



Dissertation on

“Heart Disease Prediction Using ML Models”

Submitted in partial fulfilment of the requirements for the award of degree of

**Master of Technology in
Computer Science & Engineering**

UE21CS7A1B – Project Work Phase - 2

Submitted by:

SHARADA.A.

PES1PG21CS034

Under the guidance of

Dr. Priyanka H

Associate Professor, CSE
PES University

January - May 2023

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India



PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING CERTIFICATE

This is to certify that the dissertation entitled

‘Heart Disease Prediction Using ML Models’

is a bonafide work carried out by

SHARADA A

PES1PG21CS034

in partial fulfilment for the completion of Fourth semester Project Work Phase - 2 (UE21CS7A1B) in the Program of Study - Master of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan 2023 – May 2023. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 4th semester academic requirements in respect of project work.

Signature
Dr. Priyanka H
Associate Professor, CSE

Signature
Dr. Shylaja S S
Chairperson

Signature
Dr. B K Keshavan
Dean of Faculty

External Viva

Name of the Examiners

Signature with Date

1. _____

2. _____

DECLARATION

We hereby declare that the Project Phase - 2 entitled “**Heart Disease Prediction Using ML Models**” has been carried out by us under the guidance of Dr .Priyanka H, Assoc. Professor and submitted in partial fulfilment of the course requirements for the award of degree of **Master of Technology in Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2023. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES1PG21CS034

Sharada.A.

Signature

ACKNOWLEDGEMENT

I would like to express my gratitude to Dr.Priyanka H, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE21CS7A1B - Project Work Phase – 2.

I am grateful to the project coordinators, Prof. Revathi G P for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Shylaja S S, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

Heart disease, one of the major causes of mortality worldwide, can be mitigated by early heart disease diagnosis. A clinical decision support system (CDSS) can be used to diagnose the subjects' heart disease status earlier. This study proposes an effective heart disease prediction model (HDPM) for a CDSS which consists of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect and eliminate the outliers, a hybrid Synthetic Minority Oversampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution and XGBoost to predict heart disease. Two publicly available datasets (Statlog and Cleveland) were used to build the model and compare the results with those of other models (naive bayes (NB), logistic regression (LR), multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), and random forest (RF)) and of previous study results.

The results revealed that the proposed model outperformed other models and previous study results by achieving accuracies of 95.90% and 98.40% for Statlog and Cleveland datasets, respectively. In addition, we designed and developed the prototype of the Heart Disease CDSS (HDCDSS) to help doctors/clinicians diagnose the patients'/subjects' heart disease status based on their current condition. Therefore, early treatment could be conducted to prevent the deaths caused by late heart disease diagnosis.

Contents

DECLARATION	1
ACKNOWLEDGEMENT	2
ABSTRACT	3
CHAPTER-1	1
1.1 Motivation	2
1.2 Scope of the Project	3
1.3 Objectives of the Project	3
CHAPTER-2	4
PROBLEM STATEMENT	4
CHAPTER-3	5
LITERATURE SURVEY	5
CHAPTER-4	14
PROJECT REQUIREMENT SPECIFICATION	14
CHAPTER-5.....	19
METHODOLOGY	19
Train your model	23
Step 4: Evaluate your trained model	23
Step 5: Model Consumption	24
CHAPTER-6.....	38
RESULTS AND DISCUSSION	38
CHAPTER-7.....	48
CONCLUSION AND FUTURE WORK	48

LIST OF FIGURES

Figure No.	Title	Page No.
1.1	Methodology which is implemented in this project	2
3.1	The Proposed system of HDPM and HDCDSS	5
4.1	Use-case diagram	15
5.1	Architecture of the Proposed Model	19
5.2	Building an ML Model involves the following high-level steps	20
5.3	sequence diagram of heart disease predictor	25
5.4	ER diagram of Heart disease predictor	27
5.5	Architecture of the system	28
5.6	Data flow diagram (level 0)	28
5.7	Data flow diagram (level 1)	29
5.8	data flow diagram (level 2)	30
5.9	output of dataset by using logistic regression	31
5.10	output of extracting dependent and independent variables of x and y	32
5.11	output of test set	33
5.12	Output of training set	33
5.13	Example of implementing XGBOOST Classifier	34

LIST OF TABLES

Table No.	Title	Page No.
5.1	sample example for ML.NET component	21
5.2	Cleveland dataset database description	26
5.3	confusion matrix and predictive measures of both logistic regression and XGBOOST classifiers.	35
7.1	Comparing the accuracies both with confusion matrix and another comparison study	49

CHAPTER-1

INTRODUCTION

To find and treat cardiac problem is extremely challenging till the upcoming technology was not existed. The effectiveness in searching problem and giving accurate medication can safe-guard the life of the people. Consistent of EU public for cardiac disease are identified other countries cardiac problem. Identifying this problem is naturally identifying in medical files for subject, appearance identification docs & analyzing for concerned effects of a Doctor. As the output got from this detection method aren't tolerant but, costly challenging research. The ML predictive models needs accurate docs. Many Pre-processing processes such as removal of unwanted value instances .etc..

The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machinelearning algorithms. Heart is one of the most extensive and vital organs of human body so the care of heart is essential.

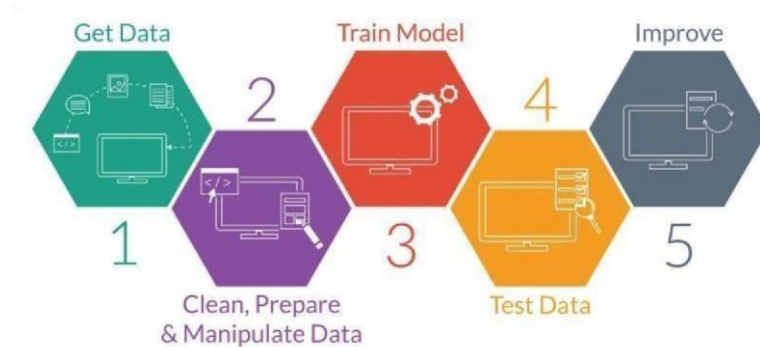


Fig 1.1: Methodology which is implemented in this project

The above fig 1.1 is a great visual of a machine learning project A to Z. The first step before we began coding is to understand the problem we are trying to solve and get the available data. In this project, it is used the heart dataset which is already available in github and Kaggle too.

1.1 Motivation

In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. The purpose of this data exploration and predictive analysis is to better understand which health factors affect a patient's risk for heart disease. To accomplish this, an introduction to the data will be made, along with a graphical analysis of the health factors in the dataset. Heart disease can be predicted based on various symptoms such as age, gender, pulse rate etc. Data analysis in healthcare assists in predicting diseases, improving diagnosis, analyzing symptoms, providing appropriate medicines, improving the quality of care, minimizing cost, extending the life span and reduces the death rate of heart patients.

1.2 Scope of the Project

An Effective Heart Disease Prediction Model for a Clinical Decision Support System. Heart disease, one of the major causes of mortality worldwide, can be mitigated by early heart disease diagnosis. A clinical decision support system (CDSS) can be used to diagnose the subjects' heart disease status earlier.

1.3 Objectives of the Project

The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease and 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: vlaues
Positive result = 1, the patient will be diagnosed with heart disease.

CHAPTER-2

PROBLEM STATEMENT

This Cardiac disease can be managed efficiently by combining the changing in lifestyles, drugs and sometimes operations. With the right medication, troubles the HD can be reduction in cost of complicated medication and actions in the heart being improving. The calculated outcomes be used to avoid . The complete parameters to this action to be predicted correctly with some exams and values the appearance of cardiac problems. Values consider form the prior base of tests and gives proper results more or less. Big data keeps high energy of the health industry to able the models systematic find inefficienct & good habits to move forward and decrease money simultaneously could apply to as much as 30% of complete healthcare is expending. This leads to implementing in other industries and fields.

Among these fields by discovering is successful in applying the big data in hyper available fields like e-commerce, business and in heath. Thus, there will be very less of effect of analysis equipments to identify unknown relations and trending of big data in nigeria races. A Heart Disease prediction algorithm can help detect heart diseases early, and maintain a healthy lifestyle.

A group of researchers has developed an algorithm to predict the risk of heart disease based on four factors – age, gender, cholesterol levels, blood pressure. Much like any other disease, heart disease can come in many different forms. Some people may have risk factors that predispose them to health problems while others might not.

CHAPTER-3

LITERATURE SURVEY

3.1 HDPM and HDCDSS

This HDPM i.e, Heart Disease Prediction Model is explained in N. L. Fitriyani et.al,[1] and have also explained about implementation of CDSS (clinical decision support system. The Model was constructed in providing high efficiency predicting for the current or absent of cardiac problem is in the present status in the subject. The design in Fig 3.1 depicts development of required Model.Initially, the cardiac problem datasets will be collecting.Next, file will be sendd for preprocessing and feature selection. Next, an XG Boost Algo is used to study of a trained dataset & get the model.Lastly, the performance metrics are executed to calculated the working of the required one and constructed Model will be deployed into CDSS.

