

Project Report
on
**PREDICTION OF HEART FAILURE DISEASE USING STATISTICAL
ANALYSIS**

Submitted

In partial fulfilment of the requirement for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

REVATHY.T.V.

(Register No. SM22AS015)

(2023-2024)

Under the supervision of

Ms. PARVATHY T S



DEPARTMENT OF MATHEMATICS AND STATISTICS (SF)

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI-682011

MAY 2024

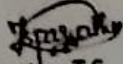


CERTIFICATE

This is to certify that the dissertation entitled, **PREDICTION OF HEART FAILURE DISEASE USING STATISTICAL ANALYSIS** is a Bonafide record of the work done by **REVATHY.T.V.** under my guidance as partial fulfilment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 29/04/24

Place: Ernakulam


Parvathy TS


Assistant Professor,

Department of Mathematics and Statistics

St. Teresa's College (Autonomous),

Ernakulam





Mrs Nisha Oommen

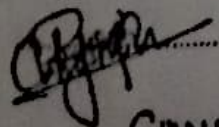
Assistant Professor & HOD,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous),


Ernakulam.

External Examiners

1. 

CHINU JOSEPH

29/4/2024.

2. 
LAKSHMI SURESH.

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **PARVATHY T S**, Assistant professor, Department of mathematics and statistics, St. Teresa's College (Autonomous), Ernakulam and has been include in any other project submitted previously for the award of any degree.

Ernakulam

Date: 29/04/24



Revathy.T.V.

SM22AS015

ACKNOWLEDGEMENTS

I take this opportunity to thank everyone who has encouraged and supported me to carry out this project.

I am very grateful to my project guide Ms. Parvathy T S for her immense help during the period of work.

In addition, I acknowledge with thanks to the Department for the all the valuable support and guidance that has significantly contributed to the successful completion of the project.

I would also like to thank the HOD for her valuable suggestions and critical examinations of the project.

Place: Ernakulam

Date: 29/04/24

Revathy.T.V.

SM22AS015

ABSTRACT

Because of the Heart failure disease became one of the primary choices for humans. Day to day I heard about the heart failure after the CORONA after 2020, due to that decrease the work out, walking etc. Evaluation of the major factors affecting heart disease should be consider to improve the health conditions

In this project, this study uses patients details and no heart failure results as the research object. First reduce the dimension of the dataset using principal component analysis and find the important factors using factor analysis. Using machine learning algorithms to predict heart failure disease.

This study referenced for the predicting heart failure disease have no heart failure of the patients and find the important factors leading to no heart failure.



ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM
Certificate of Plagiarism Check for Dissertation

Author Name	REVATHY.T.V.
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. PARVATHY T S
Department	Post Graduate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	PREDICTION OF HEART FAILURE DISEASE USING STATISTICAL ANALYSIS
Similarity	6% AI 7%
Paper ID	1690475
Submission Date	2024-04-24 12:49:19

Signature of Student

Signature of Guide

Checked By
College Librarian

Contents

CERTIFICATION	1
DECLARATION	2
ACKNOWLEDGMENT	3
ABSTRACT	4
CONTENT	5
1.INTRODUCTION	6
1.1 Introduction	
1.2. Objectives	
2.LITERATURE REVIEW	9
3.METHODOLOGY	12
3.1 About data	
3.2 Software used	
3.3 Methodology	
3.3.1 Principal component analysis	
3.3.2 Factor analysis	
3.4 Machine learning	
3.4.1 Decision tree	
3.4.2 Logistic regression	
3.4.3 Support vector classifier	
3.4 Machine learning	
3.41 Logistic regression	
3.42 Decision tree	
3.43 Support vector classifier	

4. DATA ANALYSIS	21
4.1 Introduction	
4.2 Principal component analysis	
4.3 Factor analysis	
4.4 Data visualization	
4.5 Data distribution graphs	
5. MODEL PREDICTION	32
5.1 Logistic regression	
5.2 Decision tree	
5.3 support vector classifier	
5.4 Prediction result	
6. CONCLUSIONS	38
7. REFERENCE	39

CHAPTER 1

INTRODUCTION

Heart disease is a condition that affected the heart and blood vessels. It is also called cardiovascular disease. Heart disease include various disorders such as coronary artery disease (CAD), Heart attack (myocardial infarction), Heart failure, Arrhythmias, Heart valve disease, Congenital heart defect.

Risk factors for heart disease include high blood pressure, high cholesterol, diabetes obesity, smoking, unhealthy diet, excessive alcohol consumption and age. These risk factors are modifiable through the lifestyle changes and medical treatments.

Symptoms of heart disease is varying depending on the certain condition but may include chest pain, shortness of breath, swelling of legs and fainting.

Prevention of heart disease lifestyle modifications such as quitting smoking, eat a healthy diet, exercising regularly etc. and some cases surgical interventions.

Early detection and treatment of heart disease are essential for preventing and improving outcomes.

In the case of heart failure the heart muscle can't pump blood . when it happens, the blood often backs up and fluid can build up in the lungs, causing shortness of breath.

Certain heart situation gradually leaves the heart too week and pump blood properly. This situation include narrowed arteries in the heart and high blood pressure.

Proper treatments help to improve the symptoms of heart failure and it helps some people live longer. Life changes also can improve the quality of life. Try to weight loss, exercise, reduce the amount of salt and manage stress.

OBJECTIVES

- To reduce the dimensionality of the data using PCA
- To identify underlying factors that contribute to heart disease use by factor analysis
- To predict heart disease using logistic regression and support vector classifier
- To compare the accuracy and selected the best model for prediction

CHAPTER 2

LITERATURE REVIEW

- Levy et.al (2006)- Study aimed to develop and validate a multivariate risk model for predicting 1,2, and 3, year survival in heart failure patients, using easily available clinic, devices, and laboratory characteristic. Heart failure a valuable tool for clinical in assessing patient and optimizing treatment strategies.
- Brian Steinhart et.al (2009)-The study aims to derive and validate a prediction model by using N-terminal pro-B-type natriuretic peptide and clinical variables to improve the diagnosis heart failure.500 patients details in the data set. A diagnostic prediction model for AHF that incorporates both clinical assessment and NT-proBNP has been derived and validated and has excellent diagnostic accuracy, especially in the case with indeterminate likelihood for AHF.
- Melissa Jehn et.al (2009)- This study aimed to use accelerometers to measure daily walking performance in chronic heart failure (CHF) patient to determine its associate class and exercise capacity. 50 CHF patients are accelerometers for 7 days, with results showing that total walking time (TWT) correlated strongly with maximal and functional exercise capacities. Patients with higher NYHA classes had lower TWT and spent less time in fast walking mode. TWT and fast walking time were strong predictors of moderate CHF. Monitoring daily walking performance may help detect disease progression and improve clinical outcomes in CHF.
- K Ramini et.al (2014)- study to find about the risk prediction models for patient with heart failure , they only focus to identify consistent predictors of risk across model. The variables of the data are age, renal function, blood pressure, blood sodium level, left ventricular ejection fraction, sex, brain natriuretic peptide level etc. They use 64 model for analysis. The conclusion is several clinical useful for death prediction model with heart failure.

- Javed butler et.al (2015)-To examine the prevalence of heart failure is expected to rise significantly unless high-risk patients are effectively screened and appropriate prevention interventions are implemented. A systematic review evaluated published heart failure risk prediction models up to August 2014, revealing 28 models across 13 publications. These models demonstrated acceptable-to-good discriminatory ability in derivation samples but lacked external validation. Overall, incident heart failure risk prediction remains in its early stages, with further research needed to assess cost-effectiveness, population screening suitability, and clinical impact.
- Tiwasker et.al (2018)-This study focuses on early detection of heart failure in India, where it leads to death. The aim is to classify records into presence or absence of heart failure risk using Statistical, Machine Learning, and Data Mining techniques. The study compares the performance of statistical evaluation, Decision Tree classifier, Random forest classifier and convolution Neural Network (CNN) here CNN shows the highest accuracy of 93%.
- Adler et.al (2020)-Study to examine a machine learning algorithm to enhance mortality prediction accuracy in patients with heart failure. The dataset of 5822 heart failure patients and 8 key variables are blood pressure, haemoglobin level between low and high risk and mortality rate in low and high risk etc and the decision tree model is used. To identify the potential of machine learning techniques to improve risk assessment in HF patients and other challenging clinical sectors.
- PK sahu and P jeripothula (2020)-The study to aim the heart disease during physical stress and mental disorder. The busy world affected by stress due to some reasons, which has led to an increase in heart disease particularly in urban areas. Heart disease including coronary artery disease(CAD) this lead to death. Predicting heart disease has become difficult requiring extensive health records and sometime genetic disorders. The dataset has 13 attributes, using SVM, Naïve Bayes, Logistic Regression, Decision

Prediction of heart failure disease using statistical analysis

Tree and KNN algorithm. SVM showing the highest accuracy at 85.2%. This model to ensure the reliability of the proposed system in real-world.

