

Prediction of Heart Attack using Supervised machine learning technique

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Engineering Degree in Computer Science and
Engineering

By

Rubesh S (Reg. No.38110468)

Raj Kumar N D (Reg. No.38110446)



**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING
SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMIL NADU**

AUGUST 2022



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Rubesh S(Reg. No.38110468)Raj Kumar N D(Reg. No.38110446)** who carried out the project entitled "**Prediction of Heart Attack using Supervised machine learning technique**" under my supervision from November 2021 to March 2022.

Internal Guide

Dr. S. Prayla shyry M.E., PhD.,

Head of the Department

Dr. L. Lakshmanan, M.E., Ph.D.,

Dr.S.Vigneshwari M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

We Rubesh S, Raj Kumar N D hereby declare that the Project Report entitled **Prediction of Heart Attack using Supervised machine learning technique** done by me under the guidance of **Dr. S. Prayla shyry M.E., Ph.D.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

Rubesh S

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the Board **of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D., Dean**, School of Computing , **Dr.S.Vigneshwari M.E., Ph.D.,** and **Dr.L.Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S. Prayla shyry M.E., Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

Abstract

Heart is the main organ that pumps blood to the body and for proper functioning of the body. Heart disease is a fatal human disease increasing globally in both developed and undeveloped countries and consequently, causes death. Normally, in this disease, the heart fails to supply a sufficient amount of blood to other parts of the body in order to accomplish their normal functionalities. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. The proposed method is to build a machine learning model capable of classifying whether the person has heart disease or not. Different algorithms are compared and the best model is used for predicting the outcome.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	i
	LIST OF FIGURES	iii
1	INTRODUCTION	1
	1.1 Outline of the Project	1
	1.2 Existing system	2
	1.3 Proposed System	3
2	LITERATURE SURVEY	4
	2.1General	4
	2.2 Review of Literature Survey	4
3	METHODOLOGY	8
	3.1MACHINE LEARNING	8
	3.2Feasibility study	10

	3.3Construction of a Predictive Model	11
	3.4 Project Requirements	12
	3.5 System Architecture	14
	3.6 Module Documents	17
	3.7Data Validation/ Cleaning/Preparing Process	18
	3.8 Exploration data analysis of visualization	19
	3.9 Prediction result by accuracy	20
	3.10 Used Python Packages	23
4	RESULTS AND DISCUSSION	25
5	CONCLUSION AND FUTURE WORK	28
	5.1CONCLUSION	28
	5.2 Future work	28
	APPENDIX	29
	A.SCREENSHOTS of the application	29
	B. Reference	26

CHAPTER 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.

The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

Data Scientist:

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to

mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data.

Required Skills for a Data Scientist:

- **Programming:** Python, SQL, Scala, Java, R, MATLAB.
- **Machine Learning:** Natural Language Processing, Classification, Clustering.
- **Data Visualization:** Tableau, SAS, D3.js, Python, Java, R libraries.
- **Big data platforms:** MongoDB, Oracle, Microsoft Azure, Cloudera.

