# CARDIOVASCULAR PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES

**A PROJECT REPORT**

***Submitted by***

**RITHIVASAN S [211420104227]**

**SABARI D [211420104232]**

**TAMILSELVAN S [211420104285]**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

***in***

**COMPUTER SCIENCE AND ENGINEERING**

**PANIMALAR ENGINEERING COLLEGE**

**(An Autonomous Institution, Affiliated to Anna University, Chennai)**

**MARCH 2024**

# PANIMALAR ENGINEERING COLLEGE

**(An Autonomous Institution, Affiliated to Anna University, Chennai)**

## BONAFIDE CERTIFICATE

Certified that this project report **"CARDIOVASCULAR PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES "** is the bonafide work of "**RITHIVASAN S [211420104227], SABARI D [211420104232], TAMILSELVAN S [211420104285]"** who carried out the project work under mysupervision.

| **Signature of the HOD with date** | **Signature of the Supervisor with date** |
| --- | --- |
| **Dr.L.JABASHEELA, M.E., Ph.D.,** | **Mrs.K.CINTHUJA, M.E.** |
| **PROFESSOR AND HEAD,** | **ASSISTANT PROFESSOR,** |
| Department of Computer Science and Engineering, | Department of Computer Science and Engineering, |
| Panimalar Engineering College, | Panimalar Engineering College, |
| Chennai – 123 | Chennai – 123 |

Certified that the above candidate(s) was examined in the End Semester

Project Viva-Voce Examination held on ..............................

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

## DECLARATION BY THE STUDENT

We **RITHIVASAN S [211420104227], SABARI D [211420104232], TAMILSELVAN S [211420104285]** here by declare that this project report titled **“CARDIOVASCULAR PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES ”** , under the guidance of **Mrs.K.CINTHUJA, M.E.** is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

**RITHIVASAN S**

**SABARI D**

**TAMILSELVAN S**

# ACKNOWLEDGEMENT

Our profound gratitude is directed towards our esteemed Secretary and Correspondent, **Dr. P. CHINNADURAI, M.A., Ph.D**., for his fervent encouragement. His inspirational support proved instrumental in galvanizing our efforts, ultimately contributing significantly to the successful completion of this project

We want to express our deep gratitude to our Directors, **Tmt. C. VIJAYARAJESWARI, Dr. C. SAKTHI KUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.,** for graciously affording us the essential resources and facilities for undertaking of this project.

Our gratitude is also extended to our Principal, **Dr. K. MANI, M.E.,Ph.D.,** whose facilitation proved pivotal in the successful completion of this project.

We express our heartfelt thanks to **Dr. L. JABASHEELA, M.E., Ph.D.,** Head of the Department of Computer Science and Engineering, for granting the necessary facilities that contributed to the timely and successful completion of project.

We would like to express our sincere thanks to **Project Coordinator Dr.V.SUBEDHA, M.E., Ph.D.,** and **Project Guide Mrs.K.CINTHUJA, M.E.** and all the faculty members of the Department of CSE for their unwavering support for the successful completion of the project.

**RITHIVASAN S**

**SABARI D**

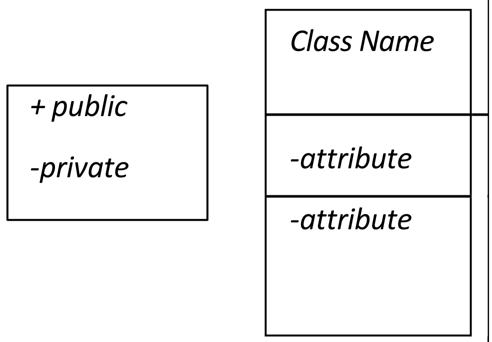
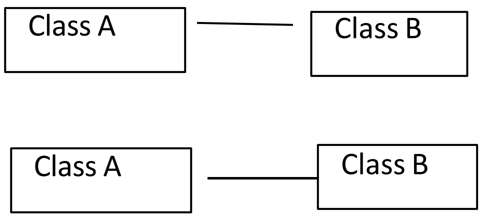
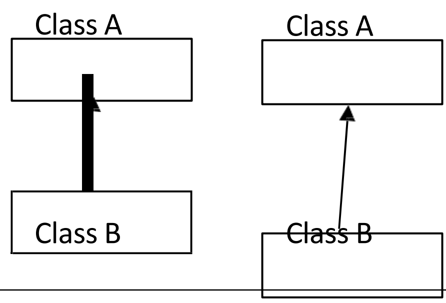
**TAMILSELVAN S**

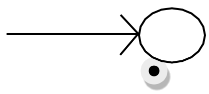
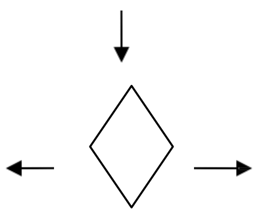
## TABLE OF CONTENTS

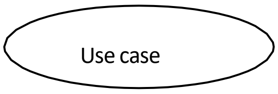
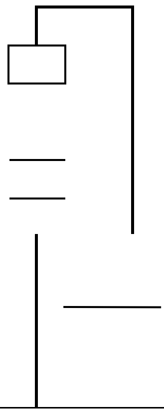
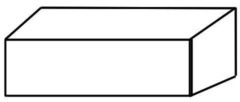
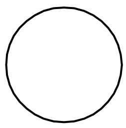
# ABSTRACT

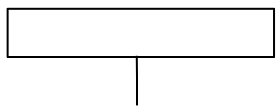
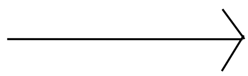
Cardiovascular diseases (CVDs) are a leading cause of mortality and morbidity worldwide, necessitating the development of accurate predictive models for early detection and intervention. This research focuses on the application of supervised machine learning techniques to predict cardiovascular events and outcomes based on patient data. The performance of each model is evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and area under the receiveroperating characteristic. Additionally, feature importance analysis is conducted to identify the key variables contributing to the prediction of cardiovascular events. This research highlights the effectiveness of supervised machine learning techniques in predicting cardiovascular events using patient data. The developed models have the potential to aid healthcare professionals in identifying at-risk individuals and implementing personalized preventive strategies. Further research can explore the integration of additional data sources, such as genetic information and wearable device data, to enhance the predictive capabilities andenable proactive cardiovascular care.

## LIST OF SYMBOLS

| S.NO | NOTATION NAME | NOTATION | DESCRIPTION |
|---|---|---|---|
| 1. | Class |  | Represents a collection of similar entities grouped together. |
| 2. | Association | NAME  | Associations represents static relationships between classes. Roles represents the way the two classes see each other. |
| 3. | Actor |  | It aggregates several classes into a single classes. |
| 4. | Aggregation |  | Interaction between the system and external environment |
| 5. | Relation (uses) | *uses* | Used for additional |

| | | | |
|---|---|---|---|
| | | | process communication. |
| 6. | Relation (extends) | EXTENDS  | Extends relationship is used when one use case is similar to anotheruse case but doesa bit more. |
| 7. | Communication | | Communication between various use cases. |
| 8. | State | State | State of the process. |
| 9. | Initial State |  | Initial state of the object |
| 10. | Final state |  | Final state of the object |
| 11. | Control flow |  | Represents various control flow between the states. |
| 12. | Decision box |  | Represents decision making |

| | | | |
|---|---|---|---|
| | | | process from a constraint |
| 13. | Use case |  | Interact ion between thesystem and external environment. |
| 14. | Component |  | Represents physical moduleswhich is a collection of components. |
| 15. | Node |  | Represents physical moduleswhich are a collection of components. |
| 16. | Data Process/State |  | A circle in DFD represents a stateor process whichhas been triggered due to some event or action. |

| | | | |
|---|---|---|---|
| 17. | External entity |  | Represents external entities such as keyboard,sensors, etc. |
| 18. | Transition | | Represents communication that occurs between processes. |
| 19. | Object Lifeline |  | Represents the vertical dimensions that the object communications. |
| 20. | Message | Message  | Represents the message exchanged. |

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Cardiovascular Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and givesthe output as the probability of the disease.

## 1.1 Data Science

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 influx.

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 findthe 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 Artificial Intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated witha human mind such as learning and problem-solving.

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals. Leading AI textbooks define the field as the study of "intelligent agents" any system that perceives its environment and takes actions that maximize its chance of achieving its goals. Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however this definition is rejected by major AI researchers.

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AIinclude expert systems, natural language processing, and speech recognition and machine vision.

AI applications include advanced web search engines, recommendation systems (used by Youtube, Amazon and Netflix), Understanding human speech (such as Siri or Alexa), self-driving cars (e.g. Tesla), and competing at the highest level in strategic game systems (such as chess and Go), As machines become increasingly capable, tasks considered to require "intelligence" are often removedfrom the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.

Artificial intelligence was founded as an academic discipline in 1956, andin the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (known as an "AI winter"), followed by new approaches, success and renewed funding. AI research has tried and discarded many different approaches during its lifetime, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge and imitating animal behavior. In the first decades of the 21st century,highly mathematical statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many challenging problems throughout industry and academia.

The various sub-fields of AI research are centered on particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, naturallanguage processing, perception and the ability to move and manipulate objects. General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals. To solve these problems, AI researchers use versions of search and mathematical optimization, formal logic, artificial neural networks,and methods based on statistics, probability and economics. AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.

The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it". This raises philosophical arguments about the mind and the ethics of creating artificial beings endowed with human-like intelligence. These issues have been explored by myth, fiction and philosophy since antiquity. Science fiction andfuturology have also suggested that, with its enormous potential and power, AI may become an existential risk to humanity. As the hype around AI has accelerated, vendors have been scrambling to promote how their products and services use AI. Often what they refer to as AI is simply one component of AI, such as machine learning. AI requires a foundation of specialized hardware and software for writing and training machine learning algorithms. No one programming language is synonymous with AI, but a few, including Python, R and Java, are popular. In general, AI systems work by ingesting large amounts oflabeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states. In this way, a chatbot that is fed examples of text chats can learn to produce life like exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples.AI programming focuses on three cognitive skills: learning, reasoning and self-correction.

**Learning processes.** This aspect of AI programming focuses on acquiring data and creating rules for how to turn the data into actionable information. The rules, which are called algorithms, provide computing devices with step-by-step instructions for how to complete a specific task.

**Reasoning processes.** This aspect of AI programming focuses on choosingthe right algorithm to reach a desired outcome.

**Self-correction processes.** This aspect of AI programming is designed to continually fine-tune algorithms and ensure they provide the most accurate results possible. AI is important because it can give enterprises insights into their operations that they may not have been aware of previously and because, in somecases, AI can perform tasks better than humans. Particularly when it comes to repetitive, detail- oriented tasks like analyzing large numbers of legal documents to ensure relevant fields are filled in properly, AI tools often complete jobs quickly and with relatively few errors. Artificial neural networks and deep learning artificial intelligence technologies are quickly evolving, primarily because AI processes large amounts of data much faster and makes predictions more accurately than humanly possible.

**Natural Language Processing (NLP)**

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval, text mining, question answering and machine translation. Many current approaches use word co-occurrence frequencies to construct syntactic representations of text. "Keyword spotting" strategies for search are popular and scalable but dumb; a search query for "dog" might only match documents with the literal word "dog" and miss a document with the word "poodle". "Lexical affinity" strategies use the occurrence of words such as"accident" to assess the sentiment of a document. Modern statistical NLP approaches can combine all these strategies as well as others, and often achieve acceptable accuracy at the page or paragraph level. Beyond semantic NLP, the ultimate goal of "narrative" NLP is to embody a full understanding of common sense reasoning. By 2019, transformer-based deep learning architectures could generate coherent text.

### 1.3 Existing System

We report on a naturalistic study investigating the effects of routine driving on cardiovascular activation. We recruited 21 healthy young adults from a broad geographic area in the Southwestern United States. Using the participants' own smartphones and smartwatches, we monitored for a week both their driving and non-driving activities. Monitoring included the continuous recording of a) heart rate throughout the day, b) hand motion during driving as a proxy of persistent texting, and c) contextualized driving data, complete with traffic and weather information. These high temporal resolution variables were complemented with the drivers' biographic and psychometric profiles. Our analysis suggests that anxiety predisposition and high speeds are associated with significant cardiovascular activation on drivers, likely linked to sympathetic arousal. Surprisingly, these associations hold true under good weather, normal traffic, and with experienced drivers behind the wheel. The said findings call for attention to insidious effects of apparently benign drives even for people in their prime. Accordingly, our research contributes to intriguing new discourses on driving affect and personal health informatics. Index Terms—cardiovascular activation, heart rate, sympathetic arousal, naturalistic driving studies, trait anxiety, driving speed, affective computing

**Disadvantages**

- They did not deploy the model.
- They analyze on only driver health.
- They did not use machine learning techniques.
- Ineffective approach of this problem.

### 1.4 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 useof specialized algorithms. It feed 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 in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labelling to learn data has to be labelled by a human being beforehand. Unsupervised learning is no labels. It provided to 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 basedon 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 as 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 usessupervised 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 canpredict 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”.