- Jing wang (2021) - To examine a prediction for this disease is one of the key approaches of decreasing its impact. In worldwide every year 550,000 patients are affected by heart failure. Both linear and machine learning models are used to predict heart failure based on various data. The results using z-score normalization and SMOTE for heart failure prediction.
- Ashir Javeed et.al (2022)- To examine ML based automated diagnostic systems developed for heart failure prediction using different types of data modalities. Heart disease is a leading global cause of mortality, often resulting from conditions like coronary artery disease and chronic heart failure. Traditional diagnostic methods such as angiography are costly and leads health risks, prompting the development of automated diagnostic systems machine learning (ML) and data mining techniques. These systems offer affordable, efficient, and reliable solutions for heart disease detection.

CHAPTER-3

MATERIALS AND METHODOLOGY

3.1 ABOUT DATA

For the purpose of the present study, data was collected from Kaggle. They consist of 500 rows and 13 variables. The table consist of Age, sex, thalach, exang, slope, cp, ca, trestbps, fibs, chol, restecg, oldpeak, thal, target.

3.2 SOFTWARE USED

SPSS is a statistical software package used for interactive or batched, statistical analysis. It was Obtained by IBM in 2009. IBM SPSS statistics is the current version brand name. The software name originally stood for statistical package for the social science, indicating the original market, then changed to statistical product and service solutions.

PYTHON PROGRAMMING

The data analysis is done using python programming. Python is a high-level. Intercepted, general-purpose Programming language. It emphasizes code readability. Python is dynamically typed, and garbage collected. It facilities multiple programming paradigms.

- NumPy
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn

3.3 Methodology

The initial and crucial step in this analysis was to study the data in detailed. The main purpose of the study was to predict heart failure disease using different machine learning models. The data for the project is multivariate data.

Dimension Reduction

Dimension Reduction refers to the process of converting a set of data having larger dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. The various methods used for dimensionality reduction include: principal component analysis, factor analysis etc. These techniques are typically used while solving machine learning problems to obtain better features for a classification or regression task.

The main benefits of Dimension Reduction are

- I. It helps in data compressing and reducing the storage space required.
- II. It fastens the time required for performing same computations.
- III. It takes care of multicollinearity that improves the model performance.
- IV. Reducing the dimensions of data to 2D or 3D may allow us to plot and visualize it precisely.

3.3.1 Principal Component Analysis

Principal component analysis is concerned with explaining the variance-covariance structured through a fewer linear combination of the original variables. Its general objectives are

1. Data reduction
2. Interpretation

Although P components are required to reproduce the total system variability, often much of this variability can be explained by a smaller k of the principal components. If so, there is almost as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on k principal components.

Definition: Let X be a p component random vector with dispersion matrix Σ . Then, the first principal component is the linear combination $l'_1 X$ that maximizes $\text{var}(l'_1 X)$ subject to $l'_1 l_1 = 1$. The second principal component is the linear combination $l'_2 X$ that maximizes $\text{Var}(l'_2 X)$ subject to $l_2 l_2 = 1$, and $\text{Cov}(l'_1 X, l'_2 X) = 0$. The i^{th} principal component is the linear combination

$l'_i X$ that maximizes $\text{Var}(l'_i X)$ subject to $l'_i l_i = 1$, and $\text{Cov}(l'_i X, l'_k X) = 0, k < i, i = 1, 2, \dots, p$. Thus, the principal components are those uncorrelated normalized linear combinations $U_i = l'_i X, i = 1, 2, \dots, p$ whose variance $\text{Var}(U_i) = l'_i \Sigma l_i$ are as large as possible.

3.3.2 Factor Analysis

Factor analysis is a reduced technique belonging to the multivariate statistical technique. Factor Analysis (FA) assumes the covariation structure among a set of variables can be described via a linear combination of unobservable (latent) variables called factors. The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.

There are three typical purposes of FA:

- I. Data reduction: explain covariation between p variables using $m < p$ latent factors
- II. Data interpretation: find features (i.e., factors) that are important for explaining covariation (exploratory FA)
- III. Theory testing: determine if hypothesized factor structure fits observed data (confirmatory FA)

Difference between Factor Analysis and Principal Component Analysis

Both factor analysis and principal component analysis have the goal of reducing dimensionality. But they differ in many respects. Some basic differences are the following:

Prediction of heart failure disease using statistical analysis

- I. In factor analysis, the original variables are expressed as linear combination of latent variables called factors. Principal components are defined as uncorrelated linear combinations of the original variables that explain maximal variance.
- II. FA refers to a statistical model, whereas PCA refers to the eigen value decomposition of a covariance (or correlation) matrix.
- III. In principal component analysis, we explain a large part of the total variance of the variables. In factor analysis, we find the covariances and correlations among the variables.
- IV. The principal components are unique (assuming distinct eigenvalues of S), where the factors are subject to an arbitrary rotation.

Types of Factor Analysis

Factor analysis are two types:

- I. Exploratory Factor Analysis (EFA)
- II. Confirmatory Factor Analysis (CFA)

Exploratory Factor Analysis (EFA) is a variable reduction technique used to explore the possible underlying factor structure of a set of observed variables without imposing a preconceived structure on the outcome. By performing EFA, the underlying factor structure is identified.

Confirmatory factor analysis (CFA) is a statistical technique used to verify the factor structure of a set of observed variables. CFA allows the researcher to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists. The researcher uses knowledge of the theory, empirical research, or both, postulates the relationship pattern a priori and then tests the hypothesis statistically.

Criteria for determining the number of factors

According to Kaiser Criterion, Eigen value is a good measure for find a factor. If eigen value is greater than 1, we consider that factor otherwise, that factor should not be considered.

Scree plot test

This method is to decide about the number of factors that should retain from the extracted factors. The test along with the plot determines which of the factors are actually contributing variance. The number of factors is plotted against proportion of variance. It extracts in the order of the extracted factors.

Kaiser Meyer Olkin Test (KMO Test)

It is the measure of sampling adequacy used to compare the magnitudes of the observed correlation coefficients in relation to the magnitude of the partial correlation coefficients. Large KMO values are good since correlation between pairs of variables can be explained by other variables. KMO value between 0.8 and 1 indicate that the sampling is adequate. KMO value is less than 0.5 indicate that the sampling is not adequate and that necessary actions should be taken. KMO value close to zero that means that there are large partial correlations compared to the sum of correlations.

3.4 MACHINE LEARNING

Machine learning is said to be a segment of artificial intelligence that is primarily concerned with the development of algorithms which permit a computer to understand from the data the past experiences on their own. The author machine learning was Arthur Samuel in 1959. From historical data machine learning system studied and it is called training data. Machine learning creates prediction model, and whenever it receive new data, predicts the output for it.

The accuracy of predicted output depends up on the amount of data. The huge amount of data support making a better model and it predicts the output more accurately.

Types of machine learning

There a lot of way to train machine learning algorithm, each of them have their own merits and demerits. To analyse the merit and demerit of each type of machine learning, look at the type of data. Labelled and Unlabelled data the two kinds of data in machine learning. Labelled data consist both the input and output parameters. They are in a completely machine-readable patterns. This disadvantage is that, it requires a lot of human labour to label the data.

Unlabelled data only contain one or more of the parameters in a machine-readable form.

This denies the need for human labour but needs complex solutions.

- Supervised learning

Supervised learning is one of the most important types of machine learning algorithm is arranged in labelled data.

- Unsupervised learning

Unsupervised machine learning is the ability to work with unlabelled data. Here human labour is not required to produce the dataset machine-readable , allow much larger datasets to be worked.

