

HEART DISEASE IDENTIFICATION USING MACHINE LEARNING CLASSIFICATION

**Sathyabama Institute of Science and
Technology (Deemed to be University)**

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

By

**GOPALAKRISHNANS(Reg. No.38110169)
KAMESVAR G(Reg. No.38110228)**



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

MARCH - 2022



SATHYABAMA

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

Accredited with Grade “A” by NAAC

(Established under Section 3 of UGC Act, 1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119

www.sathyabamauniversity.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFAIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **GOPALAKRISHNAN S(38110169), KAMESVAR G(38110228)** who carried out the project entitled “**Heart Disease Identification Method Using Machine Learning Classification**” under my supervision from November 2021 to May2022.

Internal Guide

Dr. A. Pravin

Head of the Department

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

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

We **GOPALAKRISHNAN S, KAMESVAR G** hereby declare that the Project Report entitled **Heart Disease Identification Method Using Machine Learning Classification** done by me under the guidance of **Dr. A. Pravin** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: Chennai

SIGNATURE OF THECANDIDATE

ACKNOWLEDGEMENT

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

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

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. A. Pravin** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

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

ABSTRACT

In this article, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques. The system is developed based on classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyper parameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms. The experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

INDEX

	TABLE OF CONTENTS	
CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	I
	LIST OF FIGURES	II
	LIST OF ABBREVIATIONS	IV
1	INTRODUCTION	1
2	LITERARY REVIEW	1
3	AIM AND SCOPE OF THE PROJECT	4
	3.1 AIM OF THE PROJECT	4
	3.2 SCOPE OF THE PROJECT	4
	3.2.1 PROPOSE SYSTEM	4
	3.2.2 ADVANTAGES	5
	3.2.3 DISADVANTAGES	5
4	WORKING THEORY OF THE PROJECT	5
	4.1 MACHINE LEARNING	6
	4.2 GATHERING DATA	7
	4.3 DATA PRE-PROCESSING	7
	4.4 RESEARCHING THE MODEL THAT WILL BE BEST FOR THE TYPE OF DATA	9
	4.5 TRAINING AND TESTING THE MODEL OF DATA	13
	4.6 EVALUATION	15
5	IMPLEMENTATION AND METHODOLOGY	16
	5.1 SOFTWARE REQUIREMENT	16
	5.2 HARDWARE REQUIREMENT	16
	5.3 MODULE NAME	16
	5.3.1 DATASET COLLECTION	15
	5.3.2PRE-PROCESSING	15
	5.3.3 FEATURE EXTRACTION	16

	5.3.4 MODEL TRAINING	16
	5.3.5 TESTING MODEL	17
	5.3.6 PERFORMANCE EVALUATION	19
	5.3.7 PREDICTION	19
6	RESULT AND DISCUSSION	20
7	CONCLUSION AND FUTURE WORK	22
	7.1 CONCLUSION	22
	7.2 FUTURE WORK	23
8	REFERENCE	23
9	APPENDIX	25
	A. SOURCE CODE	25
	B. OUTPUT SCREENSHOTS	31
	C. PLAGARISM REPORT	33

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
4.1	RESEARCHING THE MODEL	9
4.2	CLASSIFICATION	10
4.3	REGRESSION	11
4.4	CLUSTERED DATA	12
4.5	OVERVIEW OF MODEL	13
4.6	TRAINING AND TESTING	13
4.7	DATA SEGMENTATION	14
4.8	CONFUSION MATRIX	15
4.9	ARCHITECTURAL DIAGRAM OF OUR MODEL	16
6.1	TEST ACCURACY OF RANDOM FOREST	20
6.2	TEST ACCURACY OF KNN	21
6.3	TEST ACCURACY OF LOGESTIC REGRESSION	21
6.4	COMPARISON OF THREE ALGORITHMS	22

LIST OF ABBREVIATIONS

ML – Machine Learning

CNN – Convolutional Neural Network

SVM – Support Vector Machine

KNN – K-Nearest Neighbor

RF – Random Forest

CHAPTER – 1

INTRODUCTION

1.1 Introduction

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world .The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available. The effective diagnosis and proper treatment can save the lives of many people. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually. Most of the people in the United States are suffering from heart disease Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze. Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a result, the ratio of death decrease.

CHAPTER -2

LITERARY REVIEW

2.1 LITERARY REVIEW

Rahul Katarya2020 Predicting and detection of heart disease has always been a critical and challenging task for healthcare practitioners. Hospitals and other clinics are offering expensive therapies and operations to treat heart diseases. So, predicting heart disease at the early stages will be useful to the people around the world so that they will take necessary actions before getting severe. Heart disease is a significant problem in recent times; the main reason for this disease is the intake of alcohol, tobacco, and lack of physical exercise. Over the years, machine learning shows effective results in making decisions and predictions from the broad set of data produced by the health care industry. Some of the supervised machine learning techniques used in this prediction of heart disease are artificial neural network (ANN), decision tree (DT), random forest (RF), support vector machine (SVM), naïve Bayes (NB) and k-nearest neighbor algorithm. Furthermore, the performances of these algorithms are summarized.

Dr. M. Kavitha2021 Heart disease causes a significant mortality rate around the world, and it has become a health threat for many people. Early prediction of heart disease may save many lives; detecting cardiovascular diseases like heart attacks, coronary artery diseases etc., is a critical challenge by the regular clinical data analysis. Machine learning (ML) can bring an effective solution for decision making and accurate predictions. The medical industry is showing enormous development in using machine learning techniques. In the proposed work, a novel machine learning approach is proposed to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied. The novel technique of the machine learning model is designed. In implementation, 3 machine learning algorithms are used, they are 1. Random Forest, 2. Decision Tree and 3. Hybrid model (Hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user's input parameter to predict the heart disease, for which we used a hybrid model of Decision Tree and Random Forest.

Abderrahmane Ed-daoudy 2019 Over the last few decades, heart disease is the most common cause of global death. So early detection of heart disease and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in Internet of Things health monitoring, streaming system and others have been generating an enormous amount of data on a continuous basis. The combination of streaming big data analytics and machine learning is a breakthrough technology that can have a significant impact in healthcare field especially early detection of heart disease. This technology can be more powerful and less expensive. To overcome this issue, this paper propose a real-time heart disease prediction system based on apache Spark which stand as a strong large scale distributed computing platform that can be used successfully for streaming data event against machine learning through in-memory computations. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark ML lib with Spark streaming and applies classification model on data events to predict heart disease. The seconds uses Apache Cassandra for storing the large volume of generated data.

RahmaAtallah 2019 This paper presents a majority voting ensemble method that is able to predict the possible presence of heart disease in humans. The prediction is based on simple affordable medical tests conducted in any local clinic. Moreover, the aim of this project is to provide more confidence and accuracy to the Doctor's diagnosis since the model is trained using real-life data of healthy and ill patients. The model classifies the patient based on the majority vote of several machine learning models in order to provide more accurate solutions than having only one model. Finally, this approach produced an accuracy of 90% based on the hard voting ensemble model.