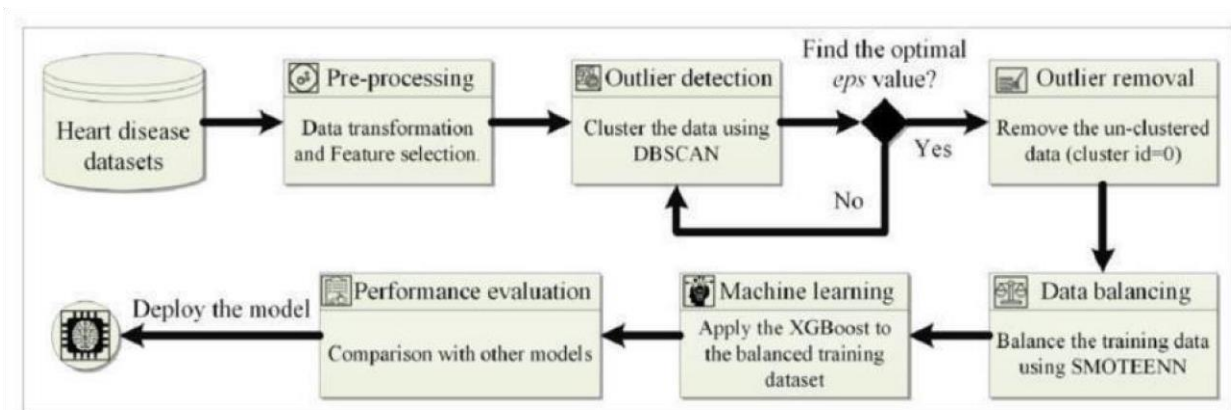


Fig 3.1: The Proposed system of HDPM and HDCDSS

3.2 GridSearchCV

In [2], G. N. Ahmad, et.al, paper, various Machine Learning algorithms such as LR, KNN, SVM, and GBC, together with the GridSearchCV, predict cardiac disease. The system uses a 5-fold crossvalidation technique for verification. A comparative study is given for these four methodologies. The Datasets for both Cleveland, Hungary, Switzerland, and Long Beach V and UCI Kaggle are used to analyze the models' performance. It is found in the analysis that the Extreme Gradient Boosting Classifier with GridSearchCV gives the highest and nearly comparable testing and training accuracies as 100% and 99.03% for both the datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle). Moreover, it is found in the analysis that XGBoost Classifier without GridSearchCV gives the highest and nearly comparable testing and training accuracies as 98.05% and 100% for both the datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle).

3.3 ECG (Echocardiogram)

In [3], D. Bertsimas, et.al, explains that how we can implement novel technology to extract ECG records. . In the last decades, there has been increasing evidence of how Machine Learning can be leveraged to detect such anomalies, thanks to the availability of Electrocardiograms (ECG) in digital format. New developments in technology have allowed to exploit such data to build models able to analyze the patterns in the occurrence of heart beats, and spot anomalies from them. In this work, we propose a novel methodology to extract ECG-related features and predict the type of ECG recorded in real time (less than 30 milliseconds). Our models leverage a collection of almost 40 thousand ECGs labeled by expert cardiologists across different hospitals and countries, and are able to detect 7 types of signals: Normal, AF, Tachycardia, Bradycardia, Arrhythmia, Other or Noisy. We exploit the XGBoost algorithm, a leading machine learning method, to train models achieving out of sample F1 Scores in the range 0.93 – 0.99. To our knowledge, this is the first work reporting high performance across hospitals, countries and recording standards.

3.4 E-Healthcare

In [4], J. P. Li, A. U. Haq, et.al, 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 such as Relief, Minimal

redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. 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 hyperparameter 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.

3.5 Hybrid ML Techniques

In [5], S. Mohan, et al., a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

3.6 SDL Technique

In [6], Mythili, T. et al. proposes a combination of three ML techniques like SVM, Decision tree and Logistic Regression. Here proposes a rule based model to compare the accuracies of applying rules to the individual results of support vector machine, decision trees, and logistic regression on the Cleveland Heart Disease Database in order to present an accurate model of predicting heart disease.

3.7 Related work

Paper 1: N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in *IEEE Access*, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511

Description: The proposed HDPM was developed to provide high performance prediction in the presence or absence of heart disease given the current condition of the subjects. First, the heart disease datasets are collected. Second, the data pre-processing for data transformation and feature selection are conducted. Third, the DBSCAN-based outlier detection method is applied to find the outlier data given the optimal parameter. Fourth, the detected outlier data are then removed from the training dataset. Fifth, the data balancing based on SMOTE-ENN method is used to balance the training dataset. Sixth, the XGBoost-based MLA is used to learn from the training dataset and generate the HDPM. Finally, the performance metrics are presented to evaluate the performance of the proposed model and the generated HDPM is then implemented within the CDSS. In this study, it has used 10-fold CV process to prevent the overlapping. This CV allows designs to study of various training sets data by repeating examples; by increasing values will be verified and possible, for useful purpose so can avoid overlapping. The prototype of the web-based Heart Disease Clinical Decision Support System (HDCDSS) was developed to provide a simple and convenient way for medical clinicians to diagnose subjects/patients based on their current condition.

Paper 2: G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in *IEEE Access*, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

Description: The main contribution of this paper was to use modern machine learning techniques to construct an intuitive medical prediction system for the diagnosis of heart disease. Different types of machine learning classifier algorithms were trained in this study, including logistic regression (LR), K-nearest neighbours (K-NN), support vector machine (SVM), and Gradient Boosting Classifier (GBC) with and without GridSearchCV, to select the best predictive model for accurate heart disease detection at an early stage. To achieve the ideal collection of attributes that strongly influenced the performance of the classifiers when predicting the target class, four model selection strategies were used, including

correlationbased feature subset evaluator and Gradient Boosting Classifier evaluator. Finally, the whole attribute set and optimal sets obtained via attribute evaluators were used to tune the hyperparameter “GridSearchCV” in the GBC classifier. One of the most challenging issues in medicine is predicting cardiac disease. It takes a lot of time and effort to figure out what’s causing this, especially for doctors and other medical professionals. Researchers used a range of algorithms, including LR, KNN, SVM, and GBC, as well as the GridSearchCV, to predict cardiac disease.

Paper 3: D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.

Description: In this manuscript is propose a novel methodology to identify heart anomalies from a newly recorded ECG. The predictive process can be summarized as: signal pre-processing, feature extraction, model training, calibration and evaluation. We design a feature extraction pipeline that crafts 110 features, which we leverage to train five different models on a collection of three datasets. Our models prove to have extremely strong performance when making prediction on unseen data, but are also able to generalize across datasets with ECGs recorded in different settings, and with population having inherently different characteristics. In addition, this approach has showed to be effective for very different kind of heart abnormalities: Normal, Atrial Fibrillation, Tachycardia, Bradycardia, Other (non-specified), Arrhythmia and Noisy. In order to further improve our models’ reliability, we calibrate our models using Temperature Scaling to minimize the Expected Calibration Error. Our work confirms that directly analyzing the characteristics of the QRS complex leads to very accurate predictions. This can have an enormous potential impact on the lives of people suffering from heart diseases. In fact, we envisioned our work to be applied in a real time setting, with a wearable device that can constantly monitor the heartbeat of the patients at risk. By designing our experiments to analyze a single lead of a common ECG, we have a good approximation of the input of a given wearable, thus achieving our initial aim without lowering the predictive power of our algorithms. We perform extensive analysis to assess the viability of our models in a real-time setting, and we find that for signals shorter than a minute (the average ECG length is 30 seconds) it takes less than 30 milliseconds from the moment in which the signal is recorded to the final model

prediction. As a result, our models prove to be a fast and reliable aid in the important task of detecting heart anomalies from the ECGs of patients who can then be directed to trained experts for further analysis.

Paper 4: J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

Description: Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. 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 such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. 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 hyperparameter 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. The suggested diagnosis system (FCMIM-SVM) achieved good accuracy as compared to previously proposed methods. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

Paper 5: C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

Description: Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give into finding predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM). This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL) Premature Ventricular Contraction (PVC)), and Second degree block (BII) to find out the exact condition of the patient in relation to heart disease.

Paper 6: Mythili, T. et al. "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)." International Journal of Computer Applications 68 (2013): 11-15.

The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. Research has attempted to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using homogenous data mining techniques. Recent research has delved into amalgamating these techniques using approaches such as hybrid data mining algorithms. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of support vector machine, decision trees, and logistic regression on the Cleveland Heart Disease Database in order to present an accurate model of predicting heart disease.

Paper 7: D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.

Heart disease is a spontaneous, treacherous, and fatal disease. It is a group of several states that result in abnormal functioning of the heart. Based on the several pathology test report heart disease is identified by a doctor. The manual heart disease prediction is time consuming and error prone. Therefore, in the present study an automated system based on the performance analysis of several machine learning techniques has been developed. First, the well-known machine learning algorithm Support Vector Machine (SVM), KNearest Neighbor (KNN), Naïve Bayes and Random Forest applied on the dataset for the prediction of heart disease. To avoid bias performance 3-fold cross validation is applied. The highest average accuracy of 87.78% is obtained by the Naïve Bayes. The performance of the model is acceptable. Further, we have applied genetic algorithm on the dataset to optimize the features. After, optimization the highest average accuracy of 96% is achieved by the naïve Base.

Paper 8: P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.242.

Description: Prediction of patterns to prevent and control diseases is a challenging and a prominent requirement in medical domain. In this paper, we propose a machine learning framework to predict the possibility of having heart disease using various algorithms. The framework is executed using five algorithms Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree (LMT). Cleveland dataset is used for training and testing the model. The dataset is preprocessed followed by feature selection to select most prominent features. The resultant dataset is then used for training the framework. The results are combined and show that Random forest gives maximum accuracy.

Paper 9: D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752947.

Description: The application of Machine learning algorithms to predict diseases is one of the finest methodology to reduce heavy work load on doctors and related medical staff. Based on the World Health Organization (WHO) report, about 85% heart disease deaths are due to Heart Attacks and Heart Strokes. In India the average death rate due to cardiovascular diseases is about 272 per 10,000 population which is greater than global average of 235 per 10,000 population. From the recent survey results, which was released by the Union Ministry of Family and Health Welfare (MoFHW), the Diabetes disease positive ratio is gradually increasing in India. 11.5 percent people were tested positive for Diabetes among urban and rural Indians who are with age 45 and above. Even there is availability of wide range of treatment methods of heart stroke patients & diabetes, Heart attack with Diabetes is the major cause of death in all parts of rural and urban areas of entire India. There are several factors causing heart and diabetes problems which include Age, Gender, Blood Pressure, Glucose levels, Skin thickness and Insulin.

Paper 10: Reddy, Kummita Sravan Kumar and K. V. Kanimozhi. "Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM." 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (2022): 1-5.

