PERFORMANCE ANALYSIS OF HEART DISEASE PREDICTION

ARYA N M (192IT117)

ASHOK KUMAR G (192IT118)

BACHELOR OF ENGINEERING

in

INFORMATION TECHNOLOGY

BANNARI AMMAN INSTITUTE AMMAN INSTITUTE OF TECHNOLOGY

ABSTRACT

Heart disease is a major cause of morbidity and mortality globally, and early detection is crucial for effective management. Models for machine learning have been created to aid in the prediction of heart disease, with LightGBM being one such model. This study aims to analyse the effectiveness of LightGBM in predicting heart disease. LightGBM was implemented using Python, and the model was trained using the drill set. The model's performance was evaluated using several metrics such as recall, accuracy, and precision F1 score, and region underneath the receiver operating characteristic (ROC) curve.This research shows that LightGBM is an effective model for predicting heart disease. Further studies could be conducted to assess the model's performance on larger datasets and to compare it to other performances machine learning mode.

One of the most common most common diseases worldwide, and many people have died as a result of it. Diseases may have an impact on people both physically and emotionally, since getting and living with an illness can change a person's outlook on life. An illness that affects several areas of an organism yet is not caused by an instant exterior damage. Diseases are frequently defined as medical disorders characterised by distinct symptoms and indicators. The most lethal illnesses in humans are arteria coronary disease (blood flow blockage), cerebrovascular disease, and lower respiratory infections. Heart disease is the most unexpected and unpredictability. With machine learning, we can anticipate cardiac disease. To get high efficiency output, we employ Convolutional Neural Network approaches.

***Keywords*:**,*Morbidity, Arteria coronary, LIGHTGBM, Respiratory infections*

CHAPTER 1

INTRODUCTION

Heart disease is a significant global public health issue. Early prediction of heart disease can help in preventing complications and improving patient outcomes. To forecast, machine learning models have been created heart health using various features like age, sex, and blood pressure, cholesterol level, and other medical history variables. Two popular developing with machine learning frameworks predictive models are LightGBM and XGBoost. Both frameworks are based on gradient boosting, which is a type of multidisciplinary ensemble learning weak models to create a strong model.

In this analysis, we will compare Namely, LightGBM's and XGBoost in predicting utilising heart disease a publicly available dataset.The collection of data utilised in this analysis is the Cleveland Heart Disease dataset, it includes 303 instances and 14 features including age, sex, the nature of the chest discomfort, resting blood pressure serum electrocardiographic findings, cholesterol, fasting blood sugar, maximal heart rate,exercise-induced angina, exercise-induced ST depression, slope the most intense exercise ST segment, the quantity of main vessels, thalassemia, and target (heart disease existence or absence).

By encoding categorical variables and dividing the data into training and testing sets, we will pre-process

the data. Finally, using the same hyperparameters, we will train and assess a LightGBM model and an XGBoost model. Using accuracy, precision, recall, and F1-score, we will assess the models. In terms of heart disease prediction, the performance study of the LightGBM and XGBoost models revealed that both models had good levels of accuracy, precision, recall, and F1-score. However, the LightGBM model outperformed the XGBoost model in terms of accuracy and F1-score. The models' precision and recall levels were comparable.

1.1Advantages of Light Gradient Boosting Machine:

1. Faster training and speed efficiency.
2. It employs a histogram-based approach, which accelerates the training process by bucketing continuous feature values into discrete bins.
3. Reduced memory use.
4. It uses less memory because it converts continuous values to discrete bins.
5. More accurate than any other algorithm for boosting.
6. By using a leaf-wise split strategy rather than a level-wise approach, which is the primary factor in, it produces far more complicated trees.
7. Support for big datasets.
8. When compared to the Extreme gradient boosting approach, it can handle huge datasets with an appreciable reduction in training time.
9. As a result, the gradient boosting 116 framework, which employs a tree-based learning algorithm, is quick, distributed, and high performing.