- Reinforcement learning

Reinforcement learning directly helps to understand how human being learn from data I their lives. This algorithm enhances from itself. It learns from new situations using a trival-and-error method. Favourable outputs are encourage or 'reinforced', and also nonavoidable outputs are discouraged or 'punished'.

3.4.1 DECISION TREE

A decision tree is a supervised ML algorithm. It can be for classification and regression tasks. It is a flow-chart or tree like structure. Each internal node provides a decision based on a feature outcome. Outcome of the decision is given by each branch, and final prediction is

Prediction of heart failure disease using statistical analysis

given by each leaf node. In customer relationship management, Credit scoring, medical diagnosis, Fraud detection, predictive maintenance, decision tree is used.

Decision tree works by:

1. Decision tree is made -up of nodes, branches and leaves. The top node is known as root node. It represents the initial decision. The internal node gives features or attributes and branches provide the possible values or outcome of those features. Leaf node provides the final predictions.
2. Splitting criteria: The algorithm divides the data into subsets on the features. To improve the purity of the subsets is the goal.
3. Tree growth: The decision tree grows by dividing the data at each node based on the chosen criteria. This will continue until a stopping criterion is met.
4. Prediction: When the decision tree is built, it can be used for prediction on unseen data. The final decision is get from the leaf node.

3.4.2 LOGISTIC REGRESSION

Logistic regression is a technique used for binary classification tasks. Using a logistic curve, it can be used for predicting the probability of occurrence of an event. In credit scoring, medical diagnosis, marketing analytics and fraud detection logistic regression is used.

Key concepts are:

Binary classification

For binary classification problems, logistic regression is used. That is where the target variable has only two possible outcomes.

- Probability Estimation

Unlike predicting the outcome directly as in linear regression, logistic regression predicts the probability of an event.

Prediction of heart failure disease using statistical analysis

- Sigmoid function

Logistic regression uses the sigmoid function to model the relation between the dependent and independent variables. It gives that the probabilities lies between 0 and 1.

- Maximum likelihood estimation

By maximizing the likelihood function, logistic regression estimates the model parameters.

- Decision boundary

To classify instances into one of the two classes, decision boundary is used. The decision boundary is set at 0.5.

- Cost function

Logistic regression lowers the cost function at the time of training. It corrects predictions and to enhance the performance, the model parameters are organized.

3.4.3 SUPPORT VECTOR CLASSIFIER

Support Vector Classifier or Support Vector Machine is a powerful and supervised ML algorithm. It is used for binary and multi class classification tasks. It used optimal hyperplane. That hyperplane separates the classes in feature space.

Key concepts are:

Margin maximization:

To find a hyperplane that maximizes the margin between the classes is the aim. The margin is distance between the hyperplane and closer data point. The closer data points are known as support vectors. In this, improves robustness and generalization to unseen data.

Kernel trick:

Prediction of heart failure disease using statistical analysis

By changing the input features into a higher dimensional space using kernel functions, Support vector classifier can handle non-linearly separable data. The classes become linearly separable in higher dimensional space and find an optimal separating hyperplane.

Regularization parameter:

SVC makes a regularization parameter(C) to balance the trade-off between maximizing the margin and reducing the classification errors.

Soft margin and slack variables:

Support vector classifier provide slack variables to allow for classification error in the cases where data is not perfectly separable. This is called soft margin classification and stops overfitting by giving some misclassifications.

CHAPTER 4

DATA ANALYSIS

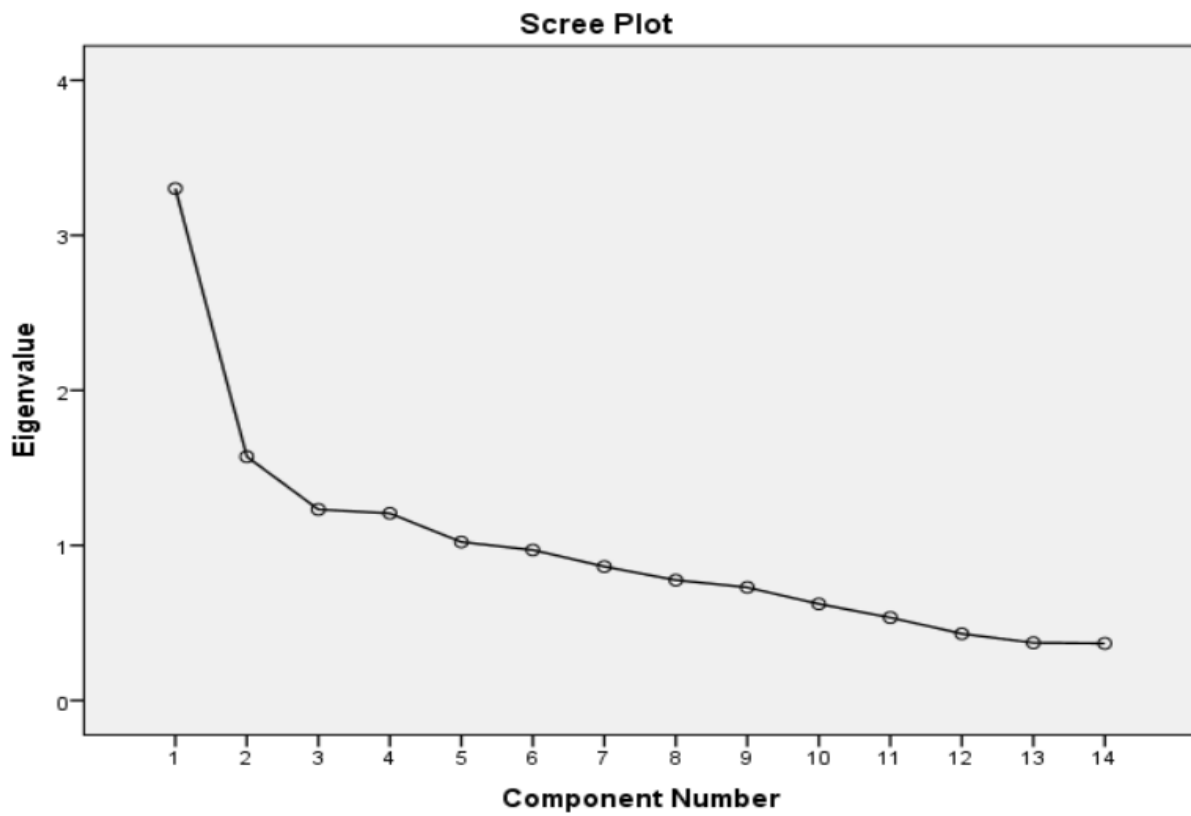
4.1 Introduction

The data is collected from kaggle containing 500 rows and 14 columns. In this project, first reduce the dimensionality of heart disease dataset and find the most influencing factors. Then predict heart failure disease using Machine learning techniques. The dependent variable is satisfaction, and the independent variables are;

- 1.Age
- 2.Sex
- 3.t halach – The person’s maximum heart rate achieved.
- 4.Exang – Exercise induced angina
- 5.Slope
- 6.cp – Chest pain
- 7.ca – Calcium
- 8.trestbps – The person’s resting blood pressure
- 9.fbs – Fasting glucose levels <70 mg/dL
- 10.chol – Cholesterol
- 11.restecg – Resting electrocardiographic measurement
- 12.Oldpeak
- 13.thal – Thalassemia
14. Target – Output class [1: heart disease, 0: normal]

4.2 Principal component analysis

The data is a set of correlated variables. The data is checked for normality, and it is found true by p-p plot in the SPSS package. Then conducted the principal component analysis on the correlation matrix and so the variables are standardized.



From the given figure, the slope of the curve is levelling off after the fifth component, which implies that the first 5 principal component effectively summarize the total variance.

Total variance explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.301	23.582	23.582	3.301	23.582	23.582	2.213	15.809	15.809
2	1.572	11.230	34.812	1.572	11.230	34.812	1.945	13.889	29.698
3	1.232	8.800	43.612	1.232	8.800	43.612	1.498	10.702	40.400
4	1.207	8.619	52.231	1.207	8.619	52.231	1.378	9.844	50.245
5	1.022	7.300	59.531	1.022	7.300	59.531	1.300	9.286	59.531
6	.970	6.930	66.461						
7	.863	6.166	72.627						
8	.776	5.544	78.170						
9	.729	5.209	83.379						
10	.623	4.448	87.827						
11	.534	3.817	91.644						
12	.430	3.072	94.716						
13	.372	2.656	97.372						
14	.368	2.628	100.000						

Extraction Method: Principal Component Analysis.