Description: In comparison to Support Vector Machine, the major goal is to forecast the Novel Intelligent model for Heart Disease prediction using Dynamic KNN (SVM). **Materials and Procedures:** Two machine learning methods, Dynamic KNN (N=92) and Support Vector Machine (N=92), are used to predict heart disease. Dynamic KNN is a simple algorithm used for disease prediction. Heart disease dataset is used for disease prediction. For each group 20 samples are taken and it is divided into training and testing dataset. **Result and Discussion:** Accuracy of Dynamic KNN is 84.44% and Support Vector Machine is 67.21%. There exists an analytical significant difference between Dynamic KNN and SVM. **Conclusion:** Dynamic KNN appears to perform significantly better than Support Vector Machine for Novel Heart Disease Prediction.

CHAPTER-4

PROJECT REQUIREMENT SPECIFICATION

Generally, specification requirements consists all requirements and specifications for the project as well as the potential contractor. This also includes the description of desired functions, features, and services, interfaces of the software solution. The contractor must poses or develop this catalog of requirements.

4.1 Project Scope

The proposed system acts as an decision support system and can convince to be an aid for physicians with the diagnosis. The algorithm, “KNN” classifies the condition of the patient into categories that either have the symptoms of the heart disease or not.

4.2 Product Perspective

Information from learning of population has helped in prediction of cardiac problems, based on blood pressure, smoking habit, blood pressure levels and cholestrol, diabetes. Researchers have used these adapted in predicted algorithms form of simplified marks card that allow subjects for evaluate of challenges facing from this cardiac diseases.

4.3 Product Features

Heart disease, one of the major causes of mortality worldwide, can be mitigated by early heart disease diagnosis. A clinical decision support system (CDSS) can be used to diagnose the subjects’ heart disease status earlier. This study proposes an effective heart disease prediction model (HDPM) for a CDSS which consists of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect and eliminate the outliers, a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTEENN) to balance the training data distribution and XGBoost to predict heart disease. The results revealed that the proposed model outperformed other models and previous study results by achieving accuracies of 95.90% and 98.40% for Statlog and Cleveland datasets, respectively.

4.4 User classes and characteristics

The main software features and information flow between system participants are covered in this section.

The user and the interagent are included among the participants. Below fig 4.1 shows the use case diagram.

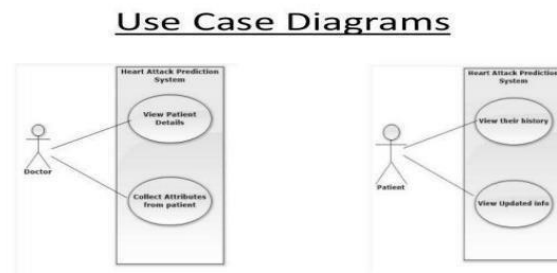


Fig 4.1 : Use-case diagram

4.5 Operating Environment

4.5.1 Software requirements

- Operating System: Windows 10 and above
- Technologies: Anaconda (Jupyter Notebook)
- Coding Languages: Python

4.5.2 Hardware requirements

- Processor: I5
- RAM: 4-16GB
- Input devices: mouse, keyboard
- Hard disk: 20 and above GB
- Speed: 1.2GB and above

4.6 General Constraints, Assumptions and Dependencies

- Here, XGBoost is implemented using the “XGBoost Version 0.81” python library. The outlier data from cardiac disease training datasets are eliminated by using the DBSCAN method, and SMOTEENN is used to balance the training dataset.
- Ultimately, XGBoost is used to study from the training dataset and generate the model also, measured 5 performance metrics to compare the performance of the proposed model with that state of art models and old observed results.
- In addition, it is ensured the applicability of the proposed model by implementation of the model into the HDCDSS to identify the patients based on their present condition. Also will employ 10fold cross validation to generate the models for all classification models, with the final performance metric being the average.
- It is implemented all the classification models in “Python Version 3.6.5” by utilizing 3 libraries:”sklearn V0.20.2”, learn-imbalanced version-0.4.3 & “XGBoost V0.81” and performs experiments in the monitor on friendly versions and model agents works also other versions and any bits processors.
- This System will be constructed on “Python V3.6.5” for using “Flask V1.0.2” as “Python Web Server Gateway Interface (WSGI)” with “Bootstrap V3.3.7” for data representation, while the proposed HDPM was loaded using Joblib V0.14.1 and XGBoost V0.81. The subjects’ data and the prediction results were stored into MongoDB by using “Pymongo V3.7.1.” MongoDB was selected since it has been widely adopted in the healthcare department .
- Classification Supervised experiments has being implemented calculated classifiers. 1st phase it is implemented all the classification models in “Python Version 3.6.5” by utilizing 3 library packages like “sklearn V0.20.2”, along with various ML library with any model and bit processors.

4.7 Risks

The risk manager must map the information items with their risk statements, citations, and policies before employing the recommendation engine. The following are the main characteristics of the recommendation engine which uses the idea of information objects to suggest compliances and hazards. By comparing the information objects of the libraries, it matches the business applications with the risk and compliance libraries and gives IT risk managers the ability to specify and map risk and compliance libraries based on the data present in the application, Shortens the time needed for risk and compliance library mapping.

4.8 Functional Requirements

An explanation of the service that the software must provide is contained in a functional requirement. It describes a piece of software or a software system. A function is nothing more than the inputs, behaviour, and outputs of the software system. A system's likely function can be determined by a computation, data manipulation, business process, user interaction, or any other specialised feature. Independent of its use, functional requirements explain how a system interacts with its surroundings. Some steps are as follows: • Making use of the algorithms to analyse train data

- Show the model's recommendations.

4.9 External interface requirements

Our recommendation system's user interface has already been developed and deployed in more extensive music streaming and downloading systems. As a result, we won't be implementing any user interfaces, but we may change our recommendation system to work with other systems and interfaces. We solely utilise a straightforward user interface to display system suggestions. Before integrating a web service with an already-existing music streaming and download online application, this user interface offers the opportunity to present recommendations. The user logs in to the main system before beginning to play and download songs that were made using the data from our recommender web service. As a result, the output of our recommendation system is displayed in the main interface of the integrated web application.

4.10 Non-functional requirements

Non-Functional Requirements define a software system's quality attribute. They assess the software system according to non-functional criteria such as responsiveness, usability, security, portability, and other criteria that are essential to the software system's success. The nonfunctional requirement "how quickly does the page load?" is an example. Systems that don't meet non-functional requirements may not be able to meet user needs. You can impose constraints or limitations on the system architecture across different agile backlogs using non functional requirements like Accuracy,Reliability, Flexibility and also QOD.

Some other non-functional requirements are as follows:

- Quality, safer, feedback time so on are performance steps.
- comes under Operating systems
- The operating system like designing,recuiting,supervising,entertaining and operabaility is described in interface constraints.
- Immediate or long-term costs is stored in economic constraints.
- The measurements words like keep up,improve capacity, port in lifecycle requirements .

CHAPTER 5

METHODOLOGY

In this chapter it will complete implementation of project coding and the steps is involved for GUI , the models of machine learning and screenshots of output and resultant of comparison study accuracy. Also, the exact design methodology which is held in the whole implementation part.

5.1 Overview

The complete project methodology will be explained below fig 5.1 i.e., architecture of proposed model the first section is collection of dataset, second section is to manipulating the data by extraction of features or attributes, later removal of data by using DBSCAN (Data based spatial clustering of Applications with Noise) again the data will be balanced by using SMOTTEENN and result will be predicted. The comparison of ML models ie., XGBOOST algorithm and Logistic regression along with involving the confusion matrix for getting better accuracy and finally the output will be displayed the one which will getting highest accuracy percentage.

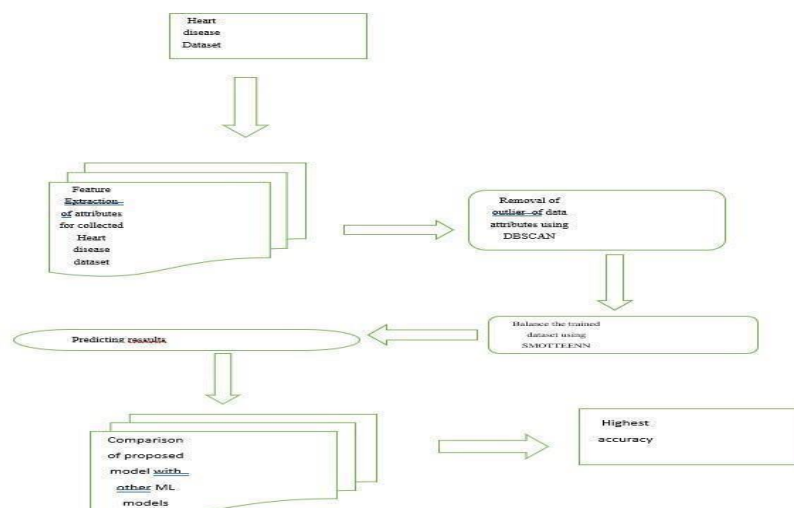


Fig 5.1: Architecture of the Proposed Model

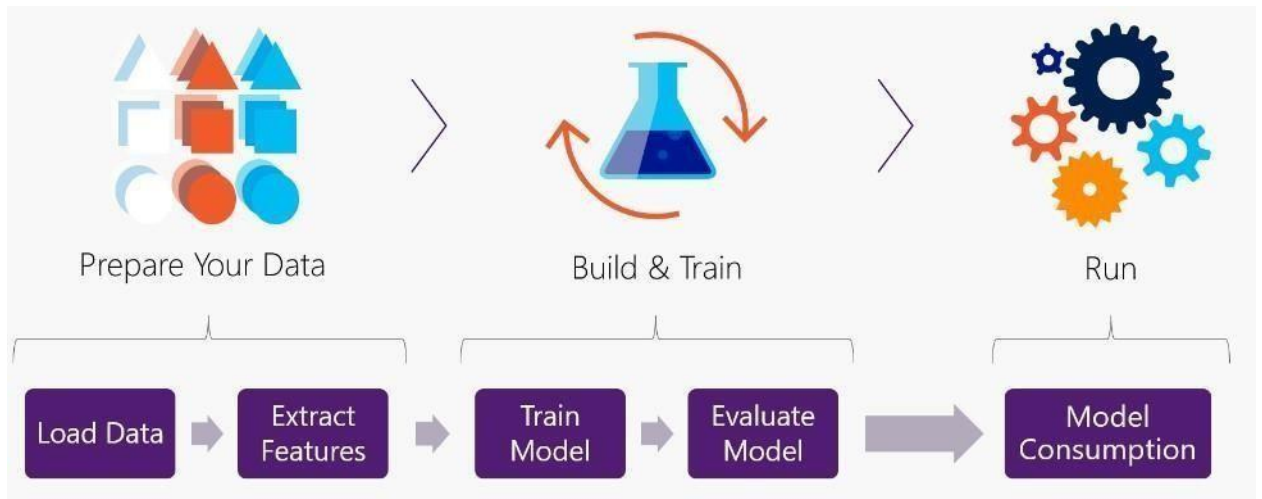


Fig 5.2: Building an ML Model involves the following high-level steps

The above fig 5.2 shows how the steps will be followed while building and training ML model with the flow.