Noor Basha 2019 Analysis and Prediction of diseases are two most demanding factors to be faced critically by the doctors and data scientist, where data analytics be very delightful issue, so in this regard, many health industries will working on variety of human syndromes, where they generate huge data. Heart disease, cancer, tumor and

Alzheimer's disease are one of the chronic human diseases, where data scientist and doctors are doing rapid and efficient analysis on these diseases using many machine learning techniques to study and predict these diseases to save and reduce human deaths.

CHAPTER - 3

AIM AND SCOPE OF THE PROJECT

3.1 AIM OF THE PROJECT

- ✓ The AIM Cardiology Solution is a cardiology benefits management program that helps ensure clinically appropriate.
- ✓ Cost-effective care for your members with heart disease.

3.2 SCOPE OF THE PROJECT

- The goal of our heart disease prediction project is to determine if a patient should be diagnosed with heart disease or not, which is a binary outcome.
- So Positive result=1, the patient will be diagnosed with heart disease.
- Negative result= 0, the patient will not be diagnosed with heart disease.

3.2.1 Propose system

- In the proposed work user will search for the heart Disease diagnosis (heart Disease and treatment related information) by giving symptoms as a query in the search engine.
- These symptoms are pre-processed to make the further process easier to find the symptoms keyword which helps to identify the heart Disease quickly.

- CFS+PSO are a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.
- This feature has been identified as the most suitable for the present system.

3.2.2 Advantages

1. It is easy to extract signatures from individual data instances, as their structures. Just collect the symptoms that enough to scaling samples.
2. Can easily predict the heart Disease level and severity easily using range level of queries.
3. The probability of vocabulary gap between diverse health seekers makes the data more consistent compared to other formats of health data.

3.2.3 Disadvantages

1. Existing systems have failed to utilize and understand the importance of misdiagnosis. A very important attribute which interconnects and addresses all these issues.
2. It varies from patient's medical history, climatic conditions, neighborhood, and various other factors.

CHAPTER 4

WORKING THEORY OF OUR PROJECT

4.1 Machine Learning

What are the 7 steps of machine learning?

7 Steps of Machine Learning

- Step 1: Gathering Data. ...
- Step 2: Preparing that Data. ...
- Step 3: Choosing a Model. ...
- Step 4: Training. ...
- Step 5: Evaluation. ...
- Step 6: Hyper parameter tuning. ...
- Step 7: Prediction.

Introduction:

In this blog, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modeling steps that can help to improve the model performance.

Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

Understanding the machine learning workflow

We can define the machine learning workflow in 3 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data

4. Training and testing the model
5. Evaluation

Okay but first let's start from the basics

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e. the trained model will provide false or wrong prediction.

4.2 Gathering Data

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done. We can also use some free data sets which are present on the internet. Kaggle and UCI Machine learning Repository are the repositories that are used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.

4.3 Data pre-processing

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

What is data pre-processing?

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects. Most of the real-world data is messy, some of these types of data are:

- 1. Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
- 2. Noisy data:** This type of data is also called outliers, this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
- 3. Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data:

1. Numeric e.g. income, age
2. Categorical e.g. gender, nationality
3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

These are some of the basic pre — processing techniques that can be used to convert raw data.

- 1. Conversion of data:** As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.
- 2. Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot of missing values in

the dataset.

3. Filling the missing values: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4. Machine learning: If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

5. Outliers detection: There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

4.4 Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data.

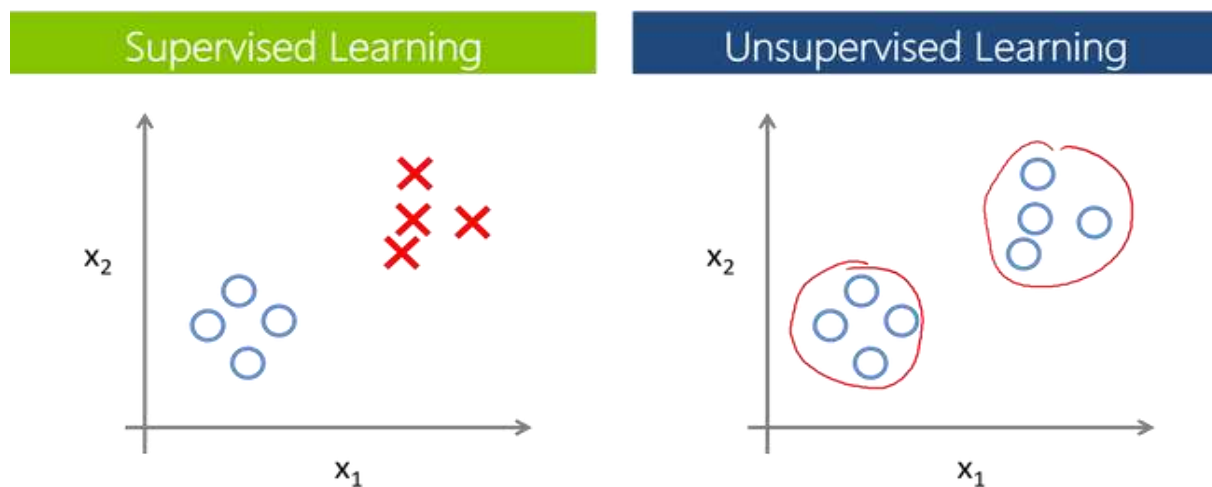


Figure 4.1: Researching the model

Supervised Learning: In Supervised learning, an AI system is presented with data which is labeled, which means that each data is tagged with the correct label.

The supervised learning is categorized into 2 other categories which are “**Classification**” and “**Regression**”.

Classification: Classification problem is when the target variable is **categorical** (i.e. the output could be classified into classes — it belongs to either Class A or B or

something else). A classification problem is when the output variable is a category, such as “red” or “blue”, “disease” or “no disease” or “spam” or “not spam”.

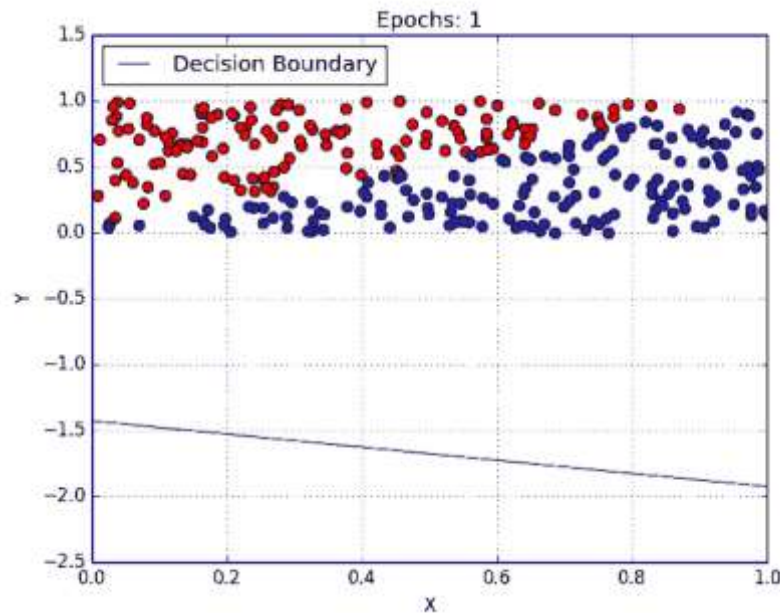


Figure 4.2: Classification

As shown in the above representation, we have 2 classes which are plotted on the graph i.e. red and blue which can be represented as ‘setosa flower’ and ‘versicolor flower’, we can image the X-axis as the ‘Sepal Width’ and the Y-axis as the ‘Sepal Length’, so we try to create the best fit line that separates both classes of flowers.

