DATA MINING IN HEALTHCARE

Goktug Cengiz

June 2019

Contents

1 Introduction

Data Mining, which is used to reveal confidential, valuable, usable information from a large amount of data and provide strategic decision support; it has created a new perspective in the use of health data and has become a method of increasing prevalence. The aim of this study is to show an example about the use of Data Mining in health.

2 Data Mining for Healthcare

There are several Data Mining specific fields such as health, business, education, etc. One of the most significant specific field is health. The basis of health system policies and managerial decisions are data and information which obtained from data. So as to be appropriate and effective for purpose of health policies and decisions depend on reliable, up-to-date and accurate data. The purpose of health information systems is to produce useful information from large amounts of health data. This information is used at the patient level for better health service delivery, better management of health institutions, effective use of resources and health policies. Health data are collected by many institutions, particularly hospitals, other health institutions, insurance companies and related public institutions. Today, the increase in the volume of digital data has created new problem areas. The main ones are; develop methods or systems for processing large amounts of multidimensional and complex data; develop methods or systems for processing new types of data; develop methods, protocols or infrastructure for processing distributed data; develop models related to the use and security of data.

The first thing that comes to mind when talking about Big Data is Data Mining. Data Mining is a process that explores patterns and relationships within the data together with the use of many analysis tools and uses them to make valid estimates. The aim of Data Mining is to create decision-making models to predict future behavior based on the analysis of past activities. Data Mining, which has been used since the 1990s in order to uncover confidential, valuable, usable information and provide strategic decision support; In addition to finding answers to these problem areas, it has created a new perspective in the use of health data and has become a method that continues to increase rapidly.

3 Data set

The chosen data set downloaded from the Kaggle which is called as Heart Disease. This data set contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with this database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Attribute Information:

- **age:** age in years
- **sex:** (1 = male; 0 = female)
- **cp:** chest pain type
- **trestbps:** resting blood pressure (in mm Hg on admission to the hospital)
- **chol:** serum cholestoral in mg/dl
- **fbs:** fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- **restecg:** resting electrocardiographic results
- **thalach:** maximum heart rate achieved
- **exang:** exercise induced angina (1 = yes; 0 = no)
- **oldpeak:** ST depression induced by exercise relative to rest
- **slope:** the slope of the peak exercise ST segment
- **ca:** number of major vessels (0-3) colored by flourosopy
- **thal:** 3 = normal; 6 = fixed defect; 7 = reversable defect
- **target:** 1 or 0

4 Implementation with R

4.1 Business Understanding

The goal is to estimate whether presence of heart disease in the patient.

4.2 Data Understanding

First, we read our data as csv, view it by means of head function and then we start to examine data type, structure and summarize it.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

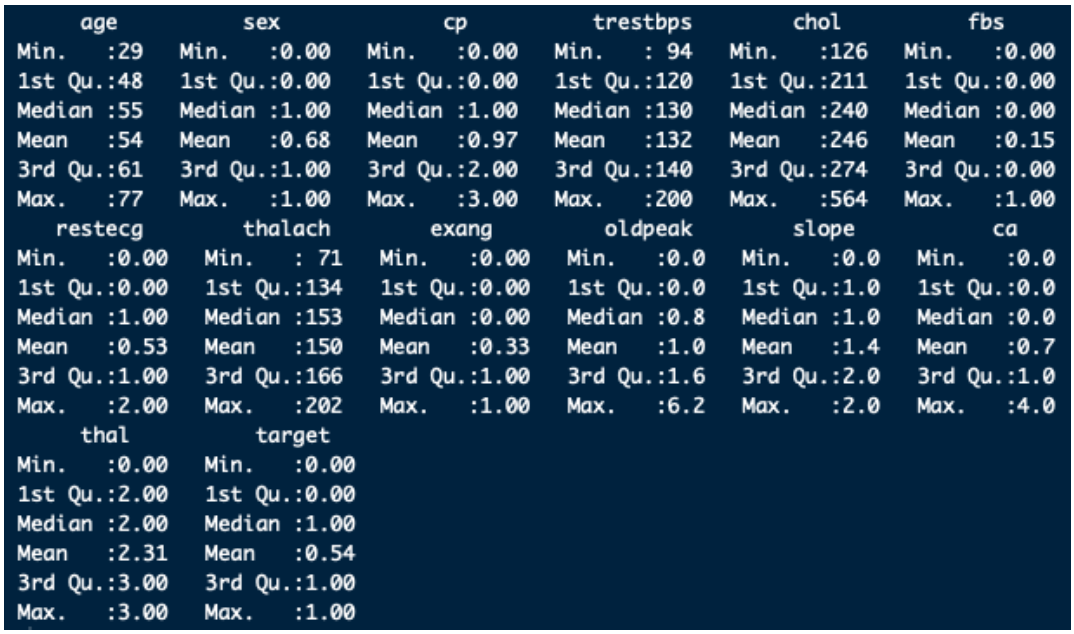
Figure 1: Head of Data Set

As you see below, our data set consist of 303 observations and 14 variables. However, some of variables' types must be changed. To exemplify, to have heart disease or not is represented by 1 and 0, but it is numeric. Therefore, it will be transformed from numeric to factor.

```
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Figure 2: Data Types

As you see below, the data set was summarized. We can get important information such as mean, median, min and max of each attributes.



