Course Number:
INSE-6220

Course Title:

Advanced Statistical Approaches to Quality

Taranjyot Singh

ID - 40120880

Principal Component Analysis for Diagnosis of Heart Disease

Concordia Institute for Information Systems Engineering
Concordia University

Abstract—Abstract — The study shows internationally that the deaths caused by heart attacks are the number one cause of death. The healthcare industry collects massive quantities of data that need to be analyzed to identify secret information for successful decision making. In this paper, we suggest a two-stage clustering approach, which is the primary component analysis and logistic regression. The main objective of this report is to analyze the machine learning algorithm to diagnose heart disease.

Index terms—PCA, Heart-disease, Logistic Regression, Principal components.

I. Introduction

The health industry generates large amounts of data every day, out of which heart diseases being the most data-intensive field. Blood vessel diseases, heart rhythm problems, heart defects all fall under the category of heart diseases. The word "heart disease" is often used interchangeably with the phrase "cardiovascular disease".(6) Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke.(2) Many cardiac disorders, such as those affecting the muscle, valves, or rhythm of the heart, are also considered types of heart disease. Heart disease is one of the most important causes of morbidity and death in the world's population. One of the most important aspects of clinical data analytics is the prediction of cardiovascular diseases.

Some common symptoms of heart attack are

- Chest pain
- Fatigue
- Nausea, Indigestion
- Pain in arms
- Sweating

Heart diseases are difficult to identify because of several risk factors like pulse rate, high cholesterol, and many other factors. Thus, Principal Component Analysis (PCA) is a good fit to proceed with the dataset. PCA is a statistical procedure to convert possibly correlated observations into a set of linearly uncorrelated observations. It is a dimensionality reduction method that uses true eigenvectors for multivariate analysis.

In this report, we evaluate the dataset using PCA. After implementing the transformation we evaluate the data on a classification algorithm Logistic Regression.

II. Heart Disease Data

We have used the Cleveland heart disease dataset from the UCI repository and it consists of 303 individual data. In this actual dataset, we have 76 features but we have only used 14 of them because of some specific reasons, they are mentioned as follows:

- age
- sex (1 = male, 0 = female)
- chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptotic) - It is lack of oxygen-rich blood in the heart muscle.
- resting blood pressure - Arteries can get damaged due to high blood pressure.
- serum cholesterol - Higher levels of bad cholesterol make you prone to heart attacks.
- fasting blood sugar - Higher blood sugar levels result in the decrement of insulin levels causing a risk of a heart attack.
- resting ECG (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
- maximum heart rate achieved - According to a study, an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%.(2)
- exercise-induced angina (1 = yes, 0 = no)
- ST depression caused by exercise compared to resting
- peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
- Number of major fluoroscopic vessels (0–3)
- Thal (3 = normal, 6 = fixed defect, 7 = reversible defect) - It displays the thalassemia.
- diagnosis of heart disease (0 = absence, 1, 2, 3, 4 = present)

Age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0

Figure1-First 20 rows of Cleveland dataset

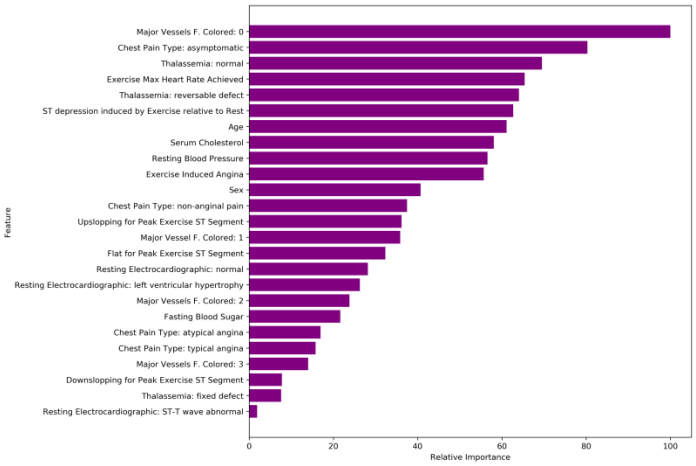


Figure2-Relative importance of selected features

III. Principal Component Analysis

Principal Component Analysis is an exploratory multivariate technique used to simplify complex datasets and orthogonally transform the dataset into a new coordinate system of linearly uncorrelated vectors. It is transformed in such a way that some projection of data of the greatest variance lies on the first coordinate (the principal component), the second-largest variance on the second coordinate, and so on. In general, data set X, which is an I × J matrix will be preprocessed before the PCA analysis. First, it will be zero centered, which means the mean of each column will be equal to zero. Then if elements of X are divided by √I-1, the analysis will be called a covariance PCA, because the matrix X^TX will be a covariance matrix.

