CLARK UNIVERSITY

BAN 5573: Project Report

Heart Failure Prediction using Machine Learning

Supervised by:

Prof. Hamidreza Ahady Dolatsara

Group Members:

Tejas Mirashi

Vikas Jangra

Dhiraj Sharma

Nisarg Thakkar

April 2022

Abstract

Heart failure is a serious condition which has become highly prevalent. With about 2% in the adult population in developed countries and more than 8% in older population over 75 years in developed countries. Almost 3-5% of hospital admissions are linked with heart failure incidents. It is essential to build an effective disease management strategy with analysis of large amounts of data, early disease detection, assessment of severity and prediction of serious events at an early stage. This will help improve the quality of life of the patients, control the growing numbers of the disease and also reduce the medical costs associated with it. The objective of this research is to present flagship machine learning methodologies applied for the prediction of heart failure and death event because of the same. More specifically, models that are build using variables that are highly correlated. The project also aims to understand whether features like smoking, high blood pressure, diabetes, etc. Result in a death event from heart failure.

List of Figures

Table of Contents

Chapter 1

Introduction

1.1. INTRODUCTION

Heart failure is not a disease but a clinical syndrome. It basically prevents the heart from fulfilling the circulatory demands of the body since it does not fill or eject blood through the ventricles. Most common symptoms of a heart failure are breathlessness, fatigue, swelling of ankles accompanied by complex issues such as jugular venous pressure, pulmonary crackles, and peripheral edema caused by structural and/or functional cardiac or non-cardiac abnormalities. It is a serious condition which is associated with high mortality and morbidity rates. The European Society of Cardiology (ESC) states that 26 million adults globally are diagnosed with heart failure, while 3.6 million are newly diagnosed every year and the remaining die within five years. Heart failure management costs are approximately 1-2% of all healthcare expenditure with most of them linked with recurrent hospital admission.

The increased prevalence, escalated healthcare costs, repeated hospitalizations, reduced quality of life and the high mortality rate have all fueled the need for early diagnosis (detection and estimation of severity) and effective treatment. For diagnosis and treatment, data is usually made available by careful history and physical examination, also supported by ancillary tests such as blood tests, chest radiography, electrocardiography (ECG), echocardiography. The combination of data produced from the above tests result in the formulation of several criteria.

Data recorded in the subject's health record, expressing demographic information, clinical history information, presenting symptoms, physical examination results, laboratory data, electrocardiogram (ECG) analysis results are employed. Our research aims to use machine learning methodologies on such patient data for early detection and prediction of a death event because of a heart failure. Patients with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

1.2. Research Question/Hypotheses Development

Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. In order to gauge this, we develop hypotheses that can answer some common behavioral myths related to HF.

H0 – Smoking has cardinal role in a cardiac arrest

H1 – Smoking has no cardinal role in a cardiac arrest

H0 – Diabetes is a key contributor to cardiac arrest

H1 – Diabetes is not the only factor that triggers a cardiac arrest

Chapter 2

2.1. Literature Review

The average rate of heart beating is around 2.5 billion times over the lifetime of a human being allowing millions of gallons of blood to pump to all the parts of the body. This flow enables the cells of the body to work carrying the major essential components needed by the organs to work. When this pumping is ceased, the essentials stop to work, fails to function.

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels including, coronary heart disease also known as heart attacks, cerebrovascular diseases (strokes), heart failure (HF), and other types of pathology. [1]

The heart works with the never-ending workload, for every living being, and it's not a wonder that it can stop working at any moment causing a risk to the life.

This project aims to analyze the patterns of the several factors that are associated with the heart and others which might have an indirect effect on it, such as the Gender, Platelets, Diabetes and many more, which helps us to mitigate the risk of heart failure or detecting it in the early stage using the concepts of Machine Learning.

Machine learning tools helps the medical practitioner to predict the survival of the patient having heart failure symptoms and to detect most important medical features.

The previous authors analyses the dataset released by the Ahmad and colleagues [2] and then described the dataset and its features. Later, they report their survival prediction performance using the classifiers they used on the two topmost important features, they are: Serum Creatinine and Ejection Fraction. [3]

Chapter 3

Methodology

3.1. Data Source and Description

The dataset used for this project was the public available dataset release by Ahmad and colleagues in July 2017 of 299 Pakistan patients of having heart failure.

