A PREDICTIVE MODEL FOR THE DIAGNOSIS OF HEART DISEASE WITH THE USE OF MACHINE LEARNING TECHNIQUES

AKPU CHUKWUMA HILARY

MATRIC NUMBER: 16010301019

BEING A PROJECT SUBMITTED IN THE DEPARTMENT OF COMPUTER SCIENCE
AND MATHEMATICS, COLLEGE OF BASIC AND APPLIED SCIENCES
IN FUFILLMENT OF THE REQUIREMENTS FOR THE
AWARD OF DEGREE OF BACHELOR OF SCIENCE
MOUNTAIN TOP
UNIVERSITY, IBAFO, OGUN
STATE, NIGERIA

2020

CERTIFICATION

This Project titled, **A PREDICTIVE MODEL FOR THE DIAGNOSIS OF HEART DISEASE WITH THE USE OF MACHINE LEARNING TECHNIQUES**, prepared and submitted by **AKPU CHUKWUMA HILARY** in partial fulfilment of the requirements of the degree of **BACHELOR OF SCIENCE** (Computer Science), is hereby accepted

____________________________ (Signature and Date)

DR. F. A. KASALI

Supervisor

____________________________ (Signature and Date)

DR. I. O. AKINYEMI

Head of Department

Accepted as fulfilment of the requirements for the degree of BACHELOR OF SCIENCE

(Computer Science)

____________________________ **(Signature and Date)**

Prof A. P. OLALUSI

Dean, College of Basic and Applied Sciences

Dedication

I dedicate this project to God Almighty.

Acknowledgement

I am most appreciative to almighty God for the gift of life and continuous strength for not only leading me through this project but also for the successful completion of my study. I also acknowledge the effort of my indefatigable supervisor Dr. F. A. Kasali for her constructive suggestions, understanding, motivations and useful comments. Sincere to say without her dedication, this project work would not have become a reality. My amiable lecturers, especially Dr. P.A. Idowu, Late Dr. M.O. Oyetunji, Dr. F.A. Kasali, Mr. O.J. Falana and Mrs Olutosin Taiwo and the other members of the department of computer science and mathematics for the seed you have sown in my life through your lectures and words of wisdom which has helped me so far in my journey in this institution.

I am deeply indebted to many people for their immense contributions in diverse ways towards the successful completion of this research work. My appreciation goes to my irreplaceable parents, Mr & Mrs Akpu, C.H., who sacrificed wealth and enjoyable moments for the sake of my success, I want to thank them for their relentless effort towards ensuring the best education for me and also for their moral, financial, spiritual, physical and psychological support in my life. May you live long and eat the fruits of your labour.

Abstract

Heart diseases have the highest death toll since 2000. Heart disease on its own is basically a deficiency in the heart of living things and there are multiple kinds of heart diseases such as arrhythmia, atherosclerosis, congenital heart defects, coronary artery disease among many others. This study aims to use machine learning techniques ranging from feature selection, principal component analysis, cross-validation and several machine learning algorithms.

Historical data on the distribution of heart disease among patients have been gathered and I acquired this data to be used in this research study. The predictive model for diagnosing heart diseases was developed using several machine learning algorithms.

Table of Contents

APPENDIX

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

The name Machine learning was brought up by Arthur Samuel in 1959. Mitchell (1997), gave A more systematic description of the algorithms studied in the machine learning field as " computer are said to gain experience E in relation to any class named T and output evaluation P if its performance in tasks T, as calculated by P, increases with experience. E". Alan Turing asked a question "Can machines think?' which was later replaced with "Can machines do what we (as thinking entities) can do?". In machine learning, we have three types of learning which are Supervised learning, Unsupervised learning, Reinforcement learning according to Brownlee (2019). Supervised learning involves the building of mathematical models from a set of data that contains both the inputs and the desired outputs otherwise called labelled data. In unsupervised learning, models are built based on unlabeled data and the computer finds hidden patterns in the provided data that are hard for humans to find. In reinforcement learning, algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment. Basically, most programming languages can be used in creating machine learning models but the most commonly used are python and R.

The heart is one of the most important organs in our body and it's about the size of the fist located slightly left of center in the chest. The heart is divided into the left and right sides. It has four chambers and the division into sides prevents oxygen-poor blood from mixing with oxygen-rich blood as one side takes blood and the other pumps blood. The Oxygen-poor blood returns to the heart after circulating the body. Compromising the right atrium and ventricle, collects and pumps blood to the lungs. The lungs refresh the blood with a new supply of oxygen. Oxygen-rich blood then enters the side of the heart which will be pumped through the aorta to supply tissues in the body with oxygen and nutrients.

Heart disease simply refers to deformities in the heart itself. Based on numerous research carried out by the Centre for Disease Control (CDC), heart diseases are the leading the death charts in most of the developed countries in the world which includes the United Kingdom, United States, Canada etc. Heart disease describes various conditions that affect your heart. There are several diseases which can be classified under heart disease “umbrella”, and they include; heart rhythm problems (arrhythmias), blood vessel diseases like coronary artery disease, heart defects one can be born which (congenital heart defects), dilated cardiomyopathy, Myocardial infarction, Heart failure, Hypertrophic cardiomyopathy, Pulmonary stenosis among others (MayoClinic.org, 2020).

The term “cardiovascular disease” is frequently used correspondently with the term “heart disease”. Cardiovascular disease alludes to conditions that include impeded veins that can lead to chest torment (angina), stroke, or even a coronary failure.

In spite of the fact that the indications of a coronary illness regularly rely upon which condition is affecting the person, there are some basic manifestations which includes chest agony, breathlessness, and heart palpitations. The most widely recognized chest torment type is known as angina or angina pectoris and happens when a portion of the heart does not get enough oxygen. At times, the symptoms of a heart disease may resemble indigestion. Heartburn and stomach-ache might occur too.

Cardiovascular diseases (CVDs) are the number 1 causes of death internationally, taking 17.9 million lives each year (World Health Organisation [WHO], 2020). Cardiovascular diseases are types of disorders in the heart. Four out of 5 cardiovascular disease deaths are cause by strokes and heart attacks, and 33% of these deaths happen rashly in individuals under 70 years old (Rehan, Qadeer, Bashir & Jamshaid, 2016). More than 75% of cardiovascular disease deaths happen in low- and middle-income countries in which Nigeria is situated.

1.2 Statement of the problem

Heart diseases has been identified as a global problem especially in developing countries as a result of socioeconomic factors like poverty, poor diet, illiteracy, dearth of medical experts amongst others and all these have led to an increase in the prevalence of the risk factors and mortality rate. According to WHO, Ischaemic heart disease has the highest death toll since 2000. Identifying people who are at high risk of CVD can greatly prevent premature death and this brings the need for this study. Lots of machine learning models have been developed for the prediction of this problem, but none tested the algorithms used with Principal Component Analysis which removes correlated features and keep the important features hence improves the model.

1.3 Aim and Objectives

The aim of this study is to use machine learning techniques to develop a model that can accurately predict the probability of a patient having a heart disease based on some risk factors. The research objectives are to:

i. Create a predictive classification model.

ii. Train the model.

iii. Test the model.

1.4 Scope of the study

The scope of this research is restricted to patients from kids to adults of all age groups. The data obtained for this research is not restricted by continents as it was collected for different hospitals in different parts of the world.

1.5 Significance of the study

This project aims to address the challenges posed by heart disease globally in the 21st century. The project draws on the parallels between these diseases, learning from existing challenges and the goal of connecting people to lay the foundations for a worldwide practice society. The aim is to renew and reinforce the development of scientific knowledge.

1.6 Arrangement of work

Chapter one was presented in this section during the presentation of a description of other chapters.

Chapter two contains a review of the literature and its deploring effects about heart, heart disease, application of machine learning and classification in health sectors.

Chapter three includes the techniques of studies used to create the model from data identification, collection and cleaning, model creation and algorithms to be used, and model testing.

1.7 Definition of terms

Data set: A collection of associated information sets consisting of distinct components, but which can be manipulated by a computer as a unit.

Evaluation: Judging the quantity, number, or value of something.

Validation: Action to check or prove something's validity or precision.

Model: A three-dimensional representation, typically on a lower scale than the original, of an individual or thing or of a suggested framework.

Data set- A collection of associated information sets consisting of distinct components, but which can be manipulated by a computer as a unit.

CHAPTER TWO

LITERATURE REVIEW

2.1 Heart diseases

Heart diseases are caused by disorders in the heart and blood vessels. It consists of various problems, many of which are associated to a process called atherosclerosis. Atherosclerosis is a condition which is developed when a substance called plaque builds up in the walls of the arteries. This gathering limit the supply to the arteries, making it harder for blood to course through. On the off chance that a blood coagulation structures, it can hinder the blood stream. This can cause a heart attack or even stroke.

2.1.1 Complications of heart disease

The complications of heart diseases may include (MayoClinic.org, 2020):

a. Heart Attack: This is very common among smokers of cigarettes. A blood coagulation which impedes the blood course through a vein which feeds the heart causes a cardiovascular failure, perhaps wrecking a piece of the heart muscle.

b. Chest pain(angina) or Stroke: Ischemic stroke which is a natural kind of stroke happens when the arteries to your brain are limited in which very little blood arrives at the brain. A stroke is an extreme health related crisis as brain tissues die within a couple of moments of a stroke.

c. Heart failure: This is a very common complication of heart disease. This happens when the heart cannot pump adequate blood to meet the needs of the body. This can come about from many forms and kinds of heart disease such as heart defects, valvular heart disease, heart infections etc.

d. Peripheral artery disease: This happens when the extremities – usually the legs – don’t get adequate blood flow. This can cause symptoms like claudication – leg pain when walking.

e. Aneurysm: This is a bulge in the wall of your artery. An individual may face life-threatening internal bleeding if the aneurysm should burst in the body.

f. Sudden cardiac arrest: This is the sudden unexpected stop of the heart from functioning, breathing (due to weak heart), and also consciousness, most often is caused by an arrhythmia. This is a very severe medical emergency and need to be treated immediately.