Let X a centered data Matrix (the sample mean of each column has been shifted to zero) of size (n, p), the p columns of X define the features/attributes of the data, and the n columns define different observations. A p*p orthogonal matrix A determines the transformation, so that: Z = A.X. In this, X^TX will be a correlation matrix.

Now, we will organize the dataset and calculate empirical mean and deviations from the mean. Then, we find the covariance matrix, its eigenvectors, and eigenvalues. Rearrange the eigenvectors and eigenvalues and select the subset of eigenvectors to project the z-scores of the data on the new basis.

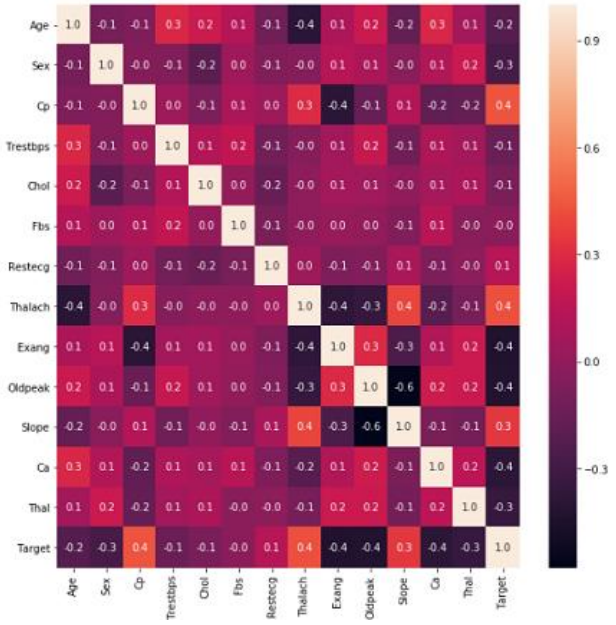


Figure3-Correlation Matrix

IV. Classification Algorithm

A useful machine learning application in the medical field would be the ability to decide whether a new patient is normal or abnormal given the 6 variables described above. This technique is called classification. The existing labelled data is used to learn about the people prone to heart-related problems considering all the factors mentioned above. The decision on which algorithm to choose would depend on the properties of data like distributions, complexity, size, the relationship between features.

The classification algorithm that we have used in this report is Logistic Regression.

Logistic Regression learns a model that predicts the probability that an observation x belongs to a certain class v, using the sigmoid function. Simple regression may be a kind of multivariate analysis wherever the number of freelance variables is one and there's a linear relationship between the independent(x) and dependent(y) variable. The linear regression algorithm finds the best possible result for a_0 and a_1. Equation for Linear Regression: Y = a_0 + a_1 * x.[(1)]

V. Experimental Result

A. Data Processing

The algorithm learns from the training dataset and then compares the predicted value column with the actual data column of the dataset.

Below is the flowchart of the approach.

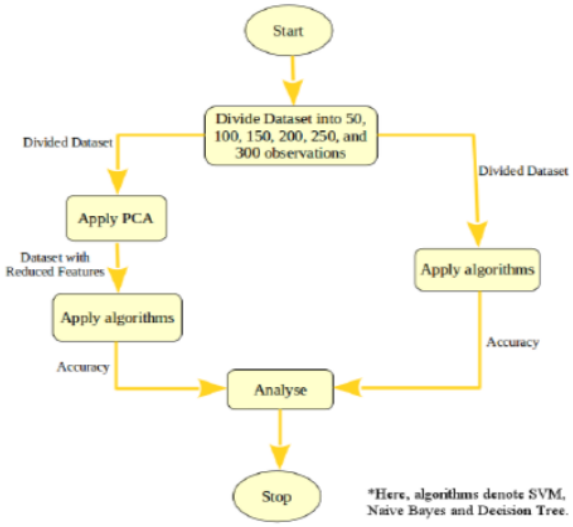


Figure4-Flowchart of the approach

It must be translated into two columns with the value 1 where the column is true, and 0 where it is false.

```
# Original Column
# |  Gender  |
# |     1    |
# |     1    |
# |     0    |

# Dummy Columns
# |  Gender_0  ||  Gender_1  |
# |     0      ||     1      |
# |     0      ||     1      |
# |     1      ||     0      |
```

Figure5-Dummies from Pandas

We have divided the age ranges into elderly, middle-aged, and young for better analysis.