Data source: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

Number of Observations: 299

Number of Columns: 13

Data Dictionary:

i. Age – age of the patient
ii. Anaemia - Decrease of red blood cells or hemoglobin (boolean)
iii. Creatinine Phosphokinase - Level of the CPK enzyme in the blood (mcg/L)
iv. Diabetes - If the patient has diabetes (boolean)
v. Ejection Fraction - Percentage of blood leaving the heart at each contraction (percentage)
vi. High Blood Pressure - If the patient has hypertension (boolean)
vii. Platelets - Platelets in the blood (kiloplatelets/mL)
viii. Serum Creatinine - Level of serum creatinine in the blood (mg/dL)
ix. Serum Sodium - Level of serum sodium in the blood (mEq/L)
x. Sex - Woman, man or binary
xi. Smoking - If the patient smokes or not (boolean)
xii. Time - Follow-up period (days)
xiii. DEATH_EVENT - If the patient deceased during the follow up period (boolean)

3.2. Preprocessing the data

The data set does not contain any null values but were containing many outliers and those were to be handled. Some features where outliers were present, were actually the true values that were used in the analysis.

3.2.1. Data Distribution

The distribution of the data plays an important role when the prediction or classification of a problem is to be done.

3.2.1.1. Death Event Distribution

The figure represents the count of the number of the death events observed.

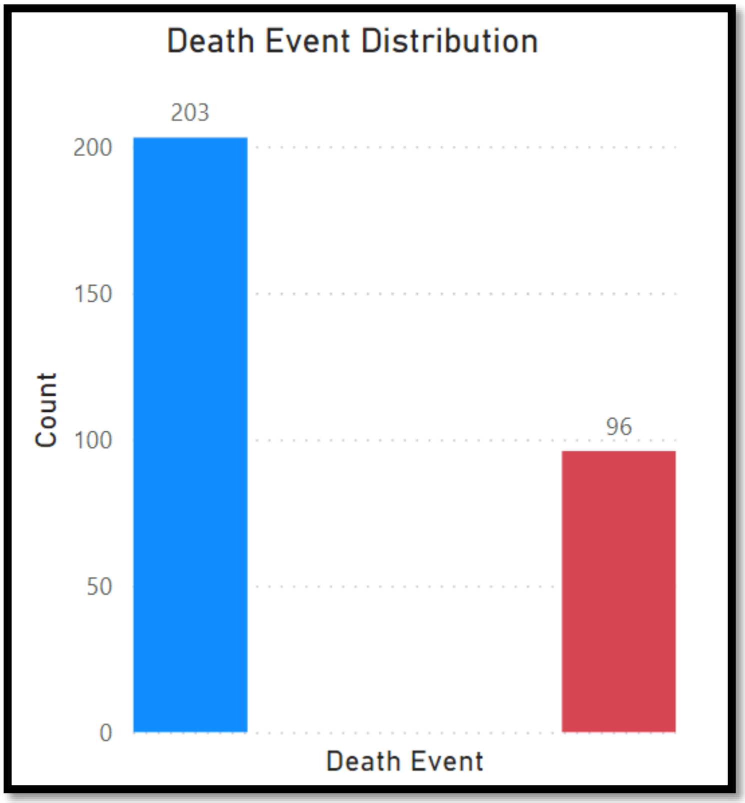


Fig 1: Death Event Distribution

3.2.1.2. Age wise Death Event Distribution

This figure represents the count of the death events per age group. We can observe that most of the death were occurred in the age group of 50-70.