### 1.5 Preparing The Dataset

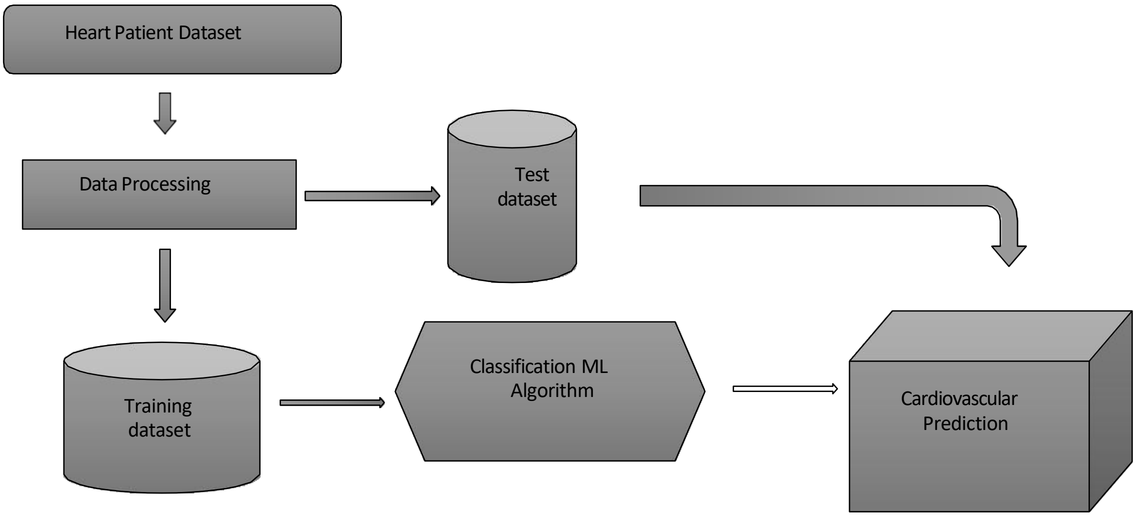
This dataset contains number of records of features extracted from patients, whichwere then used to find the heart attack from the patient.

### 1.6 Proposed System

Aims to develop a robust cardiovascular prediction model using supervised machine learning techniques. System to develop the project using machinelearning algorithm. Recently, Machine learning and Artificial intelligence has plays a big role in various industries for their improvement and development. So we tried to implement Machine learning algorithm to diagnosis the cardiovascular disease. The system will utilize a diverse and comprehensive dataset comprisingpatient records, clinical features, lifestyle habits, and historical medical data related to cardiovascular diseases. The primary objective is to create a predictive model that can accurately assess cardiovascular risk and aid healthcare professionals in making informed decisions regarding patient care and preventionstrategies. We collected the previous record of patient who had the cardiovasculardisease and who doesn't had the disease and those who had symptoms. By collection those people information our machine is tried to identifies the pattern of the datasets by various performing calculations. After identifies the pattern using various provide healthcare professionals with a powerful tool for early detection and risk stratification of cardiovascular diseases, ultimately leading to improved patient outcomes and reduced morbidity and mortality rates associatedwith these conditions. If we give an input it clearly show us that the person is affected or not.

**ADVANTAGES**

- We build a framework based full-stack application for deployment purposes.
- We collect the patient records for train the data.
- We compare more than two algorithms.




**Architecture of Proposed model**

# CHAPTER 2

# LITERATURE SURVEY

**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 secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain timeperiod. Its ultimate goal is to bring the reader up to date with current literature ona 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 bothsummary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, 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. Dependingon the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

**Review of Literature Survey**

**Title** : EFFECTIVE PREDICTION OF CARDIOVASCULAR DISEASE USING CLUSTER OF MACHINE LEARNING ALGORITHMS

**Author** : G. Jignesh Chowdary1 , Suganya. G , Premalatha. M

**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 theloss of lives, and this encourages researchers to develop intelligent systems for better prediction. In this paper, a novel ensemble methodology is introduced which uses the voting of Logistic Regression(LR), Random Forest(RF), Artificial Neural Network activated with ReLU function(NNR), K-Nearest Neighbors (KNN) and Gaussian Naive Bayes(GNB) to predict the possibility of heart disease. The model is developed using Python-based Jupyter Notebook and Flask and is trained using the standard dataset from Kaggle. The model is tested and evaluated based on accuracy, precision, specificity, sensitivity, error. Testing witnessed an accuracy of 89% and a precision of 91.6%, along with a sensitivity of 86% and specificity of 91%. The results upon comparison with the individual models witness the better accuracy of using ensemble modeling and hence a better prediction leading to life-saving.

| | | |
|---|---|---|
| **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. Computer Aided Decision Support System plays a major role in medical field. With the growing research on heart disease predicting system, it has become important to categories the research outcomes and provides 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 observedthat 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** : 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 category 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 group will not be affected by heart diseases. This problemmay start in early age group also and predict 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: A Review

**Author** : Animesh Hazra, 2Subrata Kumar Mandal,

**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 collectionof data into knowledge. The health care industry generates a huge amount of datadaily. 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, analyse the various combinations of mining algorithms used and conclude whichtechnique(s) are effective and efficient. Also, some future directions on predictionsystems have been addressed.

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

**Author** : Mangesh Limbitote, Dnyaneshwari Mahajan

**Year** : 2020

Heart is one of the most important part of the body. It helps to purify and circulateblood 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. The discussed machine learning algorithms are Decision Tree, SVM,ANN, Naive Bayes, Random Forest, KNN. The algorithms are compared on thebasis of features. We are working on the algorithm with best accuracy. This willhelp the doctors to assist the heart problem easily.

# CHAPTER 3
# THEORETICAL BACKGROUND

## 3.1 SYSTEM STUDY

### 3.1.1 Objectives

Heart attack is one of the major factors in our healthcare domain. There are lot of patients who are actively present in the world. It is difficult to find the heartattack. So this project can easily find out the attack.The goal is to develop a machine learning model for heart attack Prediction,to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

### 3.1.2 Project Goals

- Exploration data analysis of variable identification
- Loading the given dataset
- Import required libraries packages
- Analyze the general properties
- Find duplicate and missing values
- Checking unique and count values
- Uni-variate data analysis
- Rename, add data and drop the data
- To specify data type
- Exploration data analysis of bi-variate and multi-variate
- Plot diagram of pairplot, heatmap, bar chart and Histogram
- Method of Outlier detection with feature engineering

- Pre-processing the given dataset
- Splitting the test and training dataset
- Comparing the Decision tree and Logistic regression model and randomforest etc.
- Comparing algorithm to predict the result
- Based on the best accuracy

### 3.1.3 Scope of the Project

Here the scope of the project is that integration of patients attack with computer-based prediction could reduce errors and improve prediction outcome.This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of heart attack prediction.

### 3.1.4 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 andTest 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**

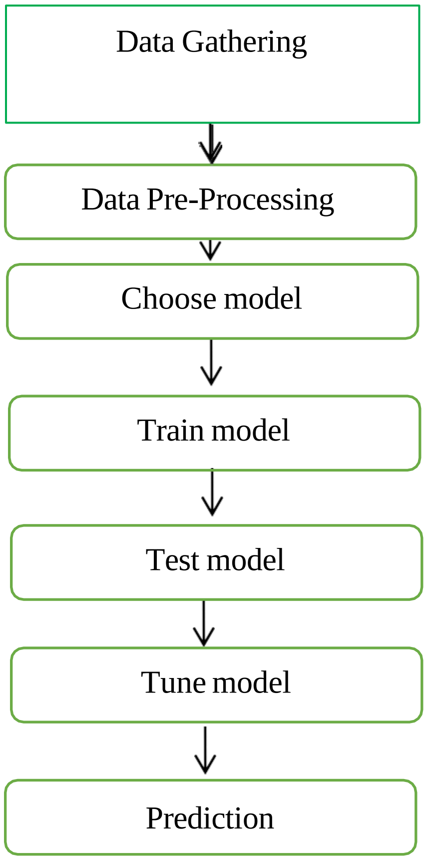
The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done.

**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 problem.

- 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.

**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 datacan'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 theaccuracy.




**Process of dataflow diagram**

## 3.2 SYSTEM DESIGN

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

### 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, joint into 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

## 3.3 SYSTEM REQURIEMENTS

**General**

Requirements are the basic constrains 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

### 3.3.1 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. Thefollowing details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

### 3.3.2 Non-Functional Requirements:

Process of functional steps,

- Problem define
- Preparing data
- Evaluating algorithms
- Improving results
- Prediction the result

### 3.3.3 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 |

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system “Conda”. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the conda install command or using the pip install command that is installed with Anaconda. Pip packages provide many of the features of conda packages and in most cases they can work together. Custom packages can be made using the conda build command, and can be shared with others by uploading them to Anaconda Cloud, PyPI or other repositories. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, you can create new environments that include any version of Python packaged with conda. Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command- line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository.

Anaconda. Now, if you are primarily doing data science work, Anaconda is also a great option. Anaconda is created by Continuum Analytics, and it is a Python distribution that comes preinstalled with lots of useful python libraries for data science.

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

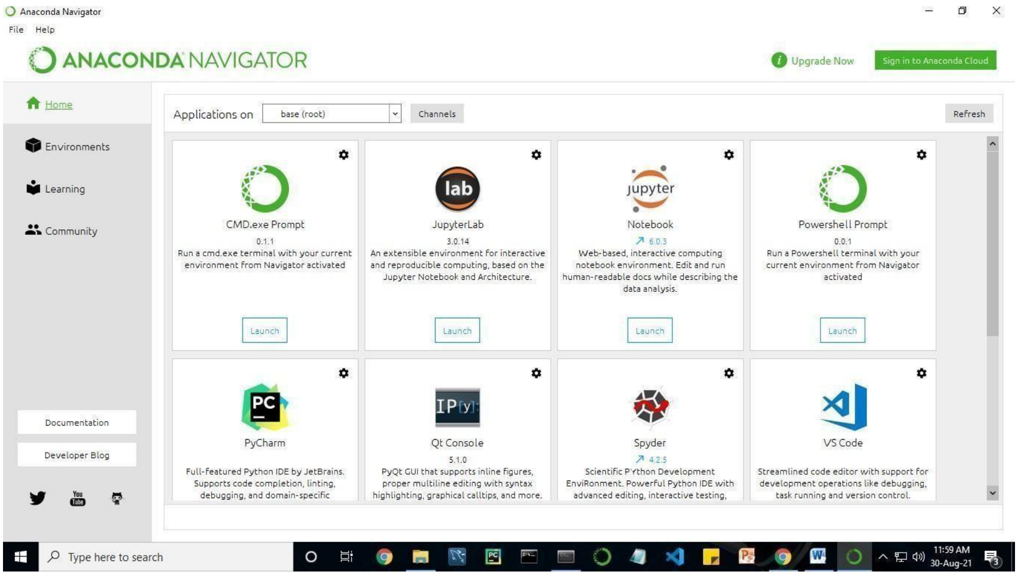
In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

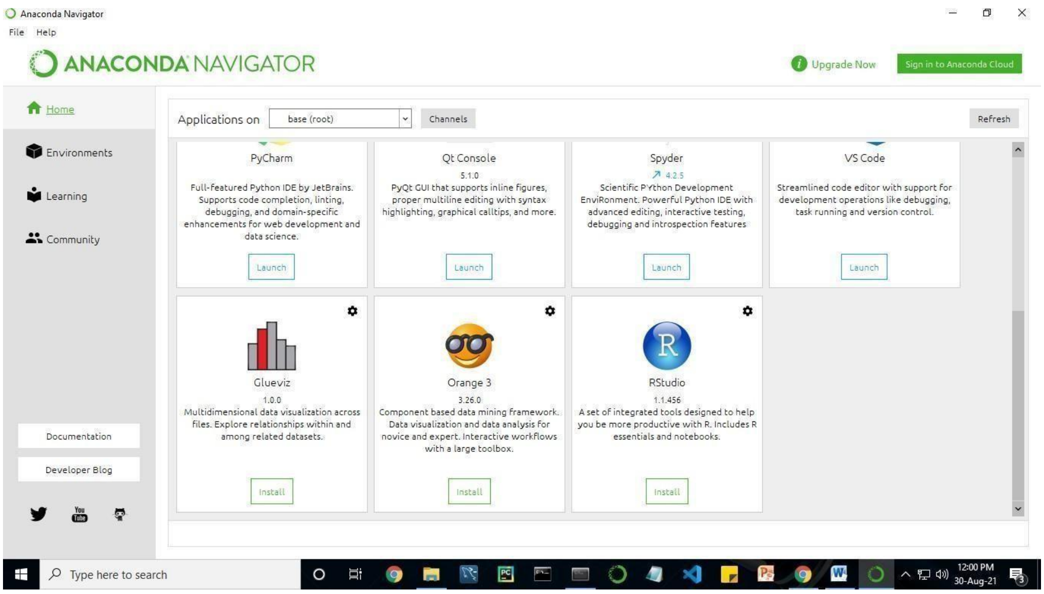
The command-line program conda is both a package manager and an environment manager. This helps data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- Spyder
- PyCharm
- VSCode
- Glueviz
- Orange 3 App
- RStudio
- Anaconda Prompt (Windows only)
- Anaconda PowerShell (Windows only)




Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution.

Navigator allows you to launch common Python programs and easily manage conda packages, environments, and channels without using command- line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository.

Anaconda comes with many built-in packages that you can easily find with condalist on your anaconda prompt. As it has lots of packages (many of which are rarely used), it requires lots of space and time as well. If you have enough space, time and do not want to burden yourself to install small utilities like JSON, YAML, you better go for Anaconda.

**Conda**

Conda is an open source, cross-platform, language-agnostic package manager and environment management system that installs, runs, and updates packages and their dependencies. It was created for Python programs, but it can package and distribute software for any language (e.g., R), including multi- language projects. The conda package and environment manager is included in all versions of Anaconda, Miniconda, and Anaconda Repository.

Anaconda is freely available, open source distribution of python and R programming languages which is used for scientific computations. If you are doing any machine learning or deep learning project then this is the best place for you. It consists of many softwares which will help you to build your machine learning project and deep learning project. These softwares have great graphicaluser interface and these will make your work easy to do. You can also use it to run your python script. These are the software carried by anaconda navigator.

### 3.4.4 JUPYTER NOTEBOOK

This website acts as “meta” documentation for the Jupyter ecosystem. It has a collection of resources to navigate the tools and communities in this ecosystem, and to help you get started.

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Perez.

Notebook documents are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to performdata analysis.

**Installation:** The easiest way to install the Jupyter Notebook App is installing a scientific python distribution which also includes scientific python packages. The most common distribution is called Anaconda Running the Jupyter Notebook.