These some most used classification algorithms:

- K-Nearest Neighbour
- Naive Bayes
- Decision Trees/Random Forest
- Support Vector Machine
- Logistic Regression

Regression:

While a Regression problem is when the target variable is continuous (i.e. the output is

numeric).

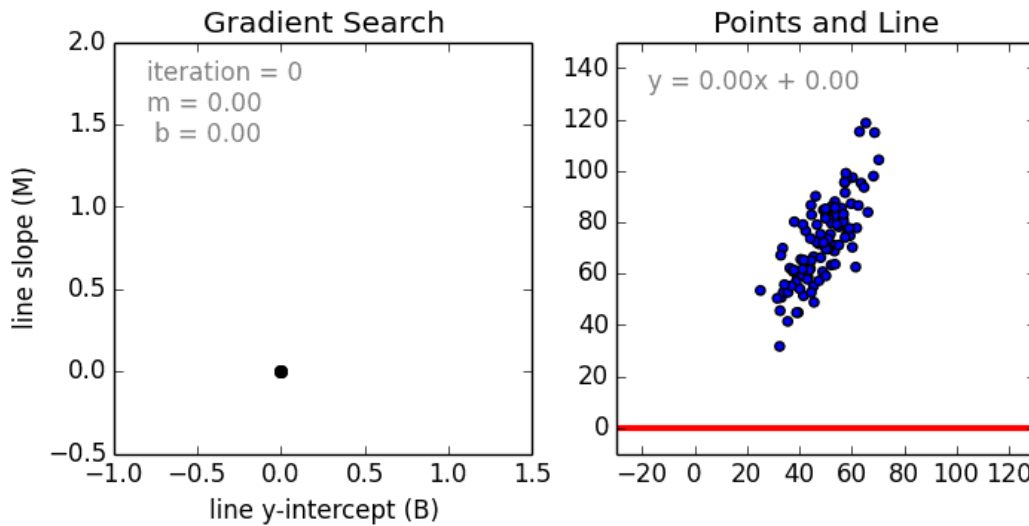


Figure 4.3:Regression

As shown in the above representation, we can imagine that the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ'. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn't present in the given data.

These some most used regression algorithms.

- Linear Regression
- Support Vector Regression
- Decision Tress/Random Forest
- Gaussian Progresses Regression
- Ensemble Methods

Unsupervised Learning:

The unsupervised learning is categorized into 2 other categories which are "Clustering" and "Association".

Clustering:

A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

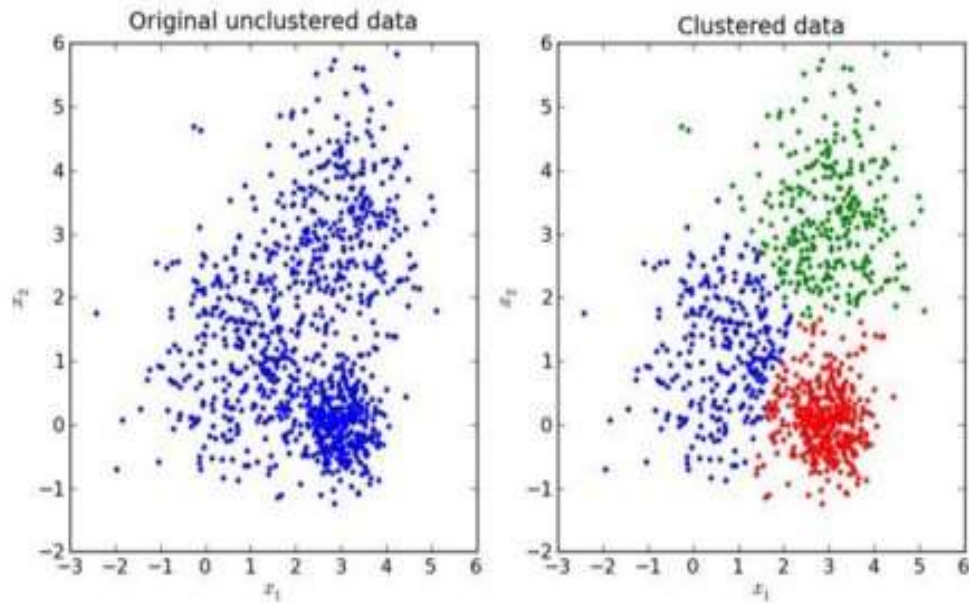


Figure 4.4: Clustered Data

Methods used for clustering are:

- Gaussian mixtures
- K-Means Clustering
- Boosting
- Hierarchical Clustering
- K-Means Clustering
- Spectral Clustering

Overview of models under categories:

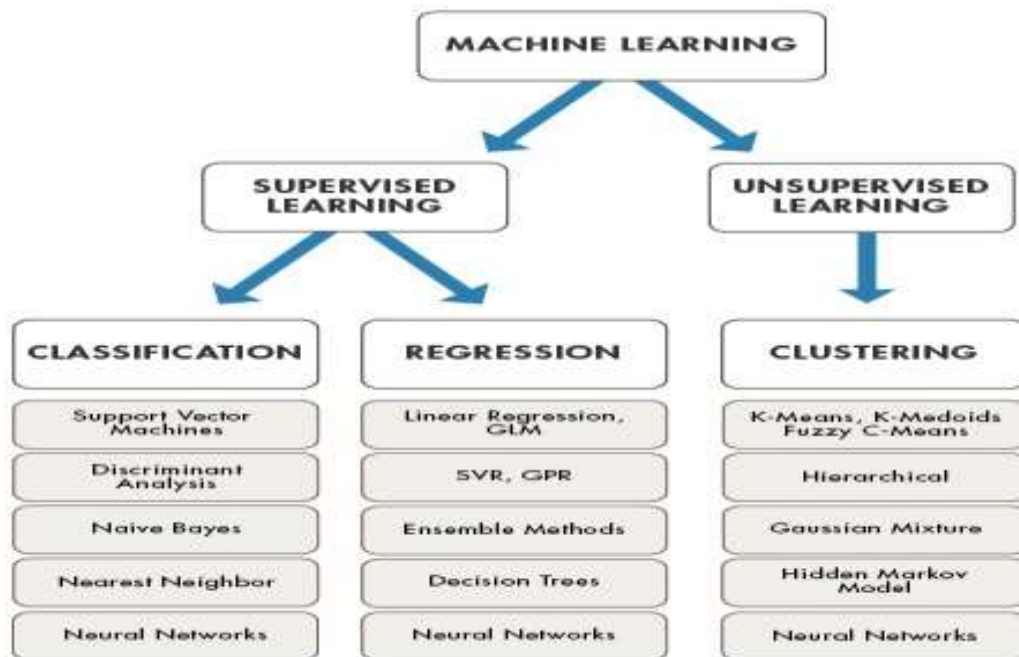


Figure 4.5: Overview of model

4.5 Training and testing the model on data

For training a model we initially split the model into 3 three sections which are 'Training data', 'Validation data' and 'Testing data'. You train the classifier using 'training data set', tune the parameters using 'validation set' and then test the performance of your classifier on unseen 'test data set'. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.