2.1.2 Risk factors

Risk factors for developing heart disease includes (MayoClinic.org, 2020):

a. Age: The older you get the higher your risk of having a damaged and narrowed artery and thickened or weakened heart muscle.

b. Sex: Women generally have low risk of a heart disease compared to men, but after menopause the risk increases.

c. Family history: Some heart disease problems can be passed down genetically especially if a parent developed it at an early age.

d. High blood pressure: Uncontrolled high blood pressure can narrow the vessels through which blood flows or harden the arteries.

e. High blood cholesterol levels: High cholesterol level in the blood can highly increase the risk of formation of plaques.

f. Diabetes: Low sugar levels raise the risk of heart failure. The two diseases share common risk factors – obesity and high blood pressure.

g. Obesity: Usually, extra weight worsens other risk factors.

h. Stress: Unrelieved stress is likely to damage your arteries and worsen other risk factors.

i. Physically inactive: Lack of exercise fuels other risk factors as well.

j. Smoking: Nicotine constricts the Blood vessels and carbon monoxide is able to harm their inner lining, tend to be very prone to atherosclerosis.

k. Heart attacks are very common for smokers.

l. Bad hygiene: Poor hygiene practice and poor dental health may contribute to heart disease.

2.2 Data Collection/Scraping and cleaning

To create machine learning models, data is needed and there are several ways to collect data.

a. Surveys: This is a method of researching which is used to collect data from predefined groups of respondents to acquire insights into various topics of interest (QuestionPro Survey Software, 2020). Data scientists carry out surveys to get attributes which they believe is key to causing or avoiding the target variable. Surveys are taken with the risk factor of the target variable in mind. Data scientists first study the risk factor of their target variables then use it as a blueprint to create surveys while also having in mind the range of people to meet, and the approximate number of samples needed.

b. Web scrapping: This is a data scraping technique which extracts data from online platforms such as websites (Wikipedia, 2020).A lot of information is on the web and web scrapping has made it easier for data scientists to have access to relevant data. This data collection technique is mostly used when text analysis is in mind.

Most data in the world are unclean. Working with an unclean data can be disastrous as it leads to wrong prediction and poor accuracy. When a data set has missing values, there are several steps someone can take:

a. Remove the entire row
b. Remove the entire column
c. Replace the missing value(s)

Replacing the missing value is the most common practice in this kind of situation. The missing value is replaced with the most frequent value in the column or with the average of all other values in the column.

2.3 Machine Learning

Machine Learning also known as ML, is an artificial intelligence branch that enables computers to use statistical and optimization techniques to learn from past examples (Quinlan, 1986; Cruz and Wishart, 2006). It is the data-based field in AI which provides systems the ability to automatically learn and improve from experience without the computer being programmed explicitly. The computer is fed with lots of data which it learns from. It finds patterns in the data which can and is more often used for prediction, clustering, or classification. The learning process first begins with observation of data. Machine learning aims to primarily allow the computer to learn on its own without the intervention or assistance of humans.

Some machine learning methods include:

Supervised learning: This is the type of learning when the data used to train the model is labelled. Beginning from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after adequate training. The algorithm used can also equate the predicted output with the right output and figure out errors which are used to modify the model accordingly.

Unsupervised Learning: In this case, the data used is not labelled. It shows how systems deduce functions to describe a hidden structure from unlabelled data. It identifies patterns in the data and can be used to cluster the attributes in the data into groups.

Reinforcement Learning: In this case, the agent interacts with its environment by producing actions and discovers errors or rewards. A reward response is mandatory so the agent can learn which action suits best and this can be referred to as reinforcement signal.

2.4 Classification

Classification is very essential in the health sector for data scientists. It splits samples of data into target classes. The target class is predicted for each data point by analysing their disease patterns. Classification is a machine learning method of classifying attributes or targets into classes. Unlike regression which predicts a continuous value, classification simply predicts if a point (as the case may be) belongs to a class or not. The data set is partitioned, and we have a partition for training the model, the other for testing the model. It involves predicting an outcome based on a specific input. To predict the result, it attempts to find the connection between characteristics/attributes/variables. The testing data set is not introduced to the model but later used to test the model's accuracy to check the out-of-sample accuracy. The testing dataset has the same attributes as the training dataset. There are several classification algorithms made available and they include:

a. K-Nearest Neighbour (K-NN): K-Nearest Neighbour Classifier is one of the simplest classifiers to detect unidentified data points using previously known data points (nearest neighbour) and classified voting data points (McGregor, Christina, Andrew, 2012). Consider that there are different objects. It would be good for us to know the characteristics of one of these artefacts to predict for its closest neighbours because similar characteristics are found in the nearest neighbouring objects. K-NN's majority vote can play a very significant part in this classifying of any new instance where k is a favourable (tiny number) integer. It is primarily known as memory-based classification because there must always be examples of runtime training in memory (Alpaydin, 1997). In case of continuous attributes, the Euclidean distance is calculated when we take the difference between the attributes and the algorithm is then referred to as K-Nearest Neighbour Regressor. Continuous attributes need to be standardized to address this major problem in order to have the same influence on distance measurement between distances (Bramer, 2007).

b. Decision Tree (DT): DT is considered one of the most famous of these classifier approaches. We can construct a decision-making tree using available data that can address the issues related to different areas of research. It is the same as the flowchart in which each non-leaf node refers to a study on a specific attribute and each branch denotes the result of that test and we can decide if or not a patient needs to be readmitted with the assistance of the medical readmission decision tree. Domain knowledge is not required to decide on any issue. Decision Tree's most common use is to calculate conditional probabilities in operational research analysis (Goharian &

Grossman, 2003). The best root-to-leaf option shows a distinctive class separation baesd on maximum information gain (Apte, 1997). Several advantages of the Tree of choice as follows; Tress of choice are self-explanatory and easy to follow when compacted.

c. Neural Networks (NN): It was developed at the beginning of the 20th century (Anderson, 1995). It was considered the best classification algorithm prior to the introduction of decision trees and the Support Vector Machine (SVM) (Obenshain, 2004). That was one of the reasons that encouraged NN in various fields of biomedicine and healthcare as the most widely used algorithm for classification (Bellazzi, 2008). NN, for instance, was commonly used as the algorithm that supports disease diagnosis including cancers (Romeo, 1998) and predicts results (Sharma, 1997). In NN, neuron or nodes are the basic elements. There are interconnections between these neurons and worked in parallel within the network to produce the functions of the output. They can make fresh findings from existing observations even in situations where certain neurons or nodes within the network fail or fall due to their ability to work in parallel. Each neuron is associated with an activation number and each edge is assigned a weight within a neural network. Neural network is mainly used to perform classification and pattern recognition tasks (Dunham, 2003). An NN's basic property is that through weight adjustment and structural modifications, it can minimize the error. Only because of its adaptive nature, it minimizes the error. NN can produce more accurate predictions. One of NN's major advantages is that it can handle noisy training data properly and can reasonably be classified as fresh data types that differ from training data. NN also has different disadvantages. Firstly, it needs many parameters, including the optimal number of parameters empirically determined hidden layer nodes, and its output in classification is highly sensitive to the selected parameters. Secondly, its process of training or learning is very slow and very expensive computationally. Another is that they do not provide any internal details about the phenomenon being investigated now. It is like a "black-box" approach for us, therefore. Bayesian methods for probabilistic method of learning Bayesian classification is used. It can be easily obtained with the assistance of the classification algorithm. (Bayes statistics theorem plays a very significant part). While attributes such as signs of patients and their condition of health are correlated with one another in the medical domain, Classifier Naïve Bayes assumes all characteristics are independent. This is Naïve Bayes Classifier's major disadvantage. If attributes are independent, the classifier ‘Naïve Bayesian’ has shown great accuracy performance. They play very significant roles in the healthcare sector. There are therefore various advantages of BBN that Researchers worldwide have used them. One of them is that it helps make the process of computing very easy. Another is that it has better speed and accuracy for huge data sets (McGregor, 2012).

2.5 Principal Component Analysis

Principal Component Analysis (PCA) is a procedure utilized in decreasing the dimensionality of huge datasets, increasing interpretability but at the same time minimizing loss of information. This is done by creating entirely new uncorrelated attributes that progressively maximizes variance. Finding such new attributes, the key segments, diminishes to unravelling an eigenvalue/eigenvector issue and the new attributes are thus defined by the data. PCA is widely used in situations where there are lots of attributes in the data with high correlation (Ian, T. J., Jorge, C., 2016).

PCA was designed by Karl Pearson in 1901 and later, it was autonomously advanced by Harold Hotelling who also named it during the 1930s. PCA is regularly used to reduce the dimension of the data by anticipating every information pointing to only the first few major components to get lower dimensional data whereas it is conserving a lot of the data's variation (Wikipedia, 2020).

Table 1: Comparison of some pattern classification algorithms (source: patel et al, 2012).