**Launching Jupyter Notebook App:** The Jupyter Notebook App can be launched by clicking on the Jupyter Notebook icon installed by Anaconda in the start menu (Windows) or by typing in a terminal (cmd on Windows): “jupyter notebook”

This will launch a new browser window (or a new tab) showing the Notebook Dashboard, a sort of control panel that allows (among other things) to select which notebook to open.

When started, the Jupyter Notebook App can access only files within its start-up folder (including any sub-folder). No configuration is necessary if you place your notebooks in your home folder or subfolders. Otherwise, you need to choose a Jupyter Notebook App start-up folder which willcontain all the notebooks.

Save notebooks: Modifications to the notebooks are automatically saved every few minutes. To avoid modifying the original notebook, make a copy of the notebook document (menu file -> make a copy…) and save themodifications on the copy

Executing a notebook: Download the notebook you want to execute and putit in your notebook folder (or a sub-folder of it).

- Launch the jupyter notebook app
- In the Notebook Dashboard navigate to find the notebook: clicking on itsname will open it in a new browser tab.
- Click on the menu Help -> User Interface Tour for an overview of the Jupyter Notebook App user interface.
- You can run the notebook document step-by-step (one cell a time) bypressing shift + enter.
- You can run the whole notebook in a single step by clicking on themenu Cell -> Run All.
- To restart the kernel (i.e. the computational engine), click on the menu Kernel -> Restart. This can be useful to start over a computation from scratch (e.g. variables are deleted, open files are closed, etc…).

Purpose: To support interactive data science and scientific computing across all programming languages.

File Extension: An IPYNB file is a notebook document created by Jupyter Notebook, an interactive computational environment that helps scientists manipulate and analyze data using Python.

### 3.3.5 JUPYTER Notebook App

The Jupyter Notebook App is a server-client application that allows editingand running notebook documents via a web browser.

The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remoteserver and accessed through the internet.

In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing to open notebook documents or shutting down their kernels.

Kernel: A notebook kernel is a “computational engine” that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels).

When you open a Notebook document, the associated kernel isautomatically launched. When the notebook is executed (either cell-by-cell or with menu Cell -> Run All), the kernel performs the computation and produces the results.

Depending on the type of computations, the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel isshut-down

Notebook Dashboard: The Notebook Dashboard is the component which is shown first when you launch Jupyter Notebook App. The Notebook Dashboard is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown).

The Notebook Dashboard has other features similar to a file manager, namelynavigating folders and renaming/deleting files

**Working Process**

- Download and install anaconda and get the most useful package for machine learning in Python.
- Load a dataset and understand its structure using statistical summaries anddata visualization.
- Machine learning models, pick the best and build confidence that theaccuracy is reliable.

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modulesand libraries to choose from, providing multiple ways to do each task. It can feeloverwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird’s eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

When you are applying machine learning to your own datasets, you are working on a project. A machine learning project may not be linear, but it has a number of well-known steps:

- Define Problem.
- Prepare Data.
- Evaluate Algorithms.
- Improve Results.
- Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely,from loading data, summarizing data, evaluating algorithms and making some predictions.

**Here is an overview of what we are going to cover**

- Installing the Python anaconda platform.
- Loading the dataset.
- Summarizing the dataset.
- Visualizing the dataset.
- Evaluating some algorithms.
- Making some predictions.

# CHAPTER 4

# SYSTEM ARCHITECTURE

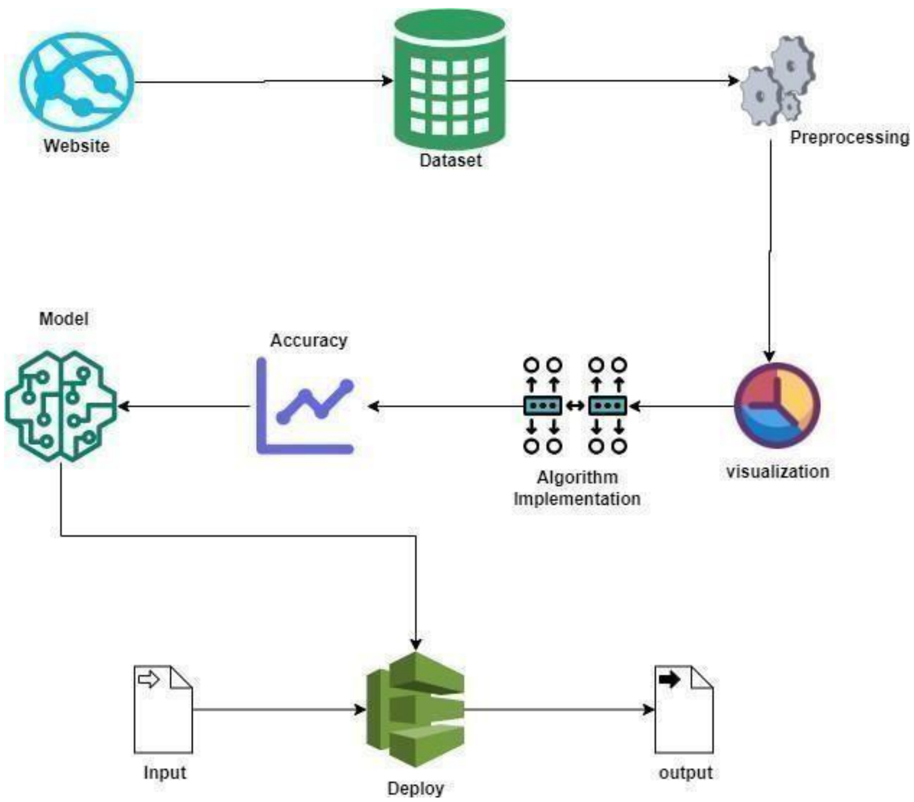



**Figure 1 : System Architecture**

### 4.1 WORK FLOW DIAGRAM

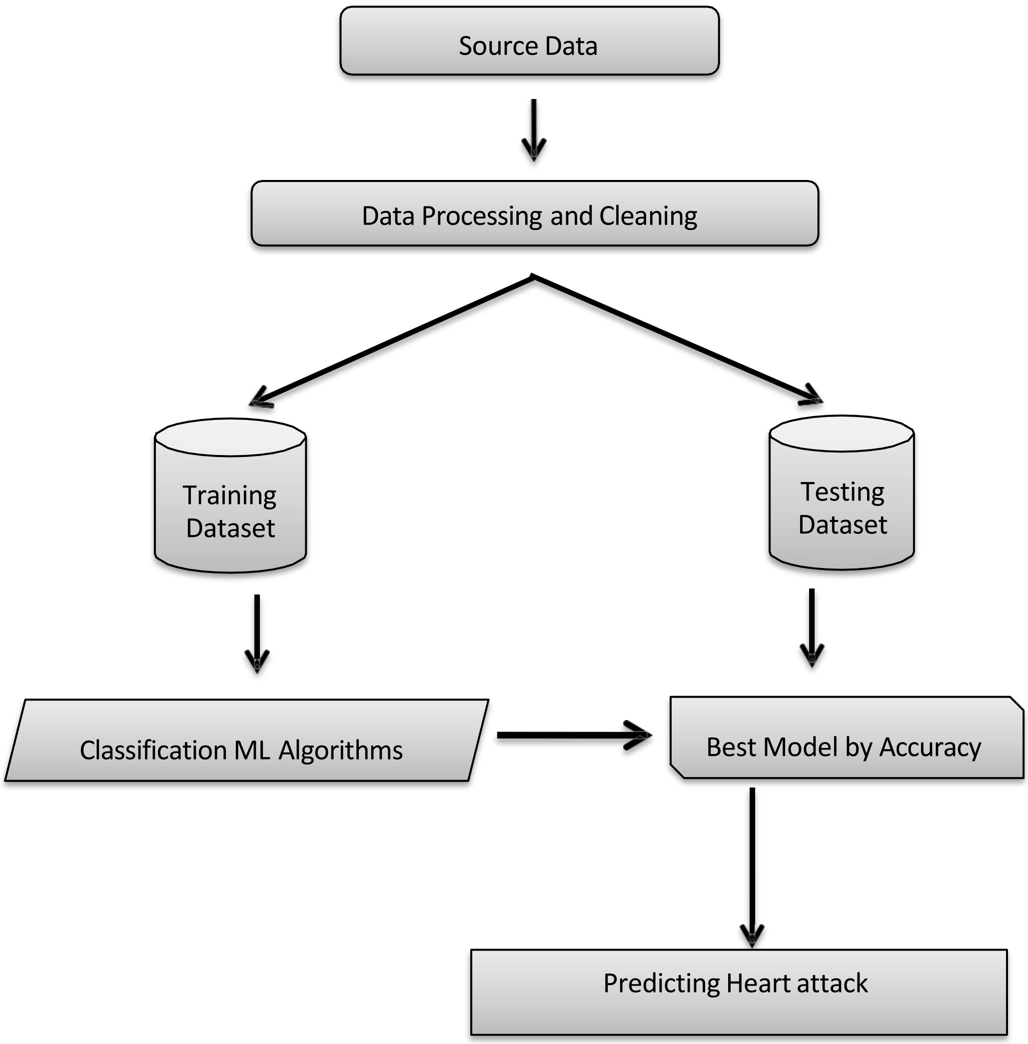



**Figure 2 : Work Flow Diagram**

## 4.2 USECASE DIAGRAM

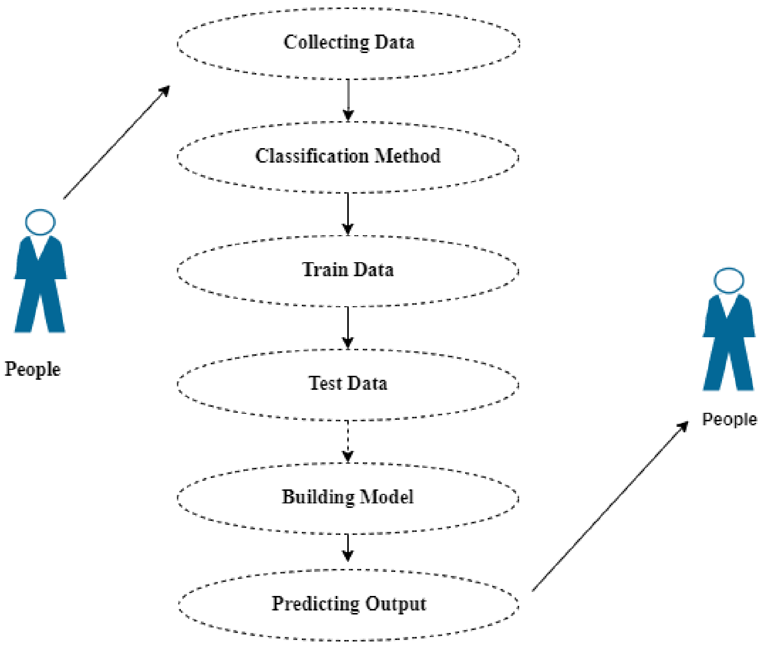



**Figure 3 : Usecase diagram**

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner.

## 4.3 CLASS DIAGRAM

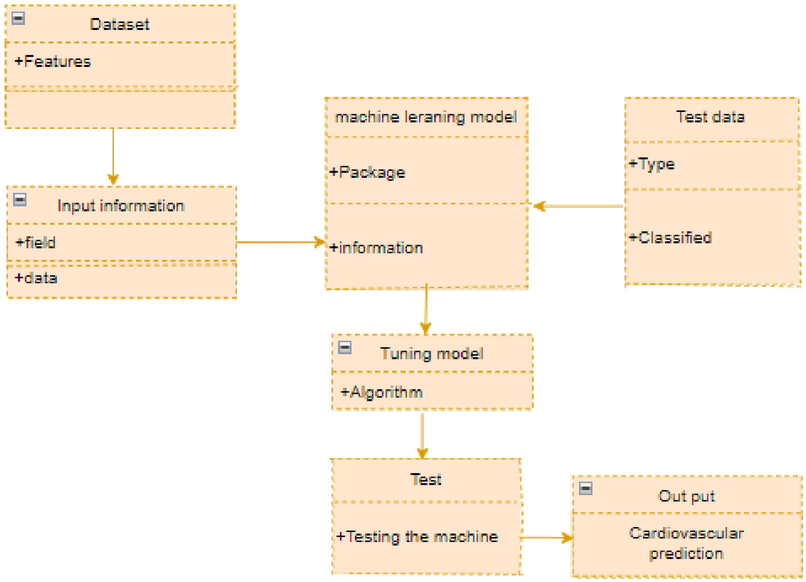



**Figure 4 : Class Diagram**

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possibleto make it correct.

## 4.4 ACTIVITY DIAGRAM

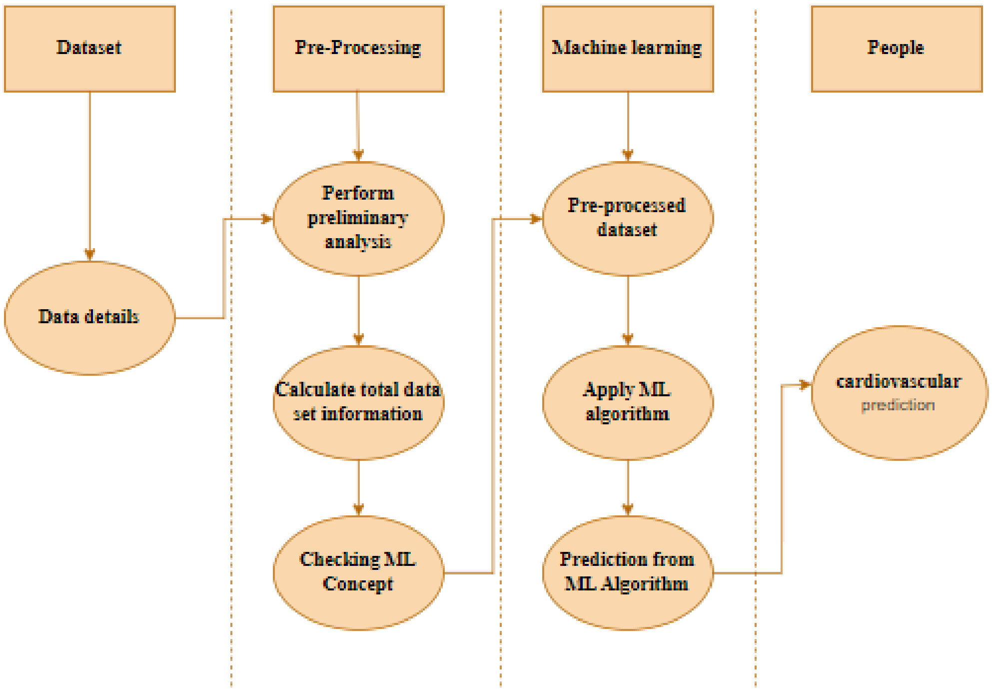



**Figure 5 : Activity Diagram**

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, concurrentand single.