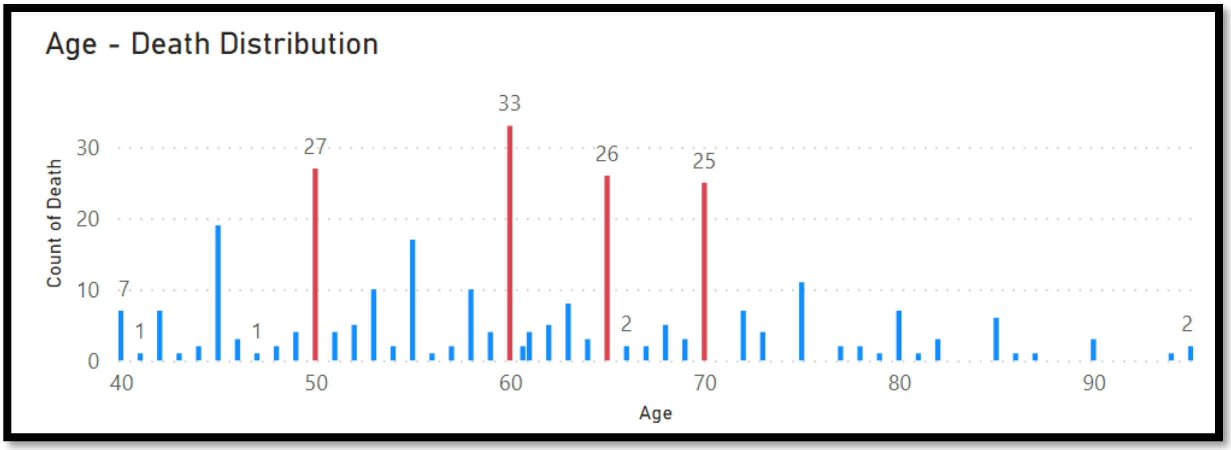


Fig 2: Count of Death Event by Age

3.2.1.3 Feature Selection

Feature selection is the most important part of the research conducted as this helps or enables us to find the correlation among our target variable (Death Event).

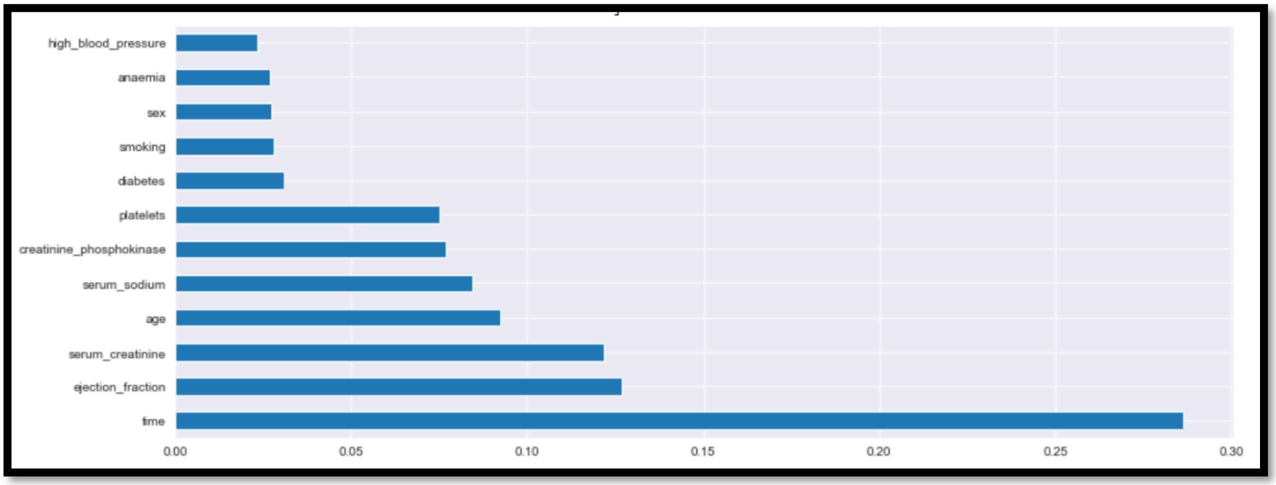


Fig 3: Important Feature Selection

3.3. Machine Learning Classifiers Proposed

The project was conducted with multiple machine learning algorithms where the dataset was analyzed initially followed by applying different machine learning algorithms consisting of linear model selection in which Logistic Regression was used. For focusing on neighbor selection technique KNeighborsClassifier was used, then tree-based technique like DecisionTreeClassifier was used, and then a very popular and most popular technique of ensemble methods RandomForestClassifier was used. Also, for checking the high dimensionality of the data and handling it, Support Vector Machine was used.

