## Prediction of heart disease

Priciple component analysis for Prediction of heart disease and regression analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | age | trestbps | chol | thalach | oldpeak |
| 1 | 63 | 145 | 233 | 150 | 2.3 |
| 2 | 37 | 130 | 250 | 187 | 3.5 |
| 3 | 41 | 130 | 204 | 172 | 1.4 |
| 4 | 56 | 120 | 236 | 178 | 0.8 |
| 5 | 57 | 120 | 354 | 163 | 0.6 |
| 6 | 57 | 140 | 192 | 148 | 0.4 |
| 7 | 56 | 140 | 294 | 153 | 1.3 |
| 8 | 44 | 120 | 263 | 173 | 0 |
| 9 | 52 | 172 | 199 | 162 | 0.5 |
| 10 | 57 | 150 | 168 | 174 | 1.6 |
| 11 | 54 | 140 | 239 | 160 | 1.2 |

# DATA of Prediction of heart disease

The data that are using here is referred from Kaggle data’s set [2]. This data consists of observation of parameter, parameter description and the value of the blood experiment of people and it is illustrated as Fig 1.

Here is a summary of the data that is used in this paper table1. As it is clear in the picture below each of this information has its special attributes. In this data different attribute has different measurement and then we should standardize them.

TABLE I. Parameter description

|  |  |  |
| --- | --- | --- |
| S no | Parameters | Parameters description |
| 1 | Age | Ages in years |
| 2 | Thestps | Resting blood sugar |
| 3 | chol | Serum cholesterol |
| 4 | Fbs | Fasting blood sugar |
| 5 | Thalach | Maximum heart rate achieved |
| 6 | Old peak | ST depression induced by exercise |

TABLE II. dataset Heart Disease UCI

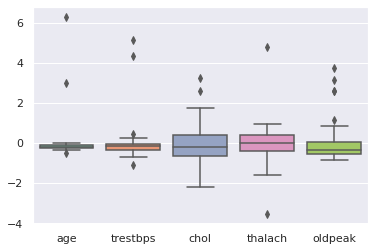


Fig 2. Boxplot of centred feature vectors.

According to the fig 2 after centering the data the distribution of the of data illustrate this box plot. Some outliers are shown in the picture for different features. According to the boxplot the biggest median in related to the thalach that is defined as maximum heart rate achieved.

The covariance matrix shows that trestbos is more related with other variables compared to the other variables. Plus, as it is shown oldpeak is negatively correlated with other variables and positively with chol. The highest positive correlation is created between age and tresbps. The positive correlation exists between another variable but there are not that much high to be mentioned Fig 3.

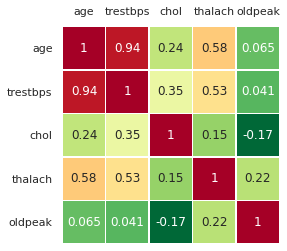


Fig 3. Covariance Matrix.

# PRINCIPLE COMPONENT ANALYSIS

Principal component analysis (PCA) is a method for achieving the principal components by using conversion of the data. Furthermore, PCA is a process that makes a data more understandable and interpretable by decreasing the dimensionality of the datasets [3].

So, we are decided to reduce the dimension of our data by decreasing the 8 vectors to r vector and it the r should be less than 3 in order to accept the data. The dimension of the data can be illustrated through the scree plot fig 4 and pareto plot fig 5.

As it shown in the scree plot approximately 90 percent of the variance can be interpreting only by three first component in the data, while other are responsible for only the rest.so the best and optimal component value for this paper 2 will be chosen, so the complexity of the data will decrease considerably.

## Explained variance

Explained variance is used in order to identify the percentage of the variance that is responsible for the principal components. It can be calculated through:



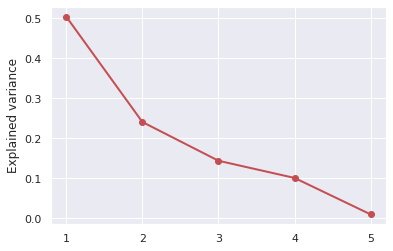
. 