Step 1: Load data

Get started

When building a model with ML.NET you start by creating an ML Context or environment. This is comparable to using Db Context in Entity Framework, but of course, in a completely different domain.

The environment provides a context for your ML job that can be used for exception tracking and logging.

```
var env = new LocalEnvironment();
```

One of the most important things is, as always, your data! Load a Dataset into the ML pipeline to be used to train your model. In ML.NET, data is similar to a SQL view. It is lazily evaluated, schematized, heterogenous. In this example, the sample dataset looks like this:

Table 5.1 :sample example for ML.NET component

Toxic (Label)	Comment (Text)
Toxic (Label)	Comment (Text)
1	-RUDE- Dude, you are rude ...
1	— OK! — IM GOING TO VANDALIZE ...
0	I also found use of the word "humanists" confusing ...
0	Ooooooh thank you Mr. DietLime ...

To read in this data you will use a data reader which is an ML.NET component as shown in above table 5.1. The reader takes in the environment and requires you to define the schema of your data. In this case the first column (Toxic or Label) is of type Boolean (meaning also the prediction) and the second column (Comment or Text) is the feature of type text/string that we are going to use to predict the sentiment on.

```
var reader = new
TextLoader(env,
new TextLoader.Arguments()
{
    Separator = "tab",
    HasHeader = true,
    Column = new[]
    {
        TextLoader.Column("Label", DataKind.Bool, 0),
        TextLoader.Column("Text", DataKind.Text, 1)
    }
});

//Load training data var trainingDataView = reader.Read(new
MultiFileSource(TrainDataPath));
```

Your data schema consists of two columns:

- A boolean column (Label) which is the sentiment (Toxic/Negative or NonToxic/Positive) and positioned as the first column.
- A text column (Text) which is a comment showing certain sentiment and is the feature we use to predict.

Note that this case, loading your training data from a file, is the easiest way to get started, but ML.NET also allows you to load data from databases or in-memory collections.

Step 2: Extract features (transform your data)

Machine learning algorithms understand featurized data, so the next step is for us to transform our textual data into a format that our ML algorithms recognize. In order to do so we create an estimator of type

TextTransform which featurizes the text converting it to numeric vectors, as shown in the following snippet:

```
var pipeline = new TextTransform(env, "Text", "Features");
```

Step 3: Train your model

Add a selected ML Learner (Algorithm)

Now that our text has been featurized, the next step is to add a learner. In this case we will use the LinearClassificationTrainer learner. For this step, you just need to append the learner to the estimators chain or flexible pipeline, while specifying what column is the feature and what column is the label or goal to predict, like in the following code:

```
var pipeline = new TextTransform(env,                                     new  
LinearClassificationTrainer.Arguments(),  
                                "Text", "Features")  
                                .Append(new LinearClassificationTrainer(env, "Features",  
"Label"));
```

The learner/trainer takes in the featurized Text (Features) and the Label as input parameters for learning from the historic data.

Once the estimator has been defined, you train your model using the Fit() API while providing the already loaded training data. This returns a model which you can use for predictions.

```
var model = pipeline.Fit(trainingDataView);
```

Note that the pipeline is a chain of estimators. An Estimator is an object that learns from data. A transformer is the result of this learning. A good example is precisely when training the model with pipeline.Fit(), which learns on the training data and produces a machine learning model which is an special case of transformer.

Step 4: Evaluate your trained model

Now that you’ve created and trained the model, evaluate it with a different dataset for quality assurance and validation with code similar to the following:

```
// Evaluate the model //Load evaluation/test data  var testDataView =  
reader.Read(new MultiFileSource(TestDataPath)); var predictions =  
model.Transform(testDataView);  var binClassificationCtx = new  
BinaryClassificationContext(env); var metrics =  
binClassificationCtx.Evaluate(predictions, "Label");  
  
Console.WriteLine($"Model's Accuracy: {metrics.Accuracy:P2}");
```

The code snippet implements the following:

- Loads the test dataset.
- Creates an additional context, since we are performing a binary classification ML task.
- Evaluates the model and create metrics.
- Shows the accuracy of the model from the metrics.

And now you have a trained and validated model for use in your applications and services.

Step 5: Model Consumption

At this point you can predict with test/sample data by consuming the model you just created and trained. The following code is a sample you would write in your “production” application when predicting something by scoring with the model:


```
// Create the prediction function var predictionFunc = model.MakePredictionFunction<SentimentIssue,
SentimentPrediction>(env);

var resultprediction = predictionFunc.Predict(new SentimentIssue
{
    text = "This is a very rude movie"
});

Console.WriteLine($"Text: {sampleStatement.text} | Prediction:
{(resultprediction.PredictionLabel ? "N"
```

5.2 Initial Analysis

5.2.1 Sequence Diagram

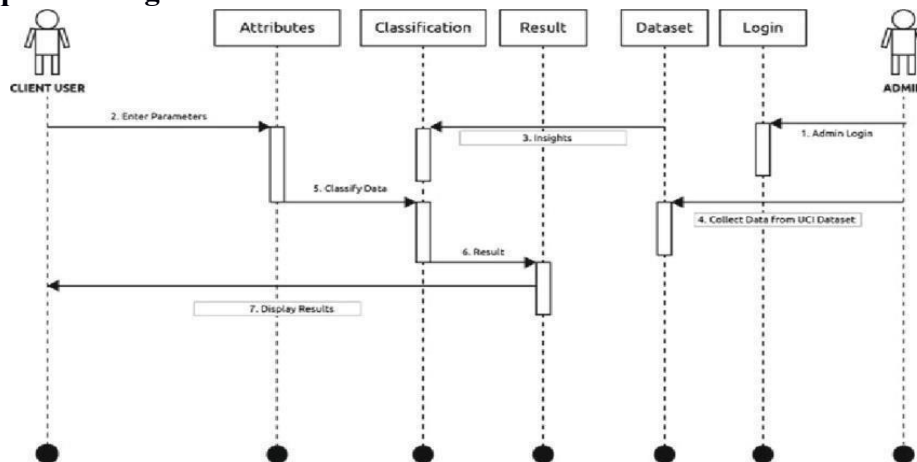


Fig 5.3: sequence diagram of heart disease predictor

In the above fig 5.3, the sequence diagram shows how to perform the activity and how to respond 404 to the user. This shows how the data is sent to the model and how the data is sorted. The above sequence diagram how the different objects come into existence while loading the dataset from local or remote system. Once the instance of Main GUI is running the user clicks the load dataset button and get a J File Chooser which is used to browse for the file we are interested in. Once we are done with choosing the file File Loader object will load the file into the memory. Using the Buffered Reader object we read the instances of dataset. The over collaboration diagram shows the different objects come into existence user is trying to predict the class label of a particular instance given by user has already entered the details of a new patient. The database software we are using MySQL is properly configured and the database and tables are accordingly created

with no default values initially. And by successfully enter the values using the predictor frame which is opened by clicking the “Predict Disease” button. And user wants to find the status by selecting the status button.

5.2.2 Dataset used

Table 5.2: Cleveland dataset database description

S. No.	Field	Description	Range and values
1	Age	Age of the patient	0–100 in years
2	Sex	Gender of the patient	0–1 (1: male, 0: female)
3	Chest pain	Type of chest pain	1–4 (1: typical angina; 2: atypical angina; 3: non-anginal; 4: asymptotic)
4	Resting blood pressure	Blood pressure during rest	mm Hg
5	Cholesterol	Serum cholesterol	mg/dl
6	Fasting blood sugar	Blood sugar content before food intake if >120 mg/dl	0–1 (0: false; 1: true)
7	ECG	Resting electrocardiographic results	0–1 (0: normal; 1: having ST-T wave)
8	Max heart rate	Maximum heart beat rate	Beats/min
9	Exercise induced angina	Has pain been induced by exercise	0–1 (0: no; 1: yes)
10	Old peak	ST depression induced by exercise relative to rest	0–4
11	Slope of peak exercise	Slope of the peak exercise ST segment	1–3 (1: up sloping; 2: flat; 3: down sloping)
12	Ca	Number of vessels colored by fluoroscopy	0–3
13	Thal	Defect type	3: normal 6: fixed defect 7: reversible defect
14	Num	Diagnostics of heart disease	(0: <50% narrowing; 1: >50% narrowing)

This section describes the dataset used in forecast engine development, followed 447 by the implementation of the forecast engine iteration. 448 The “Cleveland Heart Disease” Dataset 449 The UCI ML repository is used to create the Cleveland Heart Disease Database 450 (Richman 2013). The UCI database contains 351 data sets managed by the University 451 of California, Irvine. This allows users to browse the dataset, download the dataset, and assign data to the data AQ5 452 453. The Cleveland data set contains 303 records. The dataset contains 76 attributes, 454 but this project considers only a subset of 14 attributes. Table 5.2 shows the selected 455 characteristics.