1.2 Advantages of Extreme Gradient Boosting Machine:

1. Regularization and regularised variant of gradient boosting machine are some names for it.
2. It avoids the model from overfitting thanks to built-in L1 (Lasso Regression) and L2 (Ridge Regression) regularisation.
3. Parallel processing, which makes advantage of its power and is faster than a gradient boosting machine.
4. It executes the model over many cores.
5. XGboost handles missing values automatically and similarly while working with the testing data.
6. XGboost makes it simple to determine the precise optimum number of boosting

CHAPTER 2
LITERATURE SURVEY

Information has exploded as a result of the development of comprehensive data warehouses that combine operational data with customer, supplier, and market data. Competition necessitates quick and comprehensive analysis of data viewed as a whole. The ability of users to successfully analyse and act on the information contained in more potent data warehousing systems has, however, been expanding further in the background. Tools and services for data mining are making the advancement required to close this gap. Solutions for application servers, data warehouses, thick clients, and thin clients are suggested in a thorough architectural overview. An analysis of current developments that are pertinent to the enterprise use of data mining tools and processes finishes this essay. We believe that data mining standardisation initiatives will intensify.

To successfully conduct data mining, a variety of procedures and approaches are used. The widespread use of computers has made a vast amount of data available. Experts have struggled to extract relevant and meaningful information from the ever-growing amount of data. Data mining has resulted from this. The technique of extracting hidden, unknown, and possibly important information from sizable databases is known as data mining. Finding correlations in a sizable relational database based on the various depths from which we examine it is another way to define data mining. It is a potent tool with great potential that aids businesses or organisations in increasing sales and profiting more from the information.

CHAPTER 3
SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Finding patterns in huge data sets is a process called data mining that combines techniques from machine learning, statistics, and database systems. The overall objective of data mining, an interdisciplinary subject of computer science, is to extract information from a data collection using intelligent algorithms and turn the information into a usable structure. The analytical stage of the "knowledge discovery in databases" process is known as data mining. To help clinicians make better diagnoses for therapeutic purposes, data mining technologies have been created for the efficient examination of medical data. To treat patients with diverse disorders, medical services have advanced significantly in recent years. One of the deadliest is the problem of heart disease that cannot be detected.

3.1.1 DISADVANTAGE

- Less sensitivity
- Reduced precision
- Reduced Performance

3.2 PROPOSED SYSTEM

To address the problem of heart disease prediction, this method makes use of a variety of input attributes. We suggested a method that uses lightGBM techniques to diagnose cardiac illness in an effective and precise manner. We need to take into account different performance metrics, such as accuracy, precision, recall, F1 score, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve, in order to analyse the performance of a Heart Disease Prediction model employing LightGBM and XGBoost.

3.2.1 ADVANTAGE

- High precision.
- Heart disease prognosis that is both clever and effective.
- Improved decision-making for the early detection and treatment of heart disease

3.3 METHODOLOGY

Popular machine learning method LightGBM can be applied to the prediction of heart disease. The gradient boosting framework LightGBM, which employs tree-based learning algorithms, has demonstrated strong performance in a number of machine learning tasks, including the prediction of heart disease.

You must first prepare your data, which often include cleaning, pre-processing, and feature engineering, before using lightGBM to forecast heart disease. You can use lightGBM to train a model that can forecast the chance of heart illness after your data is prepared. the fundamental procedures for using lightGBM to forecast cardiac disease: The Gather information on pertinent heart disease risk factors, such as age, gender, blood pressure, cholesterol level, etc. To make sure the data is prepared for analysis, clean and pre-process it. Divide your data into two portions: one for the model's training.

DATA MINING IN PREDICTION OF HEART DISEASE

Supervised learning and unsupervised learning are the two basic approaches used in data mining. In supervised learning, a training set is used to learn the model parameters, whereas in unsupervised learning, like in k-means clustering, no training set is used. Although prediction models provide predictions about continuous

valued, classification models categorise discrete, unordered values or data. Classification models include things like decision trees and neural networks, whereas prediction models include things like regression, association rules, and clustering.

METHOD AND MATERIALS

For forecasting cardio vascular illnesses, many strategies are adapted. In the planned study, the risk of a heart attack is predicted using lightGBM. The dataset and methodology utilised for heart attack prediction were presented in the remaining sections of the research. When using computer-aided methods to diagnose cardiac disease, the information is gathered from various sources and assessed by software programmes. Clinical decision support systems have typically been built by computers using knowledge from medical experts, with the manual conversion of this knowledge into computer algorithms.