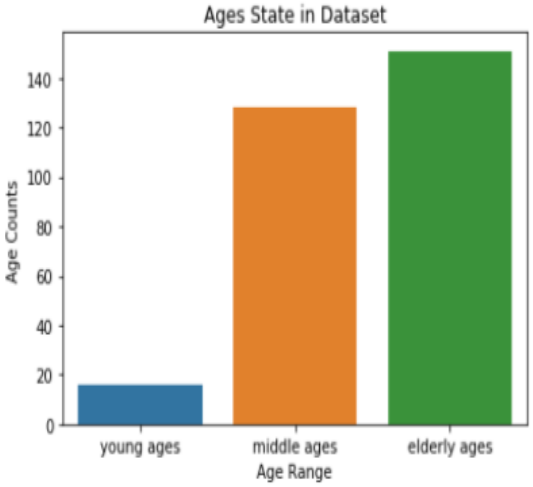


Figure6-Division of age ranges

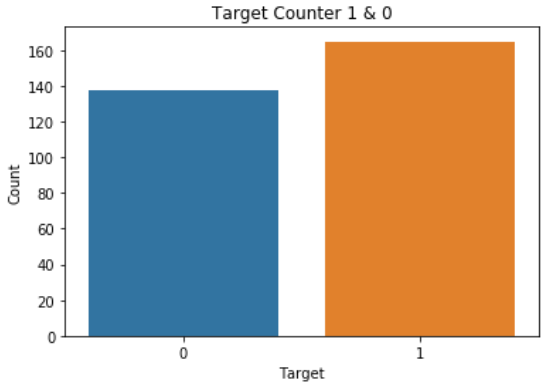


Figure7-Barplot for Target class

We are taking column Gender, with values 1 for males and 0 for females. It needs to be converted into two columns with the value 1 where the column would be true and 0 where it will be false.

B. Data Analysis

Now we look at the people's ages who are suffering from heart disease or not. In this case, target = 1 means that the person has heart disease, and target = 0 means that the person is not suffering.

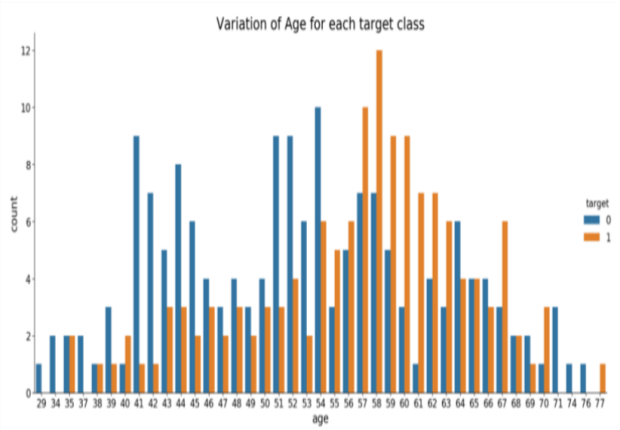


Figure8-Age Variation Chart

It seems that people aging 50+ are the ones who suffer the most from heart disease.

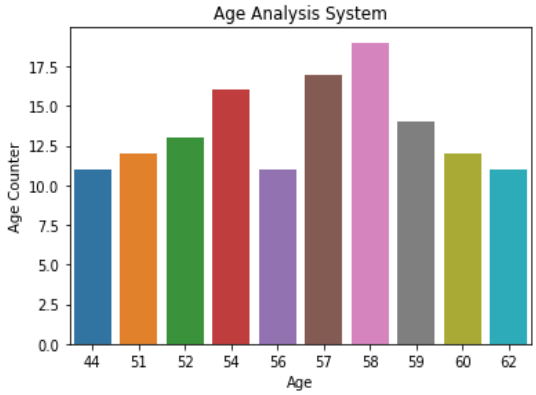


Figure9-Age analysis

We categorized chest pain with 4 values which are 1 for least, 2 for slightly distressed, 3 for medium pain, and 4 for too bad.