```
      age              sex              cp            trestbps        chol            fbs
 Min.   :29.00   Min.   :0.00   Min.   :0.00   Min.   : 94   Min.   :126   Min.   :0.00
 1st Qu.:48.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:120   1st Qu.:211   1st Qu.:0.00
 Median :55.00   Median :1.00   Median :1.00   Median :130   Median :240   Median :0.00
 Mean   :54.37   Mean   :0.68   Mean   :0.97   Mean   :132   Mean   :246   Mean   :0.15
 3rd Qu.:61.00   3rd Qu.:1.00   3rd Qu.:2.00   3rd Qu.:140   3rd Qu.:274   3rd Qu.:0.00
 Max.   :77.00   Max.   :1.00   Max.   :3.00   Max.   :200   Max.   :564   Max.   :1.00
    restecg          thalach         exang          oldpeak        slope           ca
 Min.   :0.00   Min.   : 71   Min.   :0.00   Min.   :0.0   Min.   :0.0   Min.   :0.0
 1st Qu.:0.00   1st Qu.:134   1st Qu.:0.00   1st Qu.:0.0   1st Qu.:1.0   1st Qu.:0.0
 Median :1.00   Median :153   Median :0.00   Median :0.8   Median :1.0   Median :0.0
 Mean   :0.53   Mean   :150   Mean   :0.33   Mean   :1.0   Mean   :1.4   Mean   :0.7
 3rd Qu.:1.00   3rd Qu.:166   3rd Qu.:1.00   3rd Qu.:1.6   3rd Qu.:2.0   3rd Qu.:1.0
 Max.   :2.00   Max.   :202   Max.   :1.00   Max.   :6.2   Max.   :2.0   Max.   :4.0
      thal           target
 Min.   :0.00   Min.   :0.00
 1st Qu.:2.00   1st Qu.:0.00
 Median :2.00   Median :1.00
 Mean   :2.31   Mean   :0.54
 3rd Qu.:3.00   3rd Qu.:1.00
 Max.   :3.00   Max.   :1.00
```

Figure 3: Summary of Data

4.3 Data Pre-processing

In this part, missing value detection, outlier detection and treatment and data transformation will be handled.

4.3.1 Data Transformation

The attributes which they are called as "cp","fbs","restecg","exang","ca","slope","thal" and "target" were transformed from numeric to factor and "sex" to character. Furthermore, "sex" attribute's level 0 will be represented as "female" and 1 as "male", "fbs" attribute's level 0 will be represented as "false" and 1 as "true" and "exang" attribute's level 0 will be represented as "no" and 1 as "yes".

```
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : chr  "male" "male" "female" "male" ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
 $ ca      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
 $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Figure 4: Data types after transformation

4.3.2 Missing Value Detection

There are no missing value in our data set. Even if the missing values were in the data set, it did not matter, because decision trees will be used in the modeling section and decision trees are robust to missing data.

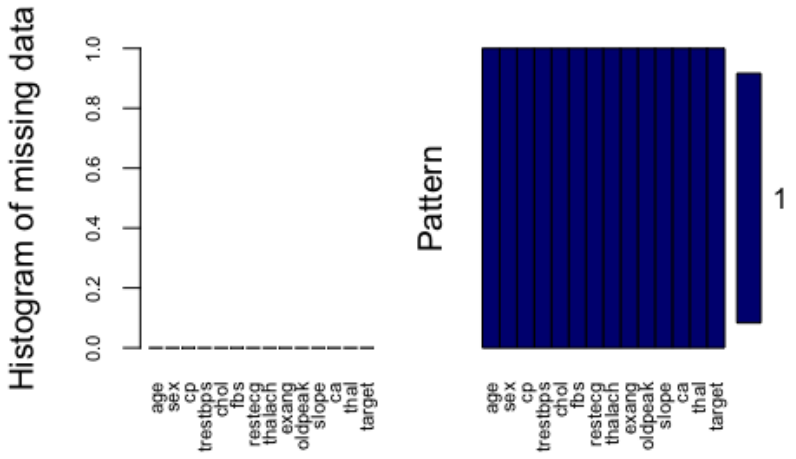


Figure 5: Missing Value Detection

4.3.3 Outlier Detection

An outlier is a data point that differs greatly from other values in a data set. These are badly affecting our model. Therefore, we have to detect and treat them. There are many methods such as boxplot, histogram etc. to detect them, but boxplot will be used. If a value exceeds the boxplot limit, it will be considered an outlier.