Fig 4 Scree Plot

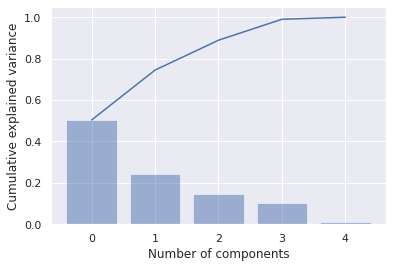


Fig. 5. Pareto diagram

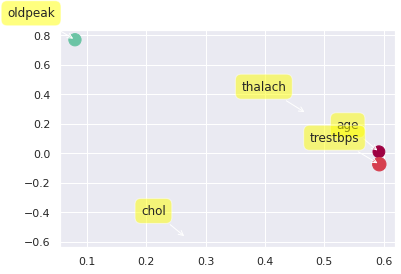


Fig 6. Scatter plot of PC2 coefficients vs. PC1 coefficients.

The scatter plot of PC2 coefficient vs. PC1 is presented in the Fig 6. This plot hep to perceive the variables that shows the same role and involvement in the data. As it is shown in the data cp, oldpeak, slop is showed the same behavior while age, chol and trsbps are the same involvement within the principal components. also, this correlation is shown in the Fig 3.

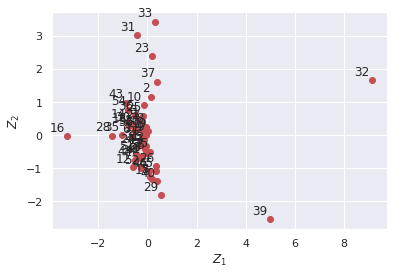


Fig 6. Scatter plot of PC2 score Vs. PC1 score

Biplot is a visualization of the principal components coefficients and principal components score for variables and observation, respectively. Each axis in biplot illustrate the principal components. Observation are shown as pointed in biplot that are each person with different blood test result that are presented by number as the first column of the data [4].

As it is presented in the biplot first principal component has 2 negative variables for the variable thalach (maximum heart rate achieved and chol (serum cholesterol) and positive coefficient with other variables. The second principal component is related to the vertical axis has a positive principal component with the chol and other play negative coefficients.

So, as it is explained above a variable represented in the biplot with a narrow angle and long direction play major role in the PC.