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A support vector machine Constructs a hyper plane or set of hyper planes in a high or limitless dimensional space that can be used for grouping, regression or other	The effectiveness of SVM lies in the selection of kernel and soft margin parameters. For kernels, different pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is	1. Highly Accurate 2. Able to model complex nonlinear decision boundaries 3. Less prone to over fitting than other methods	1. High algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks. 2. The choice of the kernel is difficult 3. The speed

	activities.	picked. Trying exponentially growing sequences of C is a practical method to identify good parameters.		both in training and testing is slow.
K Nearest Neighbour	The object is graded by a variety vote of its neighbouring nodes, the item being given to the most general class of its adjacent k neighbors (k is a positive integer). If k = 1, the item will be assigned to a	Two parameters are considered to optimize the performance of the kNN, the number k of nearest neighbour and the feature space transformation.	1. Analytically tractable. 2. Easy to enforce 3. Uses local knowledge that can generate highly adaptive behaviour 4. Quite readily lends itself to parallel implementations	1. Large storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying test tuples.

	class of its adjacent neighbour.			
Bayesian Method	Based on the rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.	In Bayes, all model parameters (*i.e.*, class priors and feature probability distributions) can be approximated with relative frequencies from the training set.	1. Naïve Bayesian classifier simplifies the computations. 2. Exhibit high accuracy and speed when applied to large databases.	1 The assumptions made in class conditional independence. 2. Lack of available probability data.
Decision Tree	Decision tree builds a binary classification tree. Each node matches	Decision Tree Induction uses parameters like a set of candidate attributes and	1. Construction does not require any domain knowledge. 2. Can handle high	1. This must be a group of the performance attribute. 2. One performance

	an attribute with the binary predicates; one branch matches the positive instances of the predicate, while the other corresponds to negative instances.	an attribute selection method.	dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	attribute is reduced. 3. Algorithms for Decision Tree are unpredictable. 4. Numerical datasets can be complicated trees built.

CHAPTER THREE

Methodology

3.1 Introduction

This section shows the methodology design used to develop the predictive model for heart disease probability in a well narrated way. The methodology comprises of a series of methods/techniques that started with the identification and collection of data needed to develop the model.

3.2 Research Design

i. Acquire relevant datasets to be used for classification algorithms.

ii. Clean and analyse the dataset to be used.

iii. Induce the dataset with classification algorithms such as logistic regression, naïve bayes, etc.

iv. Record and compare the performance of each model on the dataset.

v. Check the important attributes in the dataset that increases the accuracy of the best performing model.

v. Apply PCA on the dataset and induce the new dataset with the classification algorithms used previously.

vi. Record and compare the performance of each model on the new dataset.

3.3 Data Identification and Collection

The data collection method to be used in this project are secondary sources of data. Data collection for this project will be carried out by surfing and downloading valid datasets which should help me achieving the goal of this project.

3.4 Data Overview

I will use a dataset containing the data of patients admitted to the hospital with and without heart diseases. The data was collected from the four following locations:

1. Cleveland Clinic Foundation (cleveland.data).
2. Hungarian Institute of Cardiology, Budapest (hungarian.data).
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data).
4. University Hospital, Zurich, Switzerland (switzerland.data)

The data has attributes including:

1. Age of the patient.
2. Sex.
    a. Value 1: male.
    b. Value 2: female.
3. Chest pain type(cp).
    a. Value 0: typical angina.
    b. Value 1: atypical angina.
    c. Value 2: non-anginal pain.
    d. Value 3: asymptomatic.
4. Resting blood pressure (trestbps) in mm Hg while admitted to the hospital).
5. Serum Cholestoral (chol) in mg/dl.
6. Fasting blood sugar>120 mg/dl (fbs).
    a. Value 1: true.

b. Value 0: false.

7. Resting electrocardiographic results (restecg).

   a. Value 0: normal.

   b. 1: have ST-T batch irregularity (T batch transposal and ST advancement or depression of >0.05mV).

   c. Value 2: show certain left ventricular hypertrophy in Estes’ criteria.

8. Max heart frequency reached (thalach).

9. Exercise induced angina (exang).

   a. Value 1: yes.

   b. Value 0: no.

10. ST depression brought by workout in relation to rest (oldpeak).

11. The slant of the topmost exercise ST section (slope).

    a. Value 0: upsloping

    b. Value 1: flat

    c. Value 2: downsloping

12. Number of major vessels(ca) (0-3) coloured by flourosopy.

13. Thal.

14. Num (attribute to be predicted): diagnosis of heart disease (angiographic disease status).

    a. Value 0: <50% diameter narrowing

    b. Value 1: >50% diameter narrowing

3.5 Formulating of Model for Predicting Heart Disease probability

The predictive model for heart disease probability was developed using the algorithm for logistic regression to define and account for variables linked to heart disease probability. Supervised machine learning algorithms were used to formulate predictive models as the pattern explaining the connection between the recognized factors (independent/input variables) and the corresponding heart disease probability (dependent/target variables) was needed, the pattern can then be used as a set of rules that can help doctors make informed decisions about the probability of Nigerians having a heart disease. For any machine learning algorithm using supervised technique, proposed for the creation of a predictive model, a mapping function can be used to conveniently articulate the general expression of the predictive model for the likelihood of heart disease– resulting in most machine learning algorithms being black-box models using evaluators instead of power series / polynomial equations. Historical dataset S composed of records of people with areas representing the set of 'probability' variables I amount of variable inputs for people ; Xij alongside the corresponding target variable (probability of a heart disease) represented by the Yj variable – the probability of heart disease in the information gathered from the hospital chosen for the research for the jth person.

Equation 3.1 displays a mapping function that describes the probability factor-to-target class relationship

–heart disease probability. $\varphi\colon X \rightarrow Y$

$$defined\ as\colon \varphi(X)=Y$$

The equation demonstrates the connection between risk variables set by a vector, X composed of I risk factor values, and Y label defining heart disease probability – yes or no heart disease probability expressed in equation 3.2. Assuming the risk factor set values for an individual are represented as X={X1,X2,X3, ... ,Xi } where Xi is the value of each probability factor, i=1 to I ; then the mapping π used to represent the heart disease probability predictive model maps each individual's probability factors to their respective heart disease target according to equation 3.2.

Assuming that an individual's risk factor set values are represented as X={X1,X2,X3,...... Where Xi is the value of each risk factor, i=1 to I; then the map used to depict the predictive heart disease risk model maps the risk factors of everyone to their corresponding heart disease evaluation.

The Logistic regression model created for the danger of heart disease in people was used to suggest a set of guidelines that can be used directly to determine the probability of heart disease by observing the values of the variables recognized by the model and the sequence of occurrences.

3.6 Logistic Regression

This type of prediction (classification) is used when the dependent variable is categorical. For example,

- To predict if a tumor is malignant (1) or not (0).
- To predict if a mail is spam (1) or not (0).

In the examples above, the dependent variables are in a binary form. Logistic regression allows for multiple classification of data. For example, classifying customers into different categories based on their information of purchasing habits.

It gets its name from linear regression but unlike linear regression which is used to predict continuous values, it predicts binary or multinomial classes.

It is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

Logistic regression fits an extraordinary s-formed curve by taking the linear regression and then changing the numeric estimate into a likelihood with the following function, which is called sigmoid function σ:

$$h_{\theta}(x)=\sigma(\theta^{T}X) = e^{(\theta0+\theta1_{*}x^{1+\theta2_{*}x^{2+\ldots)}}} / 1+e^{(\theta0+\theta1_{*}x^{1+\theta2_{*}x^{2+\ldots)}}}$$

Or:

$$ProbabilityOfaClass_{1}=P(Y=1|X) =\sigma(\theta^{T}X) = e^{\theta^{TX}} /1+e^{\theta^{TX}}$$

In this condition, $\theta^{T}X$ is our regression result (the addition of the variables weighted by the coefficients), exp is the exponential function and $\sigma(\theta^{T}X)$ is the sigmoid or logistic function, which is also known as logistic curve. It is a typical "S" shape (sigmoid curve).

Logistic regression will send the inputs through the sigmoid function and then treats the output as a likelihood/probability:

The objective of logistic regression algorithm is to find the best parameters θ, for h _θ(x) = σ({θ^TX}), so that the model predicts accurately the label of each case.

3.7 Naïve Bayes Classifier

The Bayesian Classification reflects a monitored technique of teaching as well as a statistical method for model. It implies an inherent probabilistic model and enables us to catch confusion about the model in a principled manner by determining outcome probabilities. It can resolve diagnosis and predictive issues.

This ranking is named after Thomas Bayes (1702-1761), who proposed Bayes' theorem. Bayesian classification offers practical learning algorithm, and it is feasible to mix previous understanding with measured information. Bayesian classification offers a helpful viewpoint for the comprehension and evaluation of many learning algorithms. It calculates specific hypothesis probabilities and is resistant to noise in output information.

The predictive model for the risk of heart disease was formulated using the naïve Bayes' classifier – a supervised machine learning algorithm that is based on the naïve Bayes' statistical theory of conditional probability shown in the equation below.

P(Class|Attributes) = (P(Attributes|Class) * P(Class)) / P(Attributes)

Where:

P(Class) = Prior probability of class (risk of heart disease);

P(attribute) = Prior probability of training data attribute values;

P(Attribute|Class) = Probability of attribute values given the class; and

P(Class|Attributes) = Probability of Class given the attribute values.

3.8 Model Accuracy

To know how accurate your model is, you will need to test the model. This can be done using the training data. In training your model, you teach how to use given factors/ attributes to predict the class that attribute belongs to. Let us say we want to predict if a patient has heart disease by giving our model the age and sex of the patient. After training our model with the dataset, we can provide a given age and sex from the dataset to the model for it to predict if that patient has a heart disease. This can be helpful, until your model gets overfitted. Overfitting is when a model is too rigid on the training dataset and not flexible enough to accurately predict out-of-sample instances. When a model is overfitted, the accuracy in predicting the training data becomes high but the accuracy in predicting any other data is low and that is why we split our datasets.

A machine learning model can be very accurate when predicting the values of the data it was trained with but when it's deployed to the world for usage, when it meets new sets of data with different patterns, the accuracy might suffer. Splitting the data will train the model with one fraction and test it with the other which the model is not used to. This gives an insight on how the model will perform in the real world.