From the given table, the first component explains 23.582% of the variation in the data set. The second component explains 11.230% of variation. The third component explains 8.800%, fourth component explains 8.619% and fifth component explains 7.300% variation of the dataset.

Component matrix

Component Matrix^a

	Component				
	1	2	3	4	5
target	-.796				
oldpeak	.672				
thalach	-.666				
exang	.610				
slope					
cp					
ca					
thal					
age					
trestbps					
sex			.621		
fbs					
chol					
restecg					

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

From the given table, first principal component is influenced by variable target, oldpeak, thalach, exang. The third principal component is influenced by sex.

4.3 Factor analysis

Factor analysis is used to determine major factors leading to patients is normal and not for heart failure disease.

KMO and Bartlett's test

The KMO test and Bartlett's test are used in factor analysis to obtain the adequacy of the data for data analysis

KMO Test: The KMO statistics values range from 0 to 1. Values closer to 1 indicate that the variables are highly correlated, and it is suitable for factor analysis. If the KMO value is below 0.5, then the variables are not suitable for factor analysis.

Prediction of heart failure disease using statistical analysis

Bartlett's test: This test finds the null hypothesis that the correlation matrix is an identity matrix.

It indicates that there are no correlations between variables.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.733
Bartlett's Test of Sphericity	Approx. Chi-Square	775.266
	df	91
	Sig.	.000

Since the value of KMO statistic is $0.733 > 0.5$ and the significance level for the Bartlett's test is below 0.05, we can conclude that data is suitable for factor analysis.

Total variance explained

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.301	23.582	23.582	3.301	23.582	23.582	2.213	15.809	15.809
2	1.572	11.230	34.812	1.572	11.230	34.812	1.945	13.889	29.698
3	1.232	8.800	43.612	1.232	8.800	43.612	1.498	10.702	40.400
4	1.207	8.619	52.231	1.207	8.619	52.231	1.378	9.844	50.245
5	1.022	7.300	59.531	1.022	7.300	59.531	1.300	9.286	59.531
6	.970	6.930	66.461						
7	.863	6.166	72.627						
8	.776	5.544	78.170						
9	.729	5.209	83.379						
10	.623	4.448	87.827						
11	.534	3.817	91.644						
12	.430	3.072	94.716						
13	.372	2.656	97.372						
14	.368	2.628	100.000						

Extraction Method: Principal Component Analysis.

From the table, a cut off an eigen value ≥ 1 would give 5 factors. After deciding the number of factors, the next step is to intercept the factor loadings. The factor loadings are the correlation between the factors and variables. Factor rotations help us interpret the factor loadings. The most used rotation is Varimax rotation. The component transformation matrix is given by table 4.4.

Component transformation matrix

Component Transformation Matrix

Component	1	2	3	4	5
1	-.680	-.599	.333	.144	.217
2	.370	-.088	.545	-.584	.467
3	.274	.221	.530	.756	.152
4	-.502	.698	-.060	-.081	.500
5	.271	-.311	-.555	.245	.680

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Applying Varimax rotation, the values of the rotated factor loadings are given by table 4.5

Rotated complex matrix

The values which are larger in magnitude are entered.

- Factor 1 is primarily a measure chest pain and target
- Factor 2 is a measure the sex
- Factor 3 is a measure slope and old peak
- Factor 4- cholesterol
- Factor 5 – Fasting BS, coronary artery

Prediction of heart failure disease using statistical analysis

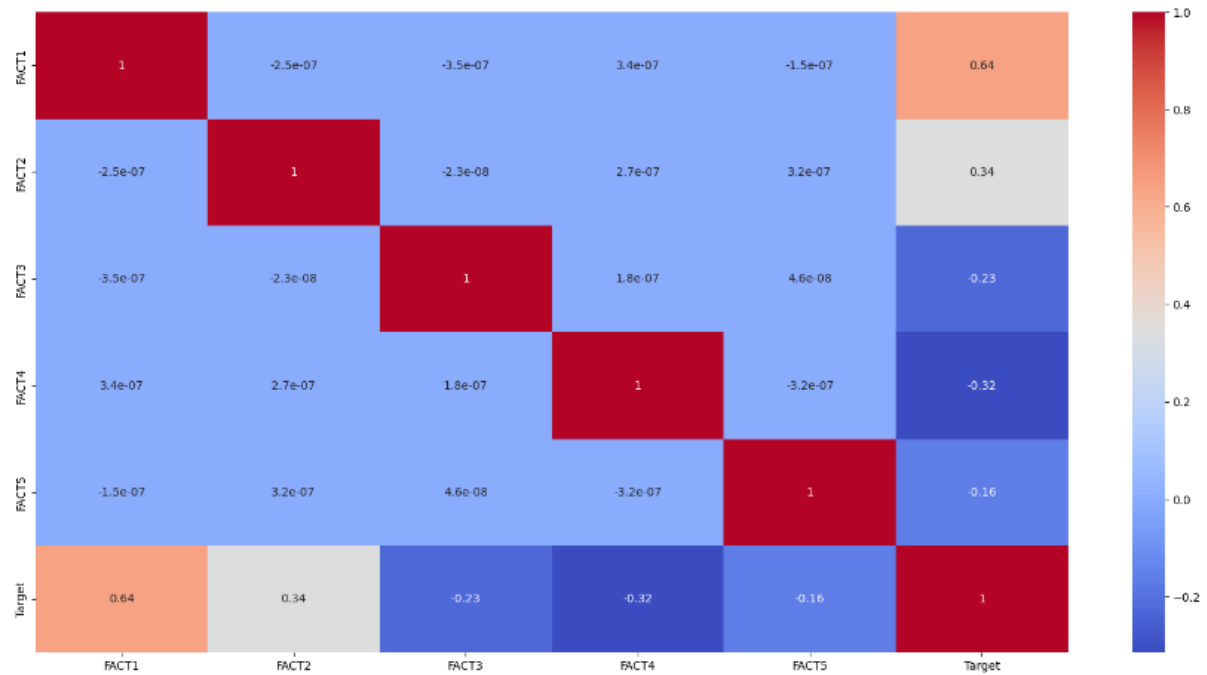
Rotated Component Matrix^a

	Component				
	1	2	3	4	5
cp	.785				
target	.641				
exang					
thalach					
slope		.865			
oldpeak		-.789			
fbs			.656		
ca			.631		
age					
sex				.809	
thal					
chol					.716
restecg					
trestbps					

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.
 a. Rotation converged in 8 iterations.