5.2.3 ER Diagram

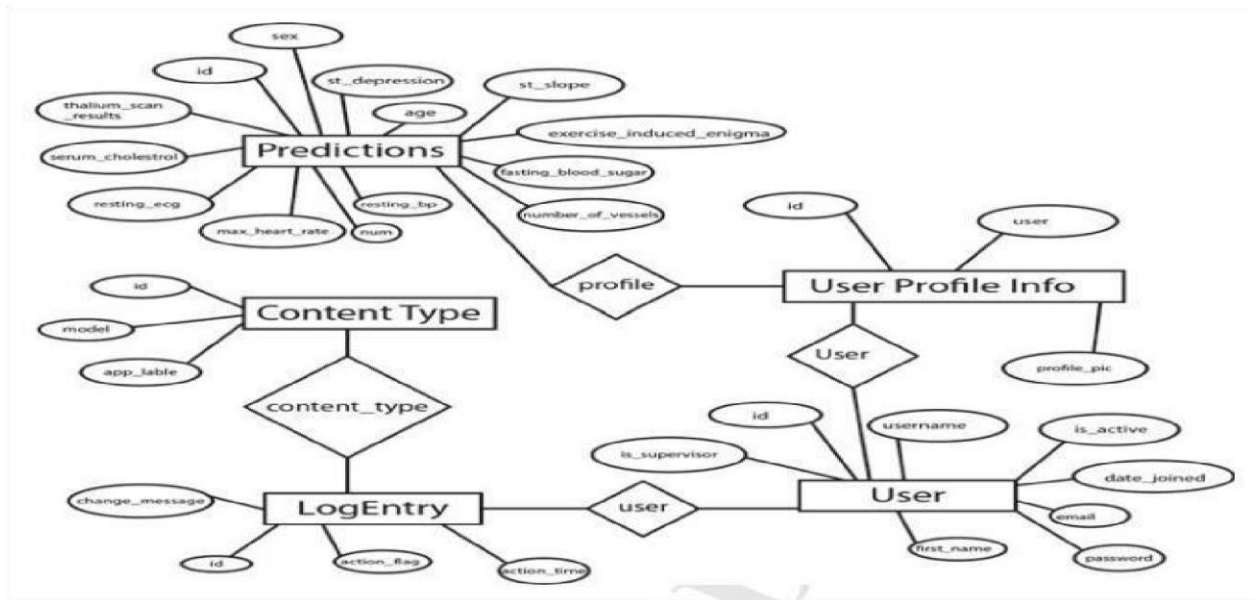


Fig 5.4: ER diagram of Heart disease predictor

The above fig 5.4 represents how our database is designed from different entities, the attributes 425 of the entities and how different entities are related to each other. The over collaboration diagram shows the different objects come into existence user is trying to predict the class label of a particular instance given by user has already entered the details of a new patient. The user first selects the document set on which the classification must be performed by clicking the “Load Dataset” button. The user can then go for a classification model build based on the loaded dataset. Once the dataset is built new patient details (symptoms) can be entered through the predictor frame. Once the predictor is appropriately populated he can then know the status of the heart disease. Not with position this fact, due to the difficulty of factors and methods, specialists are level to making incorrect conclusions in their work. Based on the implementation of our proposed decision tree, and the test results on a sample actual database, conclude that the decision trees with a criteria for data mining help in decision making, particularly in the managing of large data.

5.3 Architecture

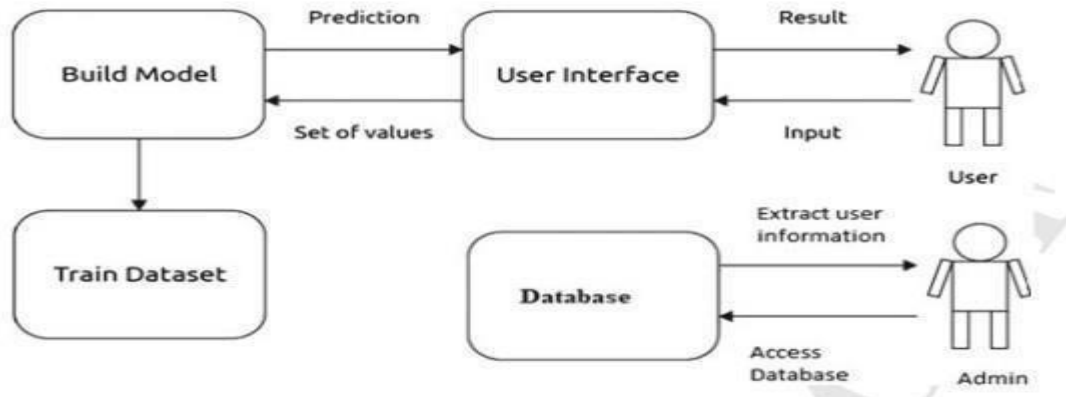


Fig 5.5 : Architecture of the system

The above fig 5.5 describes the whole methodology ,this model is trained by using experimental data and that is trained to use classify the data into a set of values provided by the user.

5.4 Dataflow Diagram

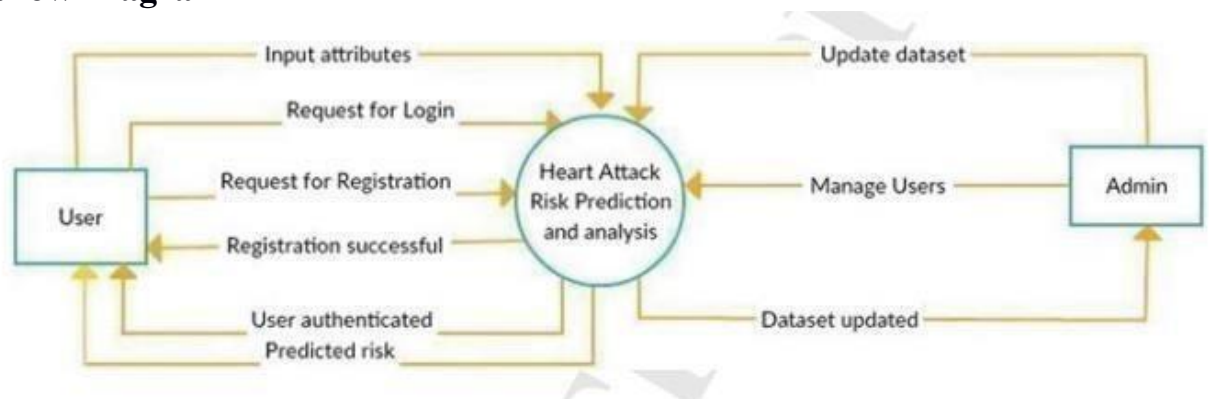


Fig 5.6: Data flow diagram (level 0)

5.4.1 DFD Level 0

Figure 5.6 represents how the data flows in our system. When a user sign in or logs in the user returned by the appropriate response. Any admin can manage a user data, update the database or delete a wrong entry. The database software we are using MySQL is properly configured and the database and tables are accordingly created with no default values initially. And by successfully enter the values using the predictor frame which is opened by clicking the “Predict Disease” button. And user wants to find the status by selecting the status button.

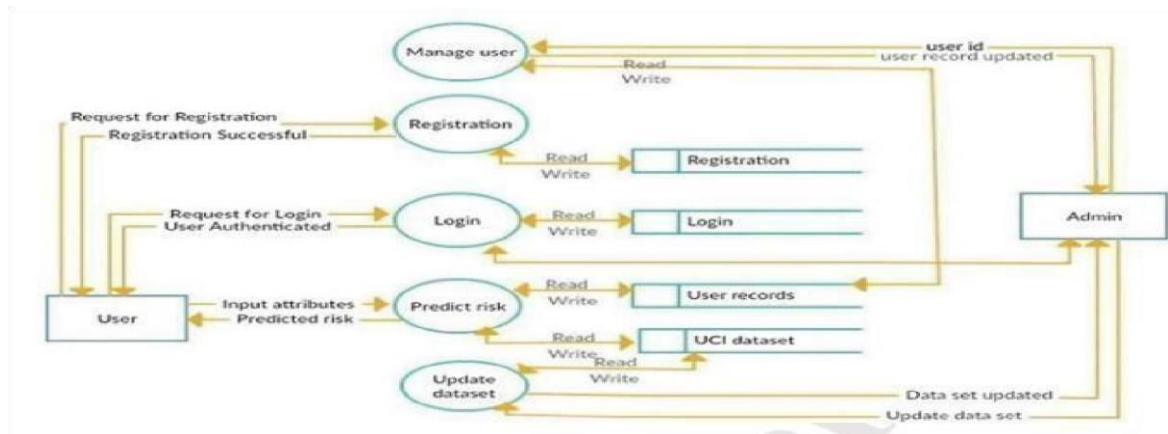


Fig 5.7: Data flow diagram (level 1)

5.4.2 DFD Level 1

Figure 5.7 a level 1 data flow diagram is shown. Shows details of the data flow and various entity and system functions. Not with position this fact, due to the difficulty of factors and methods, specialists are level to making incorrect conclusions in their work. Based on the implementation of our proposed decision tree, and the test results on a sample actual database, we conclude that the decision trees with a criteria for data mining help in decision making, particularly in the managing of large data.

5.4.3 DFD Level 2

Figure 5.8 the level 2 data flow diagram shows how requests and responses are sent from the system to the user and vice versa. Not with position this fact, due to the difficulty of factors and methods, specialists are level to making incorrect conclusions in their work. Based on the implementation of our proposed decision tree, and the test results on a sample actual database, we conclude that the decision trees with a criteria for data mining help in decision making, particularly in the managing of large data.