1.2 Existing System:

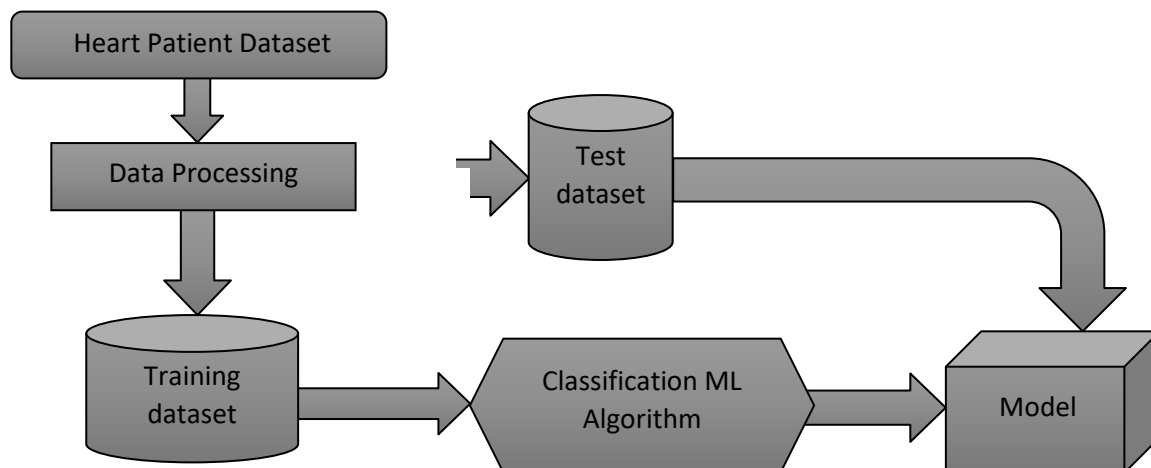
The Acoustic analysis has the potential to be a relatively low cost and non-invasive diagnostic tool for point-of-care assessment and remote monitoring of LVAD recipients. Prior work on acoustic analysis in LVAD recipients has focused on analysis of LVAD sounds to detect pump thrombosis. The analysis of heart sounds in LVAD recipients has been limited by the dominant LVAD sounds present within their precordial sounds. We have developed a novel signal processing pipeline to mitigate LVAD-generated sounds within precordial sounds recorded from LVAD recipients, potentially rendering automated heart sound analysis in this population feasible. Their analysis revealed that pump activity in LVAD recipients induces heart sounds with diverse acoustic signatures, such as S1 and S2 sounds with varying frequency ranges and relative amplitudes, and variabilities in S2 components based on changes in aortic valve mobility across different pump speeds.

Disadvantages:

1. Analyzing the sound graphs and getting the sound it's a complex process.
2. Better classification can be done by proper comparison of algorithms.

1.3 Proposed System:

The proposed method is to build a machine learning model for classification of heart disease. The process carries from data collection where the past data related to heart disease are collected. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Heart disease, if found before proper treatment, can save lives. Machine learning is now applied and mostly used in health care where it reduces the manual effort and better model makes error less which leads to saving lives. The data analysis is done on the dataset proper variable identification done that is both the dependent variables and independent variables are found. Then proper machine learning algorithms are applied on the dataset where the pattern of data is learnt. After applying different algorithms a better algorithm is used for the prediction of outcome.



Architecture of Proposed model

Advantages:

1. The data used here will not be as complex as sound graphs and analysis of data will be easy.
2. Performance metrics of different algorithms are compared, and the better prediction is done.

CHAPTER 2

LITERATURE SURVEY

2.1 General

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is a secondary source and discusses published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a reorganization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

2.2 Review of Literature Survey

Title: REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING
AND HYBRID INTELLIGENT TECHNIQUES

Author: R. Chitra¹ and V. Seenivasagam²

Year: 2013

The Healthcare industry generally clinical diagnosis is done mostly by doctor's expertise and experience. The Computer Aided Decision Support System plays a major role in the medical field. With the growing research on heart disease predicting systems,

it has become important to categorize the research outcomes and provide readers with an overview of the existing heart disease prediction techniques in each category. Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. From the study it is observed that Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system. The commonly used techniques for Heart Disease Prediction and their complexities are summarized in this paper.

Title:EFFECTIVE PREDICTION OF CARDIOVASCULAR DISEASE USING CLUSTEROF MACHINE LEARNING ALGORITHMS

Author: G. Jignesh Chowdary¹, Suganya. G ² , Premalatha. M³

Year:2020

Cardiovascular diseases are one of the diseases that account for the loss of millions of lives each year. Lack of early prediction is the primary reason for the loss of lives, and this encourages researchers to develop intelligent systems for better prediction. The heart is a muscular organ that is largely made of cardiac muscle. The heart pumps oxygenated blood throughout the body by rhythmic contraction and dilation. Cardiovascular diseases deal with the problems related to heart and blood vessels. Cardiovascular diseases are the life-threatening diseases of the present-day world. According to the World Health Organisation(WHO), 17 million people die each year due to Cardiovascular disease, which accounts for 31% of the deaths worldwide[1]., and it is among the top 10 global causes of deaths around the world [2]. It is challenging to identify cardiovascular disease because it depends on many factors. Here are some significant factors that cause cardiovascular disease; they are namely high blood pressure, excessive lipids in blood, stress, obesity, metabolic syndrome, diabetes, and many more. Sometimes family history may also be a factor for the disease. Several tests are performed before the diagnosis of Cardiovascular disease; they are the electrocardiogram(ECG), blood pressure, cholesterol, blood sugar, and auscultation.