I will use the train_test_split function from sklearn.model_selection library.

Peter Drucker had a famous saying which goes; "if you can't measure it, you can't improve it". After a model is created and trained, the accuracy of the model must be measured and there are several ways to measure the accuracy of Machine Learning models, few are mention below:

Evaluation metrics for regression models:

a. Mean Squared Error (MSE).
b. R squared

Evaluation metrics for classification models:

a. Recall.
b. F1 score.
c. Accuracy.
d. Log loss.

3.8.1 Confusion Matrix

A confusion matrix is an m x m matrix, where we have m to be the targets to be predicted.

a. Accuracy: The fraction of the overall predictions that were correct.

b. Precision: The proportion of true positives that were identified correctly.

c. Negative Prediction Value: The amount of negative cases which were actually, properly recognised.

d. Recall: The amount of actual positive cases that were properly recognised

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive			Positive precision	a/(a+b)
	Negative			Negative Precision	d/(c+d)
		Recall	Recall	Accuracy = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Table 3.1: confusion matrix

This matrix will be the basis on which I will score the models produced.

3.8.2 F1 Score

We’ve previously discussed precision and recall for classification problems. What if we are trying to get the best precision and recall at the same time? The formula for deriving F1 score is as follows:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

F1-score is the average of precision and also recall values for a classification problem.

3.8.3 Distribution plots

Distribution plots visually shows a distribution and the range of numbers which are plotted against a particular dimension. They visually gauge distribution of sample datasets by evaluating the observed distribution of the dataset with the predicted values expected from a particular distribution. I will be using distribution plots to show how well the models are predicting ‘unknown data’. The graph will show two distributions, one for the actual data, the other for the predicted data.

3.9 Principal Component Analysis

Principal Component Analysis (PCA) is a technique used to decrease dimensionality as big data sets improve interpretability while reducing data loss. This is achieved by generating completely new non-correlated variables that successively optimize variance. Finding those new variables, the main components, reduces the issue of the ownvalue/eigenvector and the new variables are then described by the dataset at hand. PCA is widely used in situations where there are lots of attributes in the data with high correlation.

CHAPTER FOUR

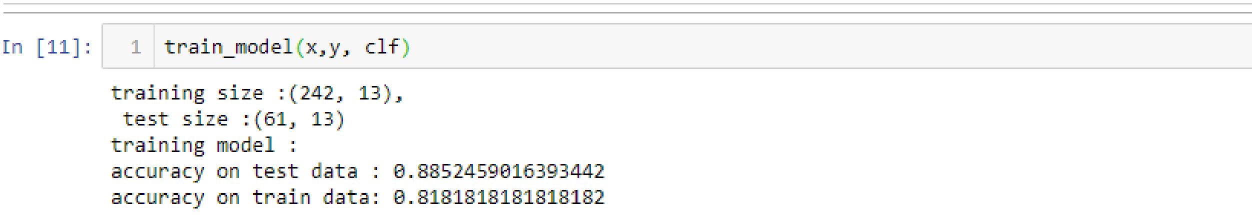
SIMULATIONS RESULTS AND DISCUSSIONS

4.1 INTRODUCTION

I created and tested several models including Support Vector Machines for classification. I then highlighted the models that performed well on testing and worked on them a little bit to improve their accuracy. The method used to improve the accuracy of the selected models was PCA (Principal Component Analysis). This helped improve the accuracy a lot. I created a function that was used to split the data, train the model on the data, print out the training accuracy and test accuracy, plot a heatmap showing the confusion matrix and a precision plot, and also plot a distribution plot of the models prediction alongside the actual data. I split the data with a set random state to enable me to recreate the same results. My test data size was 20% percent of the entire dataset. There was no need for validation data because the models were not trained using the test data

4.2 Gaussian Naïve Bayes

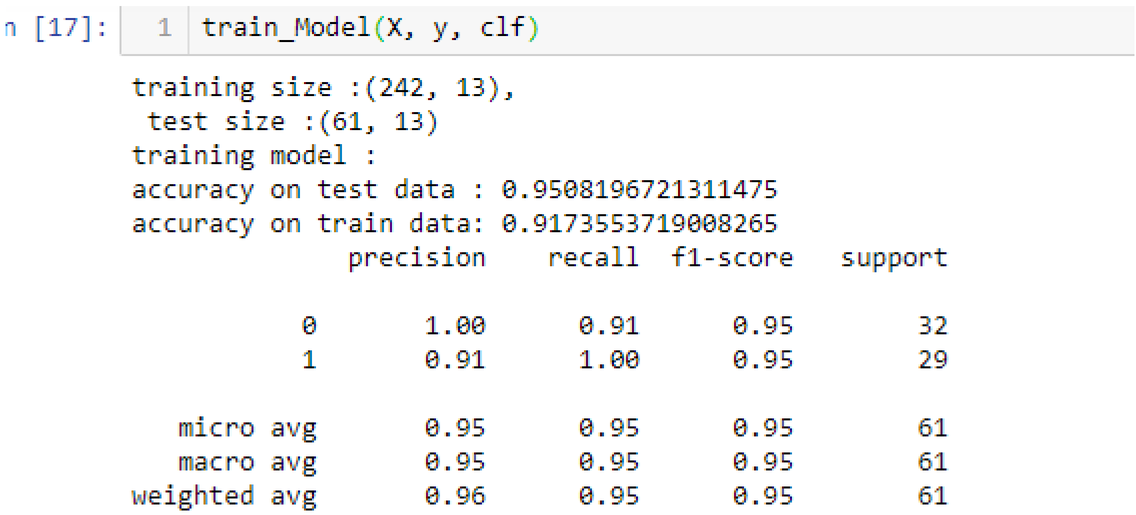
This model performed a little bit better than my decision tree model, so it was selected for further improvements. It got an accuracy of 88% on test data



```
In [11]:  train_model(x,y, clf)
```

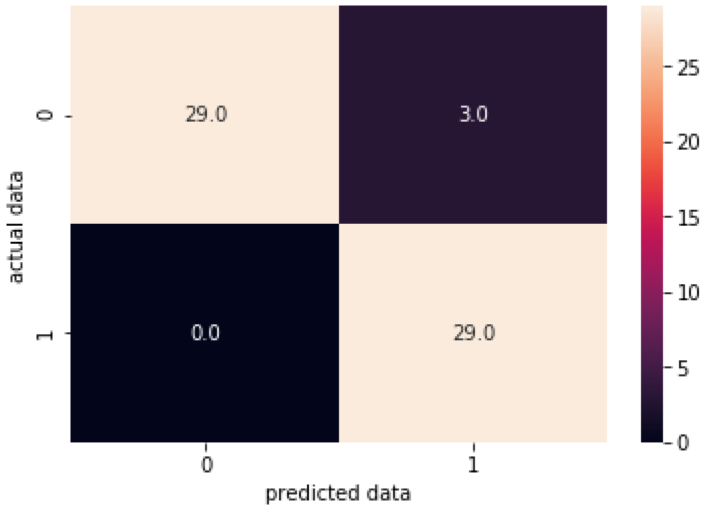
```
training size :(242, 13),
 test size :(61, 13)
training model :
accuracy on test data : 0.8852459016393442
accuracy on train data: 0.8181818181818182
```

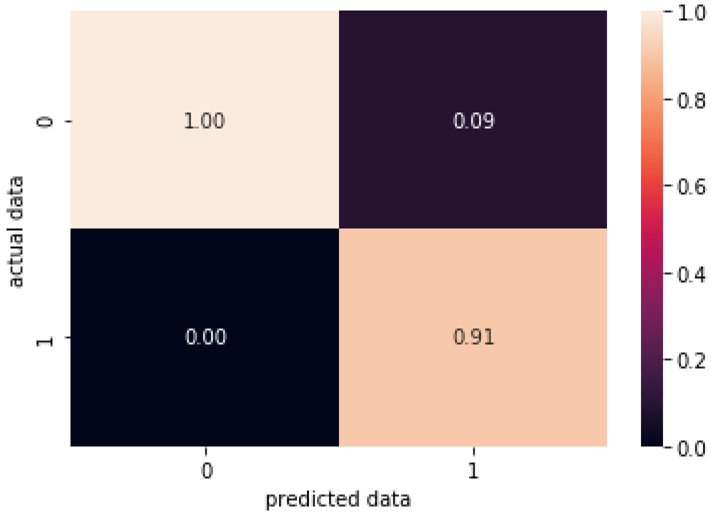
After applying PCA with 13 components, the accuracy on test data increased to 95%. And as shown in the screenshots provided, the training accuracy was less than the test accuracy to show that overfitting was not an issue.



```
n [17]:  train_Model(X, y, clf)
```