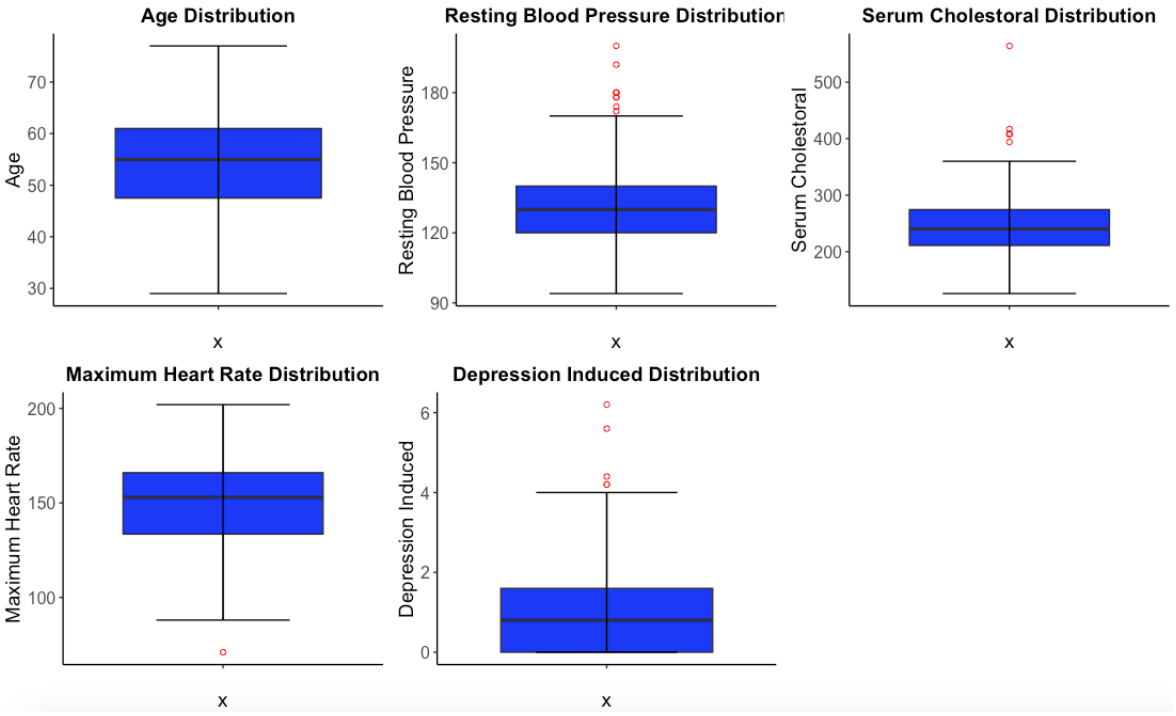


Figure 6: Outlier Detection by Box plot

As you see above, there are outliers in columns which they are called as "trestbps", "chol", "thalach", "oldpeak". In the next section, they will be treated.

4.3.4 Outlier Treatment

In this section, the values which they were detected as outliers, will be treated by means of Random Forest Algorithm. So as to do that, NA values will be assigned to outlier values and they will be predicted. Finally, predicted values will be accepted instead of outlier values.

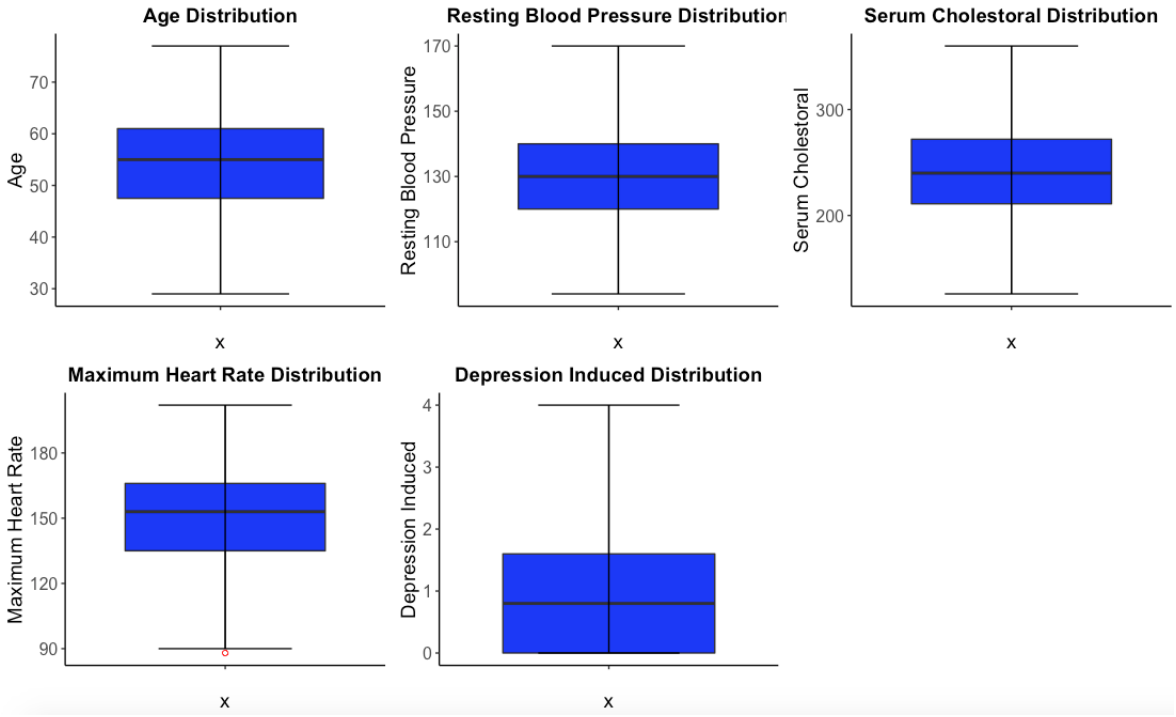


Figure 7: After Outlier Treatment

4.4 Data Visualization

In this section, our cleaned data set will be visualized by means of pie charts, bar plots, box plots, histograms and they will be interpreted.

As you see below, continuous variables are expressed on the basis of ratio histogram graph. For instance, when we interpret it we can say that the age range of people with the most heart disease is in the 55 - 65 band.