3.4 SYSTEM ARCHITECTURE

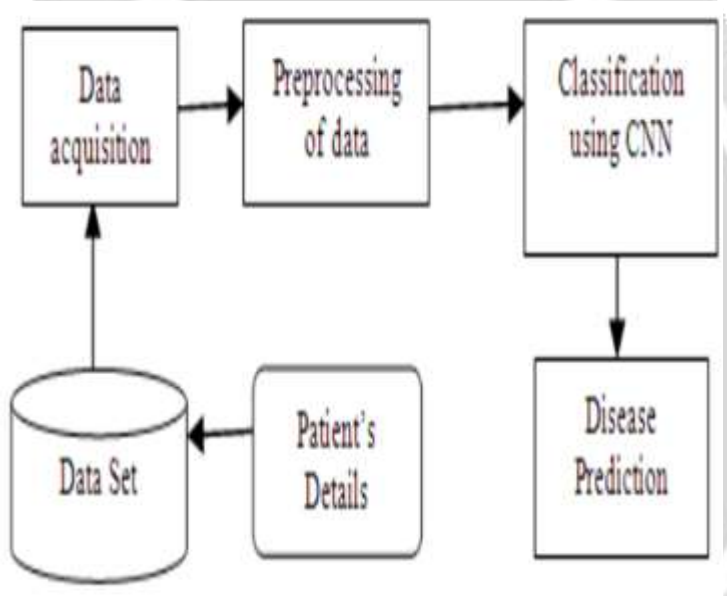


Fig.3.4 System Architecture

The patients' characteristics, such as the frequency of their chest pain and their age in years, were noted. This system makes use of 14 attributes. Other crucial measures include the latch (highest heart rate reached), blood pressure (mm Hg), serum cholesterol levels (mg/dl), and electrocardiographic result. These variables need to be examined every two hours. In the actual world, information isn't always accurate, but in the case of medical information, it's always true. Data-related discrepancies are removed using a variety of methods. Data

Light Gradient Boosting Machine(LightGBM)

LightGBM is a decision tree-based gradient boosting system that is quick, distributed, and high-performing. It is used for ranking, regression, classification, and many other machine learning problems. Because it is based on decision tree algorithms, it divides the tree leaf wise with the greatest fit, in contrast to other boosting methods that divide the tree depth- or level-wise instead of leaf-wise. The leaf-wise technique eliminates more loss when growing on the same leaf in Light GBM as a result than the level-wise strategy, leading to noticeably higher accuracy than any of the existing boosting strategies. Also, it is quite quick, which makes the word "light" appropriate.

Extreme Gradient Boosting Machine(XGBOOST)

Extreme Gradient Boosting, sometimes known as XGBoost, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting. Supervised machine learning uses algorithms.

CHAPTER 4
MODULE IMPLEMENTATION

4.1 MODULES SEPARATIONS

- Dataset Collection
- Data Pre-processing
- Data Clustering
- Classification

4.2 MODULE DESCRIPTION

4.2.1 DATASET COLLECTION

A data set (or dataset) is a collection of data, and the Collect the Heart dataset is made up of cardiac information. Most frequently, a data set correlates to the contents of a single database table or statistical data matrix, where each row corresponds to a specific member of the data set in question and each.

4.2.2 DATA-PREPROCESSING

Data filtering is the process of identifying and (or removing) missing records from a record set, table, or database. It also refers to recognising the parts of the data that are incomplete, wrong, inaccurate, or irrelevant and then replacing, changing, or deleting the soiled or coarse data.

4.2.3 DATA CLUSTERING

The task of grouping a set of items in a way that they are more similar (in some way or another) to each other than to those in other groups is known as cluster analysis or clustering (clusters). using k-Means Initial clusters in the clustering procedure are set at k. Records are assigned to the cluster with the closest centroid, or centre, at each iteration. The distance between each record and the cluster's centre is determined after each iteration.

4.2.4 CLASSIFICATION

A database or repository's procedure for categorising data uses a variety of methodologies and standards for the Product dataset. Data is categorised using LightGBM in order to be used as effectively and efficiently as possible and to forecast results. A database or business intelligence programme that offers the capacity to scan, identify, and separate data is typically used to accomplish this.

CHAPTER 5
SYSTEM SPECIFICATION

H/W SYSTEM CONFIGURATION:-

- Processor - Pentium –IV

- RAM - 4 GB (min)
- Hard Disk - 20 GB

S/W SYSTEM CONFIGURATION:-

- System of operation: Windows 7 or 8
- Application Server : python idle

5.1 PYTHON

Python is a general-purpose, interpreted programming language. Python uses garbage collection and has dynamic typing. Programming paradigms including procedural, object-oriented, and functional programming are all supported. Python is frequently called a "batteries included" language linguistics because it contains extensive standard library.