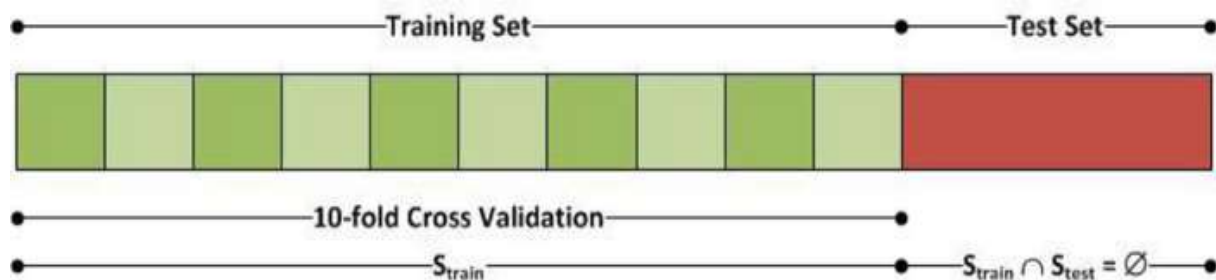


Figure 4.6: Training and testing

Training set:

The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Validation set:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

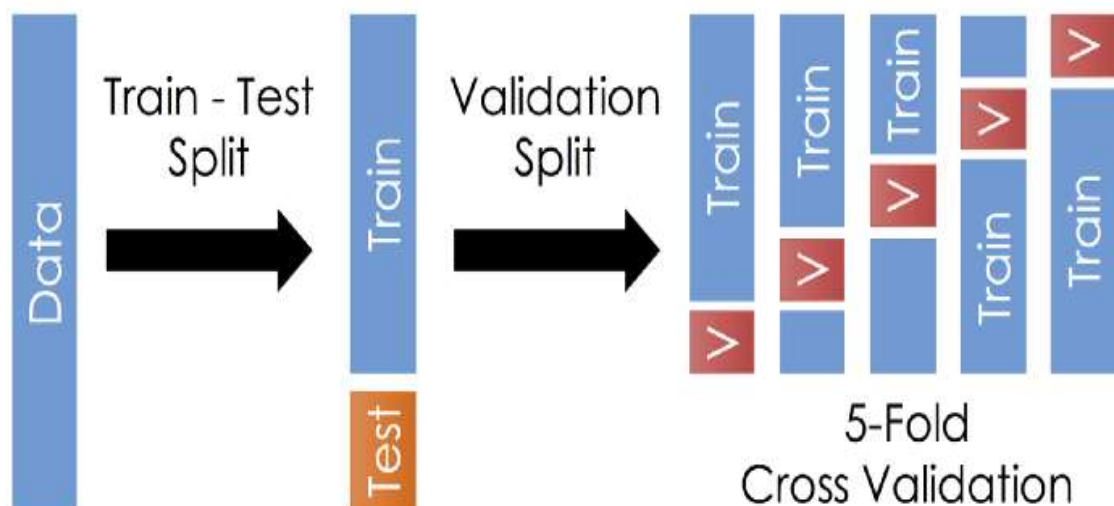


Figure 4.7: Data Segmentation

Once the data is divided into the 3 given segments we can start the training process.

In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration. The model uses any one of the models that we had chosen in step 3/ point 3. Once the model is trained we can

use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are 'True positives', 'True Negatives', 'False Positives' and 'False Negative'. We prefer that we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

n=165		Predicted: NO	Predicted: YES
Actual: NO		50	10
Actual: YES		5	100

Figure 4.8: Confusion Matrix

True positives : These are cases in which we predicted TRUE and our predicted output is correct.

True negatives : We predicted FALSE and our predicted output is correct.

False positives :We predicted TRUE, but the actual predicted output is FALSE.

False negatives :We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix. Accuracy = (True Positives + True Negatives) / (Total number of classes) i.e. for the above example: Accuracy = (100 + 50) / 165 = 0.9090 (90.9% accuracy)

4.5 Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to

increase the number of true positives and true negatives.

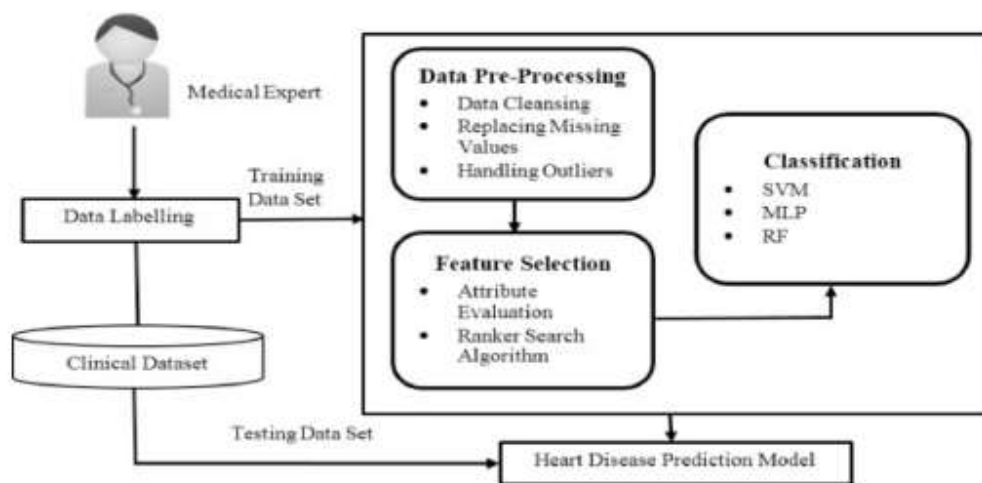


Figure 4.9: Architectural Diagram of our model

CHAPTER 5

IMPLEMENTATION & METHODOLOGY

5.1 Software Requirements:

Operating system: Windows 10.

Coding Language: Python

5.2 Hardware Requirements:

System: Pentium i3 Processor.

Hard Disk: 500 GB.

Monitor: 15" LED

Input Devices: Keyboard, Mouse

Ram: 2 GB

5.3 Module name:

- Dataset collection
- Pre-Processing
- Feature Extraction
- Model training
- Testing model
- Performance Evaluation
- Prediction

5.3.1 Dataset collection:

Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes. Predictive models are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models. The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand. For example, a loan default model would not benefit from tiger population sizes but could benefit from gas prices over time. In this module, we collect the data from kaggle dataset archives. This dataset contains the information of divorce in previous years.

5.3.2 Pre-Processing:

The Wisconsin Prognostic Cleave Land Train Dataset is downloaded from the UCI Machine Learning Repository website and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values.

5.3.3 Feature Extraction:

This is done to reduce the number of attributes in the dataset hence providing advantages like speeding up the training and accuracy improvements. In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be

informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

5.3.4 Model training:

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model. This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model. Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning. In this module we use supervised classification algorithms like linear regression to train the model on the cleaned dataset after dimensionality reduction.

