Data Mining

Heart disease dataset

Outlines:

- Goal
- Problem
- Dataset
- Graphs
- Preprocessing
- Data mining techniques
- Findings and results

Introduction

Health is the most valuable thing a person has.

Goal

Our main goal is to gather relevant information about the factors that influence the risk of gaining heart diseases. Also, apply several data mining techniques.

Problem

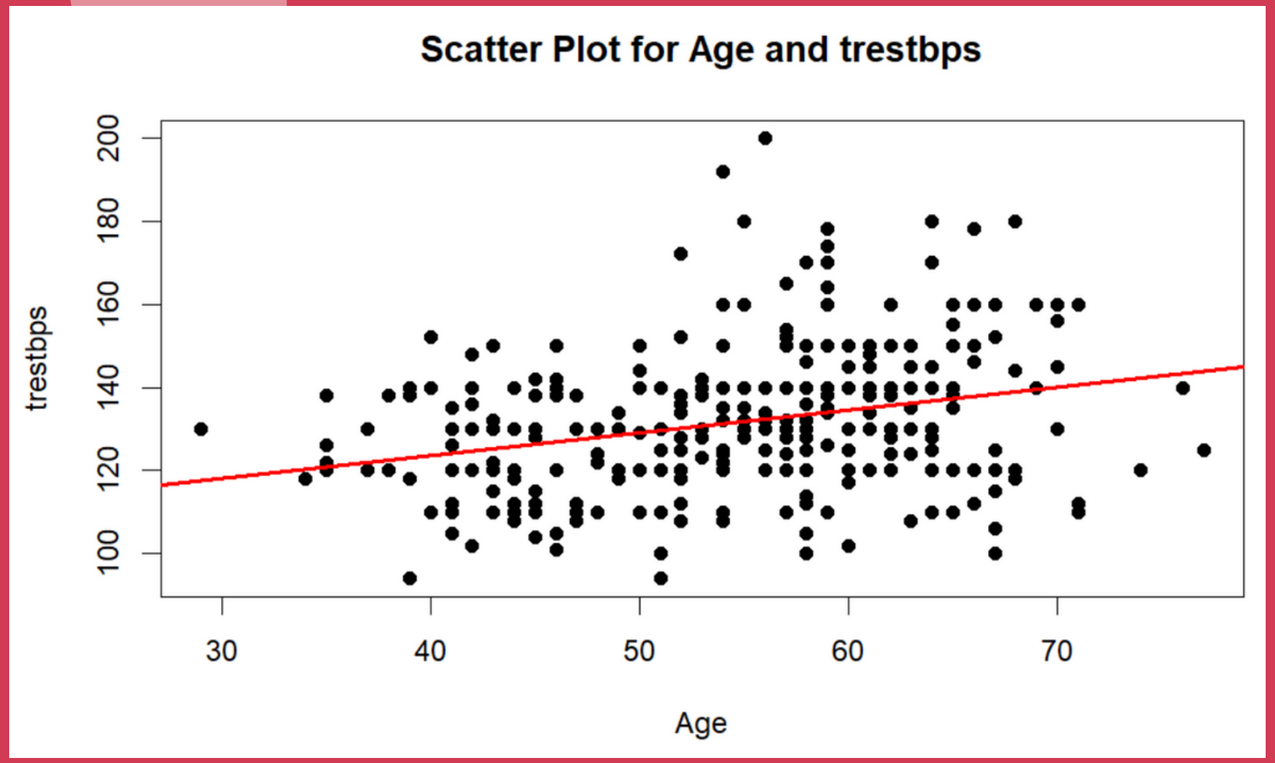