As an ABC language replacement, Python was envisaged in the late 1980s. List comprehensions and a garbage collection system that could gather reference cycles were added in Python 2.0, which was released in 2000. The 2008 release of Python 3.0 was a significant update to the language, although it is not entirely backward-compatible, and a lot of Python 2 code does not run unaltered on Python 3.

CHAPTER 6
SYSTEM DESIGN

6.1 UML DIAGRAM

Unified Modeling Language is known as UML. A general-purpose modelling language with standards, UML is used in the field of object-oriented software engineering. The Object Management Group oversees and developed the standard..

6.2 CASE DIAGRAM IN USE

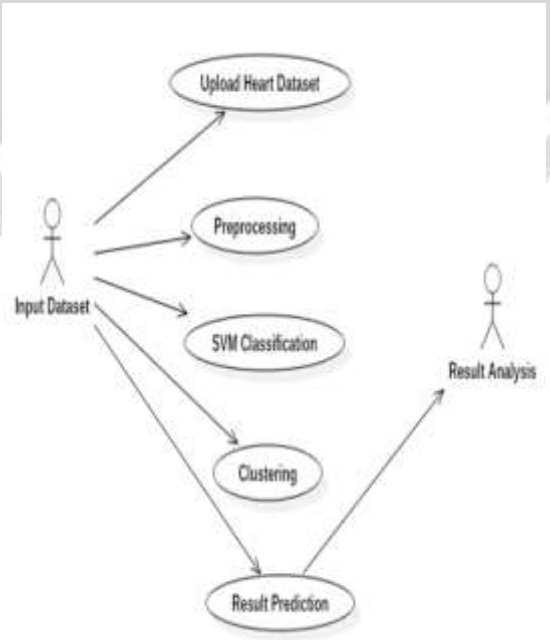


Fig.6.2 case diagram in use

6.3 SEQUENCE DIAGRAM

Developers frequently use sequence diagrams to model the interactions place when a specific use case is executed and the order in which various system components interact with one another to perform a function.