5.3.5 Testing model:

In this module we test the trained machine learning model using the test dataset. Quality assurance is required to make sure that the software system works according to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that you test the program against should be stated in the technical specification document. Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your

clients to encounter bugs after the software is released and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

5.3.6 Performance Evaluation:

In this module, we evaluate the performance of trained machine learning model using performance evaluation criteria such as F1 score, accuracy and classification error. In case the model performs poorly, we optimize the machine learning algorithms to improve the performance. Performance Evaluation is defined as a formal and productive procedure to measure an employee's work and results based on their job responsibilities. It is used to gauge the amount of value added by an employee in terms of increased business revenue, in comparison to industry standards and overall employee return on investment (ROI). All organizations that have learned the art of "winning from within" by focusing inward towards their employees, rely on a systematic performance evaluation process to measure and evaluate employee performance regularly. Ideally, employees are graded annually on their work anniversaries based on which they are either promoted or are given suitable distribution of salary raises. Performance evaluation also plays a direct role in providing periodic feedback to employees, such that they are more self-aware in terms of their performance metrics.

5.3.7 Prediction:

"Prediction" refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days.

The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be. The word "prediction" can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you're using machine learning to determine the next best action in a marketing campaign.

Other times, though, the "prediction" has to do with, for example, whether or not a transaction that already occurred was fraudulent.

In that case, the transaction already happened, but you're making an educated guess

about whether or not it was legitimate, allowing you to take the appropriate action. In this module we use trained and optimized machine learning model to predict whether the patient give the answer some questions.

CHAPTER – 6

RESULT AND DISCUSSION

By applying different machine learning algorithms and then using deep learning to see what difference comes when it is applied to the data, three approaches were used. In the first approach, normal dataset which is acquired is directly used for classification, and in the second approach, the data with feature selection are taken care of and there is no outliers detection. *e results which are achieved are quite promising and then in the third approach the dataset was normalized taking care of the outliers and feature selection; the results achieved are much better than the previous techniques, and when compared with other research accuracies, our results are quite promising