Dataset

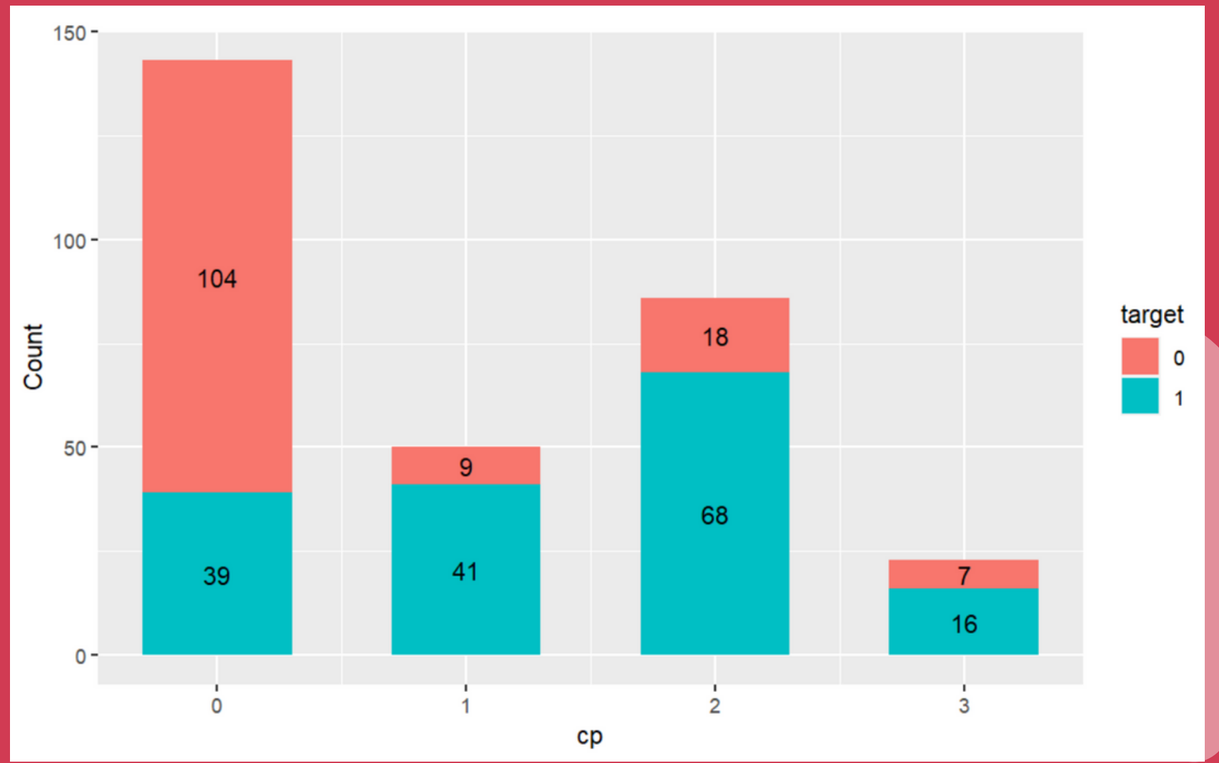
We applied our data mining tasks on data set consisting of:

- 1025 tuples
- 14 attributes which are :

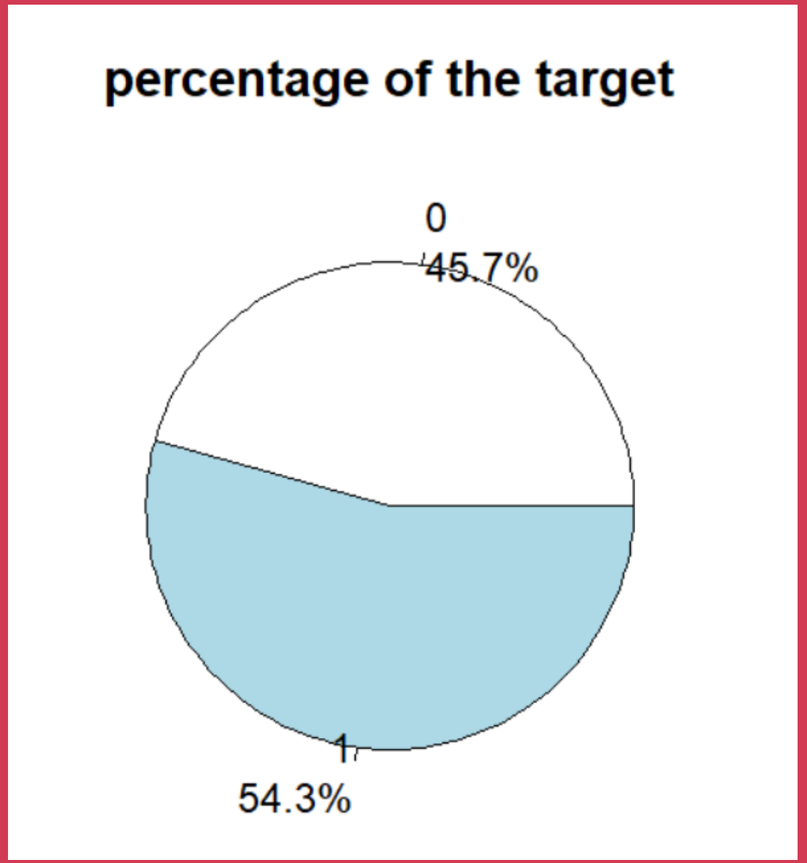
Age	Sex	Chest pain type	Resting blood pressure	Serum cholestoral
Fasting blood sugar	maximum heart rate achieved	Exercise induced angina	Thalassemia Value	Number of major vessels
Slope	oldpeak	Target	Resting electrocardiographic results	