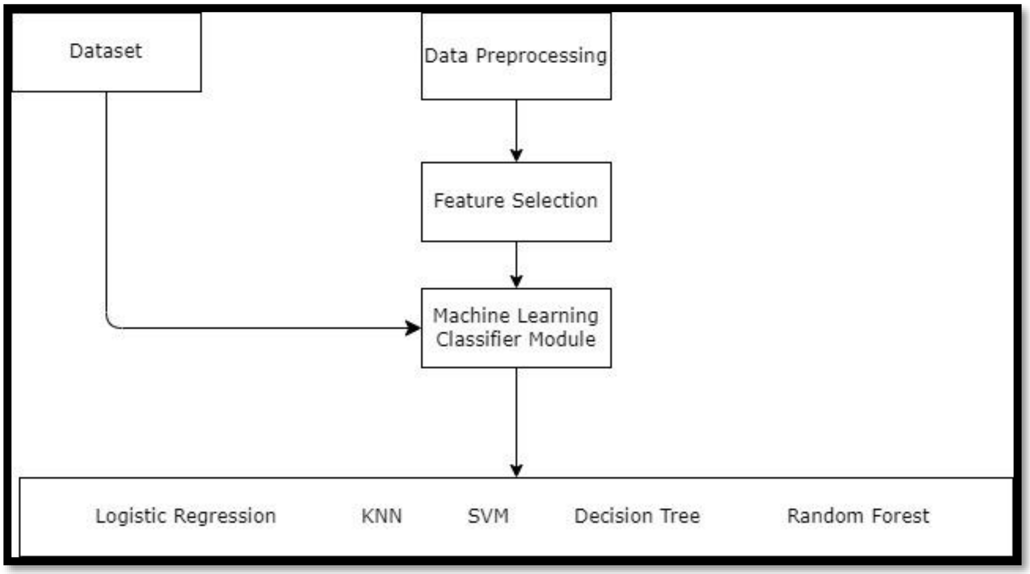
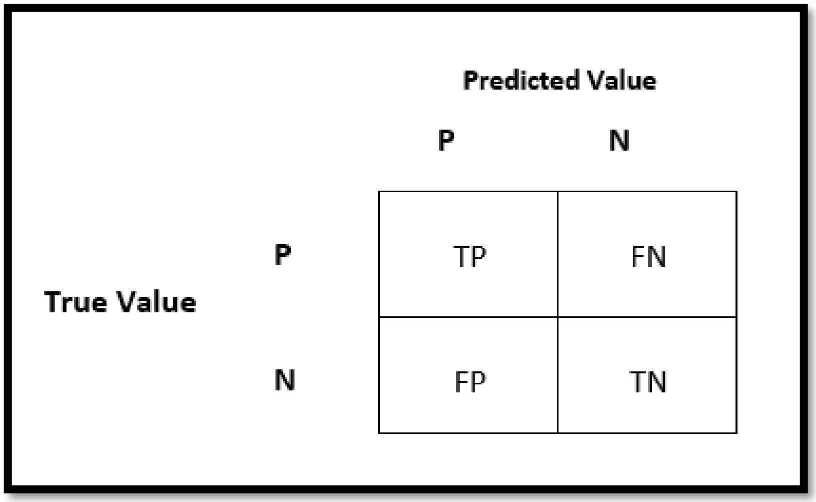


Fig 4: Project Flow Chart

3.4. Evaluation Method Used

For the evaluation, the most widely used methods was used, i.e. Confusion Matrix, and the Accuracy Score.

For the confusion matrix, we have the formulated table in the form of the Predicted vs True Value.



		Predicted Value	
		P	N
True Value	P	TP	FN
	N	FP	TN

Fig 5: Confusion Matrix Format

The confusion shows the comparison of the predicted accuracy and the actual accuracy. Where P = Positive, N = Negative, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

For measuring the accuracy, we used the accuracy score card. It is defined as the as the true positive values plus true negative values divided by true positive plus true negative plus false positive plus false negative.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.5. Analysis

The dataset was divided in 20-80 ratio for test train split. Further, the models were trained on the train dataset and tested on the remaining 80% of the dataset. Following this, the different algorithms were applied on the dataset and the prediction were completed.