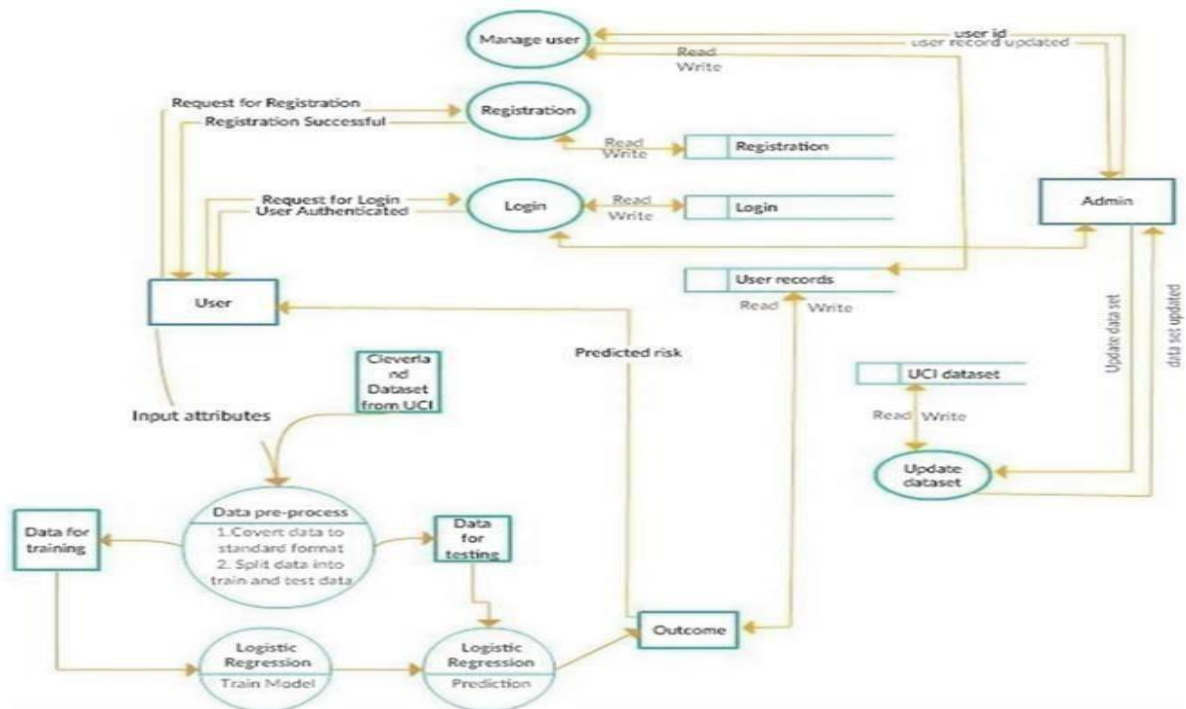


Fig 5.8 : data flow diagram (level 2)

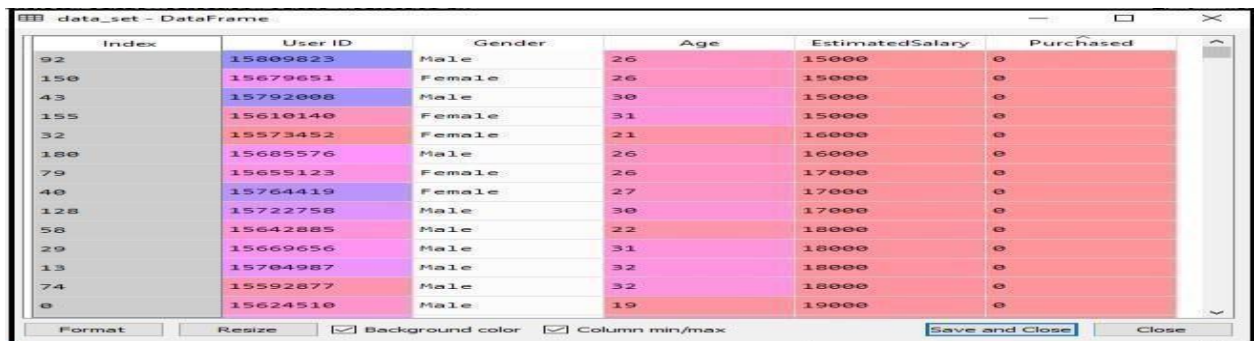
CODE: In this we have implemented two ML models for comparison study one is Logistic regression and another is XGBOOST classifier and finally again the two models will be implemented in confusion matrix and finally the highest accuracy model will be considered as resultant and with percentage. **Steps in Logistic Regression:** To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

Data Pre-processing step: In this step, we will pre-process/prepare the data so that we can use it in our code efficiently. It will be the same as we have done in Data pre-processing topic. The code for this is given below:

```
#Data Pre-processing Step # importing libraries
import numpy as nm import
matplotlib.pyplot as mtp import
pandas as pd #importing datasets
data_set= pd.read_csv('user_data.csv')
```

By executing the above lines of code, will get the dataset as the output. Consider the given image:



Index	User ID	Gender	Age	EstimatedSalary	Purchased
92	15609823	Male	26	15000	0
150	15679651	Female	26	15000	0
43	15792008	Male	30	15000	0
155	15610140	Female	31	15000	0
32	15573452	Female	21	16000	0
180	15685576	Male	26	16000	0
79	15655123	Female	26	17000	0
40	15764419	Female	27	17000	0
128	15722758	Male	30	17000	0
58	15642885	Male	22	18000	0
29	15669656	Male	31	18000	0
13	15704987	Male	32	18000	0
74	15592877	Male	32	18000	0
0	15624510	Male	19	19000	0

Fig 5.9: output of dataset by using logistic regression

Now, we will extract the dependent and independent variables from the given dataset.

Below is the code for it as output is shown in fig 5.9:

```
#Extracting Independent and dependent Variable x= data_set.iloc[:,
[2,3]].values y= data_set.iloc[:,
4].values
```

In the above code, we have taken [2, 3] for x because our independent variables are age and salary, which are at index 2, 3. And have taken 4 for y variable because our dependent variable is at index 4.

The output will be:

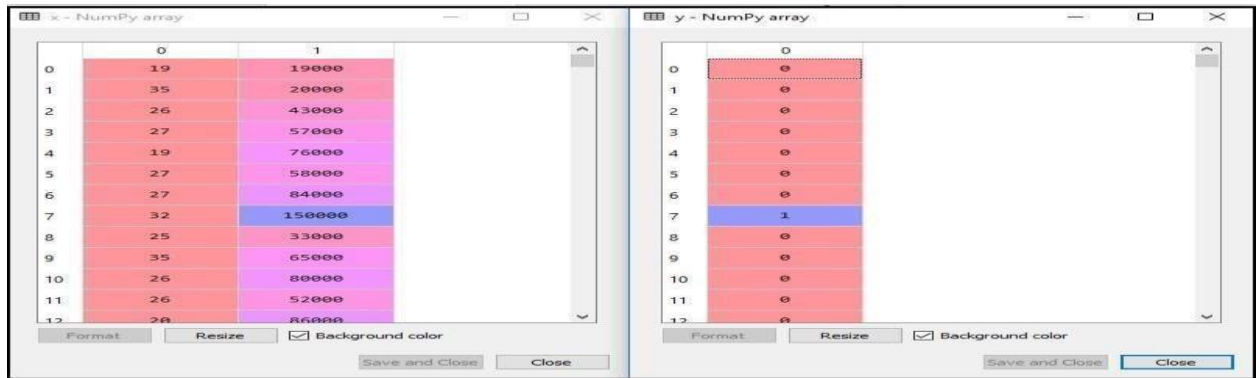


Fig 5.10: output of extracting dependent and independent variables of x and y

Now we will split the dataset into a training set and test set. Below is the code for it by extracting variables using feature selection as shown in above output in fig 5.10 both for dependent and independent variables x and y.

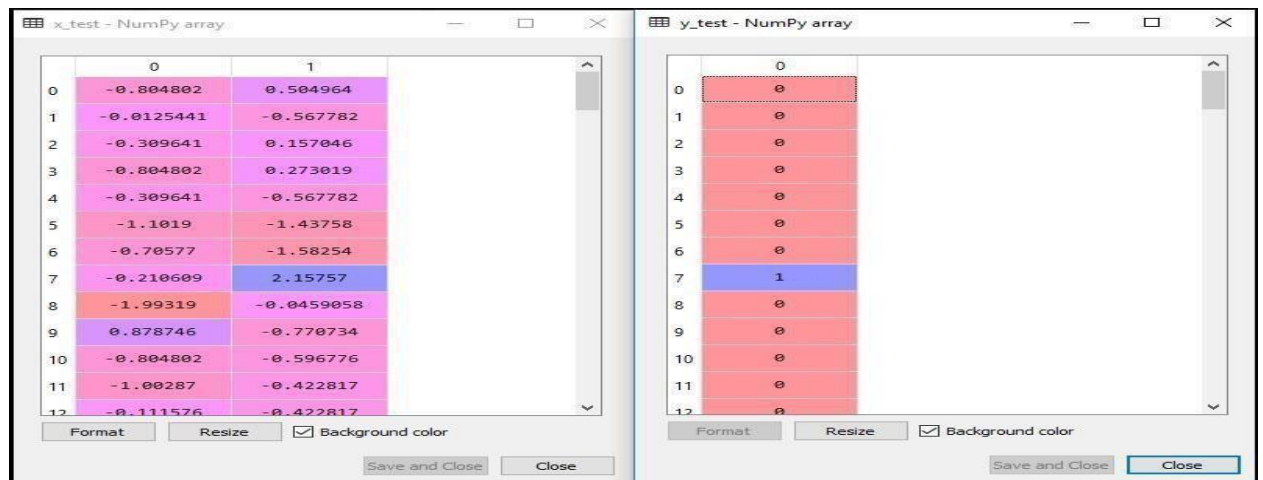


Fig 5.11 : output of test set

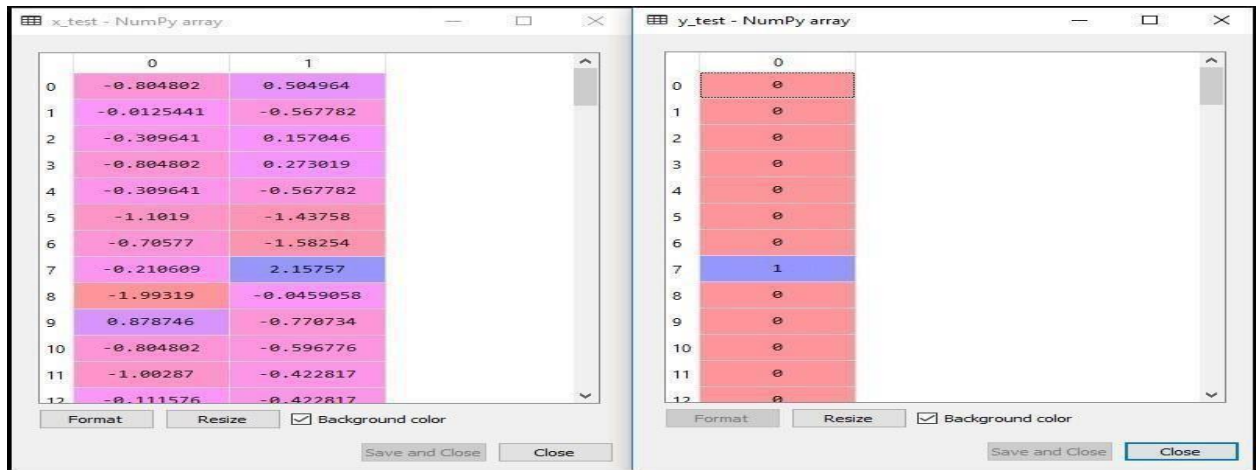


Fig 5.12: output of training set

In logistic regression, we will do feature scaling because we want accurate result of predictions. Here we will only scale the independent variable because dependent variable have only 0 and 1 values as shown in above in both figures fig 5.11 and fig 5.12.