4.4 Data visualization

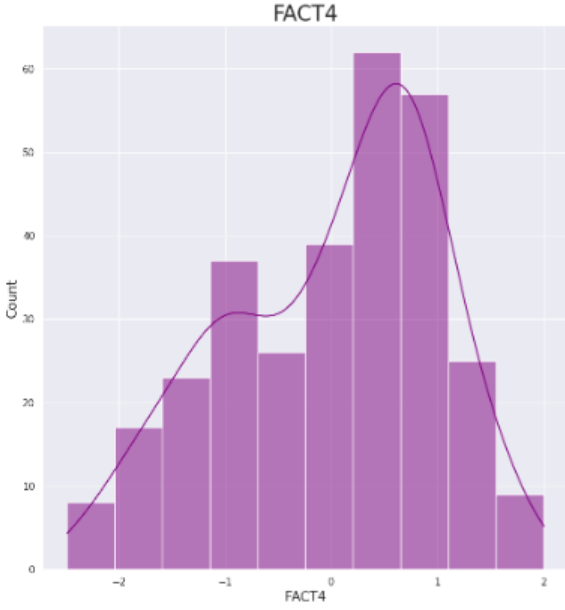
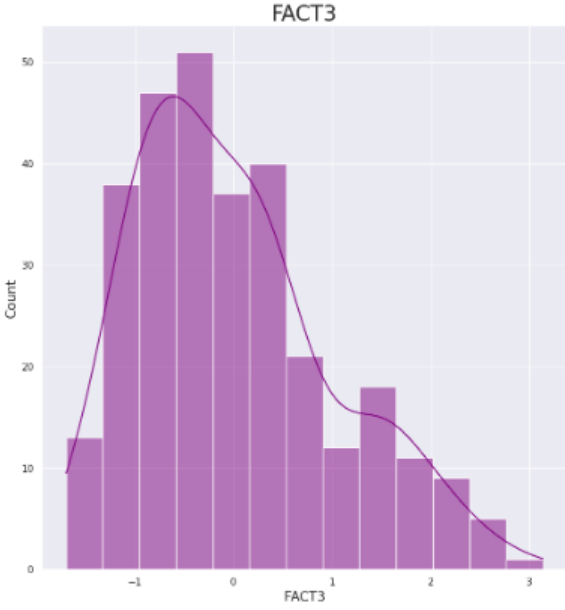
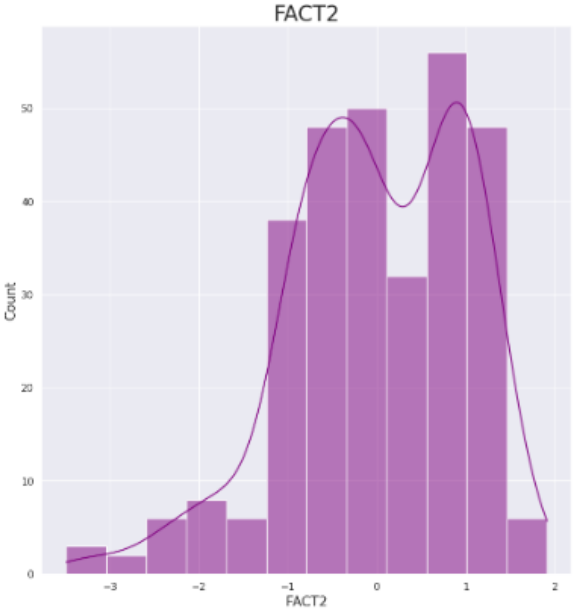
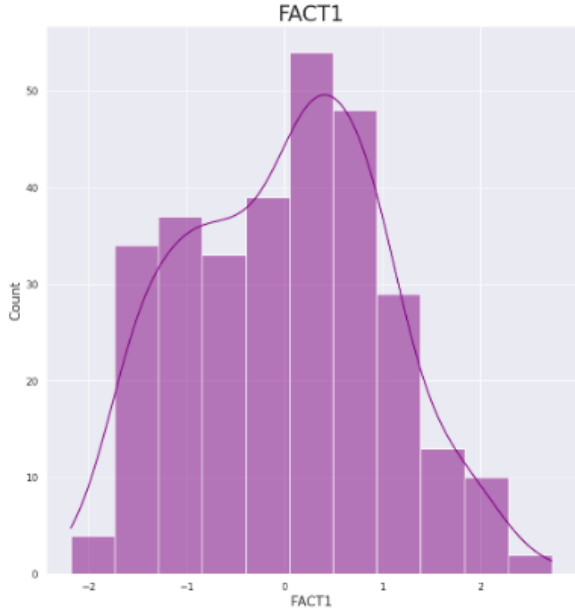
A table showing correlation coefficients between variables is called correlation matrix. Each cell in the table depicts a correlation between two variables. A correlation matrix is applied to recap data, as an input into a more advanced analysis, and as a diagnostic for advanced analysis.



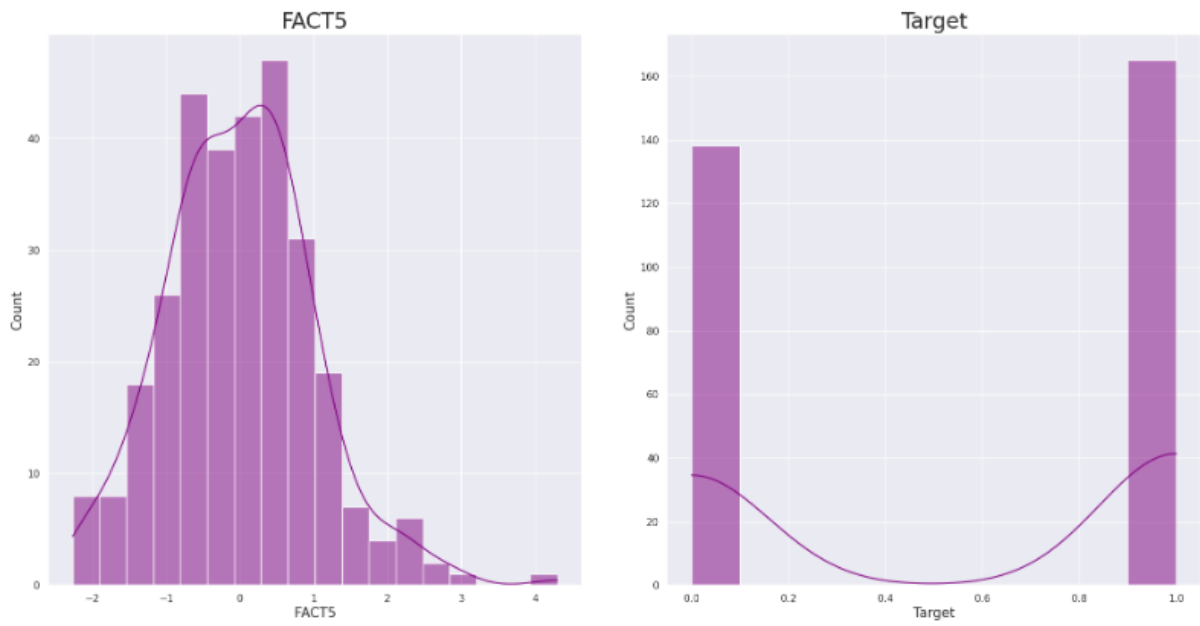
4.5 Data distribution graphs

Data distribution graphs are the visual representation that provide the distribution of data values in each dataset. These graphs give insights into the central tendency, spread and shape of the dataset.

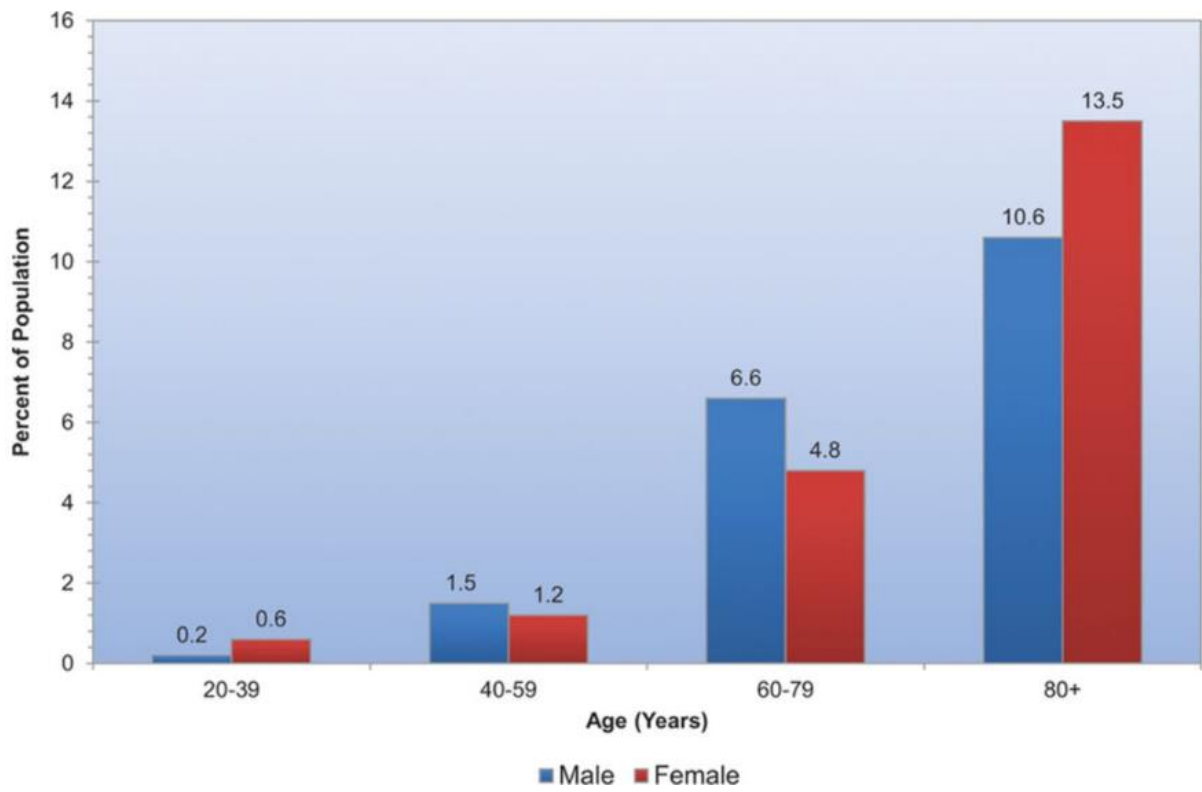
Prediction of heart failure disease using statistical analysis