The correlation between the variables were determined using the heatmap, the darker the variable, more is that correlated.

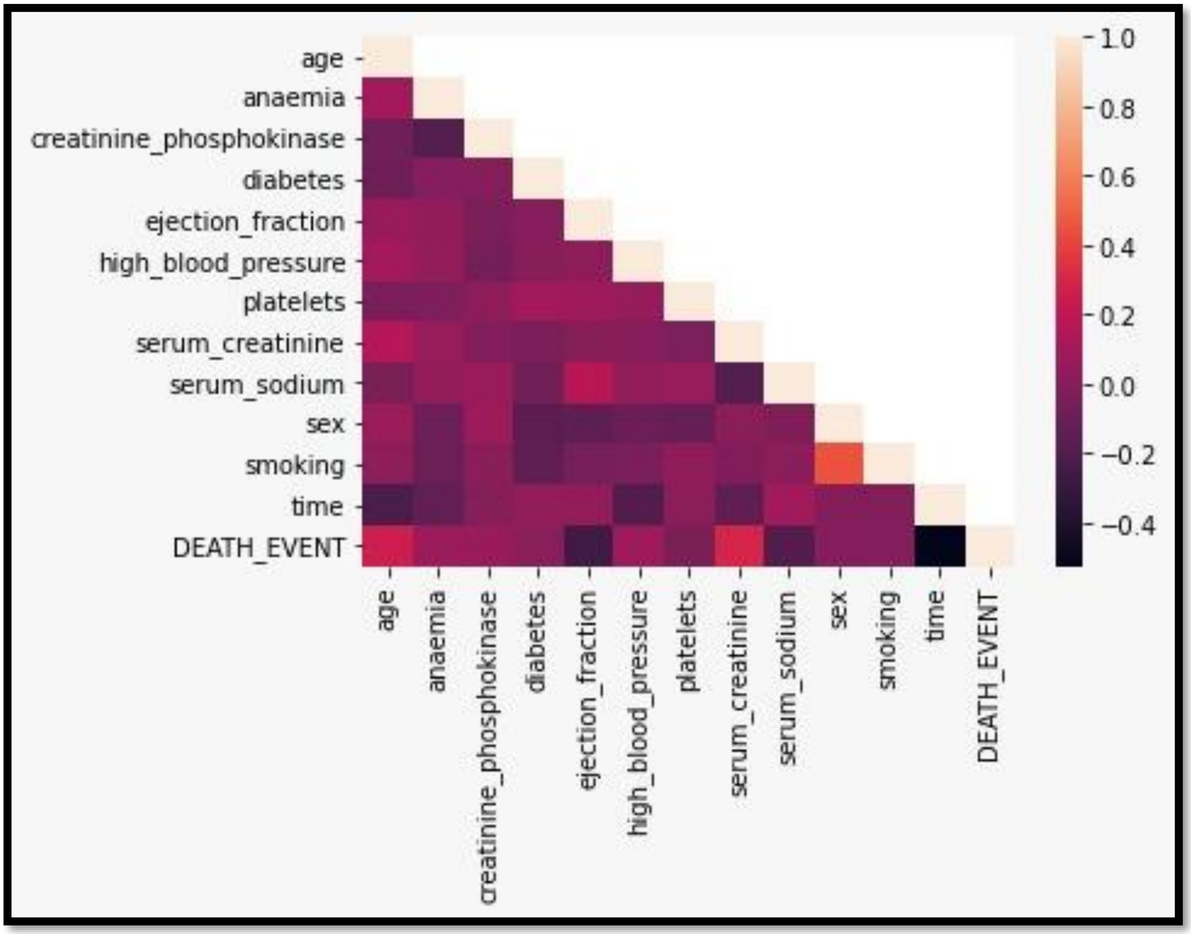
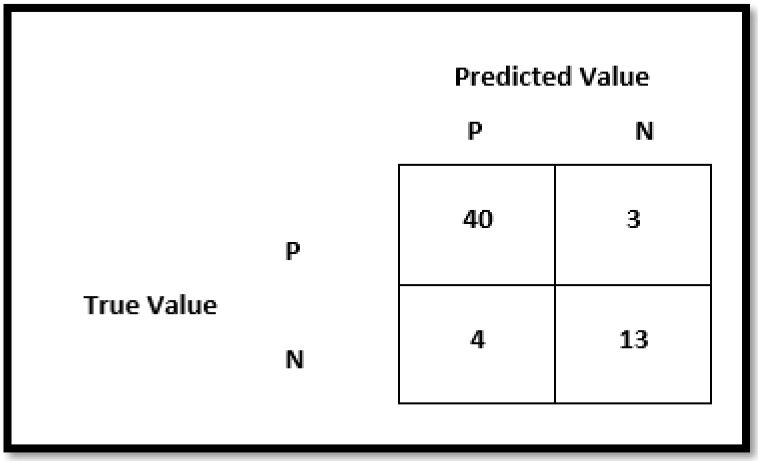