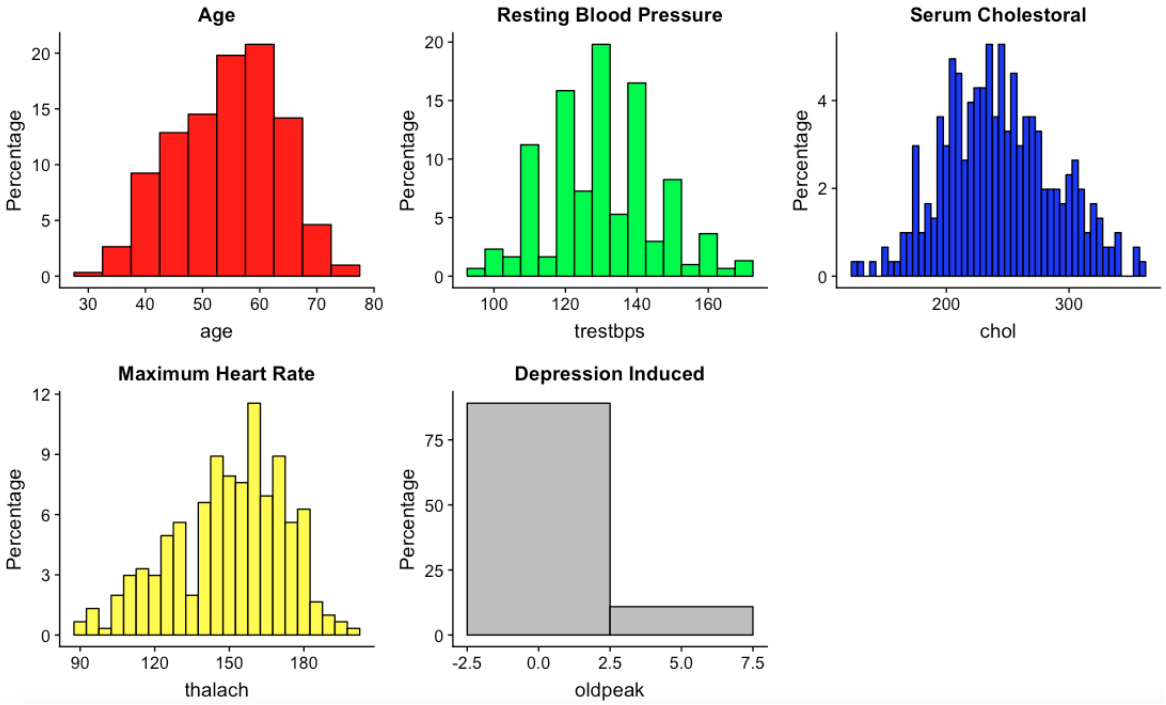


Figure 8: Histogram for Continuous Variables

According to the following graphics, categorical variables are expressed in ratio-based bar plots. It can be interpreted as there are almost no people with 4 major vessels, no thal with 0 level, male ratio is almost 3 times that of women and type zero chest pain is the most common type. Furthermore, if 2nd level of reading electro cardio, zero level of thal and 4 number of major vessels are removed, it does not affect our result of model. But nevertheless it will not be removed.