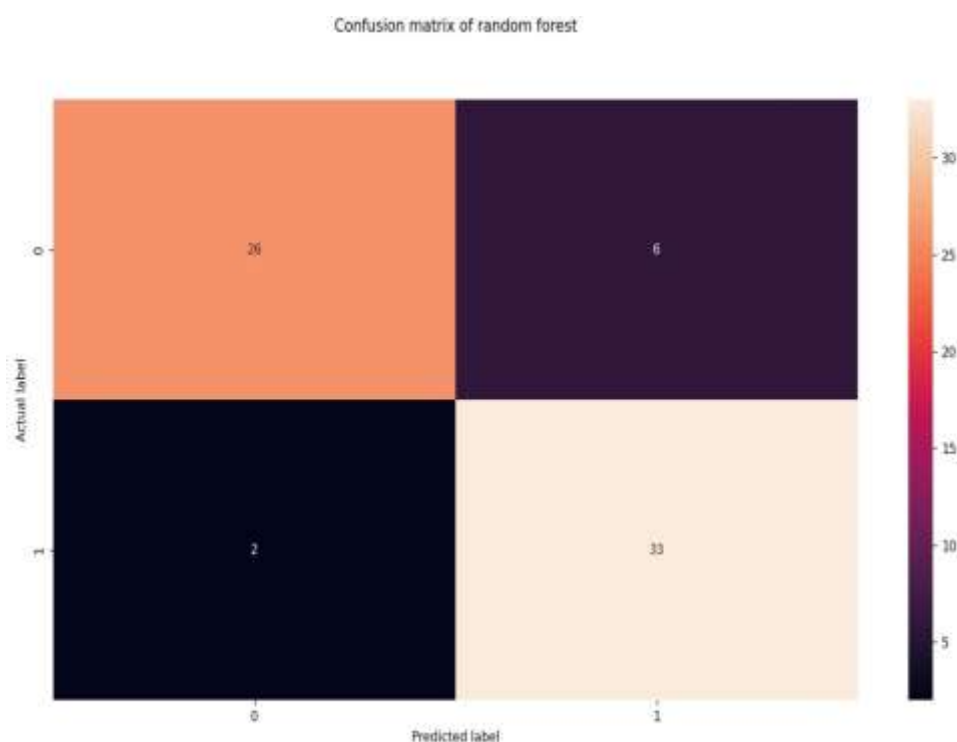


Figure 6.1: Test Accuracy of Random Forest

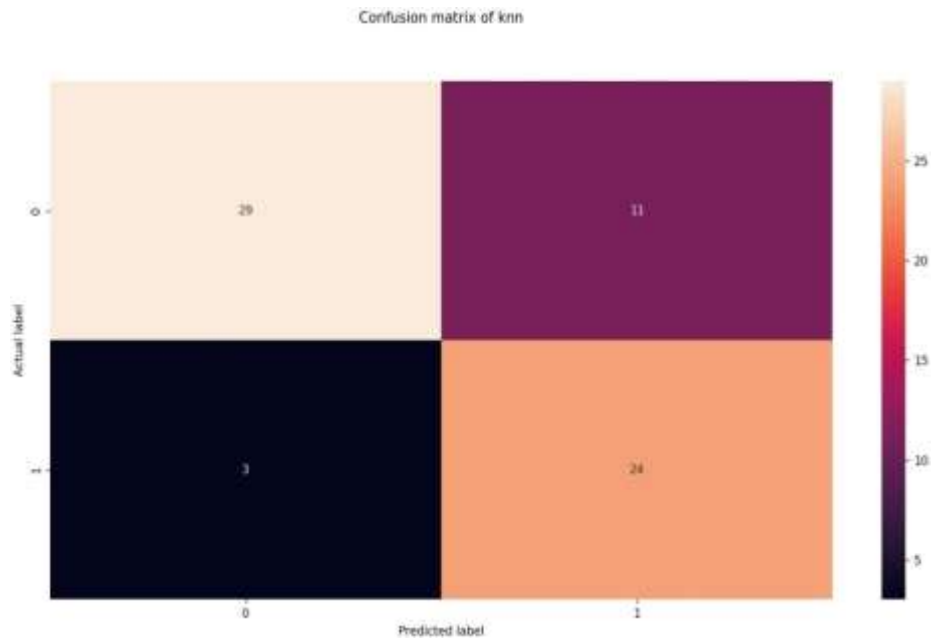


Figure 6.2: Test Accuracy of KNN

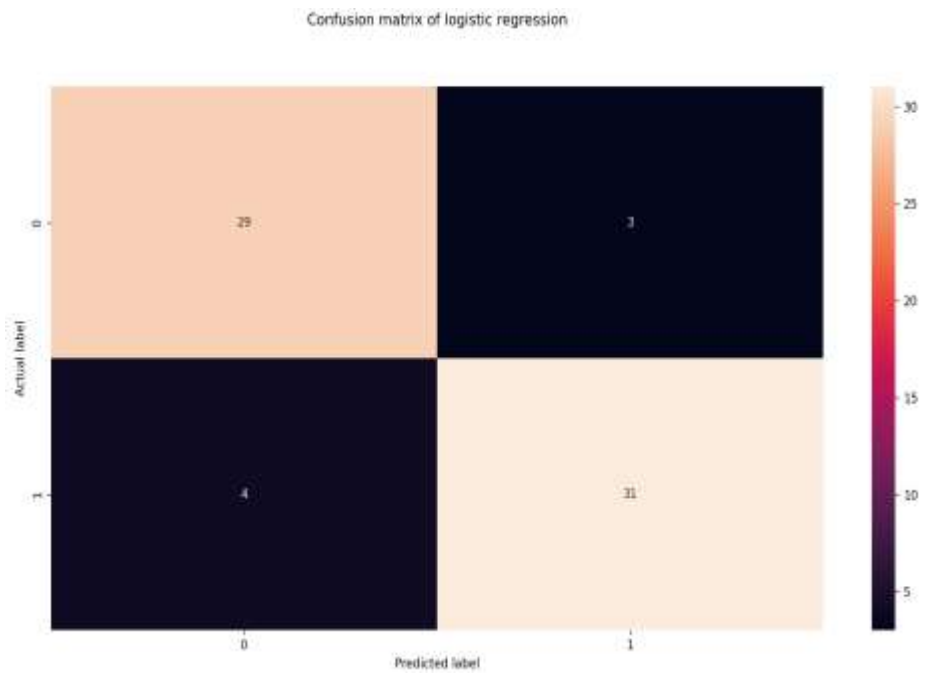


Figure 6.3: Test Accuracy of Logistic Regression

Figure 6.1: Comparison of three algorithms

```
2 KNN Score: 79.10%
Precision Score : 68.57%
Recall Score    : 88.89%
Confusion Matrix :
[[29 11]
 [ 3 24]]
Accuracy of Random Forest: 88.06%
Precision Score : 94.29%
Recall Score    : 84.62%
Confusion Matrix :
[[26  6]
 [ 2 33]]
Accuracy : 89.55%
Precision Score : 88.57%
Recall Score    : 91.18%
Confusion Matrix :
[[29  3]
 [ 4 31]]
```

Here we observe and compare the accuracy of three models namely Logistic Regression, KNN, Random Forest among these, Logistic Regression model has the best overall accuracy and F1 score. Therefore, we should use Logistic Regression algorithm to predict the heart disease.

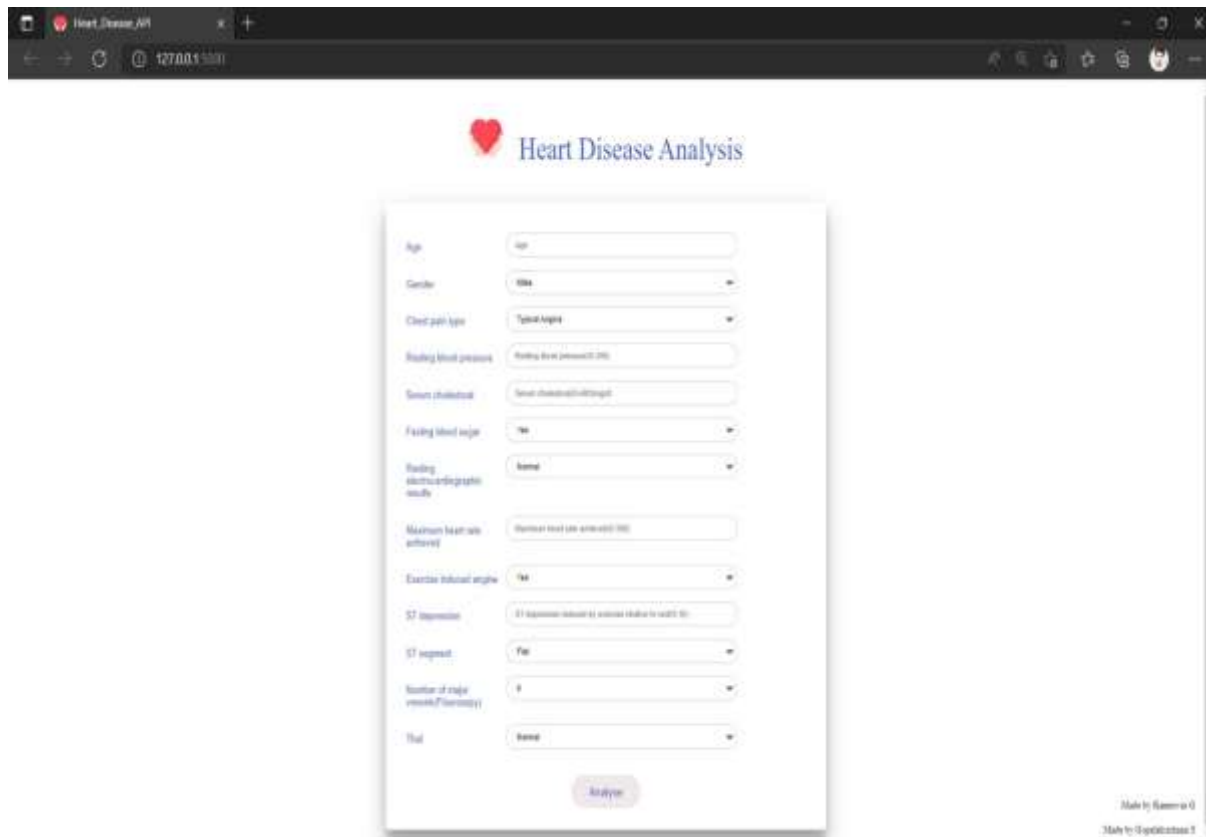


Figure 6.4 Web page interface

CHAPTER – 7

CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

Clinical finding is a significant region of exploration which assists with recognizing the event of a coronary illness. The framework, utilizing different methods referenced, will thus uncovered the root coronary illness alongside the arrangement of most plausible heart Diseases which have comparative side effects. The information base utilized is a portrayal data set so to decrease the dataset tokenization, separating and stemming is finished. The venture presents a novel mixture model to recognize and affirm CAD cases requiring little to no effort by utilizing clinical information that can be

effectively gathered at clinics. Intricacy of framework is diminished by decreasing the dimensionality of the informational collection with PSO. It gives reproducible and target finding, and subsequently can be a significant extra device in clinical practices. Results are equivalently, encouraging and along these line the proposed half and half technique will be useful in coronary illness diagnostics. Trial results exhibit the predominance of the proposed half breed technique concerning forecast precision of CAD.

7.2 FUTURE WORK

In this paper three methods in which comparative analysis was done and promising results were achieved. The conclusion which we found is that machine learning algorithms performed better in this analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this paper. The methods which are used for comparison are confusion matrix, precision, specificity, sensitivity, and F1 score. For the 13 features which were in the dataset, K-Neighbors classifier performed better in the ML approach when data preprocessing is applied.

REFERENCES

- [1] S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6S, pp. 1009–1015, 2019
- [2] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, May 2020
- [3] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *Proc. IEEE 5th Int. Conf. Conver. Technol. (ICT)*, Mar. 2019, pp. 1–4
- [4] U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, "A novel integrated diagnosis method for breast cancer detection," *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, 2020.

