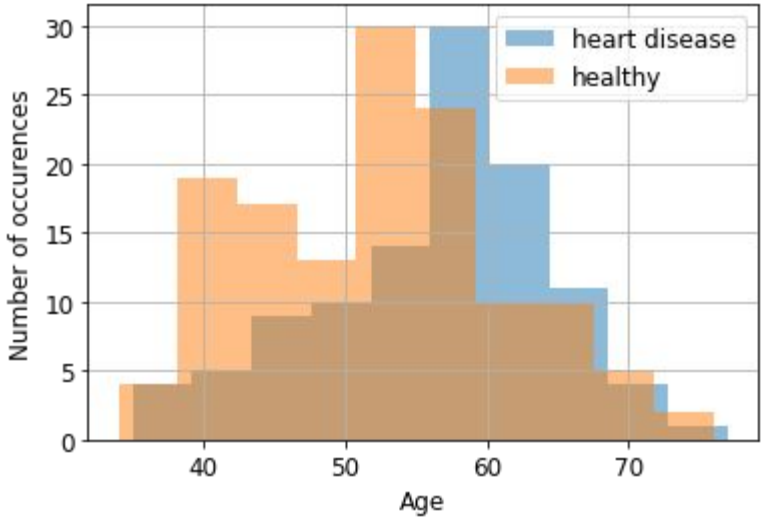


Findings: Most Important Features

Valuable information can be extracted from the classifiers such as the most important features. Out of the 13 attributes, 5 have a significant impact on the evaluation of the presence of CVD: age, heart rate, cholesterol, blood pressure at rest and ST depression (on electrocardiogram) after exercise.

The two figures below give some additional details on the impact of the age and maximum heart rate (thalach) for patients with and without CVD. The histograms show that older persons tend to get a CVD and healthy persons can reach a higher maximum heart rate.





Limitations

The findings from this work were limited by the data available.

There is a few number of volunteers (303) who accepted to share their information in the open-source dataset. With a greater amount of participants, more variety in the data would be collected for an improved prediction accuracy.

Considering the limitation of the time spent on the project, the machine learning algorithms developed for the course could be further improved looking at methods for features selection, optimizing the training process to avoid overfitting, ...

Conclusions

Three classifiers outperformed the others on the testing set with an accuracy of **85.2%** classifiers: the **Neural Net**, **Gaussian Process** and **Linear SVM**. The classifiers could accurately predict if the patient presented a CVD or not using the 13 attributes provided.

5 attributes out of the 13 provided in the study were important in the detection of CVD: **age, heart rate, cholesterol, blood pressure** at rest and **ST depression** (on electrocardiogram) after exercise. An analysis showed that older persons tend to get a CVD and patients with CVD cannot reach as high maximum heart rate as healthy ones.

Acknowledgements

This analysis was performed thanks to the open-source data shared by the University of California Irvine on Kaggle. The heart disease dataset contains valuable information describing the health of human subjects (age, sex, heart rate, cholesterol, chest pain, blood pressure at rest...) and the presence ornot of a heart disease in the patient.

I unfortunately did not get any feedback on my work.