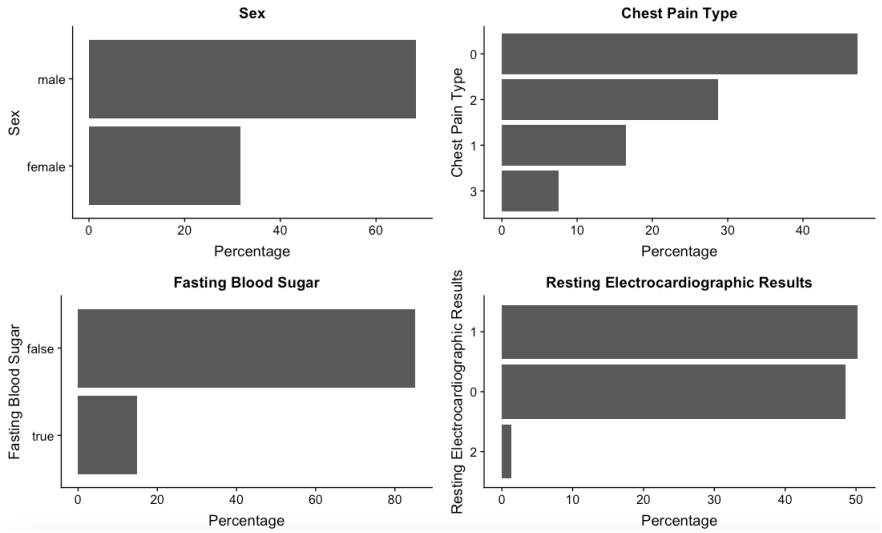


Figure 9: Bar plots of Categorical Variables

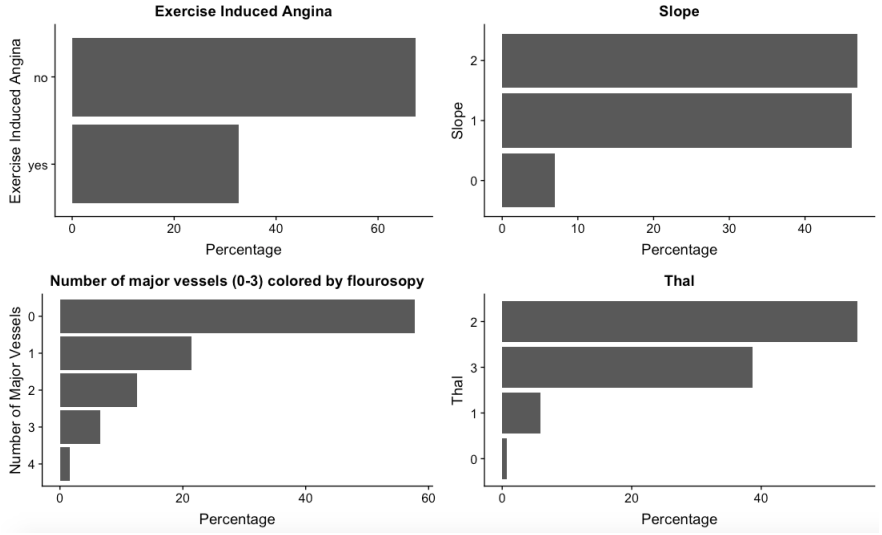


Figure 10: Bar plots of Categorical Variables 2

The graphs that you see below have only two differences from the above bar plots that they are expressed with a single pie chart and colored.

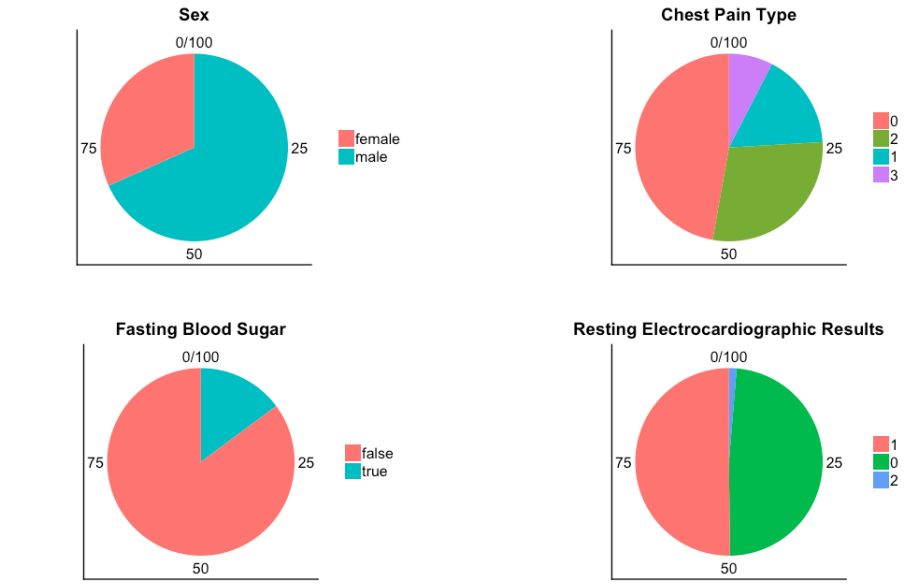


Figure 11: Pie Charts of Categorical Variables

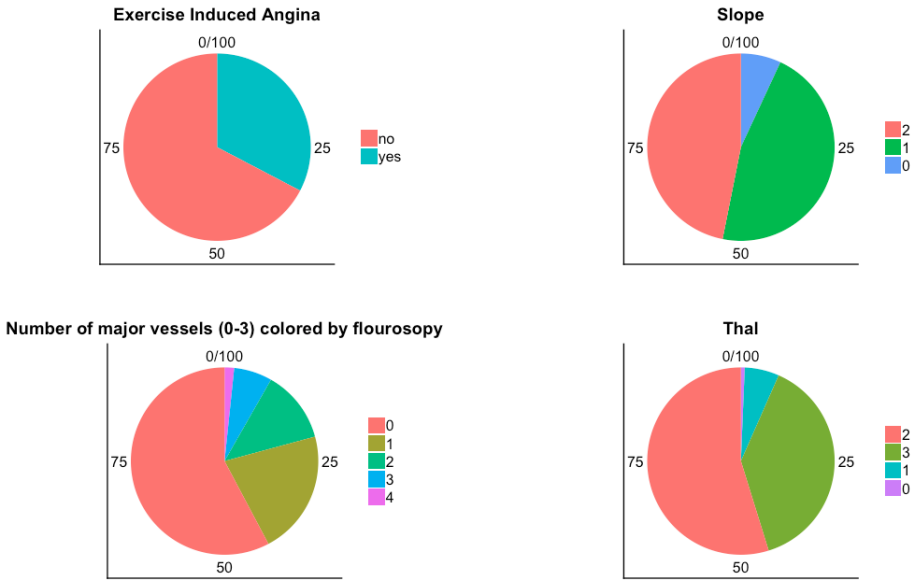


Figure 12: Pie Charts of Categorical Variables 2

In the following figures, the correlation between the continuous values is shown in numerical and then graphical form. Only a significant correlation between thalach and oldpeak was observed.

	trestbps	chol	thalach	oldpeak
trestbps	1.000	0.099	-0.080	0.164
chol	0.099	1.000	-0.035	0.011
thalach	-0.080	-0.035	1.000	-0.344
oldpeak	0.164	0.011	-0.344	1.000