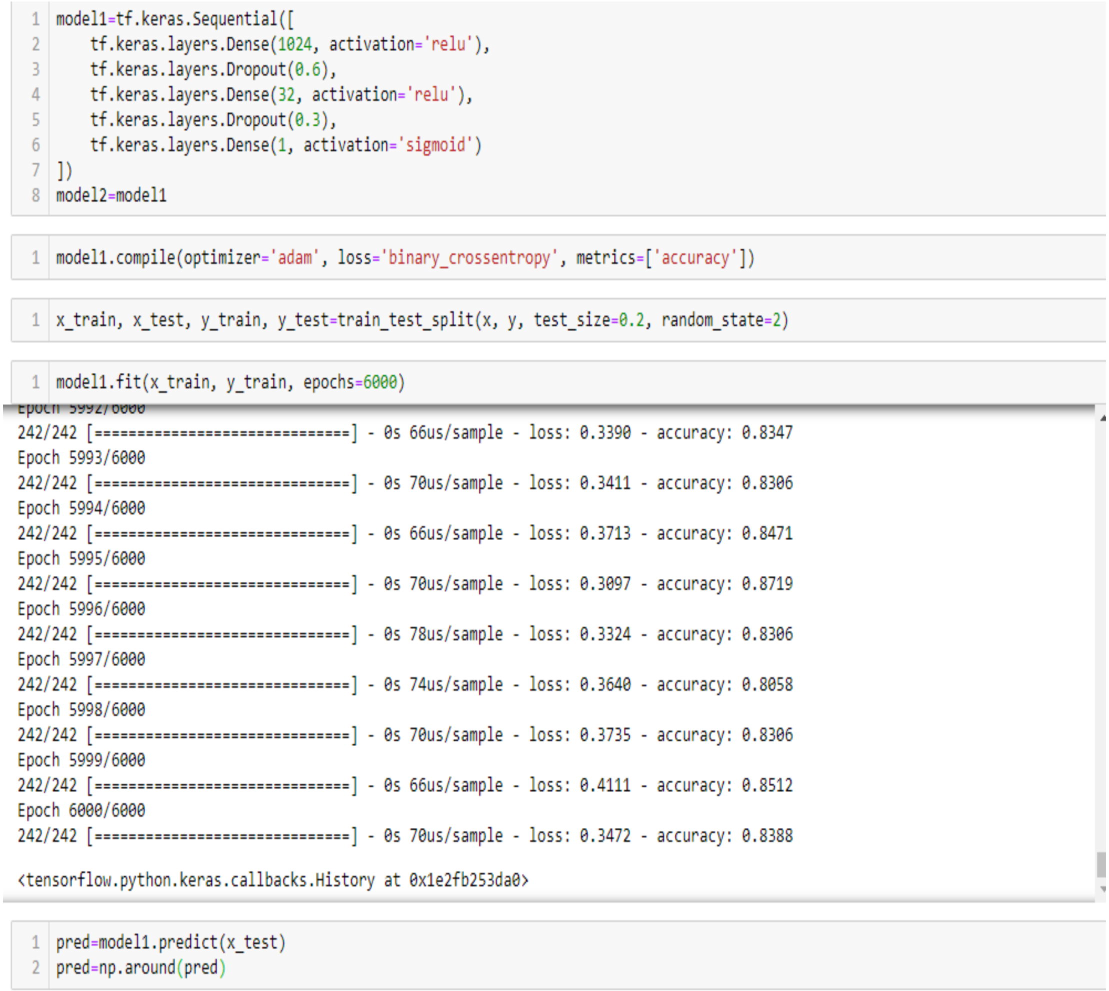
```
training size :(242, 13),
 test size :(61, 13)
training model :
accuracy on test data : 0.9508196721311475
accuracy on train data: 0.9173553719008265
              precision    recall  f1-score   support

           0       1.00      0.91      0.95        32
           1       0.91      1.00      0.95        29

   micro avg       0.95      0.95      0.95        61
   macro avg       0.95      0.95      0.95        61
weighted avg       0.96      0.95      0.95        61
```





4.3 Neural Networks

Neural Network is not always a go-to model when the performing not so complex classifications with not much data cause simpler models often perform better. Surprisingly, the neural network I created performed excellently well. It needed much training; I trained the model for 6000 Epochs and I added adequate dropouts to prevent overfitting on the training data. Below is a screenshot of the structure of the neural network created.



```
model1=tf.keras.Sequential([
    tf.keras.layers.Dense(1024, activation='relu'),
    tf.keras.layers.Dropout(0.6),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dropout(0.3),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model2=model1
```

```
model1.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

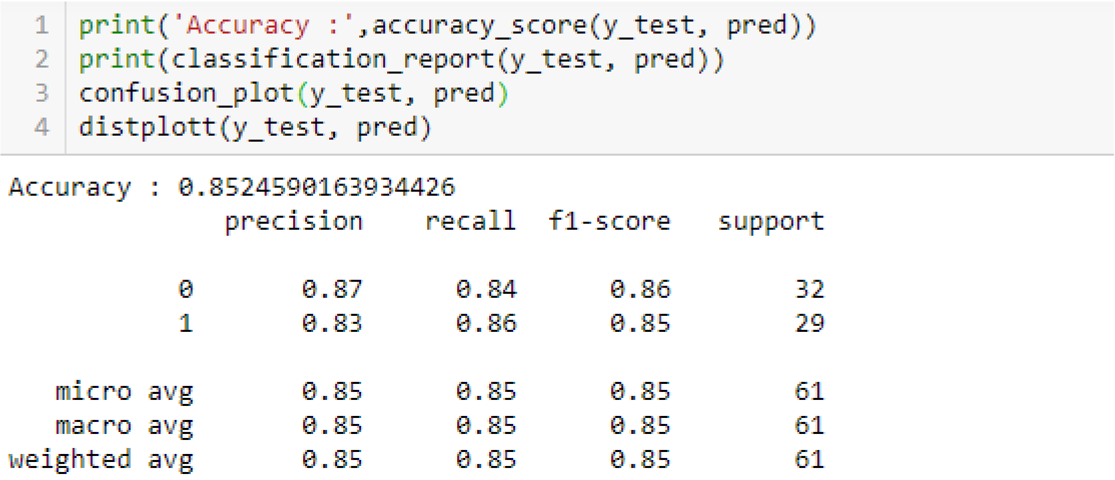
```
x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.2, random_state=2)
```

```
model1.fit(x_train, y_train, epochs=6000)
```

```
Epoch 5992/6000
242/242 [==============================] - 0s 66us/sample - loss: 0.3390 - accuracy: 0.8347
Epoch 5993/6000
242/242 [==============================] - 0s 70us/sample - loss: 0.3411 - accuracy: 0.8306
Epoch 5994/6000
242/242 [==============================] - 0s 66us/sample - loss: 0.3713 - accuracy: 0.8471
Epoch 5995/6000
242/242 [==============================] - 0s 70us/sample - loss: 0.3097 - accuracy: 0.8719
Epoch 5996/6000
242/242 [==============================] - 0s 78us/sample - loss: 0.3324 - accuracy: 0.8306
Epoch 5997/6000
242/242 [==============================] - 0s 74us/sample - loss: 0.3640 - accuracy: 0.8058
Epoch 5998/6000
242/242 [==============================] - 0s 70us/sample - loss: 0.3735 - accuracy: 0.8306
Epoch 5999/6000
242/242 [==============================] - 0s 66us/sample - loss: 0.4111 - accuracy: 0.8512
Epoch 6000/6000
242/242 [==============================] - 0s 70us/sample - loss: 0.3472 - accuracy: 0.8388

<tensorflow.python.keras.callbacks.History at 0x1e2fb253da0>
```

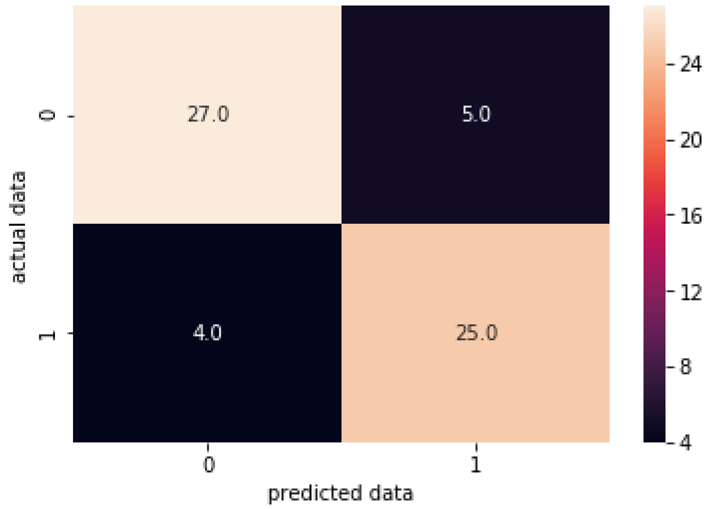
```
pred=model1.predict(x_test)
pred=np.around(pred)
```

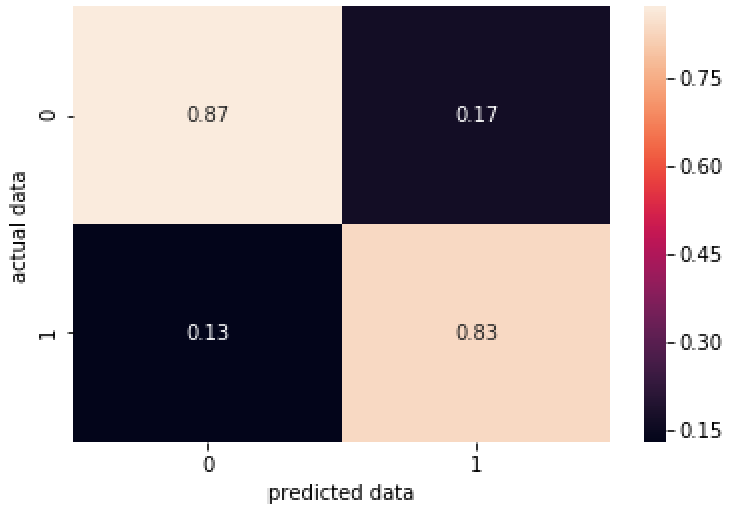
I used a sigmoid activation function on the final layer which is the output layer. This produces outputs from 1 to 0 meaning it produces an output close to 1 if it predicts the output is one otherwise it produces an output close to zero and sometimes beyond (negative). I had to round the predictions up so it could be in a binary form because we are dealing with a classification and not a regression problem. I did this using np.around as shown in the image above. After successful training, the model was tested and had an accuracy of 85% on the test data.



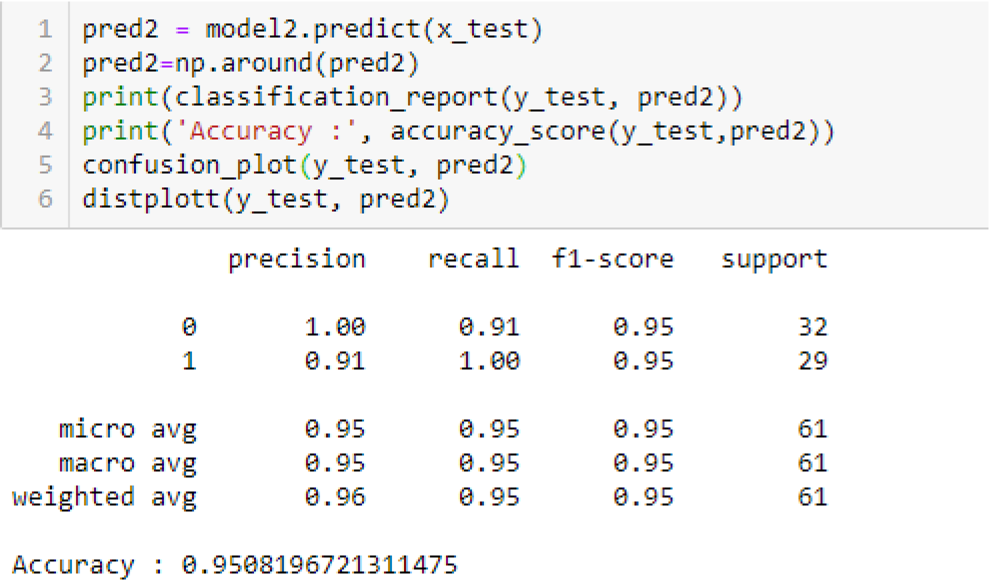
```
print('Accuracy :',accuracy_score(y_test, pred))
print(classification_report(y_test, pred))
confusion_plot(y_test, pred)
distplott(y_test, pred)
```