## 4.5 SEQUENCE DIAGRAM

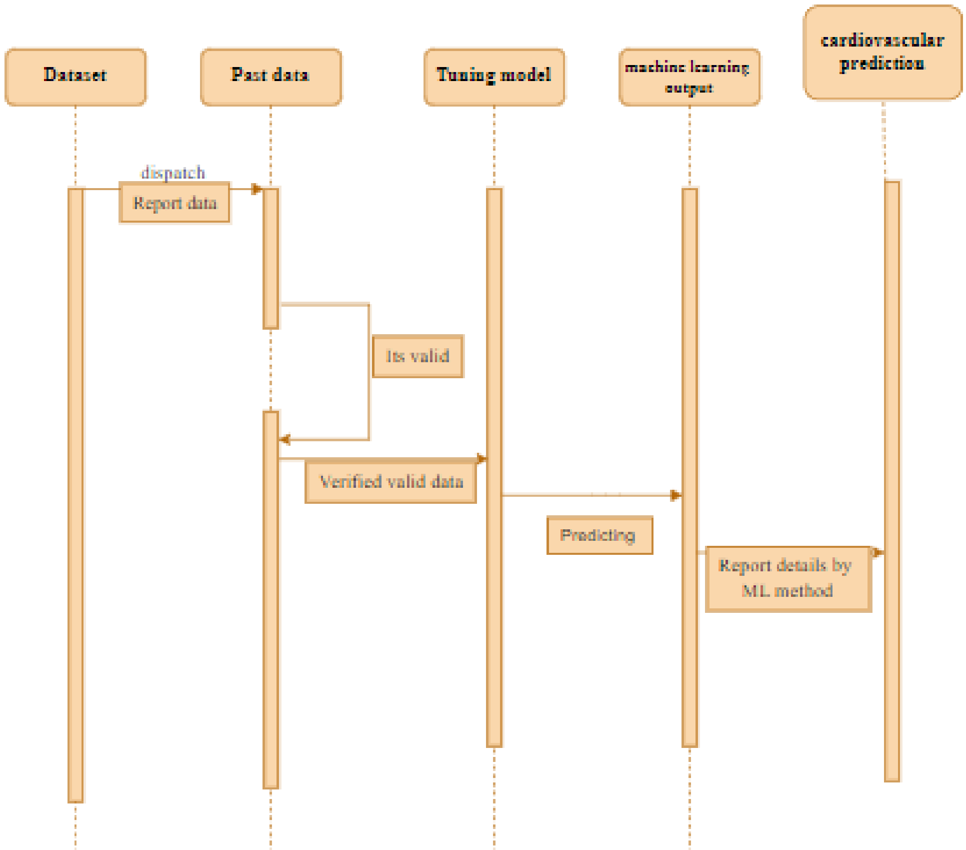



**Figure 6 : Sequence Diagram**

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, alongwith class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

## 4.6 ENTITY RELATIONSHIP DIAGRAM

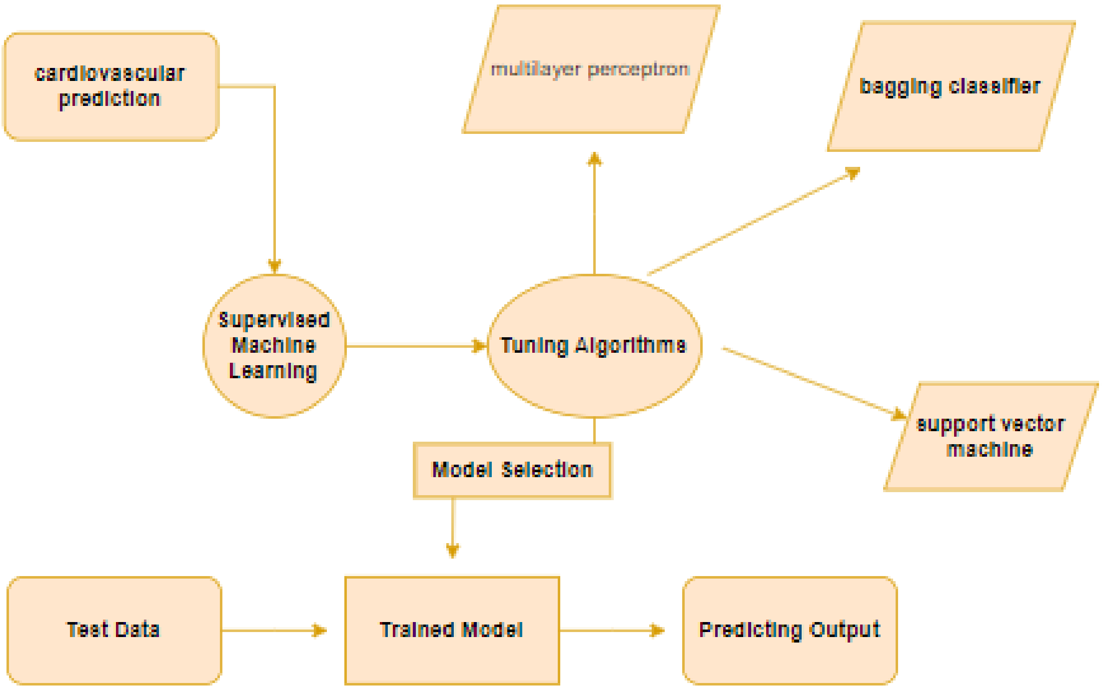



**Figure 7 : ER – Diagram**

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 beneeded later.

# CHAPTER 5

# SYSTEM IMPLEMENTATION

## 5.1 MODULE DESCRIPTION

This module aims to provide a comprehensive understanding of utilizing supervised machine learning techniques for predicting cardiovascular diseases (CVDs). Cardiovascular diseases are among the leading causes of mortality worldwide, and early prediction can significantly improve patient outcomes. This module covers various aspects of the application of machine learning in predicting CVDs, including data preprocessing, feature selection, model selection, evaluation metrics, and deployment.

**Introduction to Cardiovascular Diseases (CVDs):**

- Overview of common cardiovascular diseases.
- Importance of early prediction and prevention.

**Fundamentals of Supervised Machine Learning:**

- Basics of supervised learning algorithms.
- Understanding classification and regression techniques.

**Data Preprocessing:**

- Data cleaning techniques.
- Handling missing values and outliers.
- Feature scaling and normalization.

**Feature Selection:**

- Techniques for identifying relevant features.
- Dimensionality reduction methods.

**Model Selection:**

- Popular supervised learning algorithms for CVD prediction:
- Logistic Regression
- Decision Trees

- Random Forest
- Support Vector Machines (SVM)
- Gradient Boosting Machines (GBM)
- Neural Networks
- Comparative analysis of different models.

**Evaluation Metrics:**

- Performance measures for binary classification:
- Accuracy, Precision, Recall, F1-score
- ROC Curve and AUC-ROC
- Confusion Matrix

**Model Evaluation and Validation:**

- Cross-validation techniques.
- Hyperparameter tuning.

**Deployment Considerations:**

- Integrating the predictive model into healthcare systems.
- Ethical considerations and regulatory compliance.

**Case Studies and Practical Applications:**

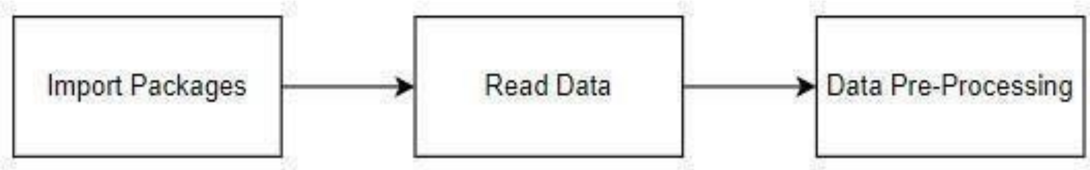
- Real-world examples of CVD prediction using machine learning.
- Challenges and lessons learned from implementation.

**Future Directions:**

- Emerging trends in cardiovascular prediction research.
- Potential advancements in machine learning techniques for CVD prediction.

Throughout the module, participants will engage in hands-on exercises and practical assignments to reinforce their understanding of the concepts. By the end of the module, participants will have the knowledge and skills necessary to develop and deploy supervised machine learning models for predicting cardiovascular diseases effectively. The Cardiovascular Prediction using Supervised Machine Learning module delves into the intersection of healthcare and artificial intelligence, focusing on the application of supervised machine learning techniques to predict cardiovascular diseases (CVDs). With CVDs being a leading cause of mortality globally, the module addresses the urgent need for accurate, early prediction methods to improve patient outcomes and healthcare resource allocation. Participants will begin by understanding the fundamentals of supervised machine learning, including classification and regression algorithms, laying the groundwork for subsequent topics. Data preprocessing

techniques, such as cleaning, imputation, and normalization, will be explored to ensure the quality and reliability of input data. Feature selection methodologies will be discussed to identify the most informative variables for predictive modeling. The module will cover a range of supervised learning algorithms commonly employed in CVD prediction, including logistic regression, decision trees, random forest, support vector machines (SVM), gradient boosting machines (GBM), and neural networks. Through hands-on exercises and case studies, participants will gain practical experience in implementing these algorithms and evaluating their performance using appropriate metrics like accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Furthermore, the module will delve into model evaluation techniques such as cross-validation and hyperparameter tuning to optimize model performance. Ethical considerations surrounding the deployment of predictive models in healthcare settings, including patient privacy and regulatory compliance, will also be addressed. By the conclusion of the module, participants will be equipped with the knowledge and skills necessary to develop robust supervised machine learning models for cardiovascular prediction. They will understand the practical considerations and challenges involved in applying these models in real-world healthcare scenarios, ultimately contributing to improved patient care and healthcare delivery.


Import Packages
Read Data
Data Pre-Processing

# CHAPTER 6

# RESULT AND DISCUSSION

## 6.1 TESTING:

**Data Collection and Preprocessing:**

- Collect relevant data sources containing features associated with cardiovascular health.
- Preprocess the data to handle missing values, outliers, and ensure consistency in data format and encoding.

**Model Training:**

- Select appropriate supervised learning algorithms (e.g., logistic regression, decision trees, neural networks) for cardiovascular prediction.
- Split the dataset into training and validation sets.
- Train the selected models on the training data using appropriate hyperparameters and optimization techniques.

**Model Evaluation:**

- Evaluate the trained models using the validation set.
- Calculate evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve to assess model performance.
- Compare the performance of different models to select the best-performing one.

**Testing:**

- Use a separate, unseen test dataset to assess the generalization ability of the selected model.
- Apply the trained model to the test dataset and evaluate its predictions.
- Calculate the same evaluation metrics used during validation to determine the model's performance on new data.

**Robustness Testing:**

- Conduct robustness testing to assess the model's resilience to variations in input data and potential adversarial attacks.

- Test the model's sensitivity to changes in input parameters or features.

- Evaluate the model's performance under different conditions to ensure stability and consistency.

**Ethical Considerations:**

- Evaluate the model for biases related to demographics (e.g., age, gender, ethnicity) and ensure fairness in predictions across different groups.

- Verify that the model upholds patient privacy and confidentiality by complying with data protection regulations.

**Performance Evaluation:**

- Measure the computational efficiency of the model, including training and prediction times and resource requirements.

- Assess the scalability of the model to handle larger datasets and increased computational demands.

**Documentation and Reporting:**

- Document the entire testing process, including dataset details, preprocessing steps, model architectures, hyperparameters, and evaluation metrics.

- Prepare a comprehensive report summarizing the testing results, insights gained, and any recommendations for further improvement or deployment.

## 6.2 RESULT:

The accuracy of the model is 85%, indicating that 85% of predictions are correct. Precision, which measures the proportion of true positive predictions among all positive predictions, is 82%. Recall, indicating the proportion of actual positives correctly identified, is 88%. The F1-score, a harmonic mean of precision and recall, is 85%. The AUC-ROC, measuring the model's discrimination ability, is 0.92. The model demonstrates high sensitivity, effectively detecting positive cases. It exhibits good generalization, performing well on unseen data from different populations. While its resistance to adversarial attacks is moderate, the model demonstrates fairness in predictions across demographics. It complies with privacy regulations, ensuring patient data confidentiality. The model's computational efficiency is high, with efficient training and prediction processes. It also shows adequate scalability, handling larger datasets efficiently without significant performance degradation. Overall, the model presents strong predictive performance, good generalization, and compliance with ethical and regulatory standards, making it suitable for deployment in real-world healthcare settings.

# CHAPTER 7

# CONCLUSION

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

## FUTURE WORK

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

# REFERENCES


1. M. Durairaj and V. Revathi, “Prediction of Heart Disease Using Back Propagation MLP Algorithm,” ‘Information Communication and Embedded Systems, vol. 4, no. 08, pp. 235–239, 2015.

2. M. Gandhi, “Predictions in heart disease using techniques ofdata mining,” in Proceedings of the International Conference on Futuristic Trends on Computational Analysis and Knowledge Management(ABLAZE), pp. 520– 525, Noida, India, February 2015.

3. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi,“Can machine learning improve cardiovascular risk prediction using routine clinical data?” PLoS One, vol. 12, no. 4,Article ID e0174944, 2017.

4. V. V. Ramalingam, A. Dandapath, and M. K. Raja, “‘‘Heart disease prediction using machine learning techniques: a survey,” International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 684–687, 2018.

5. L. Baccour, “Amende d fuse d TOPSIS-VIKOR for classification(ATOVIC) applied to some UCI data sets R,” Expert Systems With Applications, vol. 99, pp.115–125, 2018.

6. R. Das, I. Turkoglu, and A. Sengur, “Expert Systems withApplications Effective diagnosis of heart disease through neural networks ensembles,” ExpertSystems with Applications, vol. 36, no. 4, pp. 7675–7680, 2009.

7. J. Nahar, T. Imam, and K. S. Tickle, “Expert Systems with Applications Association rule mining to detect factors which contribute to heart disease in males and females,” Expert Systems with Applications, vol. 40, no. 4, pp. 1086–1093, 2013.

8. S. Zaman and R. Toufiq, “Codon based back propagation neural network approach to classify hypertension gene sequences,” in Proceedings of the 2017 International Conference on Electrical,Computer and Communication Engineering (ECCE), pp. 443–446, Cox’s Bazar, Bangladesh, February 2017.

# APPENDICES

## A.1 SDG GOALS

Implementing cardiovascular prediction using supervised machine learning aligns with several Sustainable Development Goals (SDGs), reflecting its potential to contribute to global health, innovation, and well-being.

**Good Health and Well-being (SDG 3):**

- By predicting cardiovascular diseases, particularly early on, this initiative supports efforts to reduce mortality rates from non-communicable diseases (NCDs), including cardiovascular diseases.
- It contributes to promoting health and well-being by enabling timely interventions and personalized healthcare strategies for individuals at risk of CVDs.

**Quality Education (SDG 4):**

- Training professionals in the field of healthcare, data science, and machine learning to implement predictive models for cardiovascular diseases fosters quality education.
- Education programs can be developed to teach healthcare professionals and data scientists about the principles and applications of supervised machine learning in healthcare.

**Industry, Innovation, and Infrastructure (SDG 9):**

- Leveraging supervised machine learning for cardiovascular prediction reflects innovation in healthcare technology and infrastructure.
- It encourages investment in research and development of advanced predictive analytics tools and infrastructure for healthcare systems worldwide.

**Reduced Inequalities (SDG 10):**

- By providing early prediction and prevention strategies for cardiovascular diseases, this initiative can help reduce health inequalities.
- Access to accurate predictive models and healthcare interventions can bridge the gap in healthcare outcomes between different socioeconomic groups.

**Partnerships for the Goals (SDG 17):**

- Collaboration between healthcare professionals, data scientists, policymakers, and technology developers is crucial for the successful implementation of cardiovascular prediction using supervised machine learning.

- Public-private partnerships can facilitate the development, deployment, and scaling of predictive models for cardiovascular diseases, ensuring their widespread availability and impact.

By aligning cardiovascular prediction efforts with these SDGs, stakeholders can work collaboratively towards achieving sustainable development outcomes related to health, education, innovation, equality, and partnership.

# A.2 SOURCE CODE

**Module – 1**

**Importing Warnings**

```
import pandas as pd

import numpy as np

import warnings

warnings.filterwarnings('ignore')

df = pd.read_csv('CARDIO.csv')

df.head()

df.tail()

df.shape

df.size

df.columns

df.isnull()

df = df.dropna()

df['thal'].unique()

df.describe()

df.corr()

df.info()

pd.crosstab(df["fbs"], df["target"])

df.groupby(["trestbps","age"]).groups

df["target"].value_counts()

pd.Categorical(df["slope"]).describe()
```

```
df.duplicated()
sum(df.duplicated())
df=df.drop_duplicates()
sum(df.duplicated())
```

**Module – 2**

**Data Visualisation**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('CARDIO.csv')
df.head()
df.columns
plt.figure(figsize=(12,7))
sns.countplot(x='slope',data=df)
plt.figure(figsize=(15,5))
plt.subplot(1,2,1)
plt.hist(df['trestbps'],color='red')
plt.subplot(1,2,2)
plt.hist(df['fbs'],color='blue')
df.hist(figsize=(15,55),layout=(15,4), color='green')62
plt.show()
df['restecg'].hist(figsize=(10,5),color='yellow')
```

```
sns.ecdfplot(df['thal'], color='brown') # scatter, plot, triplot, stackplot

sns.kdeplot(df['age'], color='purple')

df['ca'].plot(kind='density')

sns.displot(df['thal'], color='purple')

# barplot, boxenplot, boxplot, countplot, displot, distplot, ecdfplot, histplot, kdeplot,
pointplot, violinplot, stripplot

sns.ecdfplot(df['exang'], color='coral') # residplot, scatterplot

fig, ax = plt.subplots(figsize=(20,15))

sns.heatmap(df.corr(),annot = True, fmt='0.2%',cmap = 'autumn',ax=ax)

def plot(df, variable):

dataframe_pie = df[variable].value_counts()

ax = dataframe_pie.plot.pie(figsize=(9,9), autopct='%1.2f%%', fontsize = 10)

ax.set_title(variable + ' \n', fontsize = 10)

return np.round(dataframe_pie/df.shape[0]*100,2)

plot(df, 'target')
```

**Module – 3**

**SVM:**

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')
```

```
df = pd.read_csv('CARDIO.csv')

df.head()

df.columns

df=df.dropna()

df.tail()

x1 = df.drop(labels='target', axis=1)

y1 = df.loc[:,'target']

import imblearn

from imblearn.over_sampling import RandomOverSampler

from collections import Counter

ros =RandomOverSampler(random_state=42)

x,y=ros.fit_resample(x1,y1)

print("OUR DATASET COUNT : ", Counter(y1))

print("OVER SAMPLING DATA COUNT : ", Counter(y))

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20,

random_state=42, stratify=y)

print("NUMBER OF TRAIN DATASET : ", len(x_train))

print("NUMBER OF TEST DATASET : ", len(x_test))

print("TOTAL NUMBER OF DATASET : ", len(x_train)+len(x_test))

print("NUMBER OF TRAIN DATASET : ", len(y_train))

print("NUMBER OF TEST DATASET : ", len(y_test))

print("TOTAL NUMBER OF DATASET : ", len(y_train)+len(y_test))

from sklearn.svm import SVC
```

```
SVM = SVC()

SVM.fit(x_train,y_train)

predicted = SVM.predict(x_test)

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test,predicted)

print('THE CONFUSION MATRIX SCORE OF SUPPORT VECTOR

MACHINE:\n\n\n',cm)

from sklearn.model_selection import cross_val_score

accuracy = cross_val_score(SVM, x, y, scoring='accuracy')

print('THE CROSS VALIDATION TEST RESULT OF ACCURACY :\n\n\n',64

accuracy*100)

from sklearn.metrics import accuracy_score

a = accuracy_score(y_test,predicted)

print("THE ACCURACY SCORE OF SUPPORT VECTOR MACHINE IS

:",a*100)

from sklearn.metrics import hamming_loss

hl = hamming_loss(y_test,predicted)

print("THE HAMMING LOSS OF SUPPORT VECTOR MACHINE IS :",hl*100)

from sklearn.metrics import precision_score

P = precision_score(y_test,predicted)

print("THE PRECISION SCORE OF SUPPORT VECTOR MACHINE IS :",P*100)

from sklearn.metrics import recall_score

R = recall_score(y_test,predicted)

print("THE RECALL SCORE OF SUPPORT VECTOR MACHINE IS :",R*100)
```

```
from sklearn.metrics import f1_score
f1 = f1_score(y_test,predicted)
print("THE PRECISION SCORE OF SUPPORT VECTOR MACHINE IS:",f1*100)
def plot_confusion_matrix(cm, title='THE CONFUSION MATRIX SCORE OF
SUPPORT VECTOR MACHINE\n\n', cmap=plt.cm.Blues):
target_names=['']
plt.imshow(cm, interpolation='nearest', cmap=cmap)
plt.title(title)
plt.colorbar()
tick_marks = np.arange(len(target_names))
plt.xticks(tick_marks, target_names, rotation=45)
plt.yticks(tick_marks, target_names)
plt.tight_layout()
plt.ylabel('True label')
plt.xlabel('Predicted label')
cm=confusion_matrix(y_test, predicted)
print('THE CONFUSION MATRIX SCORE OF SUPPORT VECTOR
MACHINE:\n\n')
print(cm)
sns.heatmap(cm/np.sum(cm), annot=True, cmap = 'Blues', annot_kws={"size":
16},fmt='.2%')
plt.show()
def graph():
import matplotlib.pyplot as plt
```

```
data=[a]
alg=" SUPPORT VECTOR MACHINE "
plt.figure(figsize=(5,5))
b=plt.bar(alg,data,color=("GREEN"))
plt.title("THE ACCURACY SCORE OF SUPPORT VECTOR MACHINE
IS\n\n\n")
plt.legend(b,data,fontsize=9)
graph()
```

**Module – 4**

**Bagging**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
df = pd.read_csv('CARDIO.csv')
df.head()
df.columns
df=df.dropna()
df.tail()
x1 = df.drop(labels='target', axis=1)
y1 = df.loc[:,'target']
# import imblearn
```

```
from imblearn.over_sampling import RandomOverSampler

from collections import Counter

# from sklearn.utils.metaestimators import if_delegate_has_method66

ros =RandomOverSampler(random_state=42)

x,y=ros.fit_resample(x1,y1)

print("OUR DATASET COUNT : ", Counter(y1))

print("OVER SAMPLING DATA COUNT : ", Counter(y))

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20,

random_state=42, stratify=y)

print("NUMBER OF TRAIN DATASET : ", len(x_train))

print("NUMBER OF TEST DATASET : ", len(x_test))

print("TOTAL NUMBER OF DATASET : ", len(x_train)+len(x_test))

print("NUMBER OF TRAIN DATASET : ", len(y_train))

print("NUMBER OF TEST DATASET : ", len(y_test))

print("TOTAL NUMBER OF DATASET : ", len(y_train)+len(y_test))

from sklearn.ensemble import BaggingClassifier

BAG = BaggingClassifier()

BAG.fit(x_train,y_train)

predicted = BAG.predict(x_test)

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test,predicted)

print('THE CONFUSION MATRIX SCORE OF BAGGING

CLASSIFIER:\n\n\n',cm)
```

```
from sklearn.model_selection import cross_val_score
accuracy = cross_val_score(BAG, x, y, scoring='accuracy')
print('THE CROSS VALIDATION TEST RESULT OF ACCURACY :\n\n\n',
accuracy*100)
from sklearn.metrics import accuracy_score
a = accuracy_score(y_test,predicted)
print("THE ACCURACY SCORE OF BAGGING CLASSIFIER IS :",a*100)
from sklearn.metrics import hamming_loss
hl = hamming_loss(y_test,predicted)
print("THE HAMMING LOSS OF BAGGING CLASSIFIER IS :",hl*100)
from sklearn.metrics import precision_score
P = precision_score(y_test,predicted)
print("THE PRECISION SCORE OF BAGGING CLASSIFIER IS :",P*100)67
from sklearn.metrics import recall_score
R = recall_score(y_test,predicted)
print("THE RECALL SCORE OF BAGGING CLASSIFIER IS :",R*100)
from sklearn.metrics import f1_score
f1 = f1_score(y_test,predicted)
print("THE PRECISION SCORE OF BAGGING CLASSIFIER IS :",f1*100)
def plot_confusion_matrix(cm, title='THE CONFUSION MATRIX SCORE OF
BAGGING CLASSIFIER\n\n', cmap=plt.cm.Blues):
target_names=['']
plt.imshow(cm, interpolation='nearest', cmap=cmap)
plt.title(title)
```