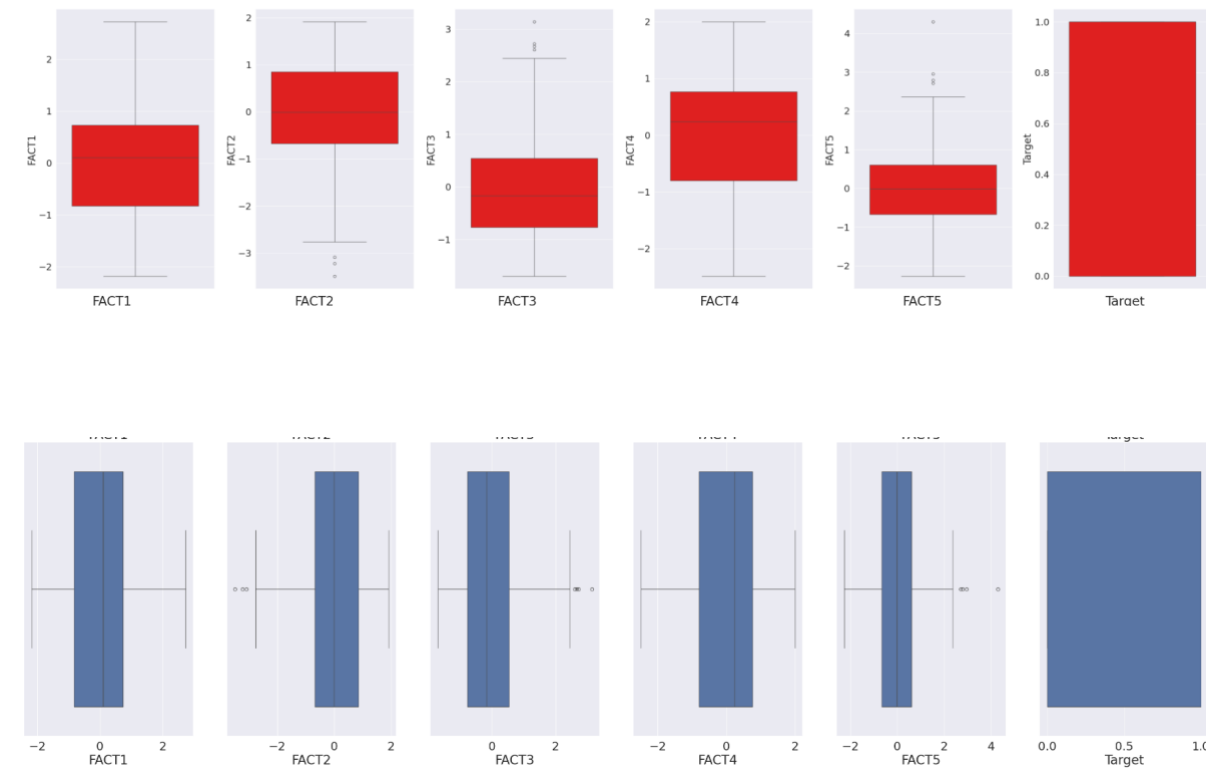
Prediction of heart failure disease using statistical analysis



Heart failure Vs Age



BOXPLOT REPRESENTATION OF FACTOR



Outlier is present in fact3 and fact5

CHAPTER 5

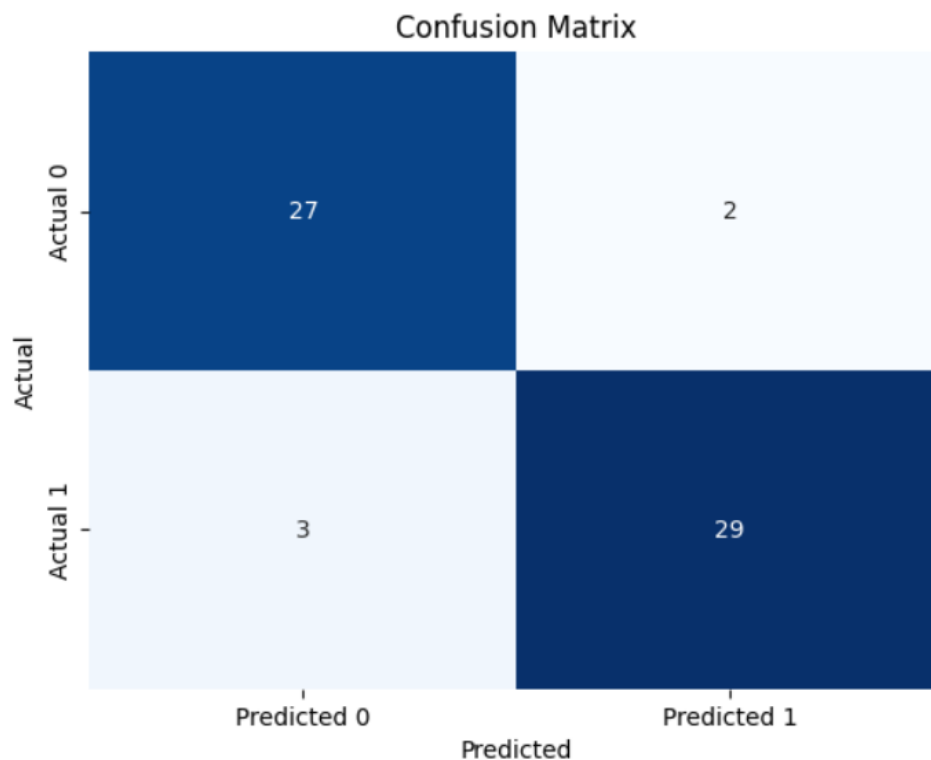
MODEL PREDICTION

Confusion matrix for the Decision tree, Logistic regression and support vector classifier are determined. Accuracy, precision, f1 score and recall are used to compare the efficiency of machine learning techniques.

5.1 Logistic Regression

Logistic Regression score
Model accuracy :0.9180327868852459
Accuracy i percentage : 91.8%

	precision	recall	f1-score	support
0	0.90	0.93	0.92	29
1	0.94	0.91	0.92	32
accuracy			0.92	61
macro avg	0.92	0.92	0.92	61
weighted avg	0.92	0.92	0.92	61