Title:A Survey on Prediction Techniques of Heart Disease using Machine Learning

Author: Mangesh Limbitote, Dnyaneshwari Mahajan, Kedar Damkondwar, Pushkar Patil

Year:2020

Heart is one of the most important part of the body. It helps to purify and circulate blood to all parts of the body. Most number of deaths in the world are due to Heart Diseases. Some symptoms like chest pain, faster heartbeat, discomfort in breathing are recorded. This data is analysed on regular basis. In this review, an overview of the heart disease and its current procedures is firstly introduced. Furthermore, an in-depth analysis of the most relevant machine learning techniques available on the literature for heart disease prediction is briefly elaborated. Heart disease is the kind of disease which can cause the death. Each year too many people are dying due to heart disease. Heart disease can occur due to the weakening of heart muscle. Also, heart failure can be described as the failure of the heart to pump the blood. Heart disease is also called coronary artery disease (CAD). CAD can occur due to insufficient blood supply to arteries.

Title: Heart Disease Prediction Using Effective Machine Learning Techniques

Author: Avinash Golande, Pavan Kumar T

Year:2019

In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and female categories and this ratio may vary according to the region. Also this ratio is considered for the people of age group 25-69. This does not indicate that the

people with other age groups will not be affected by heart diseases. This problem may start in the early age group also and predicting the cause and disease is a major challenge nowadays. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases.

Title: Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques

Author: 1Animesh Hazra, 2Subrata Kumar Mandal, 3Amit Gupta, 4Arkomita Mukherjee and 5Asmita Mukherjee

Year: 2017

A popular saying goes that we are living in an “information age”. Terabytes of data are produced every day. Data mining is the process which turns a collection of data into knowledge. The healthcare industry generates a huge amount of data daily. However, most of it is not effectively used. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes are not much prevalent. The aim of this paper is to summarize some of the current research on predicting heart diseases using data mining techniques, analyze the various combinations of mining algorithms used and conclude which technique(s) are effective and efficient. Also, some future directions on prediction systems have been addressed.

CHAPTER 3

METHODOLOGY

3.1 MACHINE LEARNING

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves the use of specialized algorithms. It feeds the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated into three categories. There is supervised learning, unsupervised learning and reinforcement learning. Supervised learning programs are both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning has no labels. It provided the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning

approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.



Process Of Machine Learning

Supervised Machine Learning **is the** majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output **is $y = f(X)$** . The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include **logistic regression, multi-class classification, Decision Trees** and **support vector machines etc.** Supervised learning requires that the data used to train the algorithm is already labelled with correct answers. Supervised learning problems can be further grouped into **Classification** problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

3.2 Feasibility study:

Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

Data collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

Preprocessing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data needs to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion needs to be done.

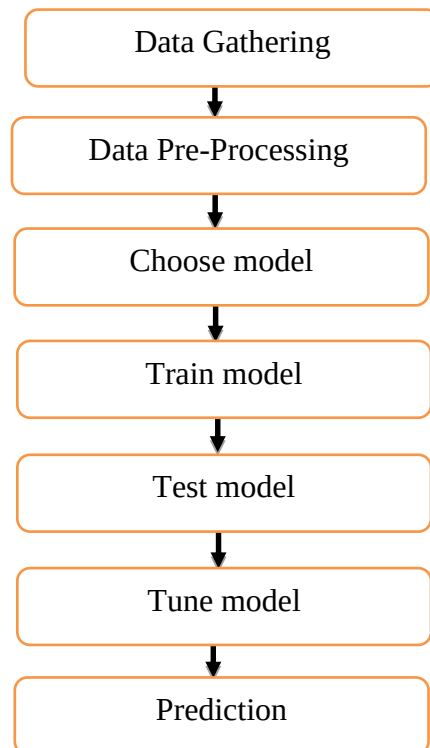
Building the classification model

The prediction of Heart attack, A high accuracy prediction model is effective because of the following reasons: It provides better results in classification problems.