```
Accuracy : 0.8524590163934426
              precision    recall  f1-score   support

           0       0.87      0.84      0.86        32
           1       0.83      0.86      0.85        29

   micro avg       0.85      0.85      0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```



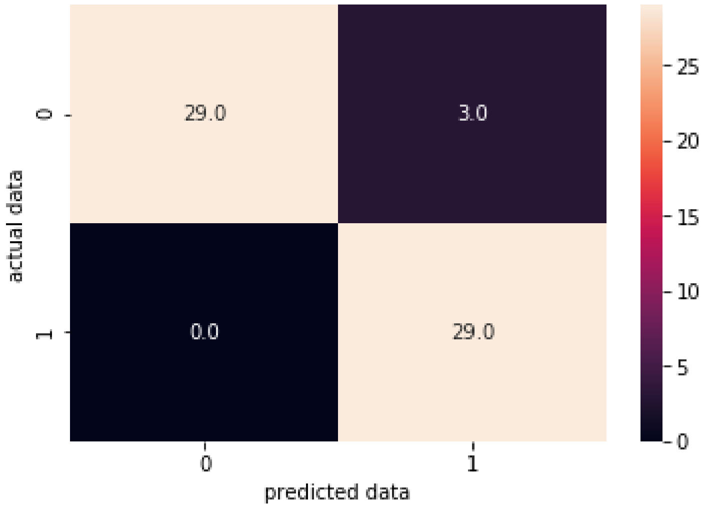


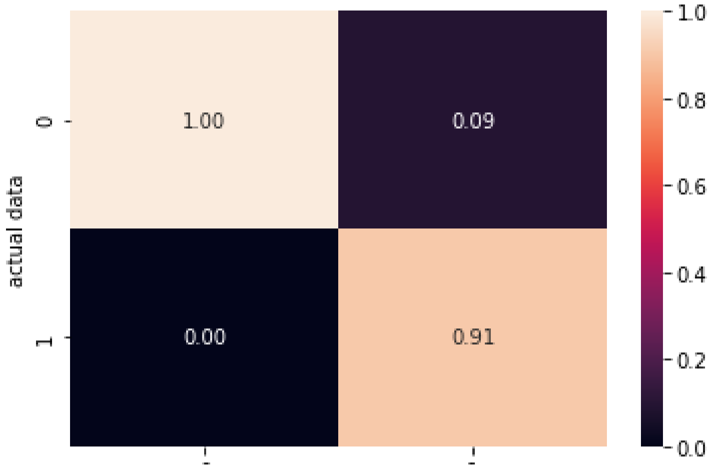
I applied pca on the data and trained the model for 4000 epochs and the accuracy remarkably increased to 95% as shown below.



```
pred2 = model2.predict(x_test)
pred2=np.around(pred2)
print(classification_report(y_test, pred2))
print('Accuracy :', accuracy_score(y_test,pred2))
confusion_plot(y_test, pred2)
distplott(y_test, pred2)
```

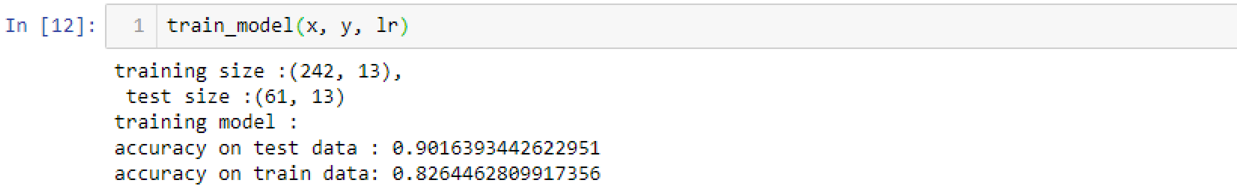
```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95        32
           1       0.91      1.00      0.95        29

   micro avg       0.95      0.95      0.95        61
   macro avg       0.95      0.95      0.95        61
weighted avg       0.96      0.95      0.95        61

Accuracy : 0.9508196721311475
```





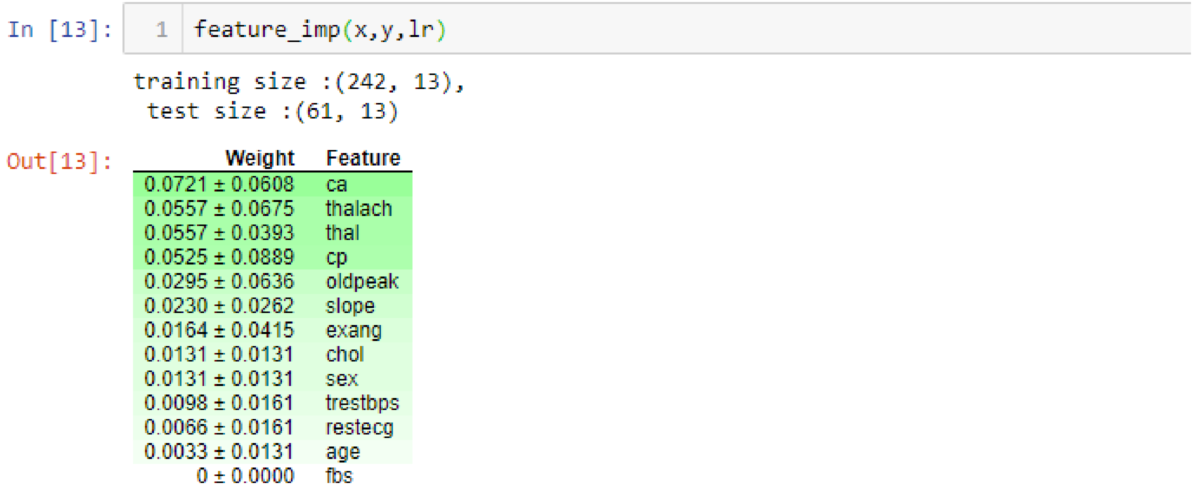
4.4 Logistic Regression

Simple models like this first comes to mind when dealing with these kinds of classification problems because it works so well even with little data. This was the best performing model with and without applying PCA. The Logistic Regression model had an accuracy of 90% on test data. No other model performed better.



```
In [12]:  train_model(x, y, lr)

training size :(242, 13),
 test size :(61, 13)
training model :
accuracy on test data : 0.9016393442622951
accuracy on train data: 0.8264462809917356
```

Before applying PCA on the dataset, there was something interesting that happened with the dataset. I plotted a graph that shows how attributes in the dataset contributes to the accuracy of the logistic regression model which is shown below.