Figure 13: Correlation between Continuous Variables 2

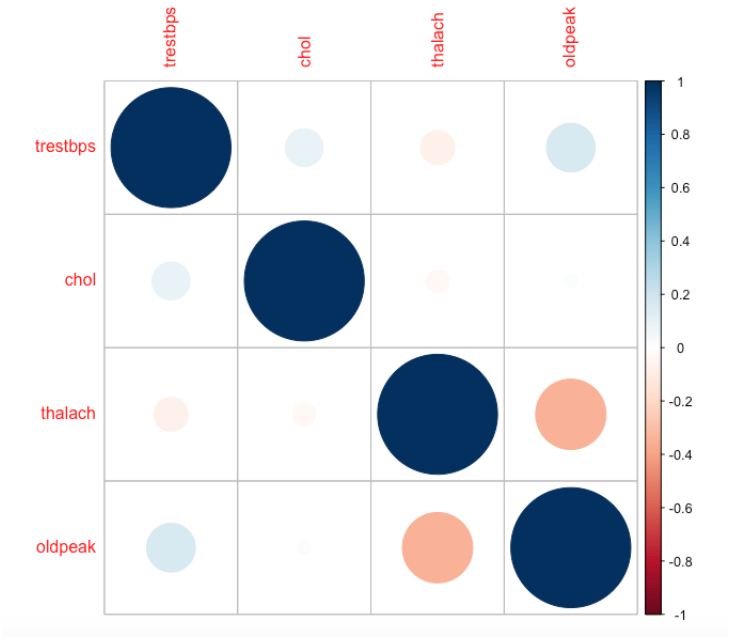


Figure 14: Correlation between Continuous Variables

In this section, the relationship between the continuous variables and the target variable is examined and visualized by box plot.

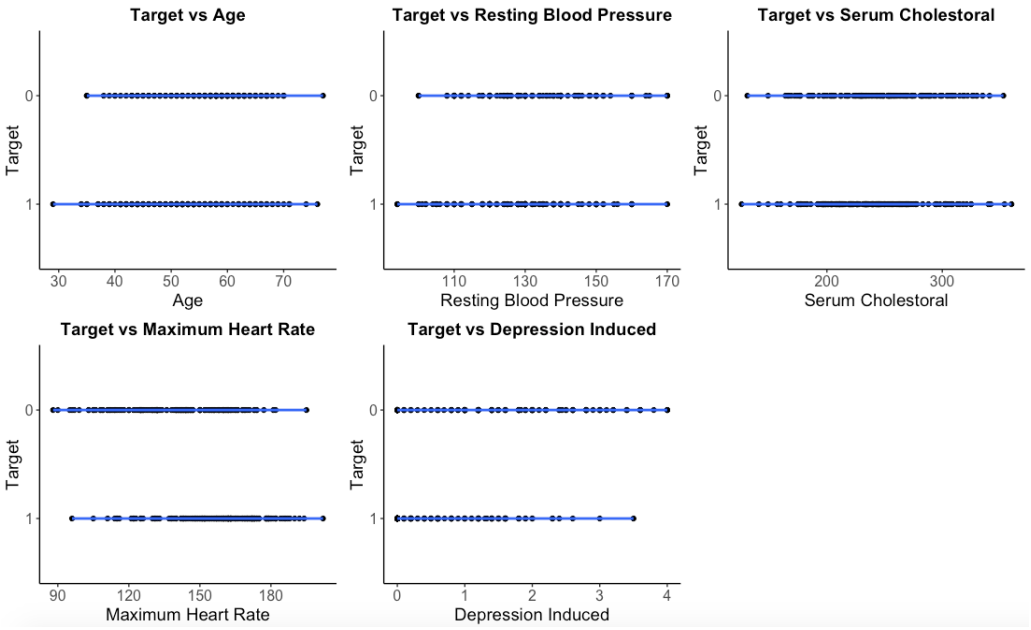


Figure 15: Target vs <rest of continuous variables>

In this section, the relationship between the categorical variables and the target variable is examined and visualized by histograms. For example, so as to interpret we can say that the rate of women with heart disease is higher than that of women without heart disease, but the proportion of men without heart disease is higher than that of men with heart disease. As a second interpretation, the possibility of fasting blood sugar being false and having heart disease gives an overwhelming to other possibilities.

