Heartdisease

Business Objective:- Objective is to cluster patient together on basis of presence of heart disease .

Constraints:-Medical records need to be accessed on a wider scale and privacy concerns might arise in step of data gathering.

|  |  |  |  |
| --- | --- | --- | --- |
| Id | Description | Type | Relevance |
| Age | Age of patient | Quantitative, Nominal | Age of patient |
| Sex | Sex of patient | Categorical,Nominal | Sex of patient |
| cp | Chest pain type | Categorical,Ordinal | 4types of chest pain ordered from 0 to 4 |
| trestbps | Resting blood pressure | Quantitative,Nominal | Resting blood pressure |
| chol | serum cholestoral in mg/dl | Quantitative,Nominal | Cholestoral |
| fbs | Fasting blood sugar | Categorical,Nominal | Whether blood sugar is above 120 mg/dl |
| restecg | Resting ecg | Categorical,Ordinal | Resting ECG ranked from 0 to 2 |
| thalach | Maximum heart rate achieved | Quantitative,Nominal | Provides maximum hear rate achieved |
| exang | exercise induced angina | Categorical,Nominal | Exercise induced angina |
| oldpeak | oldpeak | Quantitative,Nominal | ST depression induced by exercise relative to rest |
| slope | slope | Categorical,Nominal | the slope of the peak exercise ST segment |
| ca | Number of vessels | Categorical,Ordinal | number of major vessels (0-3) colored by flourosopy |
| thal |  | Categorical,Ordinal | thal: 3 = normal; 6 = fixed defect; 7 = reversable defect |
| target | target | Categorical,Nominal | Whether there is heart disease or not |

3.Data pre-processing

* Description of dataset

age sex cp ... ca thal target

count 303.000000 303.000000 303.000000 ... 303.000000 303.000000 303.000000

mean 54.366337 0.683168 0.966997 ... 0.729373 2.313531 0.544554

std 9.082101 0.466011 1.032052 ... 1.022606 0.612277 0.498835

min 29.000000 0.000000 0.000000 ... 0.000000 0.000000 0.000000

25% 47.500000 0.000000 0.000000 ... 0.000000 2.000000 0.000000

50% 55.000000 1.000000 1.000000 ... 0.000000 2.000000 1.000000

75% 61.000000 1.000000 2.000000 ... 1.000000 3.000000 1.000000

max 77.000000 1.000000 3.000000 ... 4.000000 3.000000 1.000000

* Mean of features

age 54.366337

sex 0.683168

cp 0.966997

trestbps 131.623762

chol 246.264026

fbs 0.148515

restecg 0.528053

thalach 149.646865

exang 0.326733

oldpeak 1.039604

slope 1.399340

ca 0.729373

thal 2.313531

target 0.544554

* Median of features

age 55.0

sex 1.0

cp 1.0

trestbps 130.0

chol 240.0

fbs 0.0

restecg 1.0

thalach 153.0

exang 0.0

oldpeak 0.8

slope 1.0

ca 0.0

thal 2.0

target 1.0

* Variance of dataset

age 82.484558

sex 0.217166

cp 1.065132

trestbps 307.586453

chol 2686.426748

fbs 0.126877

restecg 0.276528

thalach 524.646406

exang 0.220707

oldpeak 1.348095

slope 0.379735

ca 1.045724

thal 0.374883

target 0.248836

For some features variance is low but that’s okay since many of these features are categorical ones, for quantitiative features variance is in enough quantity.