Fig 6: Correlation Heatmap

3.5.1. Logistic Regression

It is statistical analysis to find relation between the dependent variable and one or more independent variable. Using the estimation of the probabilities, it helps in predicting the likelihood of happening of an event.

This project recorded the accuracy of 88% in the logistic regression.



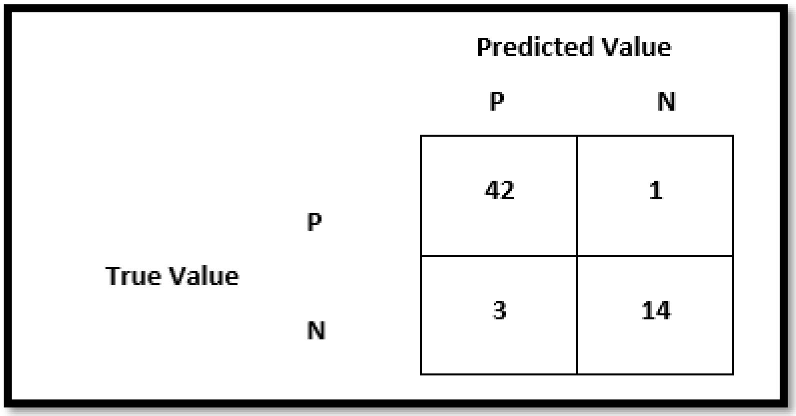
		Predicted Value	
		P	N
True Value	P	40	3
	N	4	13

Fig 7: Confusion Matrix of Logistic Regression

3.5.2. K Nearest Neighbor

KNN is a model that classifies data points based on the points that are most similar to it. In this we will use test data to make an “educated guess” on what an unclassified point should be classified as. It is a non-parametric method.

The accuracy observed was 93.3%.



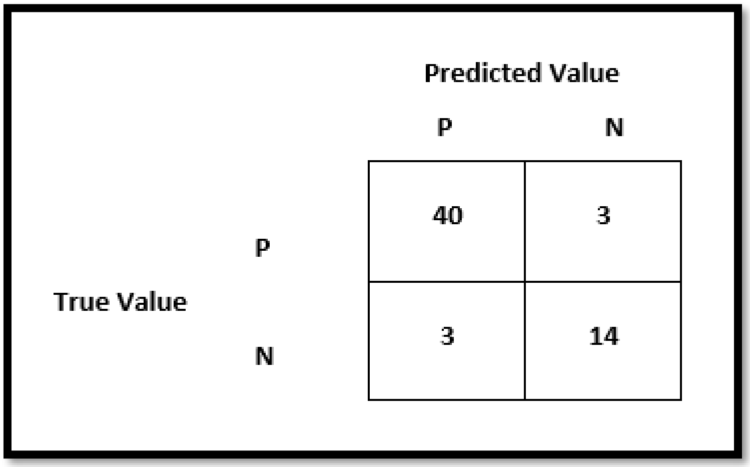
		Predicted Value	
		P	N
True Value	P	42	1
	N	3	14

Fig 8: Confusion Matrix of K Nearest Neighbor

3.5.3. Support Vector Machine

This algorithm will classify the dataset in two categories, and the hyperplane will be the deciding factor with the maximum margin distance.

The accuracy observed was 90%.