- ❑ It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
- ❑ It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

3.3 Construction of a Predictive Model

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to pre-process then, what kind of algorithm with model. Training and testing this model working and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy.



Process of dataflow diagram

List of Modules:

- ❓ Data Pre-processing
- ❓ Data Analysis of Visualization
- ❓ Comparing Algorithm with prediction in the form of best accuracy result
- ❓ Deployment Using Flask

3.4 Project Requirements

General:

Requirements are the basic constraints that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements
2. Non-Functional requirements
3. Environment requirements
 - A. Hardware requirements
 - B. software requirements

Functional requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

Non-Functional Requirements:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

Environmental Requirements:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

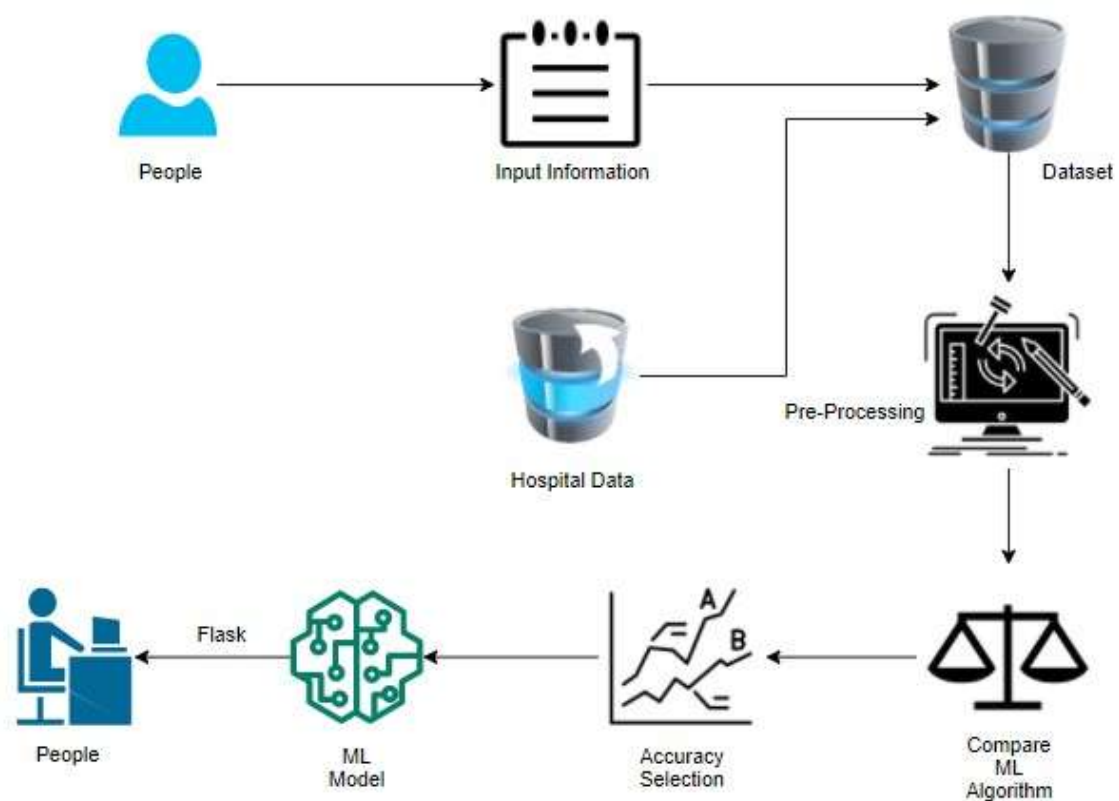
2. Hardware requirements:

Processor : Pentium IV/III

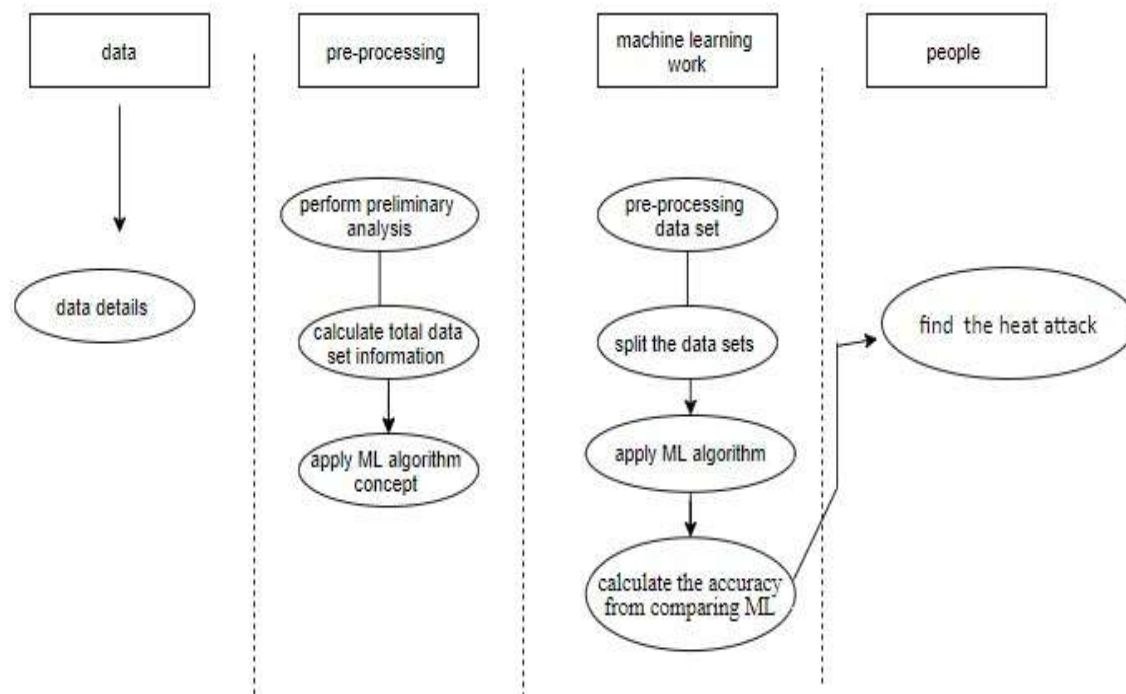
Hard disk : minimum 80 GB

RAM : minimum 2 GB

3.5 System Architecture



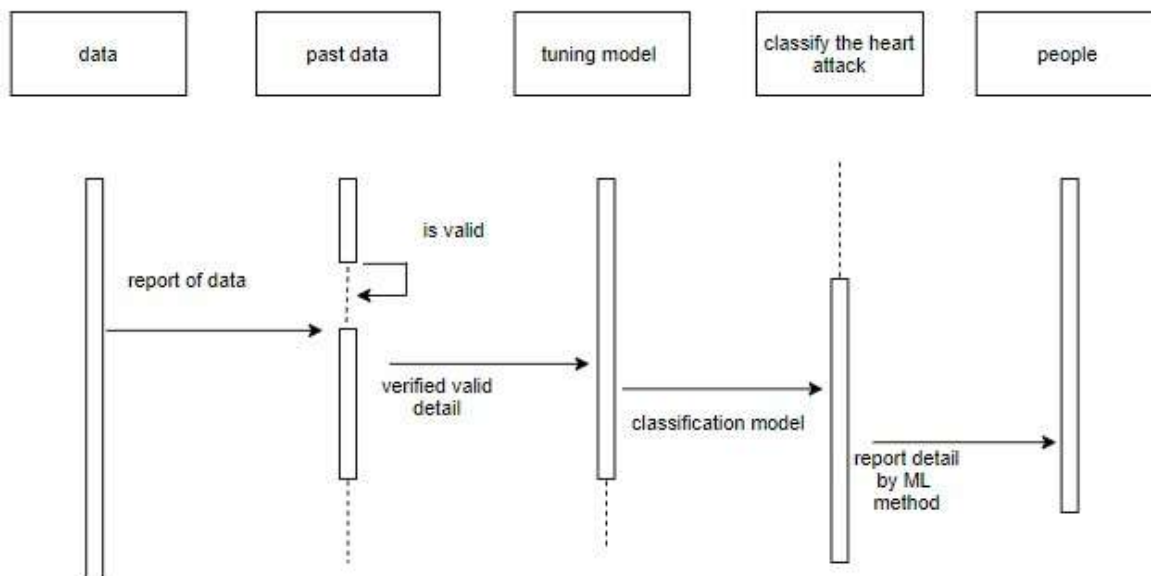
3.5.1 Activity Diagram:



Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow

chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

3.5.2 Sequence Diagram:



Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modelling, which focuses on identifying the behaviour within your system. Other dynamic modelling techniques include activity

diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

3.6 Module description:

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it

able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before joining code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- ❑ import libraries for access and functional purpose and read the given dataset
- ❑ General Properties of Analyzing the given dataset
- ❑ Display the given dataset in the form of data frame
- ❑ show columns
- ❑ shape of the data frame
- ❑ To describe the data frame
- ❑ Checking data type and information about dataset
- ❑ Checking for duplicate data
- ❑ Checking Missing values of data frame
- ❑ Checking unique values of data frame

- ❓ Checking count values of data frame
- ❓ Rename and drop the given data frame
- ❓ To specify the type of values
- ❓ To create extra columns

3.7 Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

3.8 Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- ❓ How to chart time series data with line plots and categorical quantities with bar charts.
- ❓ How to summarize data distributions with histograms and box plots.

3.9 Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

True Positive Rate(TPR) = $TP / (TP + FN)$

False Positive rate(FPR) = $FP / (FP + TN)$

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = $(TP + TN) / (TP + TN + FP + FN)$

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you

have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

$$\text{F- Measure} = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN})$$

F1-Score Formula:

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

3.10 Used Python Packages:**sklearn:**

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like `train_test_split`, `DecisionTreeClassifier` or `Logistic Regression` and `accuracy score`.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. The following are the basic steps involved in performing the random forest algorithm:

- ❑ Pick N random records from the dataset.
- ❑ Build a decision tree based on these N records.
- ❑ Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

CHAPTER 4

RESULT AND DISCUSSION

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- ❑ At the beginning, we consider the whole training set as the root.
- ❑ Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- ❑ On the basis of attribute values records are distributed recursively.
- ❑ We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

Classification report of Decision Tree Classifier Results:

	precision	recall	f1-score	support
1	0.85	0.76	0.80	45
2	0.73	0.83	0.78	36
accuracy			0.79	81
macro avg	0.79	0.79	0.79	81
weighted avg	0.80	0.79	0.79	81

Confusion Matrix result of Decision Tree Classifier is:

```
[[34 11]
 [ 6 30]]
```