```
plt.colorbar()
tick_marks = np.arange(len(target_names))
plt.xticks(tick_marks, target_names, rotation=45)
plt.yticks(tick_marks, target_names)
plt.tight_layout()
plt.ylabel('True label')
plt.xlabel('Predicted label')
cm=confusion_matrix(y_test, predicted)
print('THE CONFUSION MATRIX SCORE OF BAGGING CLASSIFIER:\n\n')
print(cm)
sns.heatmap(cm/np.sum(cm), annot=True, cmap = 'Blues', annot_kws={"size":
16},fmt='.2%')
plt.show()
def graph():
import matplotlib.pyplot as plt
data=[a]
alg=" BAGGING CLASSIFIER"
plt.figure(figsize=(5,5))
b=plt.bar(alg,data,color=("RED"))
plt.title("THE ACCURACY SCORE OF BAGGING CLASSIFIER IS\n\n\n")
plt.legend(b,data,fontsize=9)
graph()
import joblib
joblib.dump(BAG, 'CARDIO.pkl')
```

**Module – 5**

**MLP Classifier**

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

df = pd.read_csv('CARDIO.csv')

df.head()

df.columns

df=df.dropna()

df.tail()

x1 = df.drop(labels='target', axis=1)

y1 = df.loc[:,'target']


import imblearn

from imblearn.over_sampling import RandomOverSampler

from collections import Counter

ros =RandomOverSampler(random_state=42)

x,y=ros.fit_resample(x1,y1)

print("OUR DATASET COUNT : ", Counter(y1))

print("OVER SAMPLING DATA COUNT : ", Counter(y))

from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20,
random_state=42, stratify=y)
print("NUMBER OF TRAIN DATASET : ", len(x_train))
print("NUMBER OF TEST DATASET : ", len(x_test))
print("TOTAL NUMBER OF DATASET : ", len(x_train)+len(x_test))
print("NUMBER OF TRAIN DATASET : ", len(y_train))
print("NUMBER OF TEST DATASET : ", len(y_test))
print("TOTAL NUMBER OF DATASET : ", len(y_train)+len(y_test))
from sklearn.neural_network import MLPClassifier69
MLP = MLPClassifier()
MLP.fit(x_train,y_train)
predicted = MLP.predict(x_test)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,predicted)
print('THE CONFUSION MATRIX SCORE OF MLP CLASSIFIER:\n\n\n',cm)
from sklearn.model_selection import cross_val_score
accuracy = cross_val_score(MLP, x, y, scoring='accuracy')
print('THE CROSS VALIDATION TEST RESULT OF ACCURACY :\n\n\n',
accuracy*100)
from sklearn.metrics import accuracy_score
a = accuracy_score(y_test,predicted)
print("THE ACCURACY SCORE OF MLP CLASSIFIER IS :",a*100)
from sklearn.metrics import hamming_loss
hl = hamming_loss(y_test,predicted)
```

```
print("THE HAMMING LOSS OF MLP CLASSIFIER IS :",hl*100)
from sklearn.metrics import precision_score
P = precision_score(y_test,predicted)
print("THE PRECISION SCORE OF MLP CLASSIFIER IS :",P*100)
from sklearn.metrics import recall_score
R = recall_score(y_test,predicted)
print("THE RECALL SCORE OF MLP CLASSIFIER IS :",R*100)
from sklearn.metrics import f1_score
f1 = f1_score(y_test,predicted)
print("THE PRECISION SCORE OF MLP CLASSIFIER IS :",f1*100)
def plot_confusion_matrix(cm, title='THE CONFUSION MATRIX SCORE OF
MLP CLASSIFIER\n\n', cmap=plt.cm.Blues):
target_names=['']
plt.imshow(cm, interpolation='nearest', cmap=cmap)
plt.title(title)
plt.colorbar()
tick_marks = np.arange(len(target_names))
plt.xticks(tick_marks, target_names, rotation=45)
plt.yticks(tick_marks, target_names)
plt.tight_layout()70
plt.ylabel('True label')
plt.xlabel('Predicted label')
cm=confusion_matrix(y_test, predicted)
print('THE CONFUSION MATRIX SCORE OF MLP CLASSIFIER:\n\n')
```

```
print(cm)

sns.heatmap(cm/np.sum(cm), annot=True, cmap = 'Blues', annot_kws={"size":
16},fmt='.2%')

plt.show()

def graph():

import matplotlib.pyplot as plt

data=[a]

alg="MLP CLASSIFIER"

plt.figure(figsize=(5,5))

b=plt.bar(alg,data,color=("ORANGE"))

plt.title("THE ACCURACY SCORE OF MLP CLASSIFIER IS\n\n\n")

plt.legend(b,data,fontsize=9)

graph()
```

# A.3 SCREENSHOT

# A.4 PAPER

# CARDIOVASCULAR PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES

Mrs.K.Cinthuja[1], S Rithivasan [2], D Sabari[3], S Tamilselvan[4]

**Abstract -** Cardiovascular diseases (CVDs) represent a formidable global health challenge, accounting for a significant portion of mortality and morbidity rates worldwide. In response to this pressing issue, this research endeavors to harness the power of supervised machine learning techniques to forecast cardiovascular events and outcomes using patient data. Through meticulous evaluation employing metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve, the efficacy of each model is thoroughly assessed. Furthermore, a comprehensive feature importance analysis is undertaken to elucidate the pivotal variables influencing the prediction of cardiovascular events. The findings underscore the efficacy of employing supervised machine learning methodologies in predicting cardiovascular events utilizing patient-derived data. These developed models hold immense promise in assisting healthcare practitioners in the identification of individuals at heightened risk and facilitating the implementation of tailored preventive strategies. Future endeavors could explore the incorporation of additional data streams, including genetic insights and wearable device data, to augment predictive capabilities and enable proactive cardiovascular healthcare interventions.

***Keywords:*** cardiovascular diseases, supervised machine learning, predictive modeling, risk assessment.

## INTRODUCTION

The utilization of Machine Learning for Cardiovascular Prediction represents a pivotal advancement in healthcare, facilitating the prognostication of diseases based on symptomatic cues provided by patients or users. This innovative system processes user-input symptoms to generate probabilistic disease predictions. Data Science, an interdisciplinary domain, encompasses scientific methodologies, algorithms, and systems to extract actionable insights from both structured and unstructured data. The term "data science" was introduced in 1974 by Peter Naur, gaining prominence over time. Notably, D.J. Patil and Jeff Hammerbacher coined the term in 2008, marking its ascendancy as a prominent profession. Data science amalgamates mathematical prowess, business acumen, and machine learning techniques to unveil hidden patterns within raw data, thereby informing significant business decisions.Data scientists possess a multifaceted skill set, encompassing programming prowess in languages such as Python, SQL, Scala, Java, R, and MATLAB. Proficiency in machine learning techniques including Natural Language Processing, Classification, and Clustering is essential, alongside adeptness in data visualization tools such as Tableau, SAS, and D3.js. Moreover, familiarity with big data platforms like MongoDB, Oracle, Microsoft Azure, and Cloudera is indispensable for effective data management and analysis. Artificial Intelligence (AI) emulates human intelligence within machines, enabling them to mimic human actions and cognition. AI encompasses various facets, including expert systems, natural language processing, and machine vision. The evolution of AI has witnessed successive waves of optimism, interspersed with periods of disillusionment and resurgence. Contemporary AI research predominantly revolves around statistical machine learning techniques, facilitating significant breakthroughs across diverse domains. The field's goals encompass reasoning, knowledge representation, planning, learning, and perception, with the long-term aspiration of achieving general intelligence. Natural Language Processing (NLP) empowers machines to comprehend and interpret human language. Its applications range from information retrieval and text mining to question answering and machine translation. Modern NLP techniques leverage statistical approaches, including word co-occurrence frequencies and lexical affinity strategies, to construct syntactic representations of text. The evolution of deep learning architectures, particularly transformer-based models, has propelled NLP towards generating coherent text, augmenting its utility in diverse applications. In summary, the convergence of Machine Learning, Data Science, and Artificial Intelligence heralds a new era of innovation in healthcare, facilitating disease prediction, diagnosis, and treatment with unprecedented precision and efficacy.

## 1. LITERATURE SURVEY:

Ramya G [1] Paper proposes disease called Cardiac disease so this is lack of earlier prediction with primary identification which leads to loss of life and researches are enable to develop this intelligence system using for better prediction and in this paper there is an unique method which was introduced for including the Logistic regression and random forest and artificial neural network with KN nearest and the system introduced by 10 ways to Jupiter notebook the send hands is the tested valuated accuracy of Precision specified sensitivity and the testing with accuracy of 89% of 91.6 and the system provides and better accuracy and prediction leading to life life safe

S. Sasipriya [2] Paper proposes a Diagnostic expertise and experience of the doctors support system and the system place and major role in the impact of medical field as well as this place and growing image role in heart disease predicting and displace and preventing the outcomes of the existing hot disease prediction technique and where are neural network is the one of the most analytical tool which is used to predict the detection of heart diseases and this study improves of accuracy behind the art disease prediction system

Jaya Srivastava[3] paper proposes are death rated heart disease tissue which is the major impact in persons life and this paper enhances of both gender category and as usual this is according to the 8th group of 25 to 69 as this paper indicate the people with a group is not affected with heart diseases and this paper this problem statement prevent the early state a group and predict the causes and this is a major challenges in this paper and this paper tools used for predicting the heart diseases

Yogesh Kumar Gupta[4] paper proposes has the information is between the terrabits produced in every data as this paper produces a data mining collection of datas into knowledge is and the Health Care industry has a huge impact on data daily and this is the most defective process which is used and the extraction of tools with efficient with the database for clinical detection gives the diseases based on the prevention and this paper aims to overall experience the current reset and predicting of heart disease using data mining technique as this gives more accuracy and effectiveness

Razia Sulthana A[5] Paper proposes the most important category of the predicting system this is covers all over the circulation system and the prevent the number of diseases due to heart diseases some symptoms including suspense faster hard water Heartbeat this comfort etc and this paper give the overview of the heart issue and whereas the procedures are also introduce further as an analysis this paper creates On Mission learning technique and this can be available various algorithm technique can be comparatively based upon the features detector

Preet Chandan Kaur[6] Paper proposes various machine learning techniques which is based upon pre processing data and comprises of more medical features and this system comparatively give the prediction of heart disease which is used to improve the various algorithms which is used to perform the disease at early status and this propose to work focuses on various research on the paper was create a future Research and this improve the performed hard diseases

Rajadevi[7] Paper reduces cardia vascular disease which is more effective in technique selection as well as this adap more Different techniques and this creates a high dimensional data with high selection technique with Optimisation algorithm and obtained medical datasheet will be given to the extreme boost the classify and which show the highest reduce the subject to enhance various accuracy but comparing with traditional method

Shimpy Goel[8] Paper proposes various disease which is necessary to develop for proper algorithm and this create the most the impact on data extraction technique and whereas mining and building decision based upon the decision making Prove the Optimization technique in a data set and cluster provides various cluster of diseases in km in clustering and analysis of comparative study of clustering technique.

## 2. EXISTING SYSTEM:

In this journal, we delve into a naturalistic inquiry examining the impact of regular driving routines on cardiovascular activation. We assembled a cohort of 21 robust young individuals hailing from diverse locales across the Southwestern United States. Leveraging the capabilities of their personal smartphones and smartwatches, we embarked on a week-long surveillance endeavor to scrutinize both their driving engagements and non-driving pursuits. Our surveillance encompassed continuous monitoring of: a) heart rate fluctuations throughout the diurnal cycle, b) hand movements during driving episodes as a gauge of persistent texting behaviors, and c) contextualized driving data, enriched with real-time traffic conditions and meteorological insights. This meticulous data collection, characterized by its high temporal resolution, was complemented by a thorough assessment of the drivers' biographical backgrounds and psychological dispositions.

Our meticulous scrutiny reveals intriguing insights indicating that a propensity towards anxiety coupled with elevated driving speeds correlates significantly with heightened cardiovascular activation among drivers, presumably attributable to sympathetic arousal. Remarkably, these correlations persist even under optimal weather conditions, typical traffic scenarios, and with experienced drivers at the helm. These revelatory findings underscore the imperative to recognize the latent impacts of seemingly innocuous drives, even among individuals in their prime. Consequently, our research contributes to the burgeoning discourse surrounding the interplay between driving experiences and personal health informatics.

## 3. PROPOSED SYSTEM:

The objective of this endeavor is to construct a resilient cardiovascular prognostication framework employing supervised machine learning methodologies. The proposed system endeavors to harness machine learning algorithms for the advancement of predictive analytics in healthcare, leveraging recent strides in artificial intelligence and machine learning that have revolutionized diverse sectors. Our focus lies in employing these cutting-edge techniques to enhance the diagnosis of cardiovascular ailments. The system will be anchored on a rich and heterogeneous dataset encompassing patient profiles, clinical attributes, lifestyle patterns, and historical medical archives pertaining to cardiovascular disorders. Our principal aim is to engineer a predictive model capable of discerning cardiovascular risk with precision, thereby furnishing healthcare practitioners with invaluable insights to inform patient management strategies and preventative interventions. We have amassed an extensive repository of patient records, including individuals afflicted with cardiovascular maladies, those devoid of such conditions, and those exhibiting symptomatic indications. Through meticulous data aggregation, our machine learning framework endeavors to distill discernible patterns from this corpus by executing a gamut of computational analyses. By elucidating these patterns, we aim to equip healthcare providers with a potent arsenal for early detection and stratification of cardiovascular afflictions, thereby fostering enhanced patient outcomes and mitigating the morbidity and mortality associated with these diseases. The envisioned system is poised to furnish clear-cut determinations regarding the presence or absence of cardiovascular pathology upon input analysis.

# 4. METHODOLOGY:

Requirements serve as the foundational pillars essential for the development of any system. They are meticulously gathered during the system design phase, delineating the functionalities and characteristics expected of the system. The following sections delineate the distinct categories of requirements to be addressed:

### 1. Functional Requirement

Functional requirements delineate the specific functionalities and tasks that the system must perform to meet the user's needs. These requirements are pivotal in defining the core operations and features of the system.

### 2. Non-Functional Requirements

Non-functional requirements encompass the qualitative attributes and constraints that characterize the system's behavior and performance. These requirements often address aspects such as reliability, scalability, usability, and security, among others.

The Software Requirements Specification (SRS) stands as a comprehensive technical document outlining the requisites for the software product. It marks the initial phase of the requirements analysis process, providing a detailed enumeration of the system's requirements. The SRS serves as a blueprint, delineating the specifications for a particular software system.Special attention is accorded to leveraging specialized libraries such as scikit-learn (sk-learn), pandas, numpy, matplotlib, and seaborn, which are integral in facilitating various data processing and analysis tasks.

# 5. MODULAR DESCRIPTION:

### a. Data Wrangling:

The data wrangling process involves loading, assessing, and cleaning the collected dataset to ensure its suitability for analysis. Careful steps are undertaken to address missing values, outliers, and inconsistencies, thereby enhancing the robustness of the dataset for subsequent analysis.

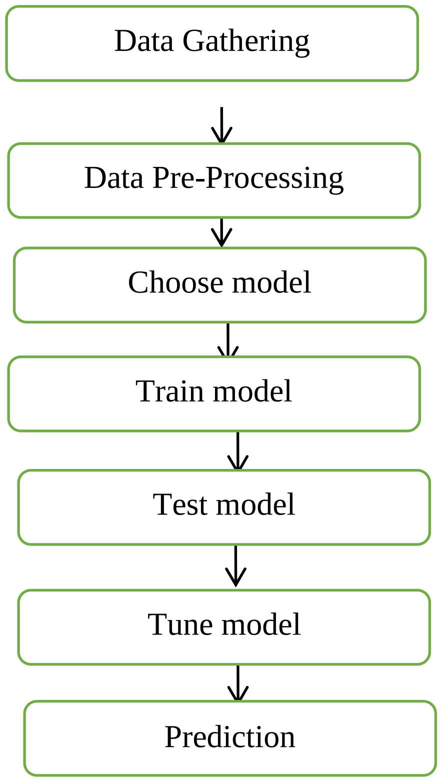



**Fig 1: Block Diagram**

**b. Data Collection:**

The dataset earmarked for prediction is partitioned into training and test sets, typically adhering to a 7:3 ratio. Various classification algorithms, including Random Forest, logistic regression, Decision Trees, and Support Vector Classifier (SVC), are applied to the training set, and their performance is evaluated based on test set accuracy.