Graphs


Scatter Plot for Age and trestbps
trestbps
100
120
140
160
180
200
30
40
50
60
70
Age


150
100
50
0
Count
104
39
9
41
18
68
7
16
0
1
2
3
cp
target
0
1

Graphs



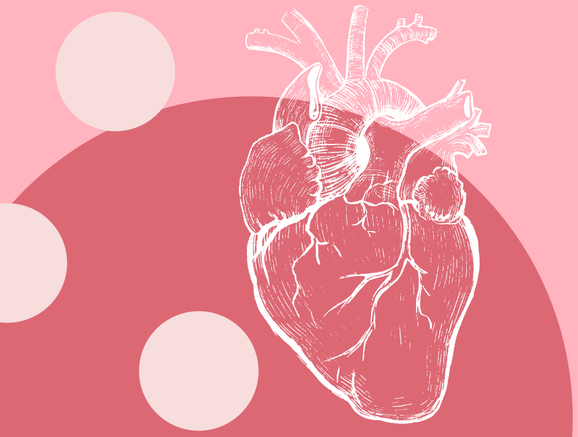
Data Preprocessing:

To have the best possible accuracy results we applied several preprocessing techniques that improve the efficiency of the data.

Preprocessing techniques used:

Data cleaning

Data transformation



Data Preprocessing:

Data cleaning

- ***Handling wrong values***
- ***Handling all outliers***

Data transformation

- ***Normlization :***
  1. min-max normalization
  2. encoding
- ***Discretization***

Data Preprocessing:

Before preprocessing:

	age <int>	sex <int>	cp <int>	trestbps <int>	chol <int>	fbs <int>	restecg <int>	thalach <int>	exang <int>	oldpeak <dbl>	slope <int>	ca <int>	thal <int>	target <int>
1	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
2	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
3	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
4	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
5	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
6	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1

After preprocessing :

	age <fctr>	sex <int>	cp <int>	trestbps <fctr>	chol <dbl>	restecg <int>	thalach <dbl>	exang <int>	oldpeak <dbl>	slope <int>	ca <int>	thal <int>	target <int>
1	2	1	0	2	0.2955326	1	0.7404580	0	0.1612903	2	2	3	0
2	2	1	0	4	0.2646048	0	0.6412214	1	0.5000000	0	0	3	0
3	3	1	0	4	0.1649485	1	0.4122137	1	0.4193548	0	0	3	0
4	3	1	0	4	0.2646048	1	0.6870229	0	0.0000000	2	1	3	0
5	3	0	0	3	0.5773196	1	0.2671756	0	0.3064516	1	3	2	0
6	2	0	0	1	0.4192440	0	0.3893130	0	0.1612903	1	0	2	1

Data mining techniques

We applied both supervised and unsupervised learning techniques on our data set.

- Classification
- Clustering

Classification

In classification we built a model using decision trees to predict our data set class label which is whether a person is targeted to have a heart disease or not, based on the other attributes.

Classification

1 Dividing our dataset into training and testing sets.

We tried 3 different sizes to devide our data into training and testing sets:

- 70% training 30% testing
- 75% training 25% testing
- 80% training 20% testing

2 Constructing the decision trees using the training set.

We tried 3 different attribute selection measures:

- Information Gain
- Gain Ratio
- Gini Index

Classification

3 **Evaluating the model using the testing set and calculating the model's accuracy and other evaluating measures.**

Information Gain:

	70% training, 30% testing	75% training, 25% testing	80% training, 20% testing
Accuracy	80.95%	79.71%	83.02%
Error Rate	19.05%	20.29%	16.98%
Sensitivity(Recall)	91.67	82.93%	89.66%
Specificity	66.67%	75%	75%
Precision	78.57%	82.93%	81.25%

Gain Ratio:

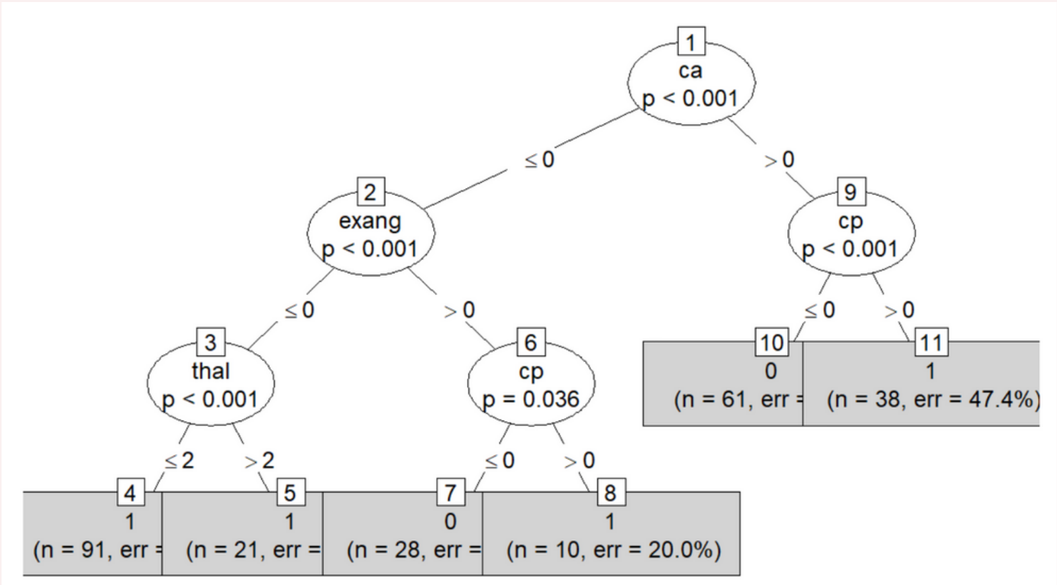
	70% training, 30% testing	75% training, 25% testing	80% training, 20% testing
Accuracy	76.19%	79.71%	79.25%
Error Rate	23.81%	20.29%	20.75%
Sensitivity(Recall)	77.08%	85.37%	82.76%
Specificity	75%	71.43%	75%
Precision	80.43%	81.40%	80%

Gini index:

	70% training, 30% testing	75% training, 25% testing	80% training, 20% testing
Accuracy	76.19%	73.91%	75.47%
Error Rate	23.81%	26.09%	24.53%
Sensitivity(Recall)	75%	73.17%	75.86%
Specificity	77.78%	75%	75%
Precision	81.82%	81.08%	78.57%

Classification

- Using information gain and partitioning the data into 80% training and 20% testing had the highest accuracy value (83.02%).



Evaluation method	value
Accuracy	83.02%
Error Rate	16.98%
Sensitivity(Recall)	89.66%
Specificity	75%
Precision	81.25%

Clustering

In clustering our model will arrange a set of objects (patients) in way that objects in the same group (cluster) are more comparable (in some sense) to those in other groups (clusters).

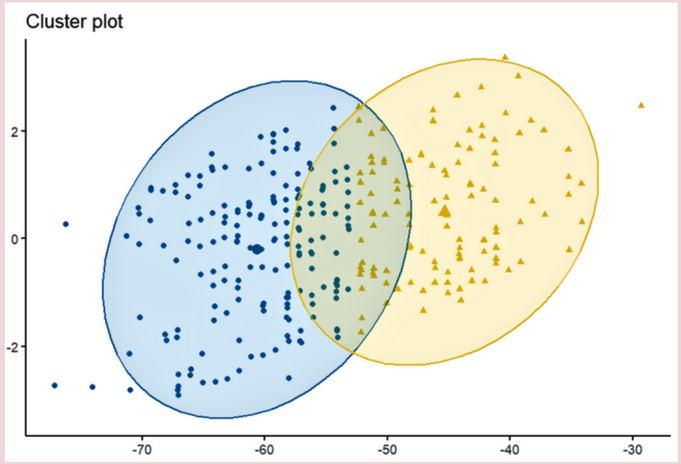
Clustering

We have two main tasks in this section :