- [5] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning,” Data Mining, Inference, and Prediction, Springer, Cham, Switzerland, 2020.
- [6] S. Marsland, “Machine learning,” An Algorithmic Perspective, CRC Press, Boca Raton, FL, USA, 2020.
- [7] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, “Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability,” IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 727–733, 2013.
- [8] M. M. A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, and R. R. Yager, “Deep learning approach for active classification of electrocardiogram signals,” Information Sciences, vol. 345, pp. 340–354, 2016.
- [9] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, “A machine learning system to improve heart failure patient assistance,” IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.
- [10] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, “Automatic methods to extract New York heart association classification from clinical notes,” in Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.
- [11] Vikas Chaurasia, Saurabh Pal, “Early Prediction of Heart Diseases Using Data Mining Techniques”, Caribbean Journal of Science and Technology, Vol. 1, pp. 208-21, December 2013.
- [12] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction", 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. ICEEICT 2016.
- [13] L. Zhu, J. Shen, L. Xie, and Z. Cheng, “Unsupervised topic hyper graph hashing for efficient mobile image retrieval,” IEEE transactions on cybernetics, vol. 47, no. 11, pp. 3941–3954, 2016.
- [14] J. Li and H. Liu, “Challenges of feature selection for big data analytics,” IEEE Intelligent Systems, vol. 32, no. 2, pp. 9–15, 2017.
- [15] R. Kannan, V. Vasanthi, “Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease”, Springer Briefs in Forensic and Medical Bioinformatics, pp. 63-72, 14 June 2018. DOI: https://doi.org/10.1007/978-981-13-0059-2_8.
- [16] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. , “ Heart disease prediction using data mining techniques”, 2017 International Conference on Intelligent Computing and Control (I2C2).
- [17] Dr. B. Umadevi, “A Survey on Prediction of Heart Disease Using Data Mining

- Techniques”, International Journal of Science and Research (IJSR), 2015.
- [18] Ambekar, S., &Phalnikar, R., “Disease Risk Prediction by Using Convolutional Neural Network”, Fourth International Conference on Computing Communication Control and Automation (ICCUBE), 2018. doi:10.1109/iccubea.2018.8697423.
- [19] Gavhane, A., Kokkula, G., Pandya, I., &Devadkar, P. K., “ Prediction of Heart Disease Using Machine Learning. 2018 Second International Conference on Electronics”, Communication and Aerospace Technology (ICECA), 2018.doi:10.1109/iceca.2018.8474922.
- [20] Jabbar, M. A., &Samreen, S., “Heart disease prediction system based on hidden naïve bayes classifier”, 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), 2016.doi:10.1109/cimca. 2016.8053261.
- [21] Ritika, Chadha, Mayank, Shubhankar, “Prediction of heart disease using data mining techniques”, Springer 2016. DOI: <https://doi.org/10.1007/s40012-016-0121-0>.
- [22] Shaikh, S., Sawant, A., Paradkar, S., &Patil, K., “ Electronic recording system- heart disease prediction system”, 2015 International Conference on Technologies for Sustainable Development (ICTSD), 2015.doi:10.1109
- [23] YogitaSolanki, Sanjiv Sharma, "A Survey on Risk Assessments of Heart Attack Using Data Mining Approaches", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.4, pp. 43-51, 2019. DOI: 10.5815/ijieeb.2019.04.05
- [24] So wmiya, C., &Sumitra, P., “Analytical study of heart disease diagnosis using classification techniques”, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017.doi:10.1109/itcosp.

APPENDICES

A.SOURCE CODE

```
#importing the libraries
import warnings
```

```

import pickle
import pandas          as pd
import numpy           as np
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose   import ColumnTransformer
from sklearn.model_selection import train_test_split, cross_val_score,
RandomizedSearchCV, GridSearchCV

from sklearn.ensemble   import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors  import KNeighborsClassifier
from sklearn.metrics    import accuracy_score, confusion_matrix,
precision_score, recall_score
warnings.filterwarnings('ignore')

```

```

except Exception as e:
    print("Unable to import the libraries",e)

```

```

#====Data
Preprocessing=====
#loading the dataset
dataset=pd.read_csv('heart_data.csv', sep='\t' )

#to check the different unique values in the dataset
for index in dataset.columns:
    print(index,dataset[index].unique())
print(dataset.dtypes)
print('The number of missing dataset:', dataset.isnull().sum().sum())

#splitting the dataset to independent and dependent sets
dataset_X=dataset.iloc[:, 0:13].values
dataset_Y=dataset.iloc[:, 13:14].values

```

```

#columns to be encoded: cp(2), restecg(6), slope(10), ca(11), thal(12)
ct=ColumnTransformer([('encoder', OneHotEncoder(drop='first'), [2,6,10,11,12])),
remainder='passthrough')
dataset_X = ct.fit_transform(dataset_X)

#splitting data to training set and test set
X_train, X_test, Y_train, Y_test =train_test_split(dataset_X, dataset_Y, test_size=0.3,
random_state=0)

#=====Evaluation=====
=====

#scores

def scores(pred,test,model):
    print(('\\n=====Scores for {} =====\\n').format(model))
    print(f"Accuracy Score : {accuracy_score(pred,test) * 100:.2f}% ")
    print(f"Precision Score : {precision_score(pred,test) * 100:.2f}% ")
    print(f"Recall Score : {recall_score(pred,test) * 100:.2f}% ")
    print("Confusion Matrix :\\n" ,confusion_matrix(pred,test))

#=====LR_Tunned=====
=====

#logistic regression

lr = LogisticRegression()
solvers = ['newton-cg', 'lbfgs', 'liblinear']
penalty = ['l2']
c_values = [100, 10, 1.0, 0.1, 0.01]

```

```

# define grid search
grid_lr={"solver":solvers,
        "penalty":penalty,
        "C":c_values}

grid_search_lr = GridSearchCV(estimator=lr, param_grid=grid_lr, n_jobs=-1, cv=10,
scoring='accuracy',error_score=0)

#getting the best parameters
grid_result_lr = grid_search_lr.fit(X_train,Y_train)
best_grid_lr  =grid_result_lr.best_estimator_
best_grid_lr.fit(X_train, Y_train)
Y_pred_knn_t  =best_grid_lr.predict(X_test)
scores(Y_pred_knn_t,Y_test,'KNeighbors_Classifier_Tunned')

#=====KNN=====
=====

#K Nearest Neighbour

#tunning the k value
accuracy=[]
for index in range(1,20):
knn = KNeighborsClassifier(n_neighbors=index)
score=cross_val_score(knn,X_train, Y_train,cv=10)
accuracy.append(score.mean())

best_k=accuracy.index(max(accuracy))

#fitting the model with best k value
knn = KNeighborsClassifier(n_neighbors = best_k)
knn.fit(X_train, Y_train)