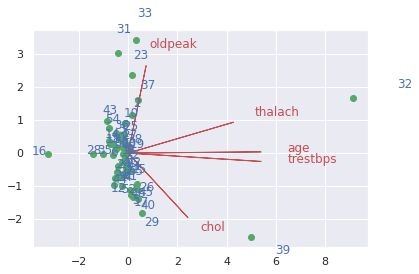


Fig 7 Biplot

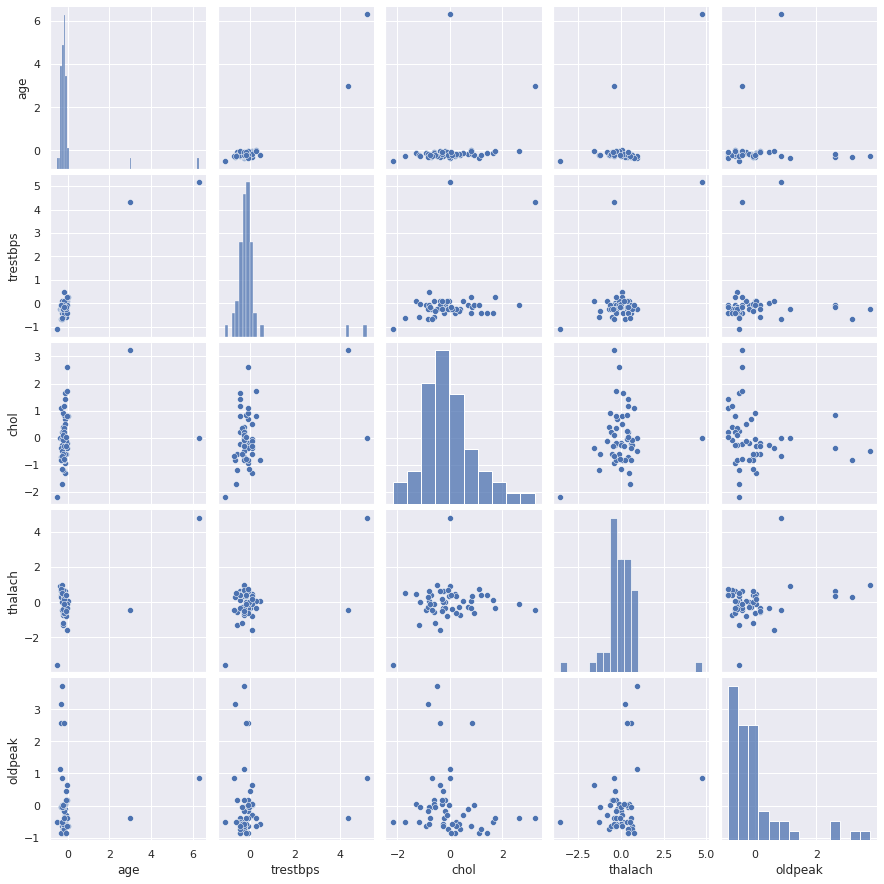


Fig 8. pair plot

In above picture the distribution of each variable is shown and the relationship between them is presented transparently. It readily will create a graphically matrix that that shows relationship between variables in the data [5]. It is very versatile mean to know how two variables would have effect on each other. For example, by increasing the age how the blood [pressure goes higher, or cholesterol goes higher while the heart rate would come considerably lower.

The eigenvector matrix in provided by Fig 9. So, the three first components can be taken:

As it is shown above in the fisrt PC is essentially a cotnacst between age, thalash and tresbps. As it is shown, there is a allposistve coefficent fot all age, trestbps, chol , thalach and oldpeak.

The second PC a cotnrast between thalach, oldpeak and chol is shown. This evidence is shown by positive coeficient for thalach anf chol and negative coefficent for oldpeak.

Finally for the third PC a contrast between chol and trestbps. As it is illustrated a negative coefficent is shown for trestbps and postitve one for chol

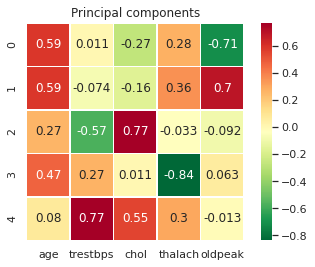


Fig 9. Principal components

# Linear Regression

## Regression Analysis

Regression analysis is the prediction of dependent variable according to the value of the independent variable. Also, it describes the effect of the variation in an independent variable within the dependent variable. The data which is used to visualize the linear regression analysis is given form Kaggle data center [6].

The r = 0.994 indicates that x and are positively related. The value of the r\_squared explained that only approximately .988percent of the total variation in blood sugar is explained by age of the persons and 0,01 percent are not explained.

The least square regression line is:

beta0hat = 0.38, beta1hat = 0.99

r\_value = 0.994, r\_squared = 0.988

The value of the beta0hat gives the value of for that gives the y for a x with no value. The value beta1hat illustrate the variation in according to a change in x, showing that for each x the y is increased by .99.

The below Anova table for linear regression illustrates df (degree of freedom, regression of mean square Anova F statistics and error.

TABLE II. Anova Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | sum\_sq | mean\_sq | F | PR(>F) |
| X | 1.0 | 71834.0034 | 71834.0034 | 7984.66 | 5.6839e-95 |
| Residuals | 97.0 | 872.6603 | 8.9964 | NaN | NaN |
| Alpha | 0.05 | 3.94 |  |  |  |

The scatter plot and regression line are illustrated in the Fig 10 and they are illustrating the negative relation between the dependent and independent variables. Parameter estimate are retrieved and illustrate In Fig11.