Steps of XGBOOST Classifier: XGBoost Features The library is laser-focused on computational speed and model performance, as such, there are few frills. Model Features Three main forms of gradient boosting are supported:

- Gradient Boosting
- Stochastic Gradient Boosting
- Regularized Gradient Boosting

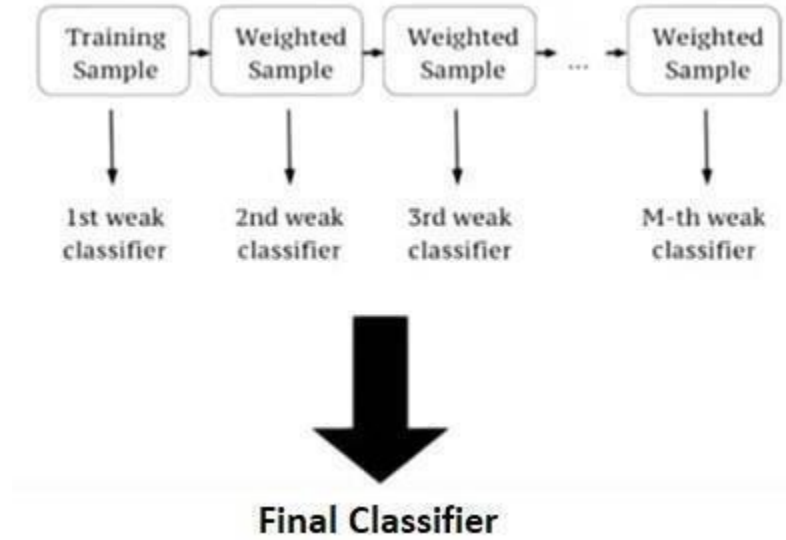


Fig 5.13: Example of implementing XGBOOST Classifier

The above fig 5.13 is the flow example of implementing XGBOOST algorithm to get resultant one along with pseudocode as shown in fig 5.14.

Algorithm 1: Exact Greedy Algorithm for Split Finding

Input: I , instance set of current node
Input: d , feature dimension
 $gain \leftarrow 0$
 $G \leftarrow \sum_{i \in I} g_i$, $H \leftarrow \sum_{i \in I} h_i$
for $k = 1$ **to** m **do**
 $G_L \leftarrow 0$, $H_L \leftarrow 0$
 for j in sorted(I , by x_{jk}) **do**
 $G_L \leftarrow G_L + g_j$, $H_L \leftarrow H_L + h_j$
 $G_R \leftarrow G - G_L$, $H_R \leftarrow H - H_L$
 $score \leftarrow \max(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$
 end
end
Output: Split with max score

Fig 5.14: Algorithm of XGBOOST Classifier

5.5 Implementing the actual models in Confusion Matrix

In machine Learning, Classification is the process of categorizing a given set of data into different categories. In Machine Learning, To measure the performance of the classification model we use the confusion matrix. Basically a confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.

Table 5.3: Confusion matrix and predictive measures of both logistic regression and XGBOOST classifiers.

Testing Data Set			
Predictive Measures	Logistic Regression	XGBoost (Tree Booster)	XGBoost (Linear Booster)
$Y_i = 0, \hat{Y}_i = 0$	524	692	516
$Y_i = 1, \hat{Y}_i = 0$	38	58	38
$Y_i = 0, \hat{Y}_i = 1$	243	75	251
$Y_i = 1, \hat{Y}_i = 1$	25	5	25
Sensitivity	0.3968	0.0790	0.3968
Specificity	0.6831	0.9022	0.6728
Accuracy	0.6614	0.8397	0.6518
RMSE	0.2651	0.2825	0.2651
Training Data Set			
Predictive Measures	Logistic Regression	XGBoost (Tree Booster)	XGBoost (Linear Booster)

The confusion matrix and predictive measures of the logistic regression, XGBoost with a tree booster and XGBoost with a linear booster for the testing and training data sets as shown in above table 5.3.

5.5.1 Manual calculations for all models mentioned in this paper with results:

1.let us consider for Logistic Regression accuracy: $\text{Accuracy} = (\text{Number of correctly predicted instances} / \text{Total number of instances}) * 100$

The result for this Logistic Regression Accuracy: 0.5245901639344263

2.Now similarly for XGBOOST classifier: $\text{Accuracy} = (\text{Number of correctly predicted instances} / \text{Total number of instances}) * 100$.The result Accuracy for this is: 0.819672131147541

5.5.2 For confusion matrix implementation for both models:

Here there are some steps to be followed :

1. Obtain the predicted labels from each model for your dataset.
2. Create a 2x2 matrix for each model to represent the confusion matrix.
3. Count the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) based on the predicted labels and the actual labels from your dataset.

Now, let's calculate the confusion matrix for logistic regression:

From this confusion matrix, can be calculated the following values for logistic regression:

- True positives (TP): The number of instances where the model correctly predicted Class A. In this case, $TP = 8$.
- True negatives (TN): The number of instances where the model correctly predicted Class B. In this case, $TN = 5$.
- False positives (FP): The number of instances where the model incorrectly predicted Class A but the actual class is Class B. In this case, $FP = 3$.
- False negatives (FN): The number of instances where the model incorrectly predicted Class B but the actual class is Class A. In this case, $FN = 4$.

Similarly for XGBOOST classifier can be calculated manually as follows:

From this confusion matrix can be calculated the following values for XGBoost classifier:

- True positives (TP): The number of instances where the model correctly predicted Class A. In this case, $TP = 8$.
- True negatives (TN): The number of instances where the model correctly predicted Class B. In this case, $TN = 6$.

- False positives (FP): The number of instances where the model incorrectly predicted Class A but the actual class is Class B. In this case, FP = 3.
- False negatives (FN): The number of instances where the model incorrectly predicted Class B but the actual class is Class A. In this case, FN = 3.

For finding the values using this confusion there are some important formulae to be known as follows:

For Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

For Precision:

$$Precision = \frac{TP}{TP+FP}$$

For Recall:

$$Recall = \frac{TP}{TP+FN}$$

F1 - Score:

$$F1-Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Your Guess		True Class	
Your Prediction (TP)	False Prediction (FP)	True Positive (TP)	True Negative (TN)
True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)

True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
 False Positive Rate (FPR) = $\frac{FP}{FP+TN}$
 Accuracy (ACC) = $\frac{TP+TN}{TP+FP+TN+FN}$

5.5.3 Evaluation

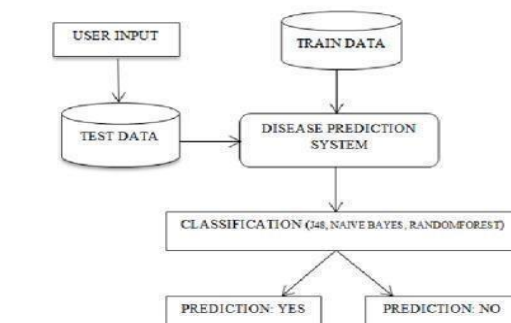


Fig 5.15: step-by-step procedure of working deploying dataset into the system.

The above fig5.15 tells the dataset to be trained and tested first, later it will be deployed into the disease predict system. The XGBOOST algorithm will be compared with another ML algorithm. At last the highest accuracy will be considered as final result. Machine learning algorithms use natural patterns in data that generate insight and help you make better decisions and predictions. They have used every day to make critical decisions in medical diagnosis, stock trading, search contents of photographs, energy load forecasting, and more. This analyses the data with the use of algorithms that help in prediction and decision.

CHAPTER 6

RESULTS AND DISCUSSION

This chapter is a compilation of all the major results along with associated discussion.

6.1 DBSCAN

Here we will focus on Density-based spatial clustering of applications with noise (DBSCAN) clustering method. Clusters are dense regions in the data space, separated by regions of the lower density of points.