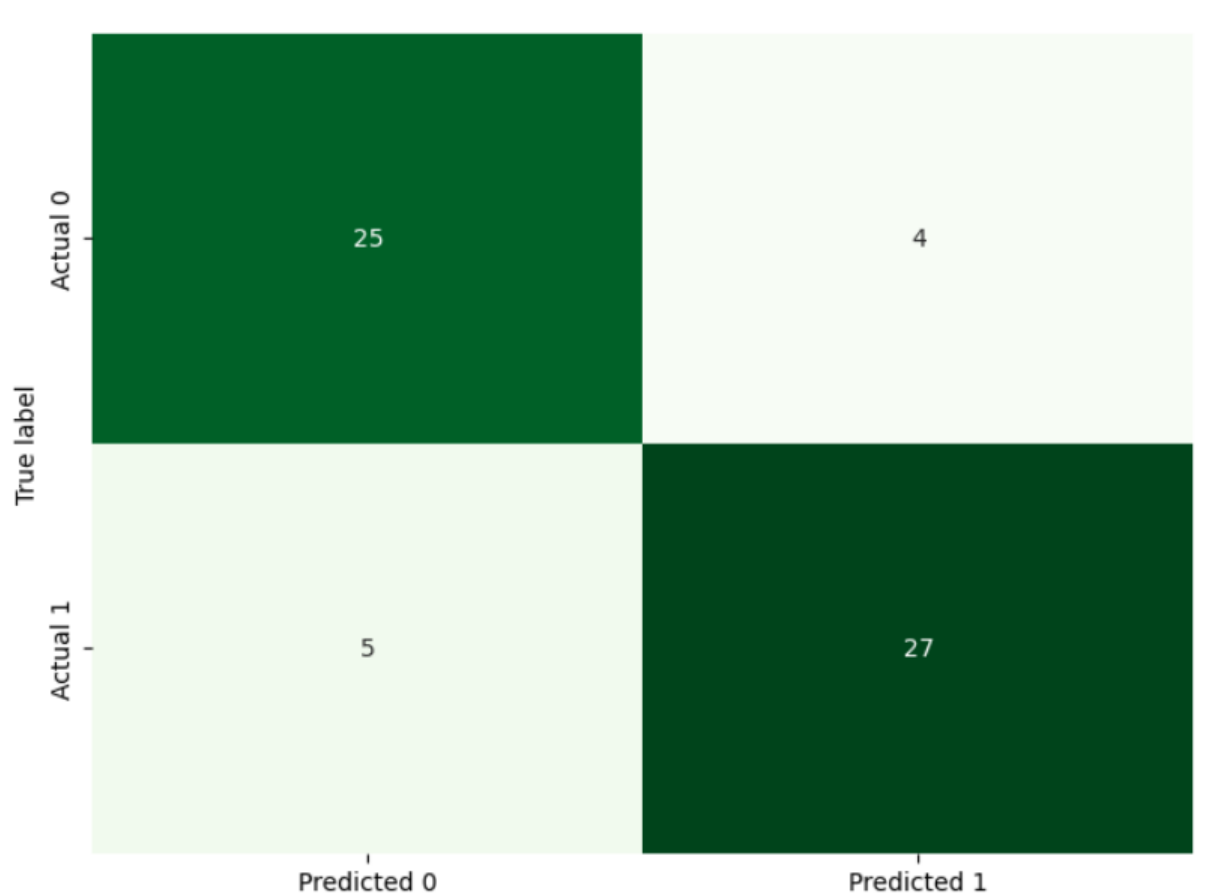
5.2 Decision Tree

Decision Tree Classifier score

Model accuracy : 0.8524590163934426

Accuracy in percentage : 85.2%

	precision	recall	f1-score	support
0	0.83	0.86	0.85	29
1	0.87	0.84	0.86	32
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61



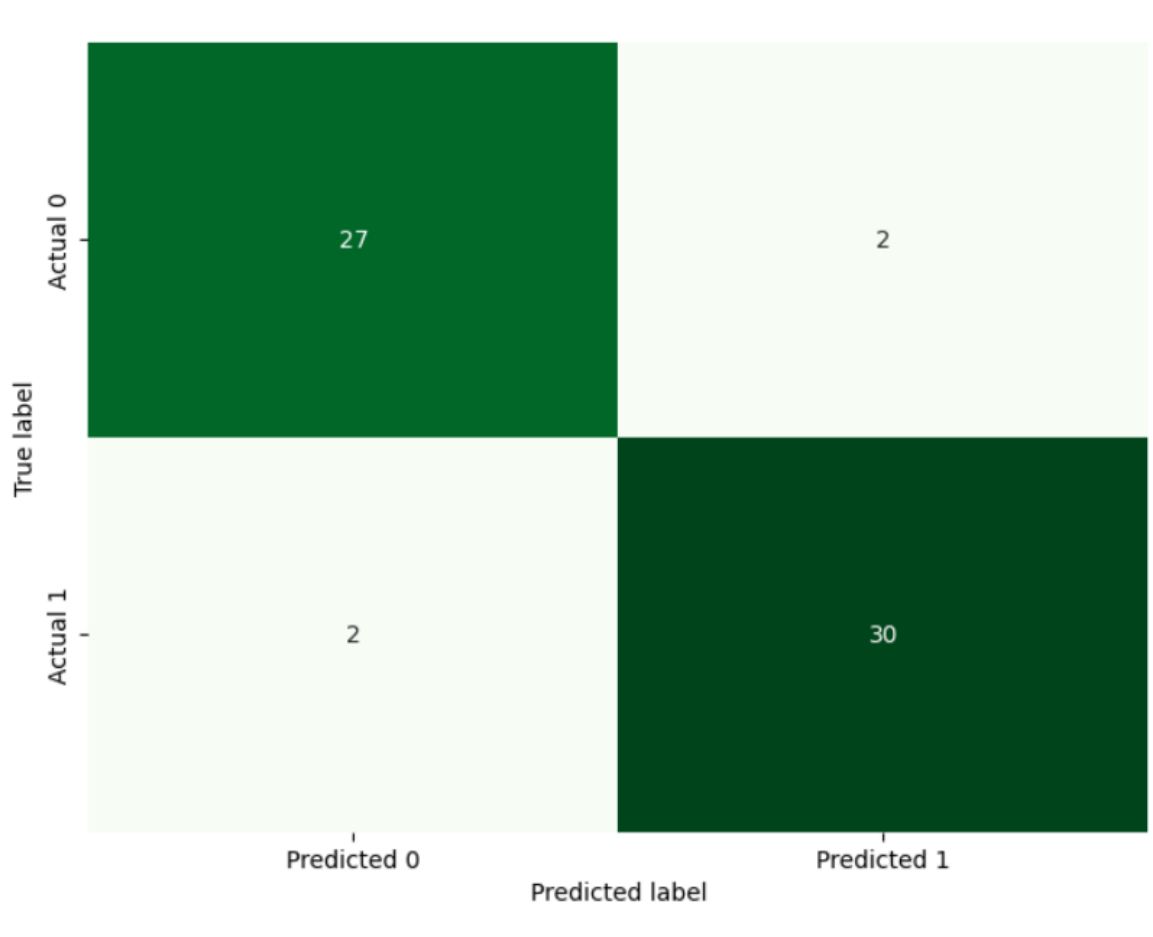
5.3 Support vector classifier

SVC Classifier scores

Model Accuracy : 0.9344262295081968

Accuracy in percentage : 93.4%

	precision	recall	f1-score	support
0	0.90	0.93	0.92	29
1	0.94	0.91	0.92	32
accuracy			0.92	61
macro avg	0.92	0.92	0.92	61
weighted avg	0.92	0.92	0.92	61



By comparing the accuracy of the three machine learning techniques, it is that Support Vector Classifier is the best model for predicting Heart failure prediction.

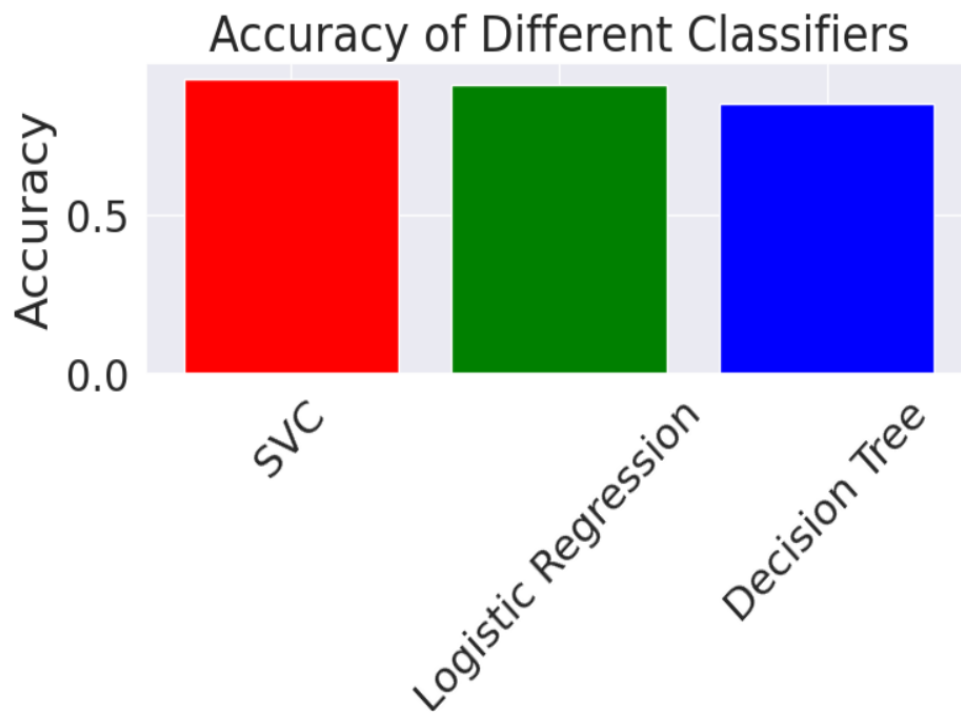
Prediction of heart failure disease using statistical analysis

Therefore, a support vector classifier is used in this project for the model prediction.