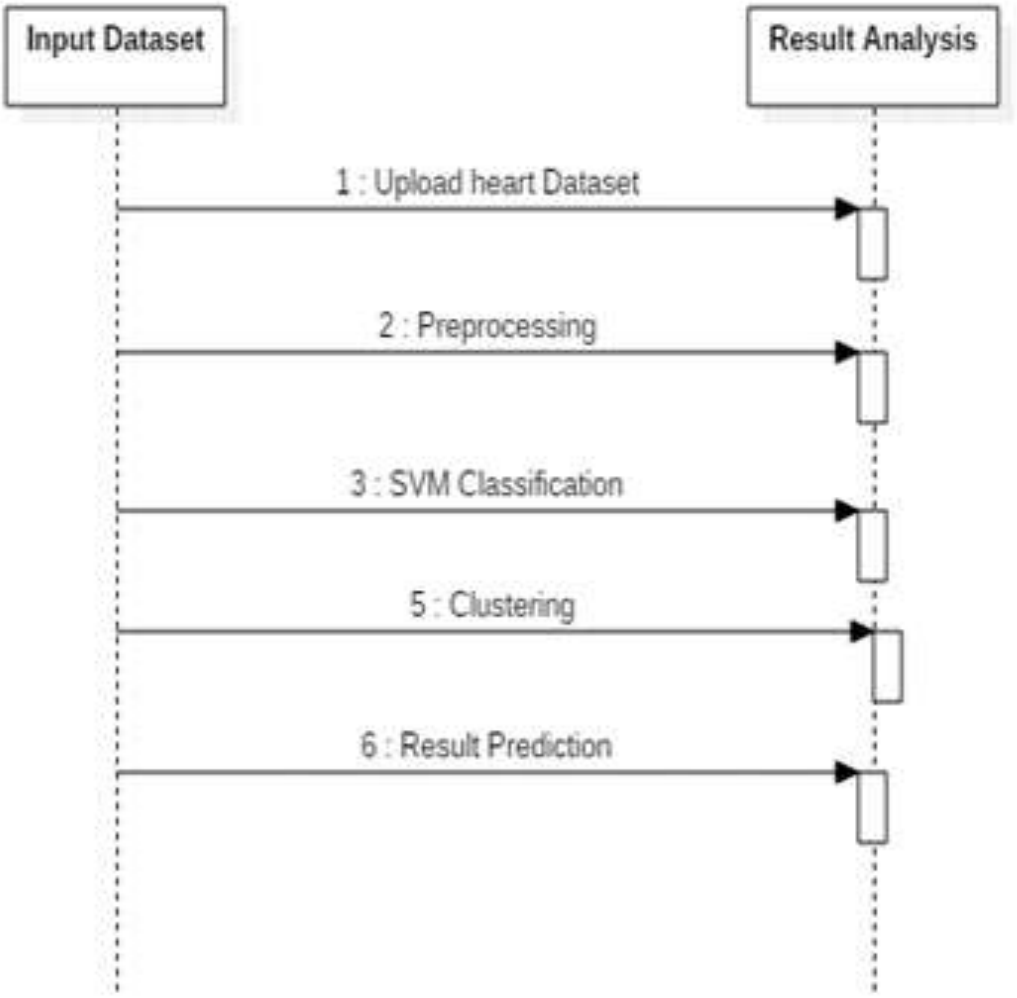
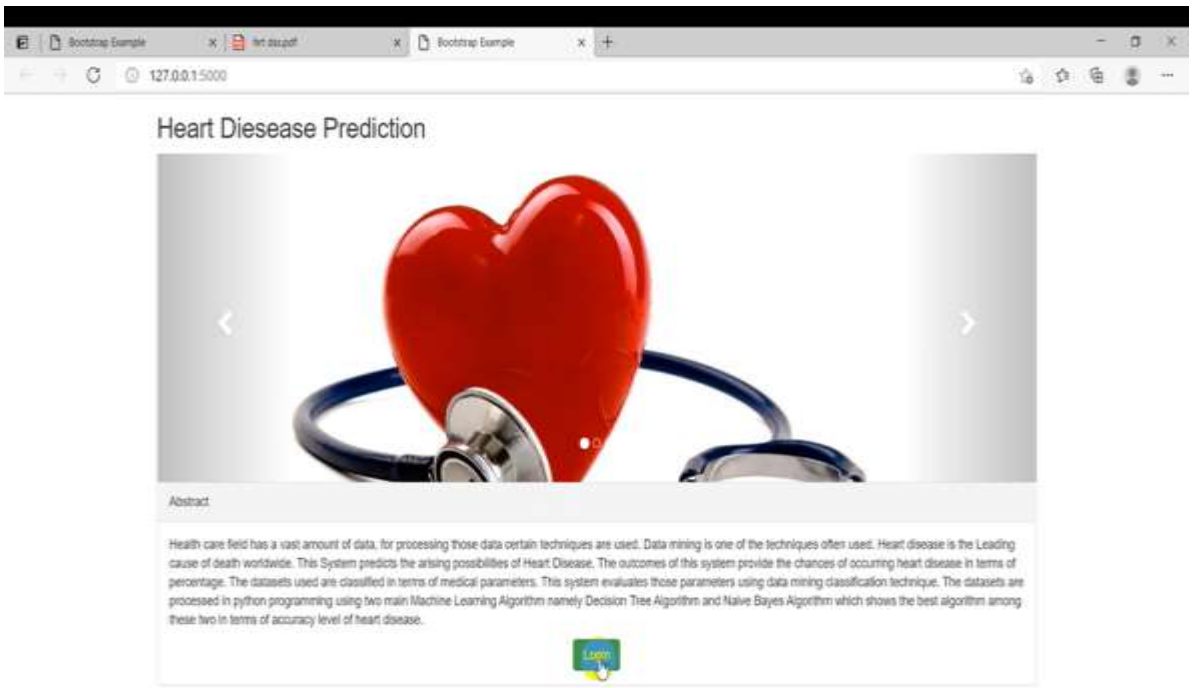