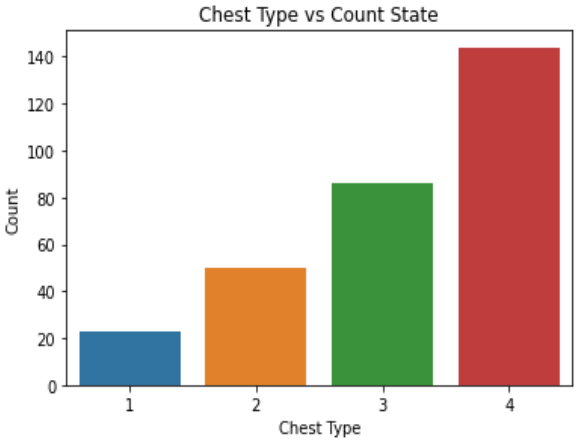


Figure10-Chest type chart

Now, we will compare the thal values with the targets.

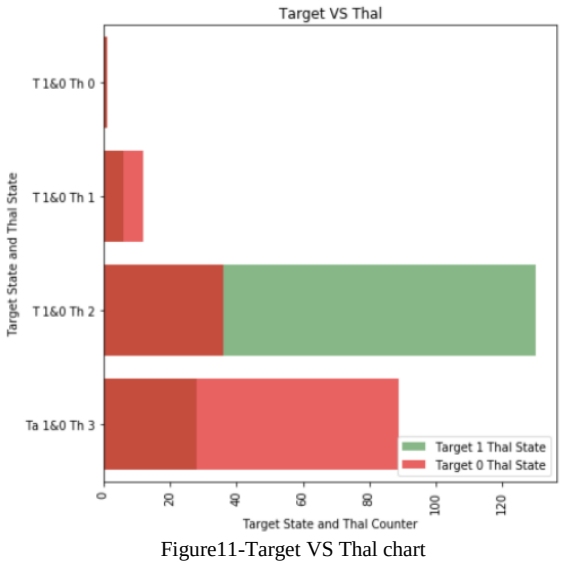


Figure11-Target VS Thal chart

C. Data Pre-Processing

The dataset is composed of 14 columns and 303 rows. Let us check the null values.

```
Age          0
Sex          0
Cp           0
Trestbps     0
Chol         0
Fbs          0
restceg      0
Thalach      0
Exang        0
Oldpeak      0
Slope        0
Ca           0
Thal         0
Target       0
dtype: int64
```

Figure12-Null values each column of data

We can either remove them or impute them since the null values are very fewer.
We have divided the data set into an 80% train data set and a 20% test dataset of the whole dataset respectively.

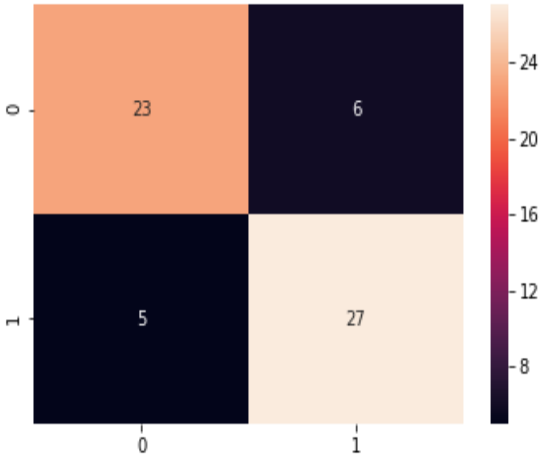


Figure13-Confusion Matrix

Other useful criteria to compare two models, is using the confusion matrix, which allows the visualization of the performance of an algorithm.[5] The confusion matrix displays the correct as well as incorrectly predicted values by a classifier.
In order to apply the logistic regression classification algorithm we have to separate dependant and non-dependant variables within our datasets. Now we can perform scale operations with z-score or normalization process. We have chosen the trade-off parameter as C=0.1. It determines the strength of regularization.
Below are the all correlation values between the data.