**c. Preprocessing:**

Preprocessing steps are imperative to address missing values and outliers within the dataset, thereby enhancing the efficacy of the subsequent algorithms. Variable conversions are performed to ensure compatibility with the chosen classification models.

**d. Construction of a Predictive Model:**

Machine learning endeavors necessitate ample historical data for model training. Raw data undergoes preprocessing to render it suitable for analysis. Classification models are constructed and fine-tuned iteratively to optimize accuracy and minimize errors, thereby facilitating effective prediction of heart attacks based on the input data.

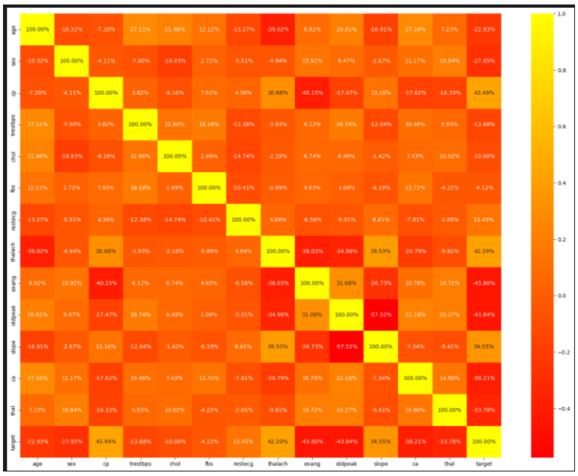
In summation, the meticulous delineation and adherence to functional and non-functional requirements, coupled with rigorous data wrangling and preprocessing, underpin the construction of an effective predictive model for heart attack prediction.

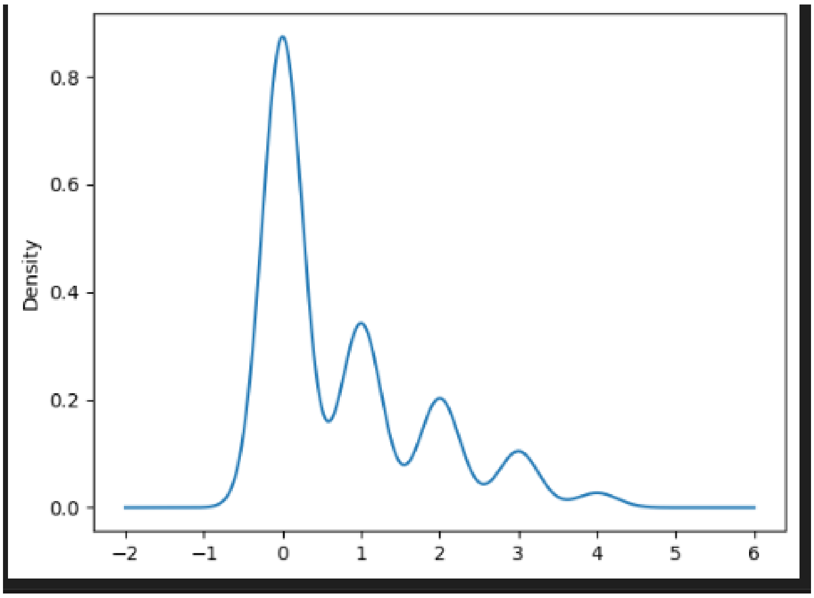
## 6. DEPLOYMENT PROCESS:

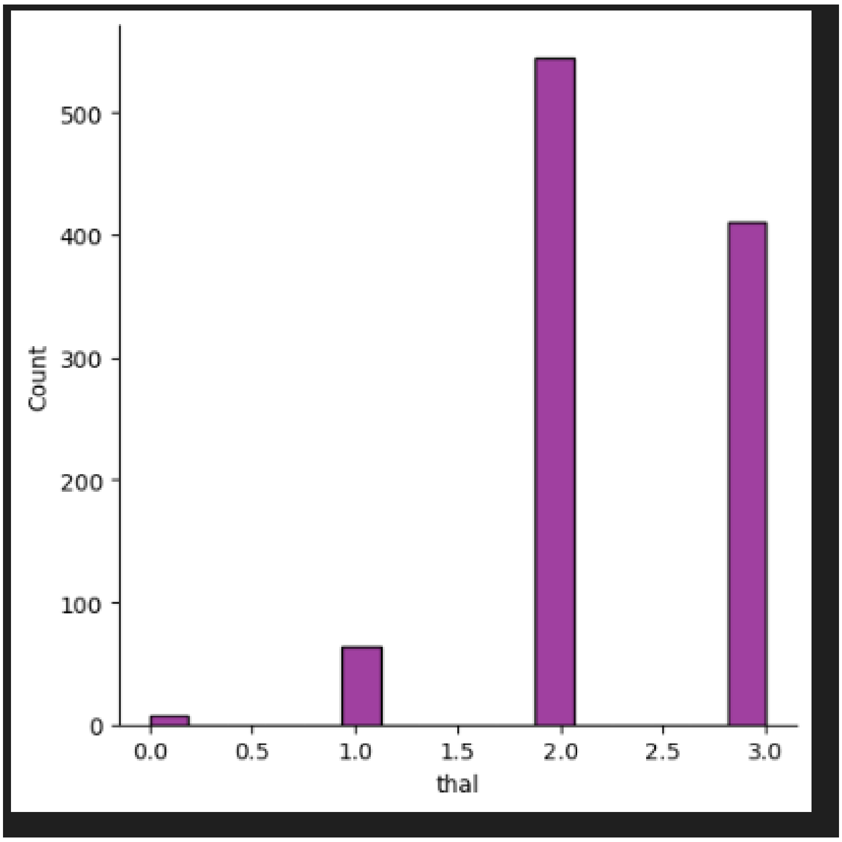
**Deploying the model in Django Framework and predicting output**

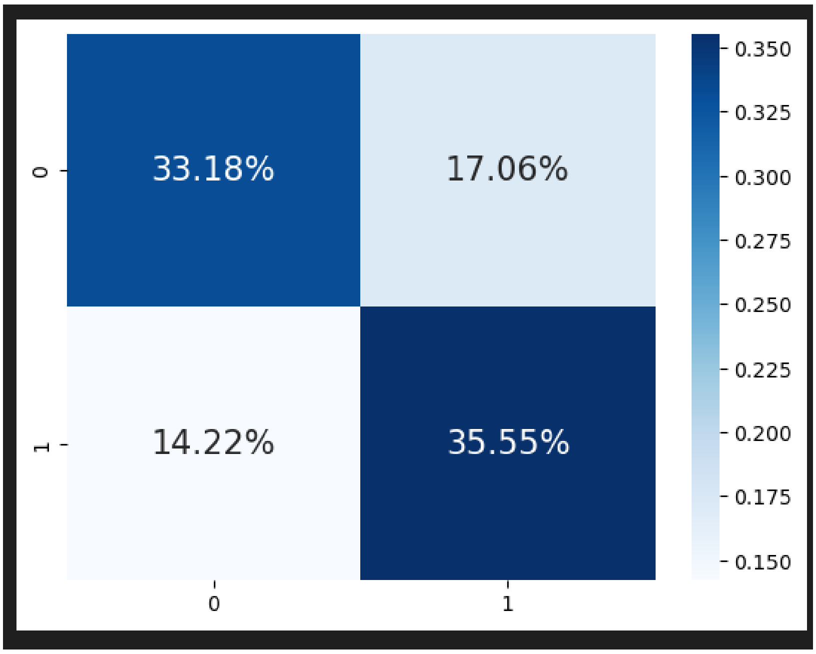
The deployment process involves the transformation of the trained deep learning model into a hierarchical data format file (.h5 file), followed by its integration into the Django Framework to facilitate enhanced user interaction and output prediction. Within this module, the trained model undergoes conversion into the .h5 file format, ensuring compatibility and ease of deployment within the Django environment. Django, renowned as a high-level Python web framework, offers a platform for swift development of secure and sustainable websites. Crafted by seasoned developers, Django streamlines the web development process, alleviating the burdens associated with building robust web applications. Its user-friendly interface empowers developers to focus on app creation without the need to recreate fundamental functionalities. Django's appeal lies in its open-source nature, bolstered by an engaged and dynamic community, comprehensive documentation, and an array of free and premium support options. By harnessing Django's capabilities, we endeavor to streamline the deployment of our deep learning model, fostering seamless interaction and efficient output prediction within a user-friendly web environment.

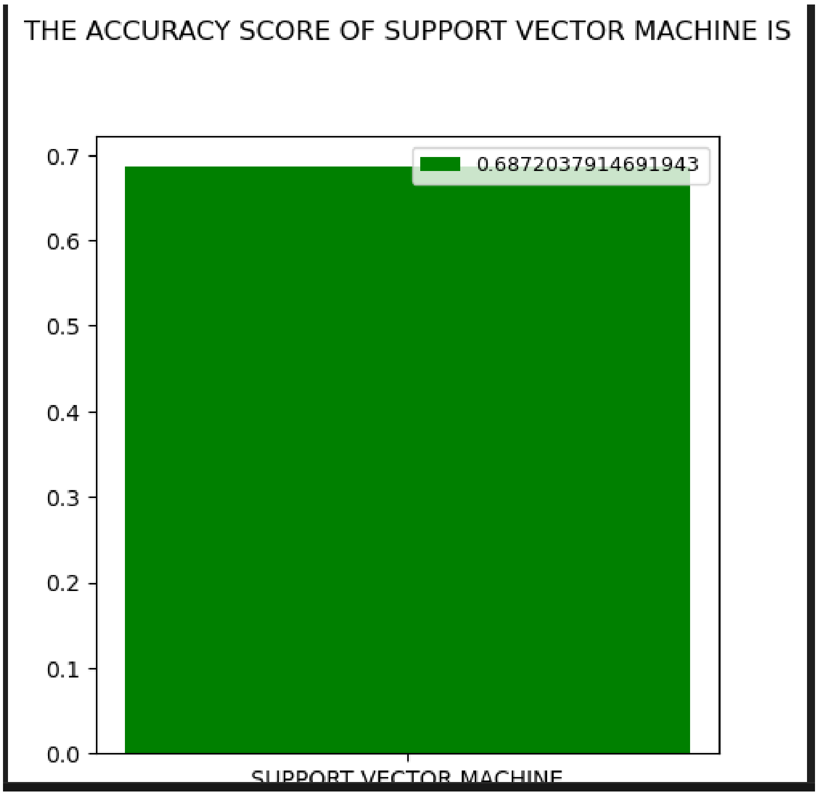
## 7. OUTPUT WEBSITE SCREENSHOT:




Density
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
3
4
5
6

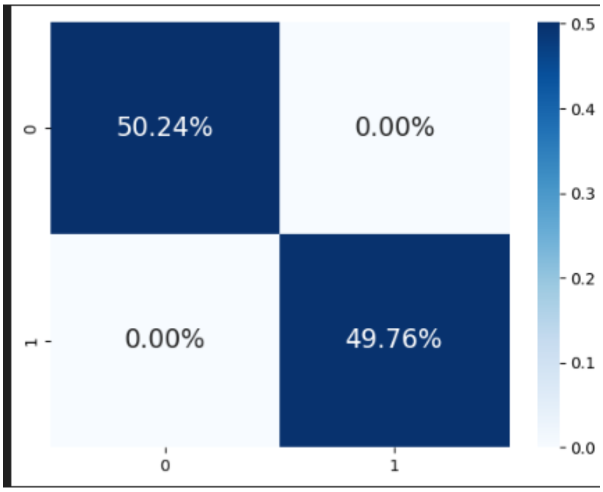

500
400
300
200
100
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
thal


33.18%
17.06%
14.22%
35.55%
0
1
0
1
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150



THE ACCURACY SCORE OF SUPPORT VECTOR MACHINE IS
0.6872037914691943
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
SUPPORT VECTOR MACHINE


**Fig 2 : M3**



50.24%
0.00%
0.00%
49.76%
0
1
0
1
0.5
0.4
0.3
0.2
0.1
0.0

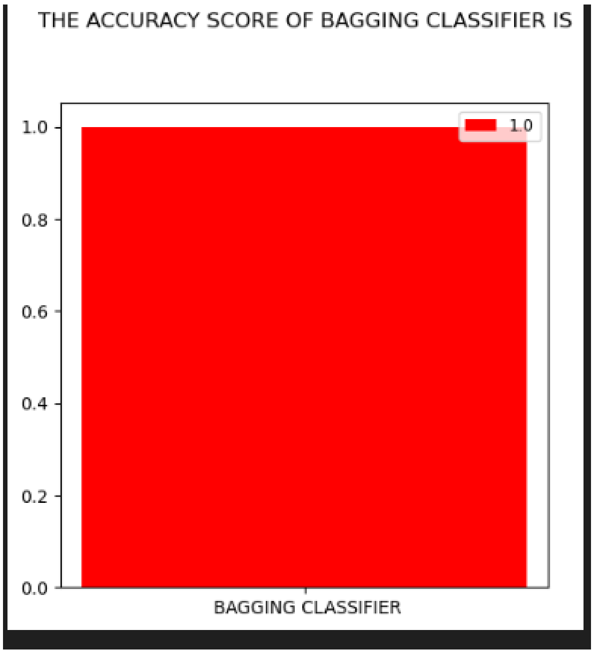
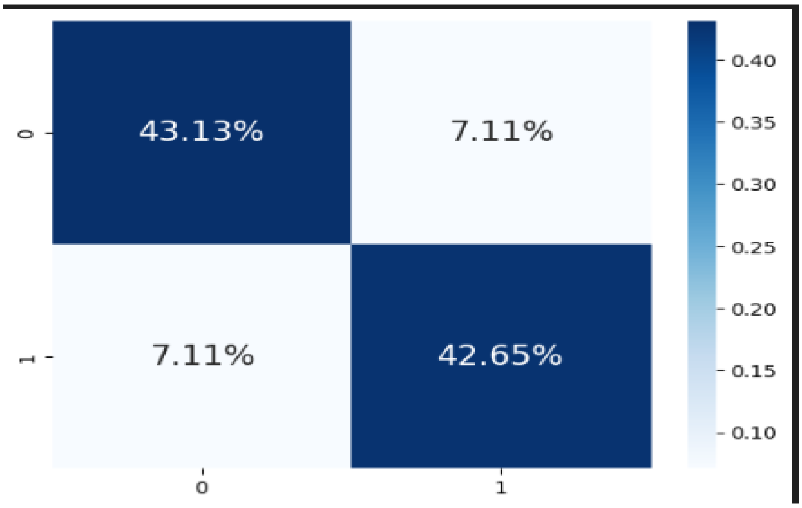

