

**FINAL PROJECT REPORT**

Machine Learning

Exploratory Data Analysis and Predictions of Heart Disease with UCI Heart Disease Data Set

Broniewski, Adam

Jahangir, Khushnur Binte

**Supervisor**

Coma-Puig, Bernat

**Universitat Politècnica de Catalunya**

June 2022

Table of Contents

[1 Abstract 3](#_Toc105407801)

[2 Introduction 3](#_Toc105407802)

[3 Related Previous Work 3](#_Toc105407803)

[4 Exploratory Data Analysis 3](#_Toc105407804)

[4.1 Pre-processing 3](#_Toc105407805)

[4.2 Feature Selection 4](#_Toc105407806)

[4.3 Visualizations 4](#_Toc105407807)

[4.4 Clustering 4](#_Toc105407808)

[5 Predictive Modeling 4](#_Toc105407809)

[5.1 Modeling Methods Considered 4](#_Toc105407810)

[5.1.1 Logistic Regression 4](#_Toc105407811)

[5.1.2 Decision Tree 4](#_Toc105407812)

[5.1.3 Random Forest 4](#_Toc105407813)

[5.1.4 Neural Network 4](#_Toc105407814)

[5.2 Validation Protocol 4](#_Toc105407815)

[5.3 Results 4](#_Toc105407816)

[5.4 Final Model Estimation and Performance 4](#_Toc105407817)

[6 Conclusions 4](#_Toc105407818)

[7 Extensions and Limitations 4](#_Toc105407819)

[7.1 Dataset: 7](#_Toc105407820)

[7.2 Problem: 7](#_Toc105407821)

[7.3 Why this dataset: 7](#_Toc105407822)

[7.4 Dataset composition: 7](#_Toc105407823)

[7.5 Other References: 7](#_Toc105407824)

[7.6 Project Title: 7](#_Toc105407825)

1. A brief description of the work and its goals, data available, and any additional information that you may have used.

2. Related previous work (if applicable)  
3. The data exploration process, including: pre-processing, feature selection/extraction, visualization,

clustering, etc.  
4. Modeling methods considered, validation protocol and the reasons why the choices were made. 5. Results obtained with each method used (along with best set of parameters), comparison of results. 6. Final model chosen and an estimation of its generalization performance.  
7. Scientific and personal conclusions  
8. Possible extensions and known limitations.

# Abstract

# Introduction

The goal of this project is to explore the [UCI - Heart Disease Data Set](https://archive.ics.uci.edu/ml/datasets/Heart+Disease) to discover trends in the data and predict whether a patient would have heart disease based on medical attributes.

The dataset has 303 instances and 14 attributes that are a combination of categorical and real values, which provides room for experimentation with different models and approaches to data pre-processing. There are also 61 other relevant papers that make use of this dataset as identified by UCI website.

# Related Previous Work

# Exploratory Data Analysis

## Pre-processing

### Reformatting

The dataset column names were renamed from the originally abbreviated version to a full naming to make it more understandable during the exploratory phase.

### Encoding

8 of the attributes are categorical and encoded with integer values. All the categorical variables use label encoding. Label encoding was maintained for some of these categories[[1]](#footnote-1) as a review of the dataset indicated an order of severity. The remaining categorical variables[[2]](#footnote-2) were treated with one-hot encoding as there is no relationship or order between each value. The target category of heart disease had increasing severities of sickness, and was simplified and encoded as either healthy or sick.

### Test and Train Data Set Split

The data set was split at the early stage of pre-processing to eliminate the risk of data leakage. All transformations to the training dataset will be applied to the test dataset independently. To test/train split is stratified to preserve that same proportion of target category examples as observed in the original data set.

It’s important to note that the only step of the process that could introduce data leakage is during scaling before the model is trained. This is mitigated with the use of a standard scaler made from the training data set.

### Missing Values

The data set is quite clean, without missing values. 6 values in major vessel count and thalassemia were found with a “?”. Given the small number of missing values, the rows were dropped.