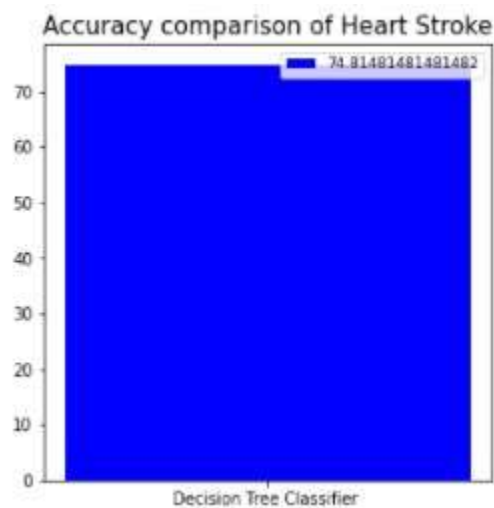
Sensitivity : 0.7555555555555555

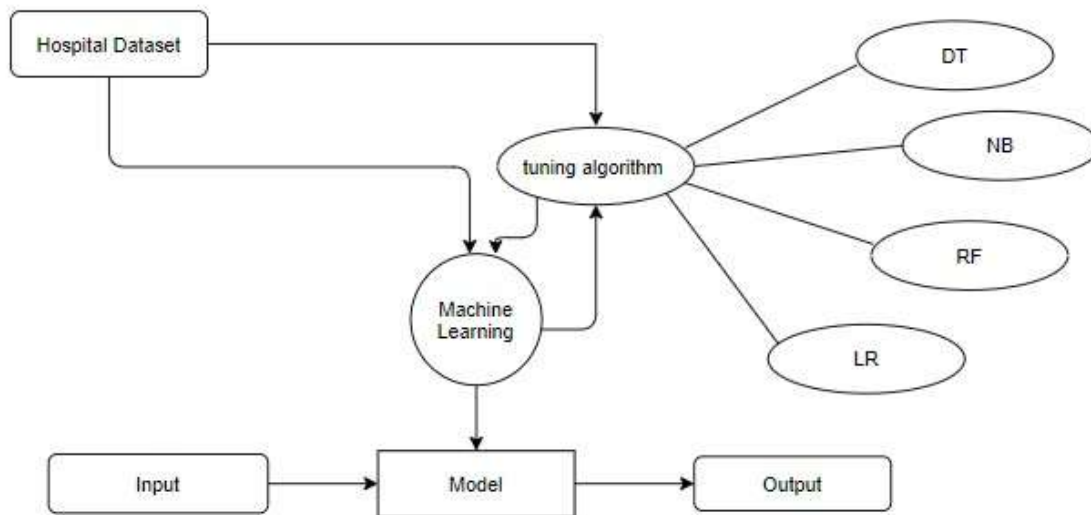
Specificity : 0.8333333333333334

Cross validation test results of accuracy:

```
[0.68518519 0.68518519 0.83333333 0.72222222 0.81481481]
```

Accuracy result of Decision Tree Classifier is: 74.81481481481482





An entity relationship diagram (ERD), also known as an entity relationship model, is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An ERD is a data modeling technique that can help define business processes and be used as the foundation for a relational database. Entity relationship diagrams provide a visual starting point for database design that can also be used to help determine information system requirements throughout an organization. After a relational database is rolled out, an ERD can still serve as a referral point, should any debugging or business process re-engineering be needed later.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on a public test set is a higher accuracy score will be found out. This application can help to find the Prediction of Heart Attack.

5.2 Future Work

- ❑ Heart Attack prediction to connect with the cloud.
- ❑ To optimize the work to implement in an Artificial Intelligence environment.

APPENDIX

A. SCREENSHOTS OF THE APPLICATION

PREDICTION OF HEART STROKE

AGE		TIME/AGE	
GENDER		THALUSION	
CP		ESANG	
BP		BLO.PDM	
CHOL		SLOPE	
		CA	

Enter the value

PREDICTION OF HEART STROKE

AGE: 70

GENDER: F - FEMALE

CP: 4 - ATOMIZED

BP: 130

CHOL: 232

THALACH: 100

RESTING: 0

OLD PEAK: 2.4

SLOPE: 2

CA:

Output prediction 1

CHOL: CHOL

FBS: FBS

ECG: ECG

SLOPE: SLOPE

CA: CA

THAL: THAL

Predict

HEART STROKE WILL OCCUR

Output prediction 2



B. Reference

1. World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.
2. American Heart Association, *Classes of Heart Failure*, American Heart Association, Chicago, IL, USA, 2020, <https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure>.
3. American Heart Association, *Heart Failure*, American Heart Association, Chicago, IL, USA, 2020, <https://www.heart.org/en/health-topics/heart-failure>.
4. S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning," *From Theory to Algorithms*, Cambridge University Press, Cambridge, UK, 2020. View at: [Google Scholar](#)
5. T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," *Data Mining, Inference, and Prediction*, Springer, Cham, Switzerland, 2020.
6. S. Marsland, "Machine learning," *An Algorithmic Perspective*, CRC Press, Boca

Raton, FL, USA, 2020.View at: [Google Scholar](#)

7. M. M. A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, and R. R. Yager, “Deep learning approach for active classification of electrocardiogram signals,” *Information Sciences*, vol. 345, pp. 340–354, 2016.View at: [Publisher Site](#) | [Google Scholar](#)
8. G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, “A machine learning system to improve heart failure patient assistance,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1750–1756, 2014.View at: [Publisher Site](#) | [Google Scholar](#)
9. R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, “Automatic methods to extract New York heart association classification from clinical notes,” in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.View at: [Publisher Site](#) | [Google Scholar](#)