Phase 1

Data Analysis and Preperation

Alexander Garcia

December 2024

Contents

1 Introduction

The goal of this phase is to explore the data by plotting its distribution, checking for outliers, null values, imbalance of the target variable and then normalizing before building a neural network.

Why heart disease? Heart disease can have an impact directly, through the person dealing with heart disease themselves, or indirectly through family members or friends. Heart disease can lead to death or massive lifestyle changes with surgeries, pacemakers, and more. If we can predict, then we can react and take necessary precautions.

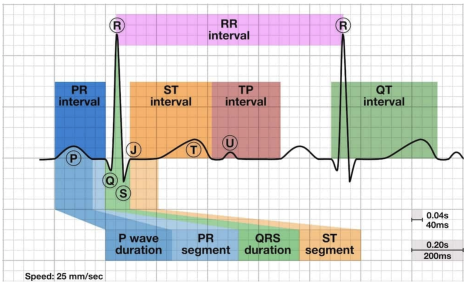
2 Dataset

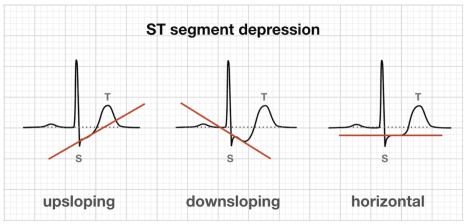
The "Heart Disease" dataset was obtained from Kaggle [Mex] and is a combination of data from Cleveland, Hungary, Switzerland, and Long Beach. The data file contains 1190 samples with 11 features and 1 target variable. The target variable is binary, where 1 represents positive for heart disease and 0 is normal. A couple data points reference, "Angina", which is a type of chest pain caused by reduced blood flow to the heart[Sta]. The data points and brief description of them:

1. Age
2. Sex
3. Chest Pain Type
4. Resting Blood Pressure
5. Serum Cholesterol (total cholesterol in blood)
6. Fasting Blood Sugar - amount of sugar in your blood after you haven't eaten for at least 8 hours
7. Resting ECG results
8. Max Heart Rate
9. Exercise Induced Angina - chest pain during exercise

10. Oldpeak is an a drop below baseline in ECG
11. ST slope - refers to slop of the ST segment in ECG
12. Target - heart disease

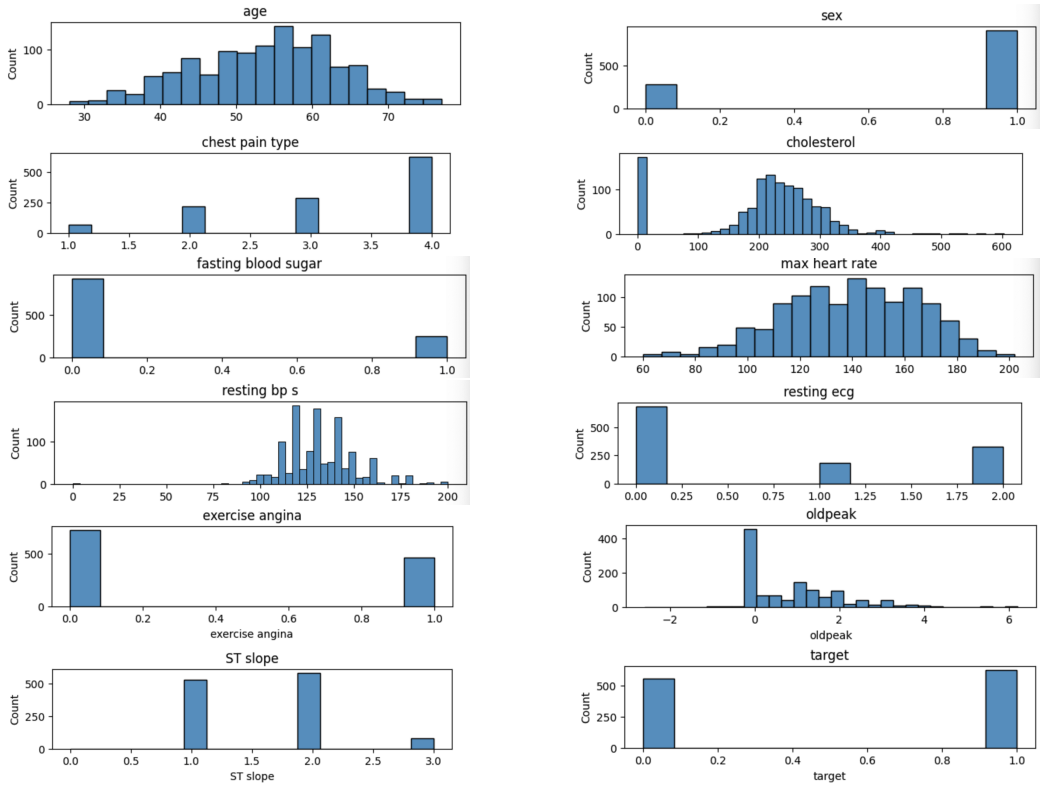
Figure 1: Sample ECG with information on the slope / depression [BB]





2.1 Data distribution

The data is in a cleaned format in the sense that it is already numerical representations, but without normalizing the values. In the charts below, the distributions show some imbalance of sex and fasting blood sugar results. However, the target variable has a nice balance.



2.2 Data Statistics

Below is the output of running the describe method on the dataframe. Most of the participants are between 50 and 60 years old. Age and maximum heart rate have nice distributions that do not require modifying or filling any null values.

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
count	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000
mean	53.720168	0.763866	3.232773	132.153782	210.363866	0.213445	0.698319	139.732773	0.387395	0.922773	1.624370	0.528571
std	9.358203	0.424884	0.935480	18.368823	101.420489	0.409912	0.870359	25.517636	0.487360	1.086337	0.610459	0.499393
min	28.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	60.000000	0.000000	-2.600000	0.000000	0.000000
25%	47.000000	1.000000	3.000000	120.000000	188.000000	0.000000	0.000000	121.000000	0.000000	0.000000	1.000000	0.000000
50%	54.000000	1.000000	4.000000	130.000000	229.000000	0.000000	0.000000	140.500000	0.000000	0.600000	2.000000	1.000000
75%	60.000000	1.000000	4.000000	140.000000	269.750000	0.000000	2.000000	160.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	4.000000	200.000000	603.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	1.000000

Figure 3: Data statistics

2.3 Data Normalization

Data normalization is the process of mutating the data, so most of the values fall between the range of 0 and 1. For this data set, the mean normalization technique was used, which calculates and then subtracts the mean for each feature in the model. The formula is as follows:

$$X_{new} = \frac{X - X_{\mu}}{X_{max} - X_{min}}$$

3 Conclusion

In summary, I have identified the preliminary steps for normalizing, cleaning, and viewing the data distributions. Data analysis and preparation are crucial parts of a neural network. Proper cleaning can help performance while addressing potential noise that could be caused by outlying or null data values.

References

[BB] Ed Burns and Robert Buttner. *The ST Segment*. URL: https://litfl.com/st-segment-ecg-library/.

[Mex] Mexwell. *Heart Disease Dataset*. URL: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset?select=heart_statlog_cleveland_hungary_final.csv.

[Sta] Mayo Clinic Staff. *Angina*. URL: https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373.