### Outliers

To detect outliers, data that was outside of 150% of the interquartile range (1.5\*IQR) was identified. Based on analysis, there were very few values that fell outside of the 1.5\*IQR. Given box plot visualizations in Figure 4‑1, only the cholesterol value above 450 was deemed to be noise and was removed from the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | |  |
|  | |  | |
| Figure ‑ Outlier detection and exploration of numerical data | | | |

## Visualizations

Visualizations are used to see if there are interesting trends or hypothesis that can be tested. This also provides insight into which models should be tested and which attributes are expected to have the most impact on the prediction.

A picture containing diagram

Description automatically generated

Figure ‑2 Boxplots of each attribute

The boxplots in Figure 4‑2, show that the data is in different ranges and will need to be normalized before training models. It also clearly shows the categorical variables that were label encoded. Resting blood pressure appears right skewed, and max heart rate achieved appears left skewed. Although this skew would not impact tree-based models, we will apply a square root and square on the right and left skewed data respectively for our other models.

A picture containing diagram

Description automatically generated

Figure 4‑3 Pairwise scatterplots and boxplots

## Clustering

## Feature Selection

Initial data exploration comparing numerical data against sex and the target category showed potential trends:

* age and resting heart rate seem correlated with heart disease
* higher max heart rate is associated with individuals without heart diseases
* cholesterol doesn't have a significant impact on heart disease
* when a participant had ST depression that is seen after exercise, they were more likely to suffer from heart disease

# Predictive Modeling

## Modeling Methods Considered

### Logistic Regression

### Decision Tree

### Random Forest

### Neural Network

## Validation Protocol

## Results

## Final Model Estimation and Performance

# Conclusions

# Extensions and Limitations

Import

- DataSplit

- final\_text vs training with stratify

Data Exploration and Analysis

- initial visualizations

- continuous variables

- distribution (normal? logarithmic?)

- comments on continuous attribute impact for target

- categorical variables

- counts vs target vs sex

- comments on categorical attribute impact for target

- deep dive into ST\_depression, which looks especially interesting

- evaluate correlation between attributes

- comment on correlation + hypothesis of which attributes will have most impact on target

Metrics

- comment on which metric we are using for choice and why

- create base dataframe for all appends

Models

- if RandomForest:

Model Contextualization + Improvement

- use RandomForest to determine feature importance

- check these features against correlation from EDA + comment

- create subset of x\_train with only features of importance

for algorithm in (RandomForest, LogisticRegression, NeuralNetwork, DecisionTree)

- build simple model with fullset

- collect metrics

for subset in (fullset, important subset)

- tune parameters of model (argument in function call + which attributes used)

- collect metrics

- choose best model (i.e. best parameters + subset used) for each algorithm

Algorithm Choice

- choose best algorithm (could end up with full vs sub set competing)

Final Test

- Test algorithm with final\_test data

- collect metrics

- discuss results

Questions:

- processing our test data -> should be done through a function similar to lab08?

- how to choose the metric used scoring and choosing the model/algorithm?

- can I tell if some of my attributes are just noise in my neural network?

## Dataset:

We will use the

## Problem:

Predict whether a patient has heart disease.

## Why this dataset:

This dataset was chosen as it meets all the requirements of the project scope. This was the most important factor for us so that we can focus on learning and understanding fundamental machine learning requirements without getting stuck on unclean data or a difficult dataset.

Interestingly, there are 14 main attributes that are used in other papers that make use of the dataset, although there are 75 attributes that can be chosen from for prediction. This provides a lot of room for experimentation and improvement on past research.

## Dataset composition:

There are several datasets available from different locations (Cleveland, Hungary, Switzerland). The UCI description indicates the “Cleveland database is the only one that has been used by ML researchers to this date”. The Cleveland database is described as being “processed”; however it is not known to what extent or what exactly this processing includes.

|  |  |
| --- | --- |
| Number of instances | 303 |
| Number of attributes | 14-75 |
| Attribute types | |
| categorical |  |
| integer |  |
| real |  |
| Analysis type | classification |

## Other References:

There are 61 other relevant papers that make use of this dataset as identified by UCI website

## Project Title:

Heart Disease Machine Learning Project 2022

1. Label Encoding: chest pain type, resting electrocardiographic, peak exercise ST slope, and thalassemia [↑](#footnote-ref-1)
2. One-hot Encoding: sex, fasting blood sugar, exercise induced angina [↑](#footnote-ref-2)