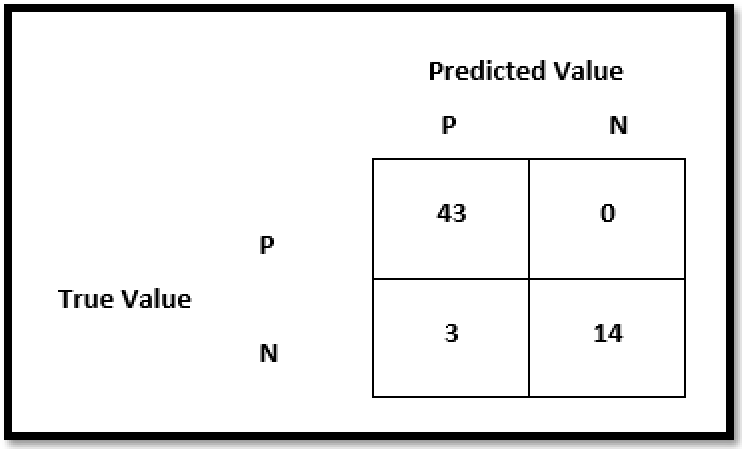
		Predicted Value	
		P	N
True Value	P	40	3
	N	3	14

Fig 9: Confusion Matrix of Support Vector Machine

3.5.4. Decision Tree Classifier

This is a predicting model in which multiple variables will be considered, and based on the depending variables, it will be classified that it is prone to a heart failure or not.

The accuracy observed was 95%.



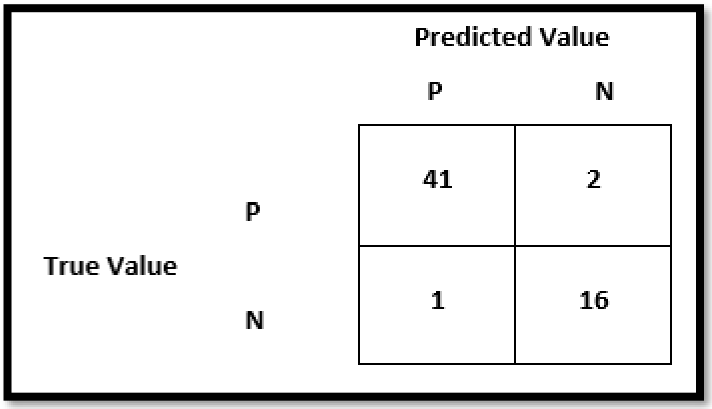
		Predicted Value	
		P	N
True Value	P	43	0
	N	3	14

Fig 10: Confusion Matrix of Decision Tree Classifier

3.5.5. Random Forest Classification

To make the prediction more accurate, random forest will also be used. The accuracy will increase when results from random trees are accumulated.

The accuracy observed was 95%.



		Predicted Value	
		P	N
True Value	P	41	2
	N	1	16

Fig 11: Confusion Matrix of Random Forest Classification

Chapter 4

Conclusion and Results

4.1. Results

We concluded that in our dataset 67.7% do not SMOKE (out of which 45.8% survived and 21.9% died) and 32.3% do SMOKE (out of which 22.2% survived and 10.1% died). This answers our hypotheses 1. Smoking has no significant impact on the death event due to a heart failure.

Similarly, in our dataset 57.9% are non-diabetic (out of which 39.4% survived and 18.5% died) and 42.1% are diabetic (out of which 28.6% survived and 13.5% died). This concludes that diabetes and death event due to heart failure have no direct relation.

4.2. Conclusion

These models can be used on datasets that have similar attributes to predict the chances of death event due to heart failure. Furthermore, doctors and healthcare organizations can use these models to gauge the chances of death due to a heart failure based on different health data available from patients. This research will also be helpful for insurance companies. Insurance companies can use these models to do a cost benefit analysis before providing insurance to patients with cardiac history and patients who fall in the age group more prone to develop cardiac problems.

References

1. [1] https://www.health.harvard.edu/topics/heart-disease
2. [2]https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0208737
3. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001
4. https://www.hindawi.com/journals/cin/2021/8387680/#literature-review
5. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
6. https://www.sciencedirect.com/science/article/pii/S2001037016300460