```

```
Y_pred_knn=knn.predict(X_test)
scores(Y_pred_knn,Y_test,'KNeighbors_Classifier')
```

```
#=====KNN_Tunned=====
=====
```

```
#K Nearest Neighbour with Hyper parameter
```

```
knn_t = KNeighborsClassifier()
n_neighbors = range(1,20)
weights = ['uniform', 'distance']
metric = ['euclidean', 'manhattan', 'minkowski']
```

```
# define grid search
```

```
grid = {'n_neighbors' : n_neighbors,
        'weights'      : weights,
        'metric'       : metric}
```

```
grid_search_knn = GridSearchCV(estimator=knn_t, param_grid=grid, n_jobs=-1,
cv=10 ,scoring='accuracy',error_score=0)
```

```
#getting the best parameters
```

```
grid_result_knn = grid_search_knn.fit(X_train,Y_train)
best_grid_knn   =grid_result_knn.best_estimator_
best_grid_knn.fit(X_train, Y_train)
Y_pred_knn_t   =best_grid_knn.predict(X_test)
scores(Y_pred_knn_t,Y_test,'KNeighbors_Classifier_Tunned')
```

```
#=====RF=====
=====
```

```
#Random Forest
```

```

n_estimators      = [int(x) for x in np.linspace(start =10, stop = 200, num = 10)]
max_features      = ['auto', 'sqrt','log2']
max_depth        = [int(x) for x in np.linspace(10, 1000,10)]
min_samples_split = [2, 5, 10,14,None]
min_samples_leaf  = [1, 2, 4,6,8,None]

random_grid = {'n_estimators'      : n_estimators, 'max_features'      :
max_features,'max_depth'          : max_depth,'min_samples_split'
:min_samples_split,'min_samples_leaf' : min_samples_leaf}

rf=RandomForestClassifier()

randomcv=RandomizedSearchCV(estimator=rf,param_distributions=random_grid,n_i
ter=300,cv=10, random_state=100,n_jobs=-1)

#getting the best parameters
randomcv.fit(X_train,Y_train)
best_grid=randomcv.best_estimator_
best_grid.fit(X_train,Y_train)
pred_2=best_grid.predict(X_test)
scores(pred_2,Y_test,'RandomForestClassifier')

#=====Saving the
models=====
#saving model to disk

pickle.dump(best_grid_lr, open('ml_model.pkl', 'wb'))
pickle.dump(ct,      open('encoder.pkl', 'wb'))

#=====Testing model
response=====

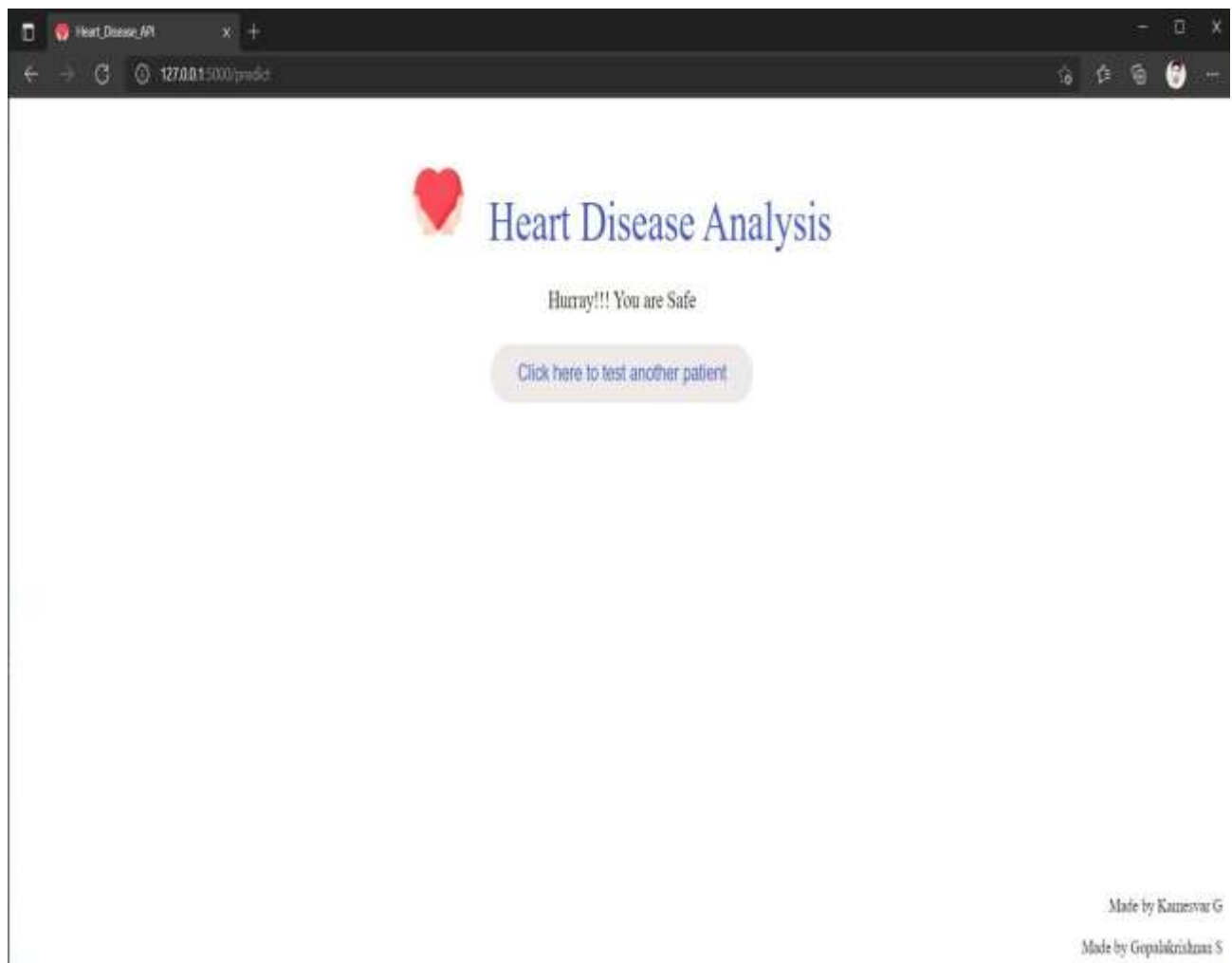
```

#test the pickle file

```
def test_model(row_number):  
    model=pickle.load(open('ml_model.pkl', 'rb'))  
  
    value,real=dataset_X[row_number,:].reshape(1,-1),dataset_Y[row_number,:]  
    print("\n The value predicted is : {} and the real value is : {}  
").format(model.predict(value), real))  
  
test_model(102)  
print("\nCompleted")
```

B.OUTPUT SCREEN SHOTS

Safe



Need Medical Attention



Made by Kameswar G

Made by Gopalakrishna S

C. PLAGARISM REPORT

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION.docx

ORIGINALITY REPORT

2 %	2 %	2 %	1 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Chino Valley Unified School District Student Paper	1 %
2	annalsofrscb.ro Internet Source	1 %
3	Mediana Aryuni, Eka Miranda, Charles Bernando, Andrian Hartanto. "Coronary Artery Disease Prediction Model using CART and SVM: A Comparative Study", 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), 2021 Publication	<1 %
4	"Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2019)", Springer Science and Business Media LLC, 2020 Publication	<1 %
5	www.frontiersin.org Internet Source	<1 %