Accuracy of different classifier

Models	Model accuracy	Model accuracy in percentage
Support vector classifier	0.934	93.4
Logistic regression	0.918	91.8
Decision Tree	0.852	85.2

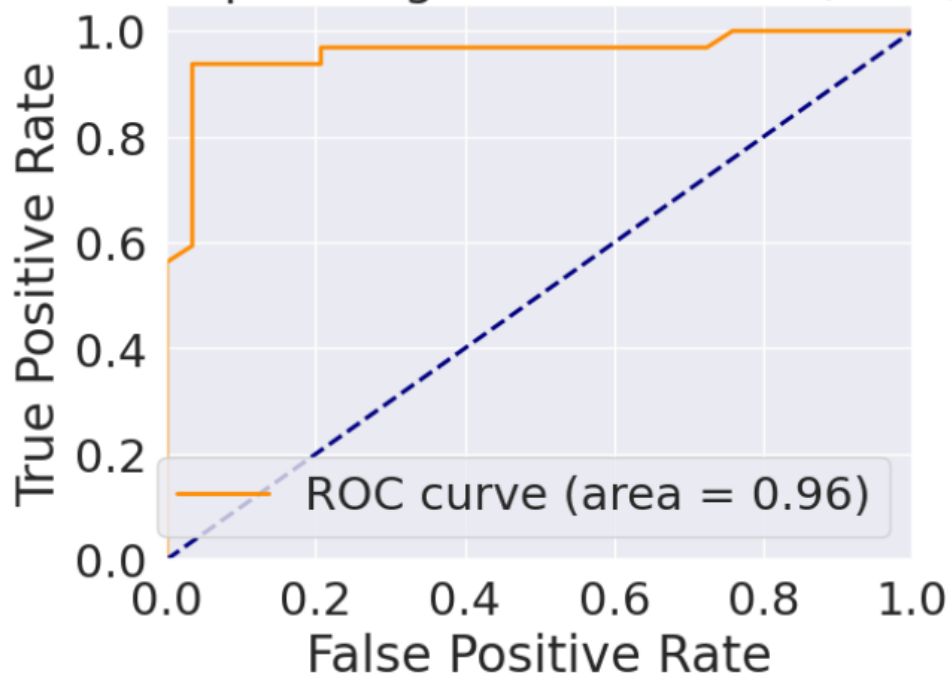
Graphical representation



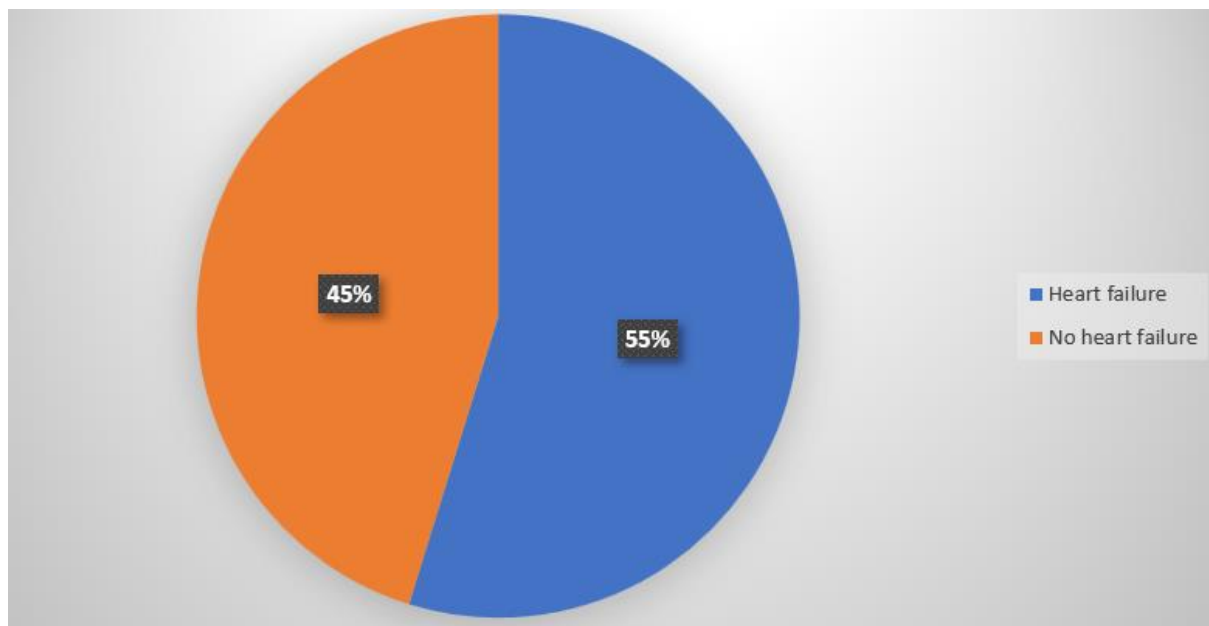
Result

We have performed 3 models to predict Heart failure disease. The above models we consider that SVC classifier model provides us with better accurate results. It has the maximum precision and f1 score. so it is selected for prediction. The ROC curve is given by

Receiver Operating Characteristic (ROC) Curve



5.5 Prediction Result



The support vector classifier is the best method for predicting airline customer satisfaction prediction. By this prediction, 55% patients are no heart failure disease provided and only 45% patients are heart failure.

CHAPTER 6

CONCLUSION

This paper presents a model to forecast heart failure disease prediction. First, reduce the dimension of the dataset using PCA, then find the major factors using factor analysis. Then they used a variety of classification machine learning algorithms to predict heart failure disease. Support vector classifier on this feature subset shows the best classification performance (accuracy:93.4%, precision:94%, recall: 91%, F1 value: 92%). By using use logistic regression model, Decision tree trained on the feature subset selected to further extract the important variables affected by the heart failure disease. So support vector classifier is the best model.

REFERENCES

- Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., ... & Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. *European journal of heart failure*, 22(1), 139-147.
- Echouffo-Tcheugui, J. B., Greene, S. J., Papadimitriou, L., Zannad, F., Yancy, C. W., Gheorghiade, M., & Butler, J. (2015). Population risk prediction models for incident heart failure: a systematic review. *Circulation: Heart Failure*, 8(3), 438-447.
- Jehn, M., Schmidt-Trucksäss, A., Schuster, T., Weis, M., Hanssen, H., Halle, M., & Koehler, F. (2009). Daily walking performance as an independent predictor of advanced heart failure: prediction of exercise capacity in chronic heart failure. *American heart journal*, 157(2), 292-298.
- Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., ... & Packer, M. (2006). The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*, 113(11), 1424-1433.
- Rahimi, K., Bennett, D., Conrad, N., Williams, T. M., Basu, J., Dwight, J., ... & MacMahon, S. (2014). Risk prediction in patients with heart failure: a systematic review and analysis. *JACC: Heart Failure*, 2(5), 440-446.
- Sahoo, P. K., & Jeripothula, P. (2020). Heart failure prediction using machine learning techniques. *Available at SSRN 3759562*.
- Steinhart, B., Thorpe, K. E., Bayoumi, A. M., Moe, G., Januzzi, J. L., & Mazer, C. D. (2009). Improving the diagnosis of acute heart failure using a validated prediction model. *Journal of the American College of Cardiology*, 54(16), 1515-1521.
- Tiwaskar, S. A., Gosavi, R., Dubey, R., Jadhav, S., & Iyer, K. (2018, August). Comparison of prediction models for heart failure risk: a clinical perspective. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCCUBEA)* (pp. 1-6). IEEE.
- Wang, J. (2021, September). Heart failure prediction with machine learning: a comparative study. In *Journal of Physics: Conference Series* (Vol. 2031, No. 1, p. 012068). IOP Publishing.