	Age	Sex	Cp	Trestbps	Chol	Fbs	restceg	Thalach	Exang	Oldpeak	Slope	Target	AgeRange
Age	1.000000	-0.097542	0.104139	0.284946	0.208950	0.118530	0.148868	-0.393806	0.091661	0.203805	0.161770	0.222853	0.806614
Sex	-0.097542	1.000000	0.010084	-0.064456	-0.199915	0.047862	0.021647	-0.048663	0.146201	0.102173	0.037533	0.224469	-0.030375
Cp	0.104139	0.010084	1.000000	-0.038077	0.072319	-0.039975	0.067505	-0.334422	0.384060	0.202277	0.152050	0.407075	0.090596
Trestbps	0.284946	-0.064456	-0.038077	1.000000	0.130120	0.175340	0.146560	-0.045351	0.064762	0.189171	0.117382	0.157754	0.222292
Chol	0.208950	-0.199915	0.072319	0.130120	1.000000	0.009841	0.171043	-0.003432	0.061310	0.046564	-0.004062	0.070909	0.132921
Fbs	0.118530	0.047862	-0.039975	0.175340	0.009841	1.000000	0.069564	-0.007854	0.025665	0.005747	0.059894	0.059186	0.130347
restceg	0.148868	0.021647	0.067505	0.146560	0.171043	0.069564	1.000000	-0.083389	0.084867	0.114133	0.133946	0.183896	0.159797
Thalach	-0.393806	-0.048663	-0.334422	-0.045351	-0.003432	-0.007854	-0.083389	1.000000	-0.378103	-0.343085	-0.385601	-0.415040	-0.299427
Exang	0.091661	0.146201	0.384060	0.064762	0.061310	0.025665	0.084867	-0.378103	1.000000	0.288223	0.257748	0.397057	0.065406
Oldpeak	0.203805	0.102173	0.202277	0.189171	0.046564	0.005747	0.114133	-0.343085	0.288223	1.000000	0.577537	0.504092	0.146949
Slope	0.161770	0.037533	0.152050	0.117382	-0.004062	0.059894	0.133946	-0.385601	0.257748	0.577537	1.000000	0.377957	0.140733
Target	0.222853	0.224469	0.407075	0.157754	0.070909	0.059186	0.183896	-0.415040	0.397057	0.504092	0.377957	1.000000	0.162808
AgeRange	0.806614	-0.030375	0.090596	0.222292	0.132921	0.130347	0.159797	-0.299427	0.065406	0.146949	0.140733	0.162808	1.000000

Figure14-Correlation values

Another very useful tool to compare two models is the Receiver operating characteristic (ROC) curve, which describes the diagnostic-ability of a binary classification model. This plot is created by plotting true positive rate against the false-positive rate. A true positive rate which is also called sensitivity is defined as follows:

Sensitivity = TP/P

where TP is the number of samples that are correctly classified as positive and P is the total number of samples with a positive label.[4]
The false-positive rate which is also called fall-out is defined as follows

Fallout = 1 – TN/N

where TN is the number of samples which are correctly predicted as negative, and N is the total number of samples with negative labels.[4]

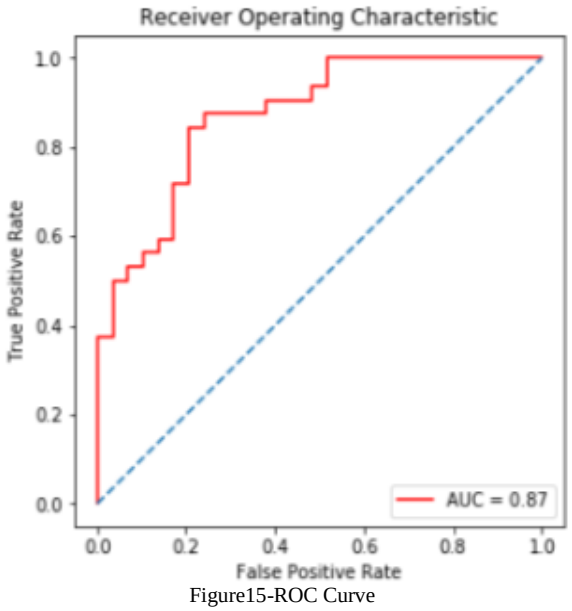


Figure15-ROC Curve

We can plot different models in a single ROC curve and the model with the bigger area under its curve is the more preferred model.

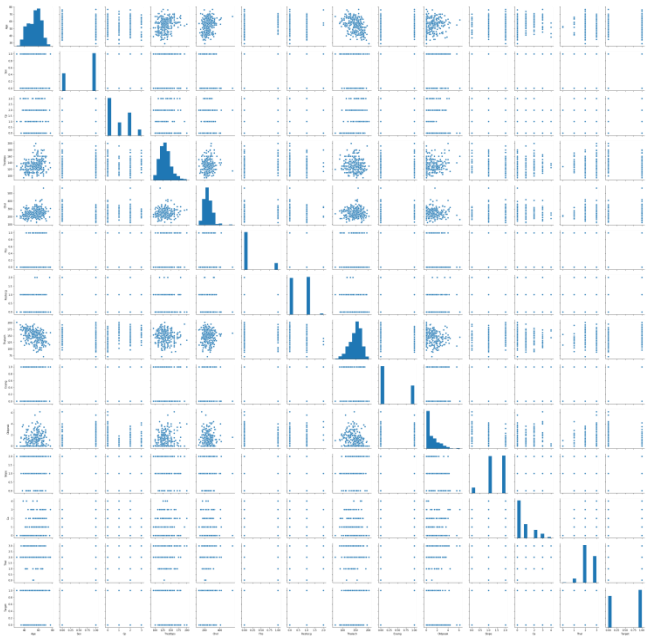


Figure16-Pairplot for Feature vectors

We know that total variance is the sum of all individual principal components and the fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance. Moreover, we know that 86% of the variance is explained by the first 5 principal components.