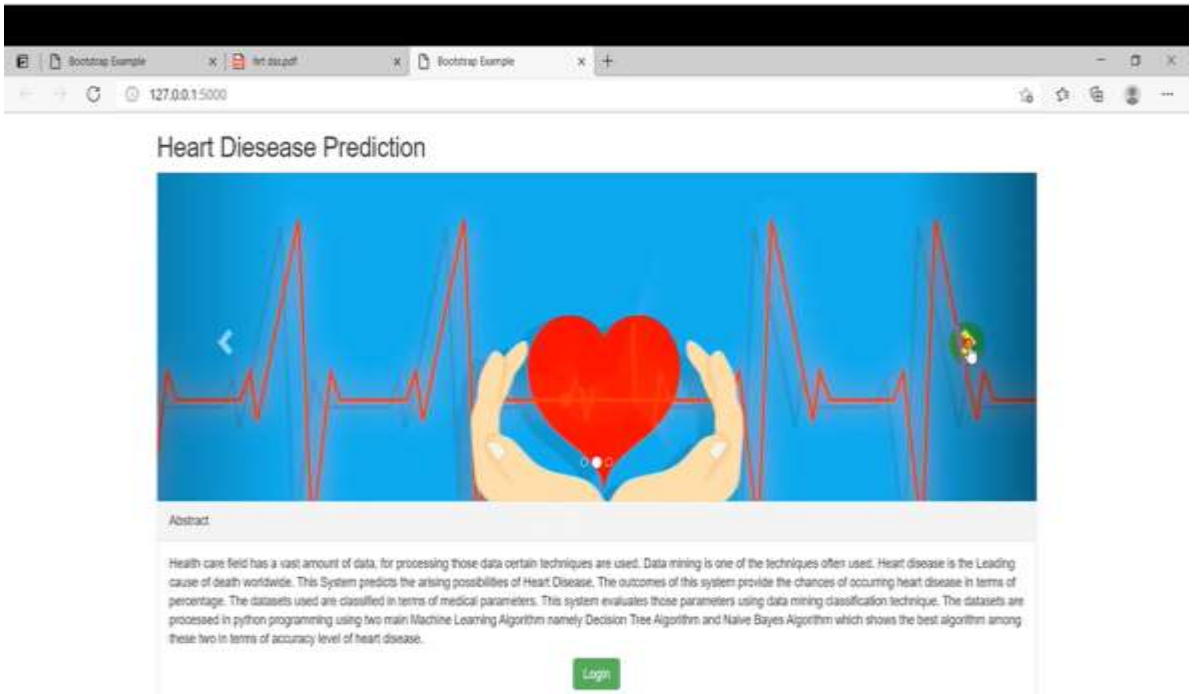
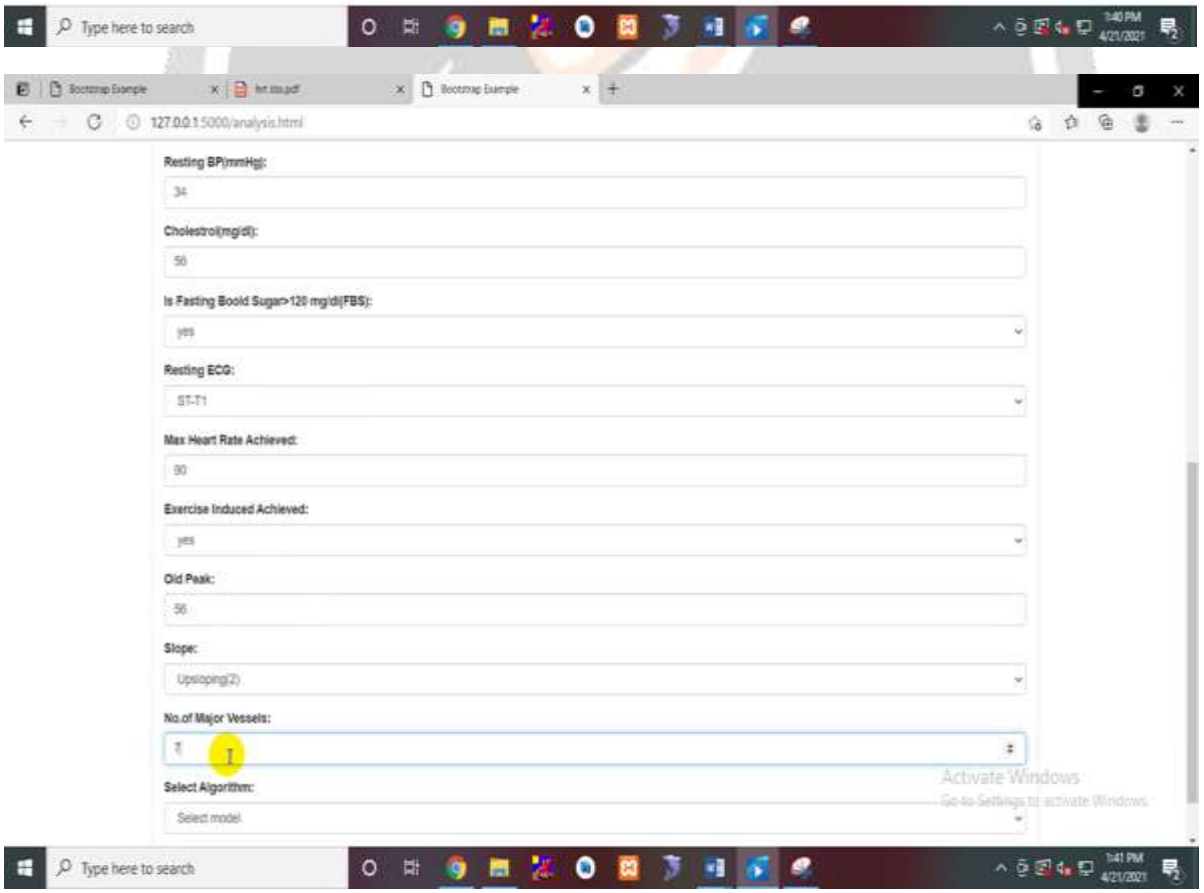