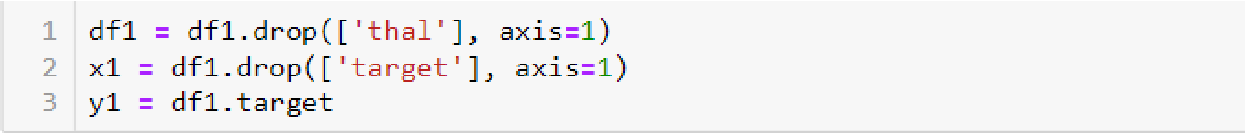
```
In [13]:  feature_imp(x,y,lr)
```

```
training size :(242, 13),
 test size :(61, 13)
```

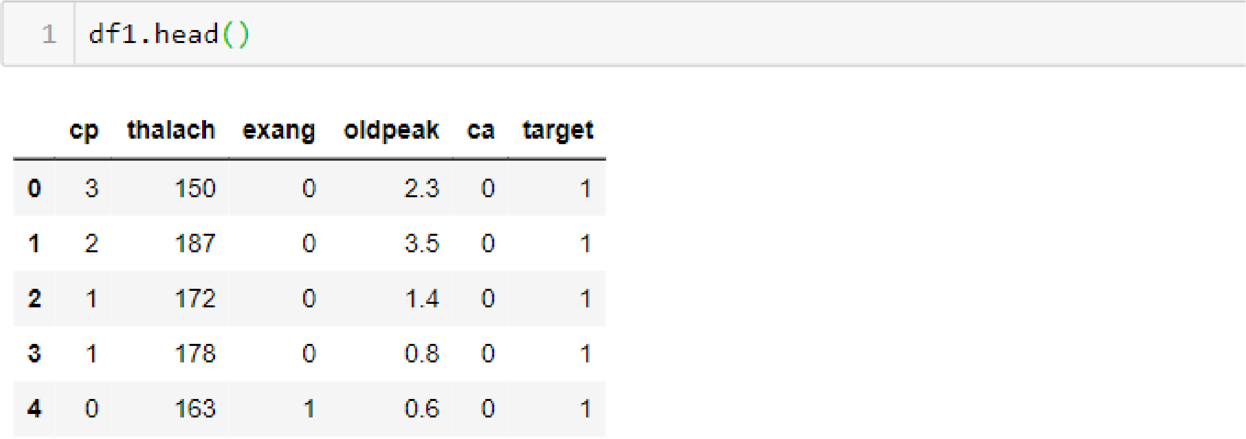
Out[13]:

Weight	Feature
0.0721 ± 0.0608	ca
0.0557 ± 0.0675	thalach
0.0557 ± 0.0393	thal
0.0525 ± 0.0889	cp
0.0295 ± 0.0636	oldpeak
0.0230 ± 0.0262	slope
0.0164 ± 0.0415	exang
0.0131 ± 0.0131	chol
0.0131 ± 0.0131	sex
0.0098 ± 0.0161	trestbps
0.0066 ± 0.0161	restecg
0.0033 ± 0.0131	age
0 ± 0.0000	fbs

The attributes highlighted green have a positive impact while the attributes highlighted red have a negative impact. I dropped attributes that were of the lowest importance iteratively. The dataset was left with just 5 attributes out of 13. These are the most important features which the logistic regression model will use to diagnose heart diseases in humans.

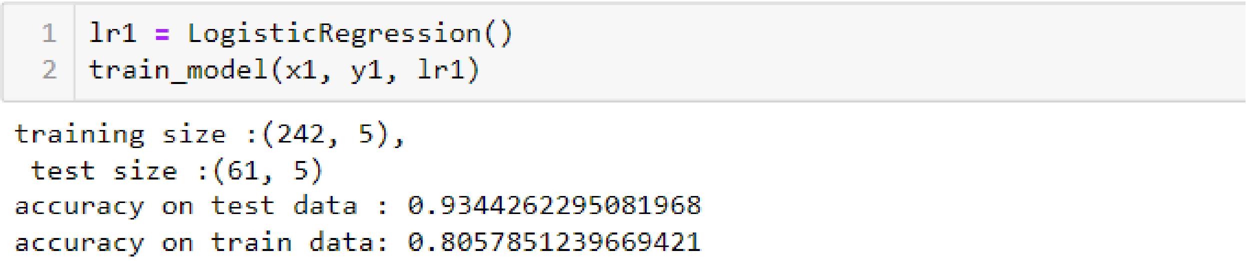


```
df1 = df1.drop(['thal'], axis=1)
x1 = df1.drop(['target'], axis=1)
y1 = df1.target
```



```
df1.head()
```

	cp	thalach	exang	oldpeak	ca	target
0	3	150	0	2.3	0	1
1	2	187	0	3.5	0	1
2	1	172	0	1.4	0	1
3	1	178	0	0.8	0	1
4	0	163	1	0.6	0	1



```
lr1 = LogisticRegression()
train_model(x1, y1, lr1)
```

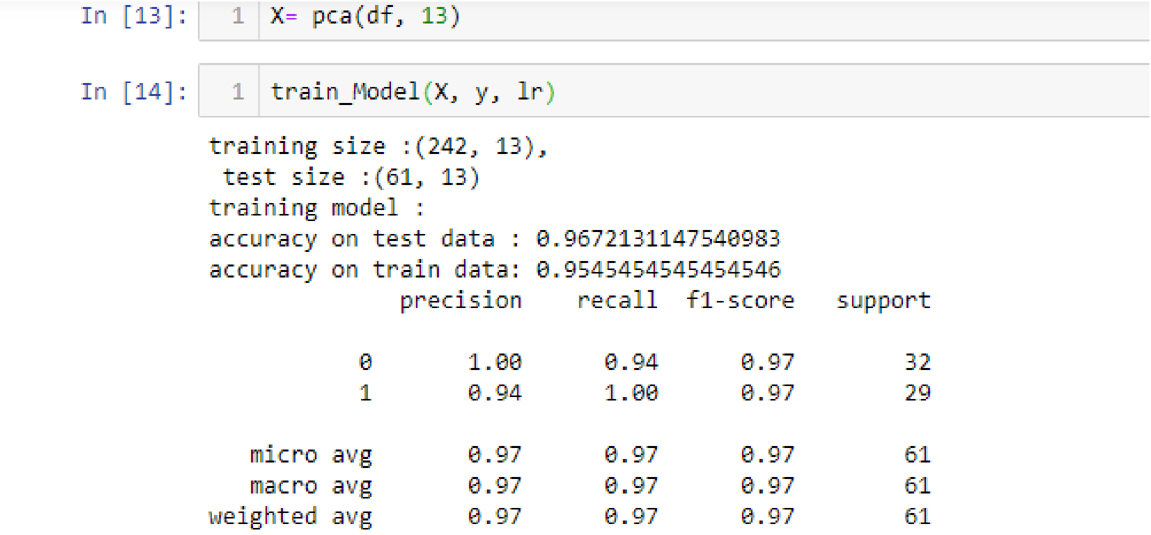
```
training size :(242, 5),
 test size :(61, 5)
accuracy on test data : 0.9344262295081968
accuracy on train data: 0.8057851239669421
```

This took a while and had lots of repeated actions, but it improved the accuracy of the model as shown above. 3% is a big deal in the health sector.

Having this new manually selected dataset, I decided to test it out on another model. I created a support vector machine classifier which I trained and tested on the original dataset. It got an accuracy of 47% on the test data and 100% on the training data (terribly overfitting). When I used the dataset in which the attributes were manually selected, the test accuracy increased significantly to 87% and accuracy on the training data reduced to 86%.

APPLYING PCA:

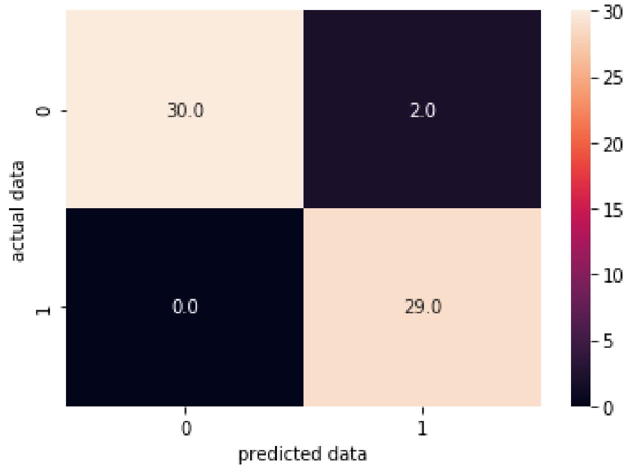
After applying PCA on the dataset, the accuracy of the logistic regression model rose to 97% on test data, a little better than previous models discussed but the training accuracy was nowhere near the previous models.