* Standard Deviation

age 9.082101

sex 0.466011

cp 1.032052

trestbps 17.538143

chol 51.830751

fbs 0.356198

restecg 0.525860

thalach 22.905161

exang 0.469794

oldpeak 1.161075

slope 0.616226

ca 1.022606

thal 0.612277

target 0.498835

* Skewness

age -0.202463

sex -0.791335

cp 0.484732

trestbps 0.713768

chol 1.143401

fbs 1.986652

restecg 0.162522

thalach -0.537410

exang 0.742532

oldpeak 1.269720

slope -0.508316

ca 1.310422

thal -0.476722

target -0.179821

* Kurtosis

age -0.542167

sex -1.382961

cp -1.193071

trestbps 0.929054

chol 4.505423

fbs 1.959678

restecg -1.362673

thalach -0.061970

exang -1.458317

oldpeak 1.575813

slope -0.627521

ca 0.839253

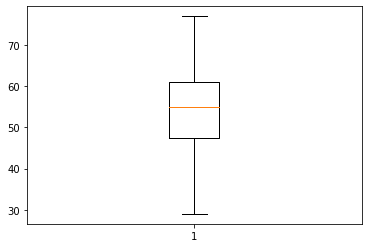
thal 0.297915

target -1.980783

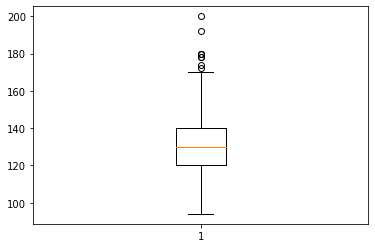
* Outlier detection

Outliers were detected using boxplots. Apart from age all other quantitative features were found to be having outliers

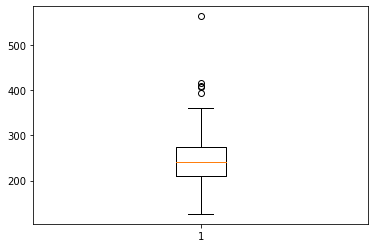
Age



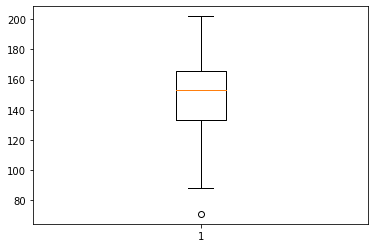
Trestbps



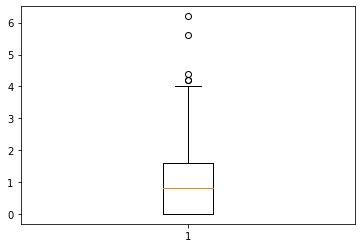
Chol



Thalach

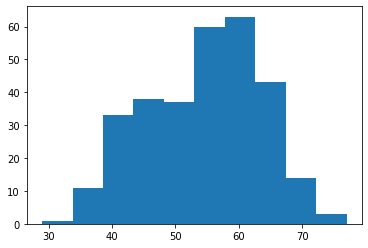


Oldpeak

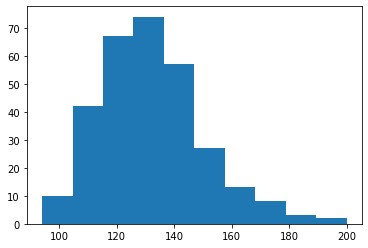


* Histogram

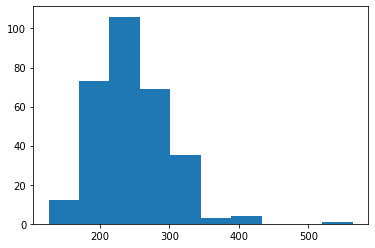
Age



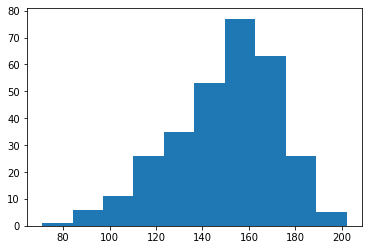
Trestbps



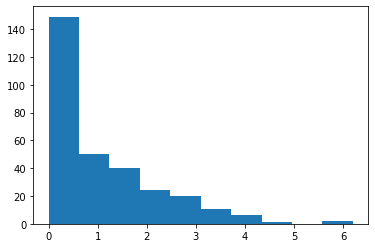
Chol



Thalach



Oldpeak



* Outliers treatment

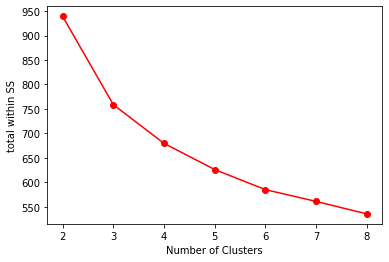
Quantitative data was separated from categorical data and winsorization was performed on it.