Fig.6.3 Sequence diagram

SCREENSHOTS:


Bootstrap Example
Bootstrap Example
127.0.0.1:5000
Heart Diesease Prediction
Abstract
Health care field has a vast amount of data, for processing those data certain techniques are used. Data mining is one of the techniques often used. Heart disease is the Leading cause of death worldwide. This System predicts the arising possibilities of Heart Disease. The outcomes of this system provide the chances of occurring heart disease in terms of percentage. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using two main Machine Learning Algorithm namely Decision Tree Algorithm and Naive Bayes Algorithm which shows the best algorithm among these two in terms of accuracy level of heart disease.
Login


Bootstrap Example
Bootstrap Example
127.0.0.1:5000
Heart Diesease Prediction
Abstract
Health care field has a vast amount of data, for processing those data certain techniques are used. Data mining is one of the techniques often used. Heart disease is the Leading cause of death worldwide. This System predicts the arising possibilities of Heart Disease. The outcomes of this system provide the chances of occurring heart disease in terms of percentage. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using two main Machine Learning Algorithm namely Decision Tree Algorithm and Naive Bayes Algorithm which shows the best algorithm among these two in terms of accuracy level of heart disease.
Login



CHAPTER 8
CONCLUSION

It is feasible to forecast the occurrence of heart disease with a high degree of accuracy based on the analysis utilising the LightGBM algorithm. The most significant predictors of heart disease, according to the feature importance analysis, were the type of chest pain, the maximal heart rate reached during exercise, and the quantity of major blood vessels coloured by fluoroscopy. This knowledge can be applied to create focused preventative plans and enhance patient outcomes. Furthermore, using machine learning algorithms like LightGBM can significantly increase the efficacy and accuracy of cardiac disease prediction, resulting in early interventions and better patient outcomes..

REFERENCES

[1] A.L. Bui, T. B. Horwich, and G. C. Fonarow, ``Epidemiology and risk prole of heart failure," Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.
[2] M. Durairaj and N. Ramasamy, ``A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255260, 2016.
[3] L. A. Allen, L.W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ``Decision making in advanced heart failure: A scientic statement from the American heart association," Circulation, vol. 125, no. 15, pp. 19281952, 2012.
[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ``Innovative artificial neural networks-based decision support system for heart diseases diagnosis," J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.
[5] Q. K. Al-Shayea, ``Artificial neural networks in medical diagnosis," Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150154, 2011.
[6] J. Lopez-Sendon, ``The heart failure epidemic," Medicographia, vol. 33, no. 4, pp. 363369, 2011.
[7] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P.W. F.Wilson, and Y. J. Woo, ``Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association," Circulation, vol. 123, no. 8, pp. 933944, 2011.
[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ``Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantication of average Parkinson's disease symptom severity," J. Roy.