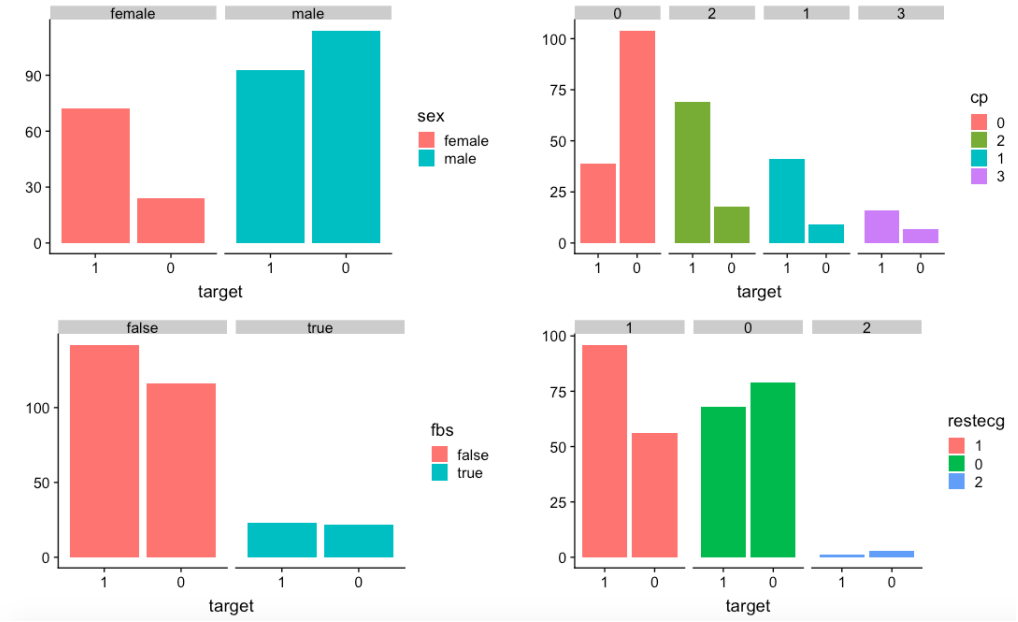


Figure 16: Target vs <rest of categorical variables>

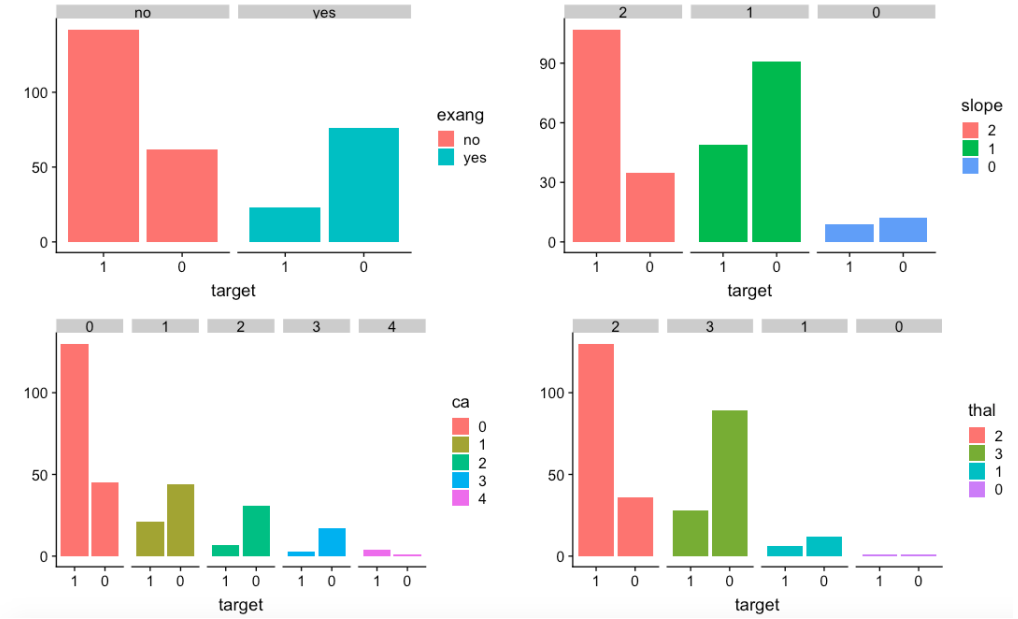


Figure 17: Target vs <rest of categorical variables>

4.5 Data Modeling with Decision Tree

Decision Tree is a non-parametric supervised learning method used for classification. Decision tree learn from data to approximate a sine curve. The deeper the tree, the more complex the decision rules and the fitter the model. Moreover, Decision tree constructs classification models in the form of a tree structure. It splits a data set into smaller and smaller subsets while also an associated decision tree is incrementally developed. In addition, A decision node has two or more branches. Finally, Decision trees can handle both categorical and numerical data.

In decision tree, there are 2 criterion which they are called as "stop" consist of pre-pruned and post-pruned and "split" includes gini index and information gain. Shortly, to obtain optimal tree we pruned our tree and obtained the result below.

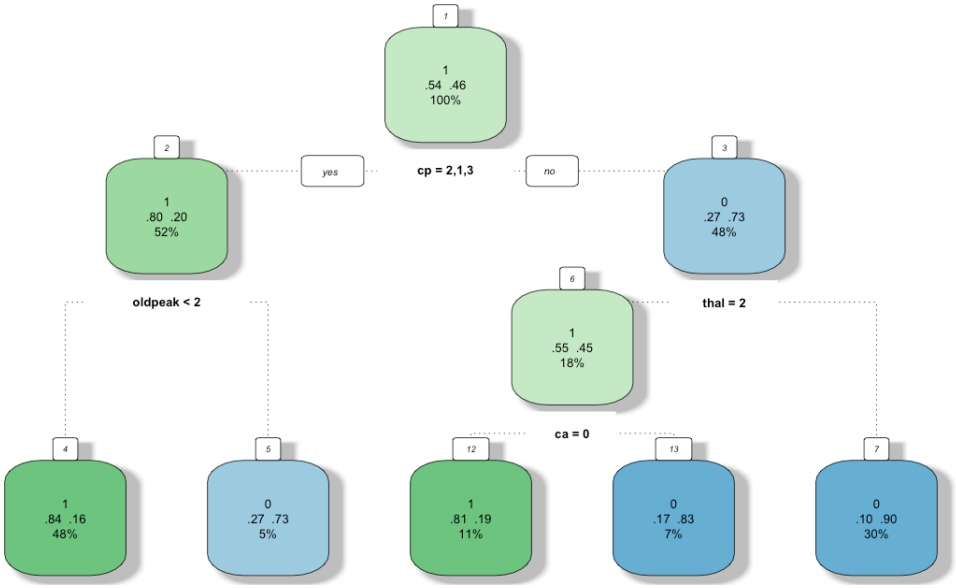


Figure 18: Optimal Tree

As you can see from the figure below, when the importance of the variables are visualized, we can see that the most important variable is cp and the least significant one is age. In fact, if age variable is removed, we can say that nothing about the result will change. If the variables are sorted by importance, we can say cp > thal > thalach > ca > exang > oldpeak > slope > sex > trestbps > chol > age.