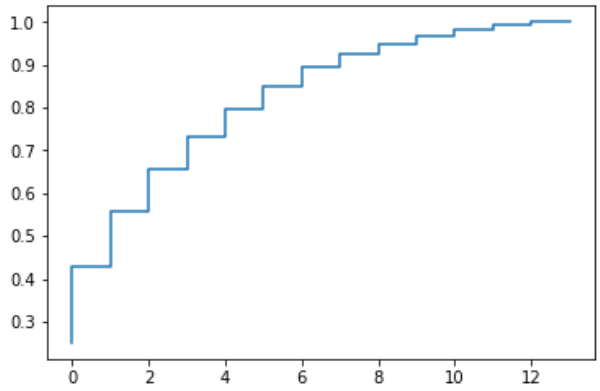
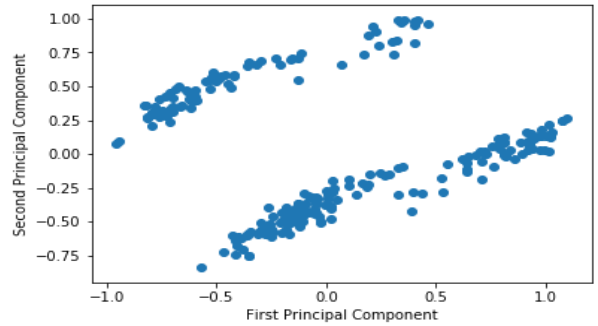


Figure17-Explained Variance ratio graph

After calculating the principal components, they can be plotted against each other to have a better insight into their relationships.

So, I have applied PCA to the data with the number of components equal to 8. You can see the reduced data on the plot below



.

Figure18-Scatter plot for PC2 vs PC1(training dataset)

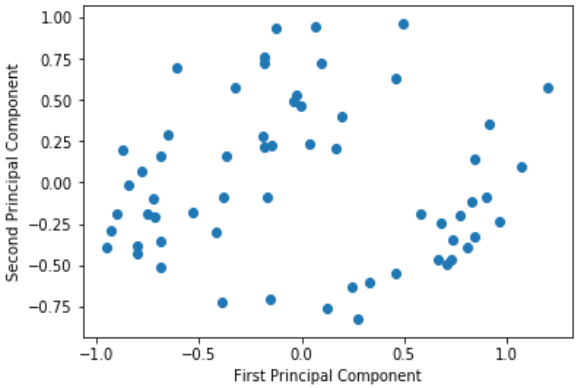


Figure19-Scatter plot for PC2 vs PC1(test dataset)

VI. Conclusion

In this report, we applied Principal Component Analysis to study the performance of clustering algorithms on the Cleveland heart disease dataset. The report is divided into two parts. In the first section, we apply PCA on a dataset of 303 people corresponding to the 14 specific features. We found that 88.3% of the explained variance was in the first 2 principal components. Moreover, we noticed that the old aged people had their values high for most of the chosen features. We used all the features from the dataset to reduce its dimensionality to get a covariance matrix. The dimension reduction of PCA is important to visualize data in higher-dimensional datasets in clustering problems. In the second section, we have used logistic regression for classifying the observations. The performance of logistic regression is considered by using all the features dataset and using the first two principal components obtained from PCA. The results show that using all features leads to higher accuracy.
The results show that PCA improves the performance of clustering methods. This will help in the development of intelligent systems in the future leading to the selection of proper treatment methods to diagnose a patient with heart disease.

VII. References

1. https://www.ijrte.org/wp-content/uploads/papers/v7i6s4/F10760476S419.pdf
2. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc
3. https://en.wikipedia.org/wiki/Principal_component_analysis
4. https://www.sciencedirect.com/science/article/pii/S2210832718301546
5. https://machinelearningmastery.com/confusion-matrix-machine-learning/
6. https://granthealth.org/heart-disease/
7. https://www.degruyter.com/view/journals/phys/17/1/article-p489.xml