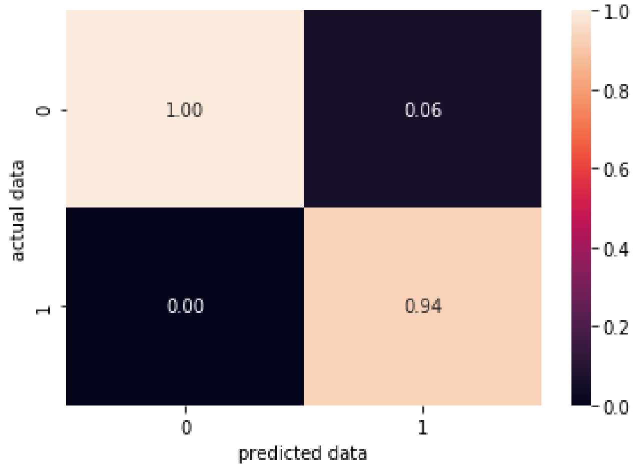


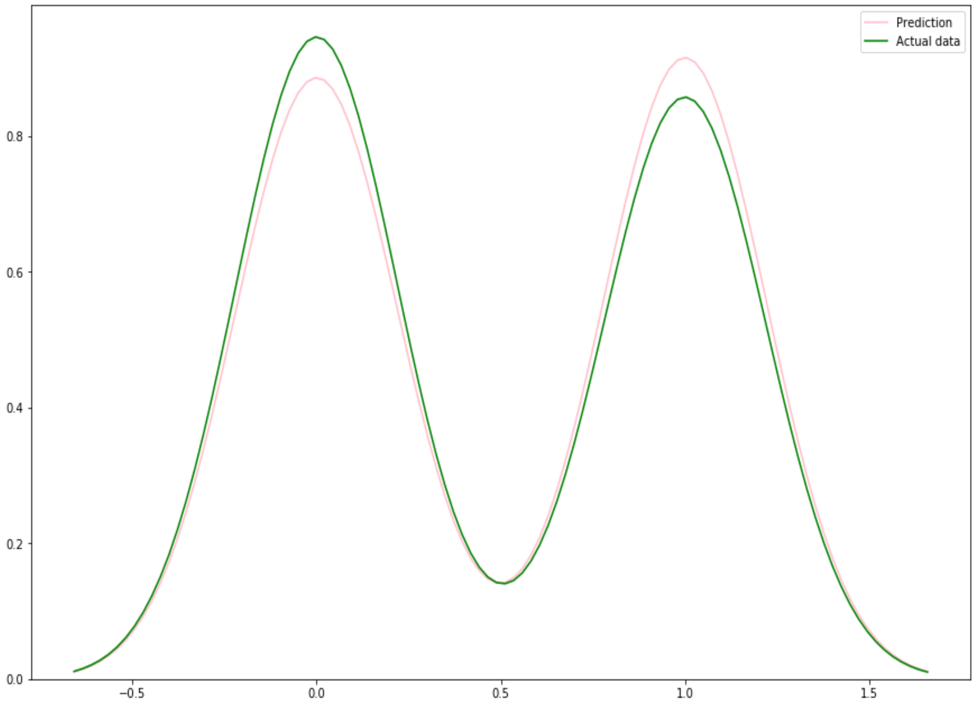
```
In [13]:  X= pca(df, 13)
```

```
In [14]:  train_Model(X, y, lr)
```

```
training size :(242, 13),
 test size :(61, 13)
training model :
accuracy on test data : 0.9672131147540983
accuracy on train data: 0.9545454545454546
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        32
           1       0.94      1.00      0.97        29

   micro avg       0.97      0.97      0.97        61
   macro avg       0.97      0.97      0.97        61
weighted avg       0.97      0.97      0.97        61
```






Prediction
Actual data
0.0
0.2
0.4
0.6
0.8
−0.5
0.0
0.5
1.0
1.5

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

To diagnose patients of various heart diseases, data on recognized risk variables of people was collected. Historical data on the distribution of heart diseases was gathered.

Using logistic regression, neural networks, and many other algorithms, the dataset comprising data on the risk variables recognised and gathered was used to formulate diagnostic predictive models for diagnosing heart diseases in humans.

The research findings disclosed the factors that were recognized as appropriate to assessing the risk of heart diseases in patients. The pattern recognition system in the logistic regression model was identified and was used to identify the major variables which were ideal to the model's accuracy in diagnosis/prediction. This pattern recognition identified in the logistic regression model helped improve the accuracy in the support vector machine classifier model.

Although feature engineering helped improve accuracy, PCA still stood as a better option as it improved the accuracy of several models. PCA greatly increased the accuracy of all models used and tested including artificial neural networks. In all cases, the logistic regression model outperformed all other model used in this research with and without applying PCA.

5.2 Conclusion

The study presented a diagnosis model for heart disease using relevant risk factors selected from a predefined set of heart disease risk factors. The risk factors used in our LR model was scaled down to five which are: chest pain type (cp); max heart rate reached; exercise induced angina; depression brought by workout in relation to rest; number of major vessels which are being coloured by fluoroscopy. Although our LR model had the best performance, all other models performed highly above average seeing nothing less than 95% on test accuracy after PCA was applied to the dataset. The logistic regression model had a test accuracy of 97% and a train accuracy of 95%.

5.3 Recommendations

A stronger understanding of the correlation between the characteristics appropriate to diagnosing heart diseases was suggested. This research was able to narrow down 5 variables which the model uses to make quality, accurate predictions. This model can be incorporated into the current Health Information System (HIS) that captures and manages clinical data that can be supplied to the predictive model for heart disease diagnosis. Going forward, only the 5 variables out of 13 identified are needed to make predictions if PCA will not be applied. It is recommended that a continuous evaluation of factors monitored for the risk of heart disease in humans be conducted to improve the quality of data appropriate to the creation of an enhanced forecast model for heart disease diagnosis using the model suggested in this research.

5.4 Limitations of the study

A major setback here is lack of funds which prevents the model from being deployed as an API on servers such as AWS servers. Without this, this research would include an online platform where the model would be deployed and can be tested on the go. For this study, several classification algorithms were used of which just 4 (gaussian naïve bayes algorithm, artificial neural networks, support vector machine classifier algorithm, logistic regression algorithm) are stated in this documentation.

References

Data Sources:

1. Hungarian Institute of Cardiology, Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleverland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor: David W. Aha(aha@ics.uci.edu) (714)856-8779. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease.

Mayo clinic staff, Disease-condition Heart diseases and symptoms. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

QuestionPro Survey Software (2020). What is a Survey – Definition, templates, methods, characteristics, and examples. Retrieved from https://www.questionpro.com/blog/surveys/.

Aqueel, A. and Shaikh, A. H. (2012). Data Mining Techniques to find out Heart Diseases: An Overview. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 1(4): 1 – 6.

Bellazzi, R., and Zupan, B. (2008), Predictive data mining in clinical medicine: current issues and guidelines. *Int. J. Med. Inform.* 77:81–97, Bramer, M. (2007). Principles of data mining: Springer

Cook, N. R. (2007). Use and misuse of the receiver operating characteristics curve in risk prediction. Circulation 115: 928 – 935.

Gauda, R. and Chahar, V. (2013). A comparative study on feature selection using data mining tools. *International Journal of advanced research in computer science and software engineering* 3(9): 26 – 33.

Goharian & Grossman. (2003). Data Mining Classification, Illinois Institute of Technology. Retrieved from http://ir.iit.edu/~nazli/cs422/CS422-Slides/DMClassification.pdf,.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 79: 340 – 349.

Hagerty, R. G., Butow, P.N. and Ellis, P.M. (2005). Communicating prognosis in cancer care: a systematic review of the literature. Ann Oncol 16: 1005 – 1053

Ibrahim, J. G., Chu, M. and Chen, M.H. (2012). Missing data in clinical studies: issues and methods. *J Clin. Oncol. 30:* 3297 – 3303.

Mitchell, T. M. (1997). Machine Learning. USA: McGraw-Hill Education.

Ian, T. J., Jorge, C. (2016). Principal component analysis: a review and recent developments. London: The Royal Society Publishing.

Wikipedia (2020). Principal component analysis. Retrieved from https://en.wikipedia.org/wiki/Principal_component_analysis.

Wikipedia (2020). Web scraping. Retrieved from https://en.wikipedia.org/wiki/Web_scraping.

World Health Organization (2020). Cardiovascular Diseases. Retrieved from who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

Rehan, F., Qadeer, A., Bashir, I., and Jamshaid, M. (2016). Risk factors of cardiovascular diseases in developing countries. *International Current Pharmaceutical Journal*, *5*(8), 69-72.

Brownlee, J. (November 11, 2019). 14 Different types of Learning in Machine Learning. Retrieved from machinelearningmastery.com/types-of-learning-in-machine-learning/

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning 1:81*-106.

McGregor, C., Christina, C., and Andrew, J. (2012). A process mining driven framework for clinical guideline improvement in critical care. Learning from medical data streams 13th conference on artificial intelligence in medicine (LEMEDS). Retrieved from http://ceur-ws.org, vol. 765.

Alpaydin, E. (1997). Voting over multiple condensed nearest neighbors. Artificial intelligence review, p. 115-132.

Apte, S., Weiss, M. (1997). Data mining with decision trees and decision rules. Watson research center. Retrieved from http://www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cover.pdf.

Anderson, J. A., and Davis, J. (1995). An introduction to neural networks. MIT, Cambride.

Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. Infect. Control hosp. Epidemiol. 25(8) :690-695.

Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D. (1998). Infrared micro spectroscopy and artificial neural networks in the diagnosis of cervical cancer. Cell. Mol. Biol. (Noisy-le-Grand, France) 44(1):179.

Sharma, A., and Roy, R. J. (1997). Design of a recognition system to predict movement during anesthesia. IEEE Trans. Biomed. Eng. 44(6):505-511.

Dunham M. H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

APPENDIX

```
def split_data(x,y):
    '''Inputs: independent variables, dependent variables
        output: a train and test portion of the data'''
    x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.2, random_state=2)
    # the random state was given a value so our results can be reproduced
    print('training size :{}, \n test size :{}'.format(x_train.shape, x_test.shape))

    return x_train, x_test, y_train, y_test
def train_model(x,y, model):
    '''Receives the data and the model to be trained as inputs and prints the accuracy on the test data and train data'''
    x_train, x_test, y_train, y_test=split_data(x,y)

    model.fit(x_train, y_train)

    print("accuracy on test data : {}".format(model.score(x_test, y_test)))

    print("accuracy on train data: {}".format(model.score(x_train, y_train)))
df=pd.read_csv(r"D:\file\location")
df.head()
```

```
lr = LogisticRegression()
#assigning variables to the independent and dependent variables
x = df.drop(['target'], axis=1)
#x=x.values this was commented out because I will later use of permutation importance.
y=df.target
y=y.values
train_model(x, y, lr)
model=tf.keras.Sequential([


   tf.keras.layers.Dense(1024, activation='relu'),
   tf.keras.layers.BatchNormalization(),
   tf.keras.layers.Dropout(0.4),

   tf.keras.layers.Dense(1024, activation='relu'),
   tf.keras.layers.Dropout(0.4),

   tf.keras.layers.Dense(1, activation='sigmoid')
])
adam = tf.keras.optimizers.Adam(lr=0.001)
```

```
model.compile(optimizer=adam, loss='binary_crossentropy', metrics=['accuracy'])

x_nn = x.values #neural network only accepts digits.

y_nn = y

x_train, x_test, y_train, y_test= split_data(x_nn, y_nn)

model.fit(x_train, y_train, epochs=200)

pred=model.predict(x_test)

pred=np.around(pred)

print('Accuracy :',accuracy_score(y_test, pred))

print(classification_report(y_test, pred))

confusion_plot(y_test, pred)

distplott(y_test, pred)

roc_plot(y_test, pred)
```