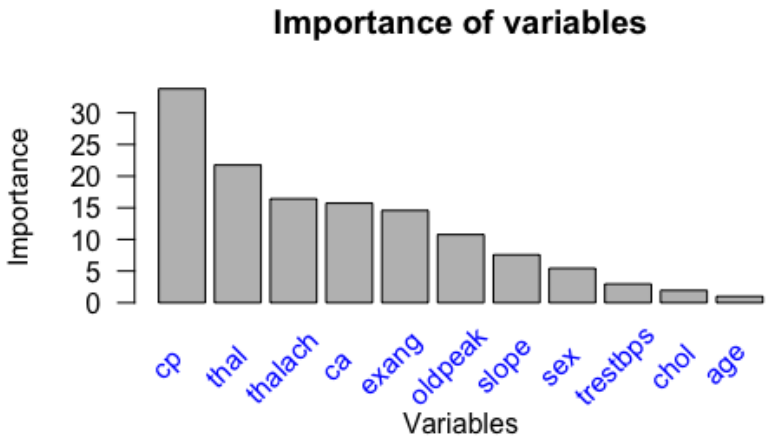


Figure 19: Importance of Variables

4.6 Evaluation

In order to summarize of prediction results on our classification problem, confusion matrix has been created. The number of correct and incorrect predictions are summarized with count values and broken down by each class. So as to calculate it, our test data set was used. Firstly, a prediction has done for each row in your test data set by using optimal tree.

In confusion matrix; top-left is called as "true-negative", top-right is called as "false-positive", bottom-left is called as "false-negative", bottom-right is called as "true-positive".

- **True positives (TP):** These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- **True negatives (TN):** These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- **False positives (FP):** These are the negative tuples that were incorrectly labeled as positive. Let FP be the number of false positives.
- **False negatives (FN):** These are the positive tuples that were mislabeled as negative. Let FN be the number of false negatives.

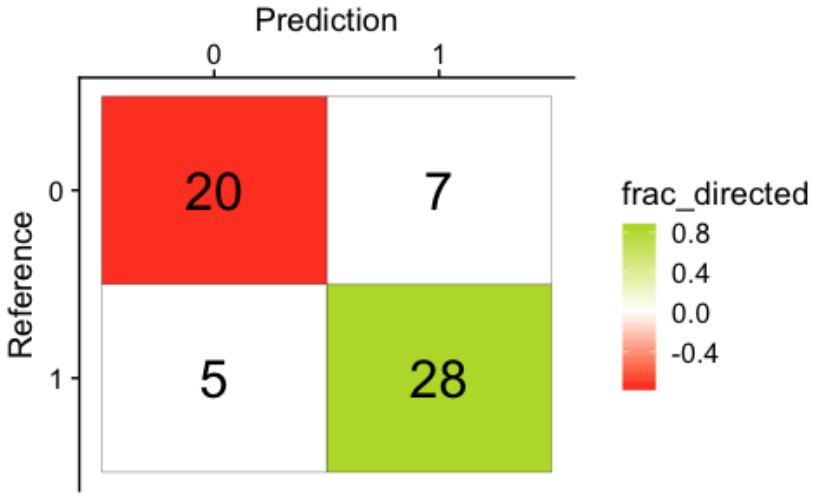


Figure 20: Confusion Matrix

So as to create a complete sensitivity / specificity report and the Roc Curve has been drawn. Before drawing the Roc curve, the value specified as AUC 'The Area under the Curve' was calculated by obtaining the probabilities of the predictions of each class and result is obtained as 0.8 which is good.

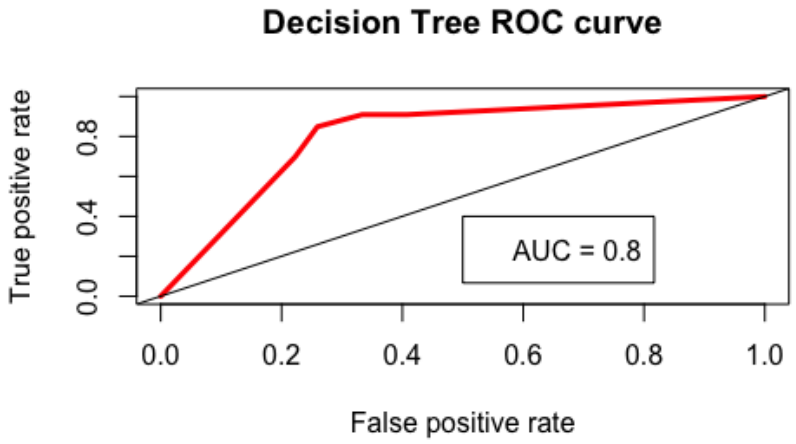


Figure 21: ROC Curve

5 Conclusion

In this brief case study, was talked about the status of data mining in the health world. A data set which called as "Heart Disease" was chosen to handle and the goal identified as to estimate whether presence of heart disease in the patient. In line with this objective, the above data science steps were performed and the data set was analyzed with data mining algorithm.