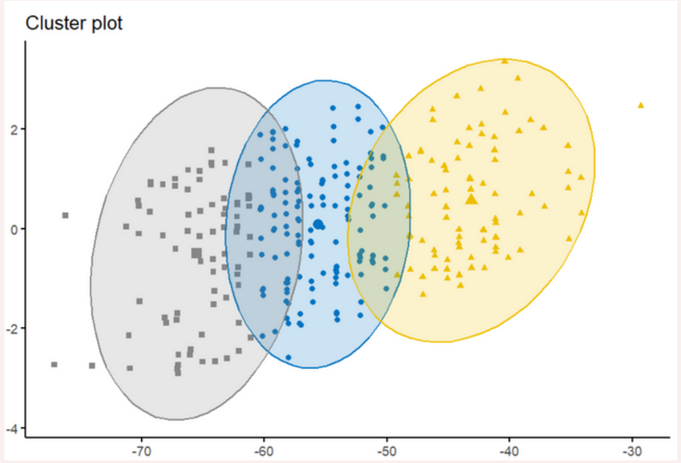
1. **Partition our data using k-mean algorithm.**
   - We tried three different k-means values which are (2,3 and 4).
2. **Cluster evaluation**
   - We will calculate the average silhouette ,total within-cluster sum of square and the BCubed (precision and recall).

Clustering

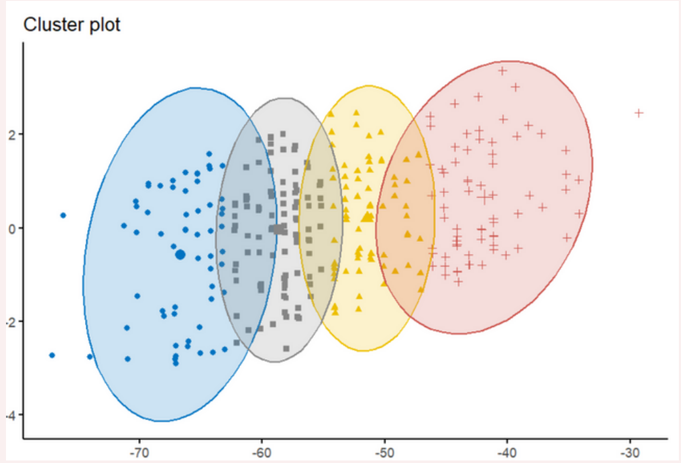
1 Partition our data using k-mean algorithm.



K=2



K=3

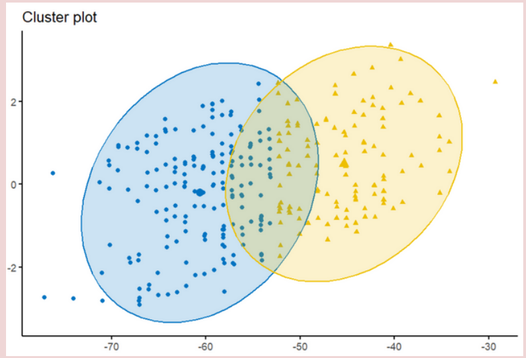


K=4

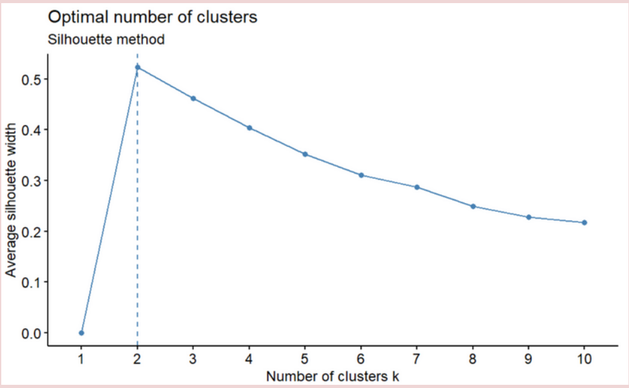
Clustering

2 Cluster evaluation

- Average Silhouette width for all cluters: 0.53
- Total within-cluster sum of square: 9211.418
- BCubed (precision): 0.5347
- BCubed (recall): 0.5523



K=2



Findings and results

From the previous information classification was considered the best option to predict the possibilities of having heart attack based on the attributes, since:

- data set includes class label which is target.
- The model had great accuracy.

References:

[1] “J48 tree in R - train and test classification,” Stack Overflow. [Online]. Available: Average Silhouette width for all cluters: 0.53

[2] “RPubs - Classification and Regression Trees (CART) in R.” [Online]. Available: Average Silhouette width for all cluters: 0.53

[3] IT 326 Data Mining lab slides.

[4]Data source: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Any Questions?

Thankyou!

Lina S. Alzeghaibi 443200923
Ghala T. Alaskar 443200657
Sarah W. Aldbasi 443200520
Reem A. Alhawass 443200461

Supervised by:
Dr. Nuha Bin Tayash