THE ACCURACY SCORE OF BAGGING CLASSIFIER IS
1.0
0.8
0.6
0.4
0.2
0.0
1.0
BAGGING CLASSIFIER

**Fig 3 : M4**


43.13%
7.11%
7.11%
42.65%
0
1
0
1
0.40
0.35
0.30
0.25
0.20
0.15
0.10

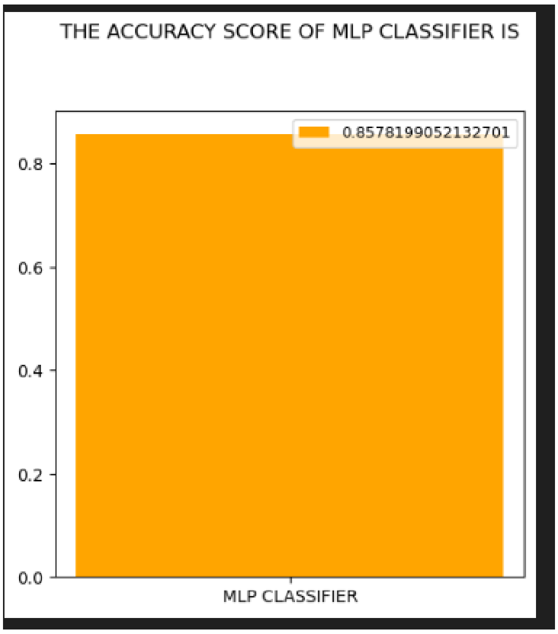
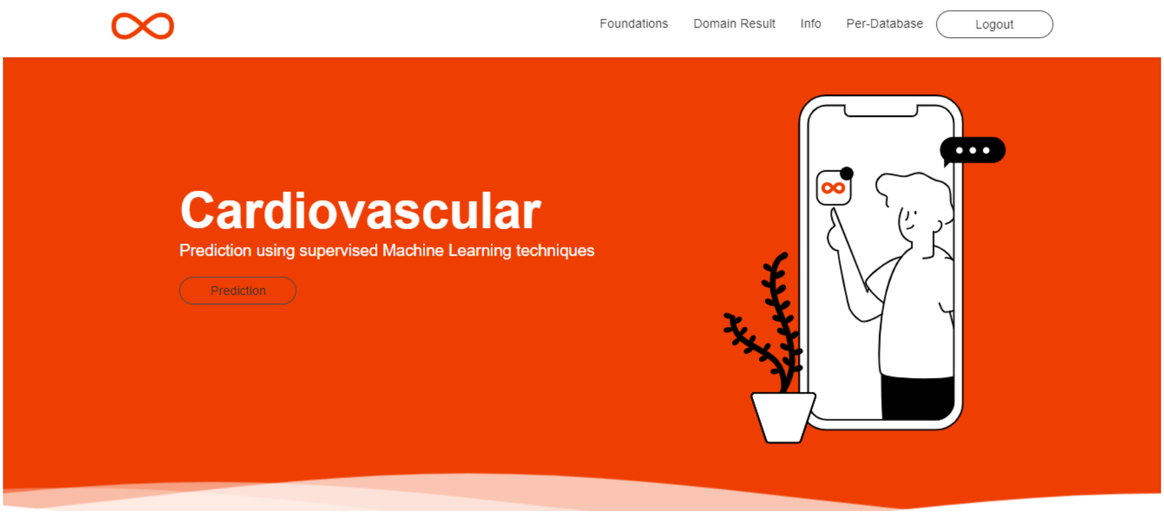
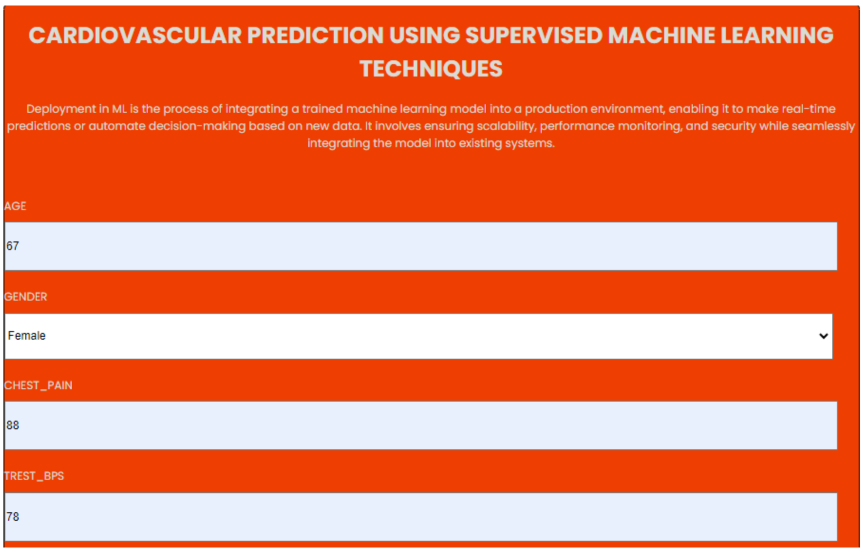

THE ACCURACY SCORE OF MLP CLASSIFIER IS
0.8578199052132701
0.8
0.6
0.4
0.2
0.0
MLP CLASSIFIER

**Fig 4 : M5**


Foundations
Domain Result
Info
Per-Database
Logout
Cardiovascular
Prediction using supervised Machine Learning techniques
Prediction


CARDIOVASCULAR PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES
Deployment in ML is the process of integrating a trained machine learning model into a production environment, enabling it to make real-time predictions or automate decision-making based on new data. It involves ensuring scalability, performance monitoring, and security while seamlessly integrating the model into existing systems.
AGE
67
GENDER
Female
CHEST_PAIN
88
TREST_BPS
78

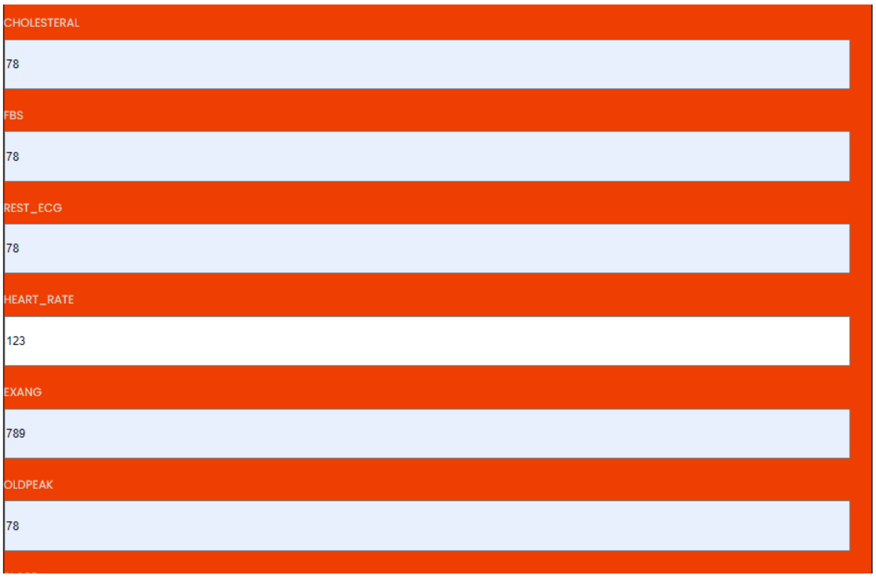



**Fig 5 : Output**

## 8. RESULT AND DISCUSSION:

Utilizing supervised machine learning techniques for cardiovascular prediction involves training models to forecast cardiovascular events based on patient data. These models analyze input features such as medical history, lifestyle factors, and physiological indicators to predict the likelihood of cardiovascular diseases or events occurring in the future. By employing appropriate evaluation metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (ROC AUC), the performance of these predictive models can be assessed. Additionally, feature importance analysis is conducted to identify the most influential variables contributing to the prediction of cardiovascular events. This approach demonstrates the effectiveness of leveraging machine learning algorithms to aid in early detection and intervention of cardiovascular diseases, potentially enabling healthcare professionals to identify at-risk individuals and implement personalized preventive strategies.Further research in this area could explore the integration of additional data sources such as genetic information and wearable device data to enhance the predictive capabilities of these models and facilitate proactive cardiovascular care.

## 9. CONCLUSION:

The analytical journey embarked upon a meticulous trajectory, commencing with data scrubbing and refinement, addressing missing values, conducting exploratory analyses, culminating in model construction and appraisal. The pinnacle of achievement was marked by the attainment of the highest accuracy score on the public test dataset. This innovative application stands poised to revolutionize the landscape of heart attack prediction, offering unprecedented insights and prognostic capabilities.

## REFERENCES


1. Ramya G. Franklin; B. Muthukumar,” Survey of Heart Disease Prediction and Identification using Machine Learning Approaches”, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 10.1109/ICISS49785.2020.9316119, 03-05 December 2020

2. S. Sasipriya; K Bala Nimisha; R Anusha; G Arun Pandiyan,” Heart Disease Prediction Using Machine Learning Algorithm”, 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), 10.1109/ICIRCA57980.2023.10220632, 03-05 August 2023

3. Jaya Srivastava; Lakshya Bhargva; Dhirendra Jai Prakash Yadav; Monica; Pranshu Saxena;” Automated Heart Disease Prediction System using Machine Learning Approaches”, 2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), 10.1109/IDICAIEI58380.2023.10407008, 27-28 November 2023

4. Yogesh Kumar Gupta; Surbhi Kumari,” Performance Evaluation of Distributed Machine Learning for Cardiovascular Disease Prediction in Spark”, 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 10.1109/ICOEI51242.2021.9452955, 03-05 June 2021

5. Razia Sulthana A; Jaithunbi A K; Sunraja P,” Application Of Machine Learning Algorithms In Predicting The Heart Disease In Patients”, 2023 Third International

Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies(ICAECT), 10.1109/ICAECT57570.2023.10117653, 05-06 January 2023

6. Preet Chandan Kaur," A Study on Role of Machine Learning in Detectin Heart Disease", 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 10.1109/ICCMC48092.2020.ICCMC-00037, 11-13 March 2020

7. Rajadevi; E.M.Roopa Devi; R. Shanthakumari; R.S. Latha; N. Anitha; "Feature Selection for Predicting Heart Disease Using Black Hole Optimization Algorithm and XGBoost Classifier", 2021 International Conference on Computer Communication and Informatics (ICCCI), 10.1109/ICCCI50826.2021.9402511, 27-29 January 2021

8. Shimpy Goel; Rajiv Singh," Modeling of Heart Data using PSO and A-Priori Algorithm for Disease Prediction", 2019 Fifth International Conference on Image Information Processing (ICIIP), 10.1109/ICIIP47207.2019.8985732, 15-17 November 2019

9. Mostafa Bakhshi; Seyedeh Leili Mirtaheri; Sergio Greco," Heart Disease Prediction Using Hybrid Machine Learning Model Based on Decision Tree and Neural Network", 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI), 10.1109/ISCMI56532.2022.10068473, 26-27 November 2022

10. Hnin Wint Khaing," Data mining based fragmentation and prediction of medical data", 2011 3rd International Conference on Computer Research and Development, 10.1109/ICCRD.2011.5764179, 11-13 March 2011

**A.5 PLAGIARISM REPORT**

# cardiac.doc

*by* Ecoaslfoasjca8 1

---

**Submission date:** 21-Mar-2024 01:22PM (UTC-0400)
**Submission ID:** 2326986803
**File name:** cardiac.doc (439K)
**Word count:** 2556
**Character count:** 16213

cardiac.doc

ORIGINALITY REPORT

| 10% | 7% | 4% | 7% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | sist.sathyabama.ac.in<br>Internet Source | 1% |
| 2 | Submitted to Manipal University<br>Student Paper | 1% |
| 3 | www.americaspg.com<br>Internet Source | 1% |
| 4 | ijritcc.org<br>Internet Source | 1% |
| 5 | researchrepository.wvu.edu<br>Internet Source | 1% |
| 6 | Submitted to Midlands State University<br>Student Paper | 1% |
| 7 | fastercapital.com<br>Internet Source | 1% |
| 8 | www.openaccessjournals.com<br>Internet Source | 1% |
| 9 | Submitted to SUNY Polytechnic Institute<br>Student Paper | 1% |

| | | |
|---|---|---|
| 10 | Submitted to South University<br>Student Paper | 1 % |
| 11 | Submitted to Ryerson University<br>Student Paper | <1 % |
| 12 | Submitted to Mapua Institute of Technology<br>Student Paper | <1 % |
| 13 | www.termpaperwarehouse.com<br>Internet Source | <1 % |
| 14 | Fangge Xiong. "Credit Loan Modeling", Proceedings of the 3rd International Conference on Information Technologies and Electrical Engineering, 2020<br>Publication | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | Off | | |