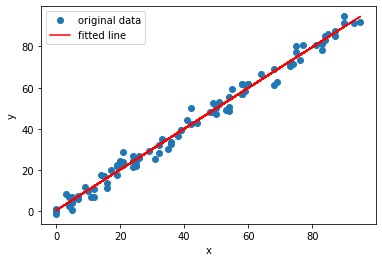
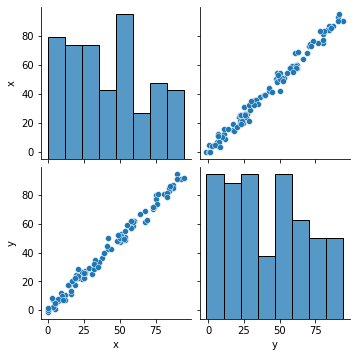


Fig 10 Scatter plot for regression line



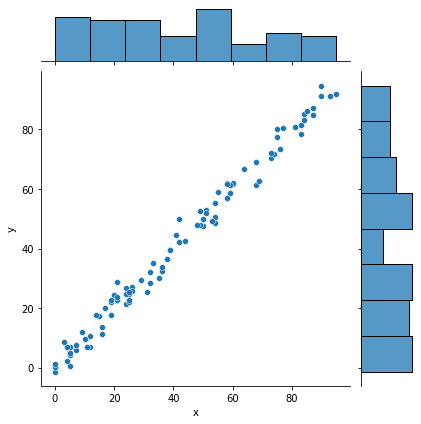
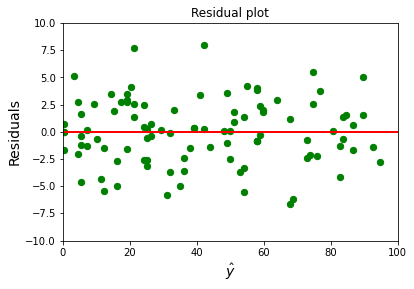


Fig 11. Retrieve the parameter estimates

## Residual analysis

Residual plot Fig 12 is visualization that illustrate residual and independent variables x and they can be plot in vertical and horizontal axis respectively [7]. As it is shown the points are randomly distributed around zero. So, it is the linear regression of model that is suitable for the data. Therefore, the residual plot does not suggest violations of the assumptions of zero means and same variance of the random errors.

Fig 11 Fig 12 Residual plot

# Conclusion

In this paper principal component analysis applied to heart disease dataset that is provided by Kaggle. And the output of the python is used to analyze the effect of the principal component on the data. By PCA is applied to data, it is concluded that 90 percent of the variance can be interpreting only by three first component in the data. but because of having problem with 3D analysis of the plot we preferred to interpret the data by 2 first components. The regression analysis is used to verify the preferability of the data in order to use in regression analysis. According to the regression plot the normal distribution is shown that verify dependent variable is depend on the independent variable. Finally, residual analysis, it is shown that points are randomly distributed around zero. So, it is appropriate to use linear regression of model that for the data.

# References

1. https://royalsoit cietypublishing.org/doi/10.1098/rsta.2015.0202https://www.statisticshowto.com/correlation-matrix/.
2. Kaggel.com.
3. <https://en.wikipedia.org/wiki/Principal_component_analysis#:~:text=Principal%20component%20analysis%20(PCA)%20is,components%20and%20ignoring%20the%20rest>..
4. <https://en.wikipedia.org/wiki/Biplot#:~:text=Biplots%20are%20a%20type%20of,matrix%20to%20be%20displayed%20graphically>.
5. <https://medium.com/@jaimejcheng/data-exploration-and-visualization-with-seaborn-pair-plots-40e6d3450f6d#:~:text=Pair%20Plots%20are%20a%20really,of%20regression%20analysis%20to%20use>.
6. <https://www.kaggle.com/tanuprabhu/linear-regression-dataset>
7. <https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate>.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published**