* Feature Scaling

Normalization and scale function of sklearn were used for feature scaling

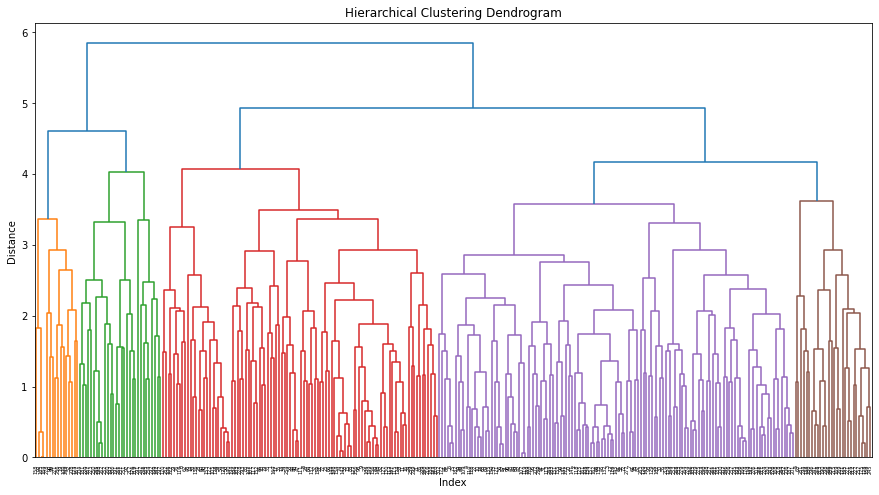
* K-Means before PCA

K-Means clustering was performed on dataset before PCA and its clusters were saved under ‘Clust1K’ column



* Hierarchical clustering before PCA

Hierarchical clustering was performed and its finding were appended to dataset under ‘Clust1H’ feature



* PCA

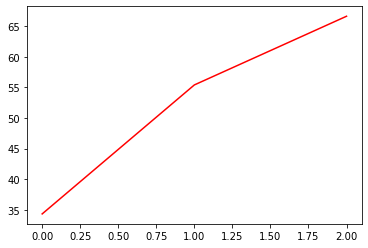
Variance of PCs

array([0.34313944, 0.21094196, 0.11229343])

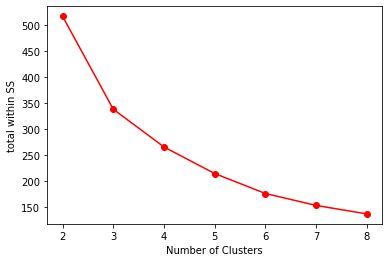
Cumulative variances of PCs

array([34.31, 55.4 , 66.63])

Variance plot for obtained PCs



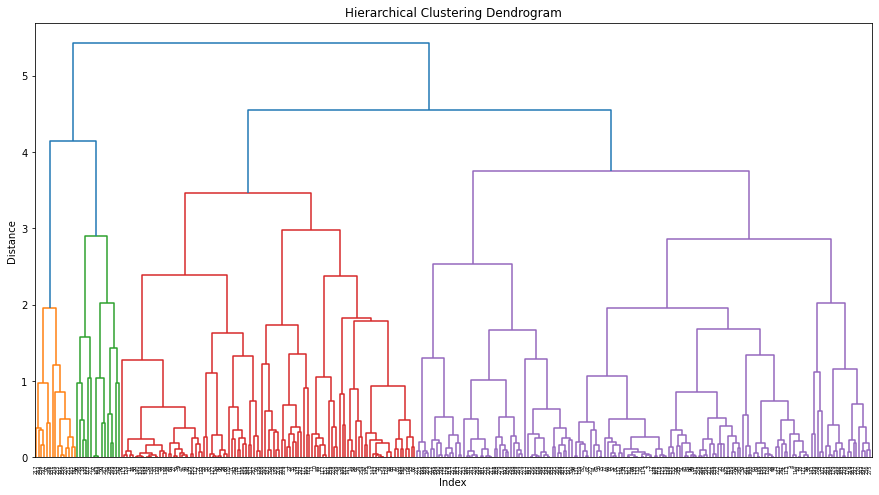
* K-means performed on PCA



Clusters were stored as ‘Clust2k’ feature

* Hierarchical clustering performed after PCA

Hierarchical clustering was performed after PCA and clusters were stored in ‘Clust2H’ feature



PCA improves the performance of model at a slight acceptable cost of accuracy, as it can be seen above, Dendrogram does get more sophesticated after PCA.This might help deciding appropriate number of clusters.