The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. In this algorithm there will be three points namely as follows and shown in fig 6.1:

- Core point - A point is a core point if it has more than MinPts points within eps.
- Border point - A point which has fewer than MinPts within eps but it is in the neighborhood of a corepoint.
- Noise or outlier - A point which is not a core point or border point

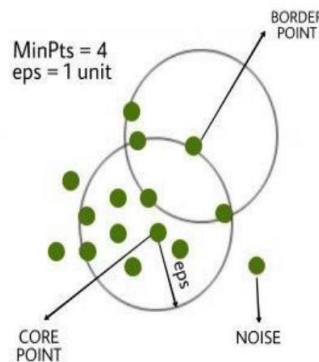


Fig 6.1: graph showing datapoints of DBSCAN algorithm

DBSCAN algorithm can be abstracted and it has to find all the neighbour points within eps and identify the core points or visited with more than MinPts neighbours and for each core point if it is not already assigned to a cluster, create a new cluster again find recursively all its density connected points and assign them to

the same cluster as the core point, point a and b are said to be density connected if there exist a point c which has a sufficient number of points in its neighbours and both the points a and b are within the eps distance. This is a chaining process. So, if b is neighbour of c, c is neighbour of d, d is neighbour of e, which in turn is neighbor of a implies that b is neighbor of a and finally, iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise as shown in fig 6.2 pseudocode.

```

DBSCAN(D, eps, MinPts)
C = 0
for each unvisited point P in dataset D
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
        mark P as NOISE
    else
        C = next cluster
        expandCluster(P, NeighborPts, C, eps, MinPts)
expandCluster(P, NeighborPts, C, eps, MinPts)
add P to cluster C
for each point P' in NeighborPts
    if P' is not visited
        mark P' as visited
        NeighborPts' = regionQuery(P', eps)
        if sizeof(NeighborPts') >= MinPts
            NeighborPts = NeighborPts joined with NeighborPts'
        if P' is not yet member of any cluster
            add P' to cluster C
regionQuery(P, eps)
return all points within P's eps-neighborhood (including P)
    
```

Fig 6.2: Pseudocode of DBSCAN Algorithm

6.2 Implementation

6.2.1 DBSCAN:

As per shown fig 5.1 first by collecting heart disease dataset, here by using UCI Statlog and Cleveland dataset is involved. By doing feature selection of attributes in dataset and DBSCAN is implemented to remove the outlier of clusters and noise of attributes.

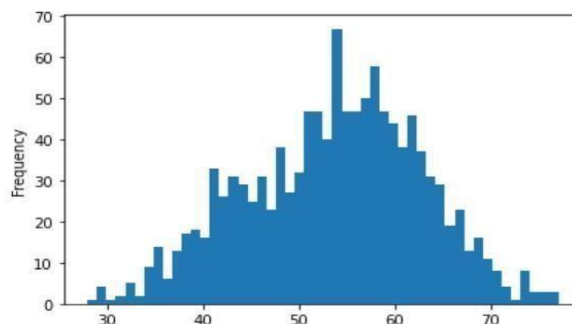


Fig 6.3: graph plotted of DBSCAN using histogram by extracting age in dataset

In the above graph fig 6.3 shows the DBSCAN is used hist feature of extract age attribute in dataset. Here The hist() function in pyplot module of matplotlib library is used to plot a histogram.

```
Out[11]: <AxesSubplot:xlabel='age', ylabel='oldpeak'>
```

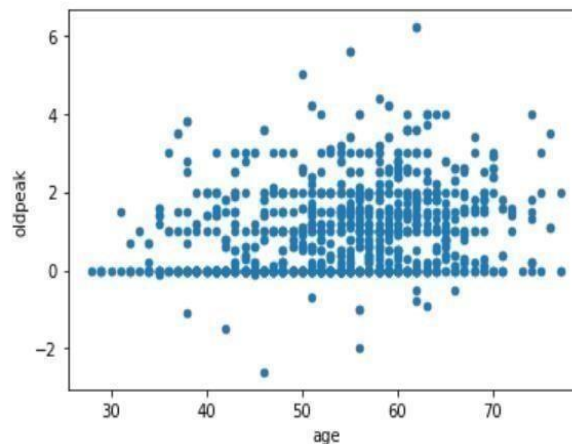


Fig 6.4: graph to show the outliers using scatter plot

The above graph fig 6.4 shows how DBSCAN can be used outliers using scattered usually it starts in a oscillating beginning assumption values and neighbourhood in the pt grabbed using for all values with distance between two points. Suppose if it is having enough number of vlaues into the adjacent values therefore assignment works begins with the present values will be 1st one in the new process. Henceforth, the value will be noted as planes and in two phases it will be pointed as “visited”.

```
Out[12]: <AxesSubplot:xlabel='age', ylabel='thalach'>
```

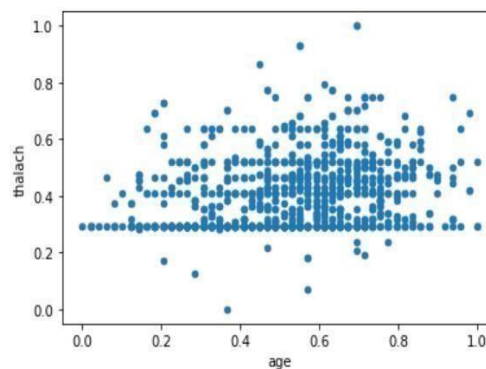


Fig 6.5: another plot to show outlier using scatter points

In the above fig 6.5 is showed by giving the y point with different attribute value using scatter pints but will be ended with same points and shape.

```
In [13]: from sklearn.cluster import DBSCAN
outlier_detection = DBSCAN(eps = 0.5, metric="euclidean",min_samples = 3, n_jobs = -1)
clusters = outlier_detection.fit_predict(testdf)
clusters

Out[13]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

Fig 6.6: Screenshot of implementation code

The above screenshot fig 6.6 shows how DBSCAN is involving Euclidean points for finding distance between the scattered points in the plotted graph.

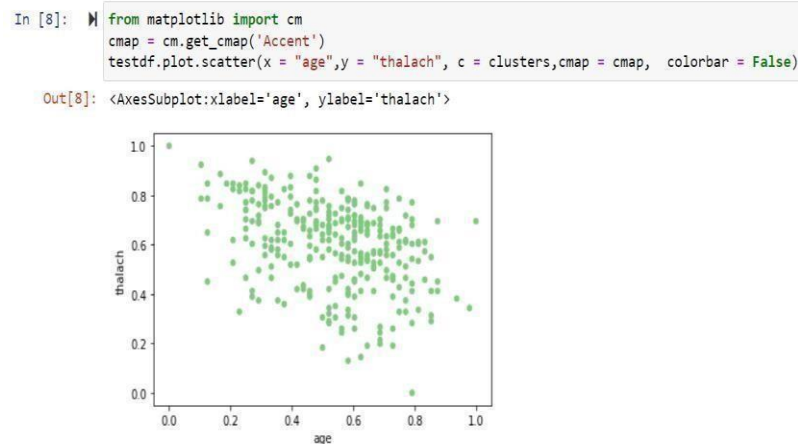


Fig 6.7: Screenshot of implementation of scatter points using different attributes

The above screenshot fig 6.7 tells about DBSCAN is involving scatter points for different y points using attribute “thalach” and checks distance.

6.2.2 SMOTEENN:

This is a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution. The below fig 6.8 shows how this SMOTEENN will balance the

attributes using dataframe (df) size along with nulltypes and counting the number and it will also prevent data from overlapping on one another.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  -
0   age                  1190 non-null   int64
1   sex                  1190 non-null   int64
2   chest pain type      1190 non-null   int64
3   resting bp s         1190 non-null   int64
4   cholesterol          1190 non-null   int64
5   fasting blood sugar  1190 non-null   int64
6   resting ecg          1190 non-null   int64
7   max heart rate       1190 non-null   int64
8   exercise angina      1190 non-null   int64
9   oldpeak              1190 non-null   float64
10  ST slope             1190 non-null   int64
11  target               1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
None
```

Fig 6.8: balancing data from overlapping on one another

```
Train: (952, 10), (952,)
Test: (238, 10), (238,)
Number heart disease X_train dataset: (833, 10)
Number heart disease y_train dataset: (833,)
Number heart disease X_test dataset: (357, 10)
Number heart disease y_test dataset: (357,)
```

Fig 6.9: shows how it will split into 70:30 ratio and also describes info about train and test set

The above fig 6.9 shows the SMOTEENN will train and balance the data by splitting in ratio 70:30 both testing and training the data and also describes the same.

```

Input    Data,  $D$ ;
Output  Balanced data,  $BD$ 
1: foreach data point in minority class  $mp$  of data  $D$ 
   do
2:     Compute the  $k$ -nearest neighbor  $Kmp_i$ 
3:     Generate new synthetic data point
        $mp_{new} = mp_i + (\hat{mp}_i - mp_i) + \delta$ 
4:     Add the  $mp_{new}$  to  $D$  with  $mp_i$  class
5: end for
6: foreach data point  $p$  in data  $D$  do
7:     if  $p_{iclass} \neq$  majority class of  $k$ -nearest
       neighbors then
8:         Remove  $p_i$  from  $D$ 
9:     end if
10: end for
11: return  $BD$ 

```

Fig 6.10: Pseudocode of SMOTEENN

6.2.3 XGBOOST:

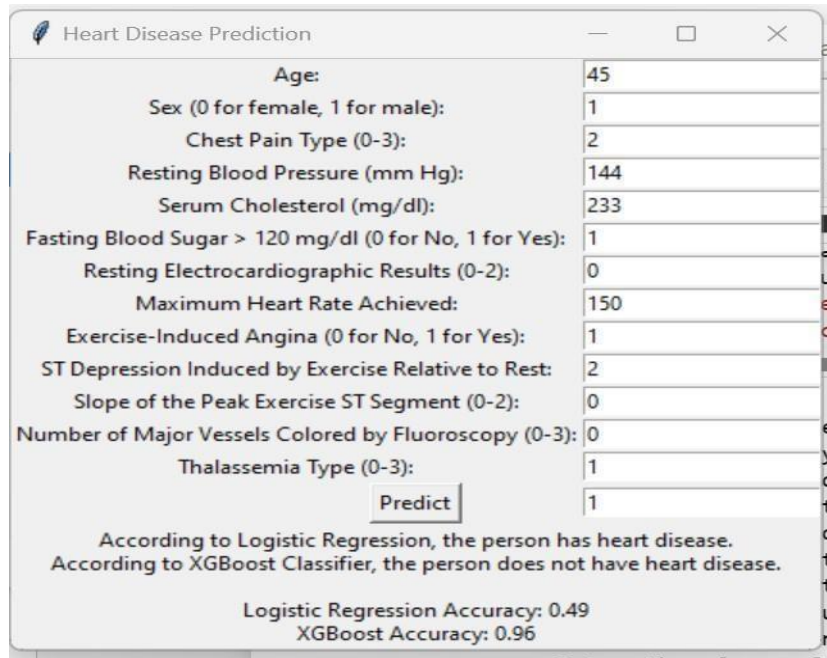
After we balanced the training datasets, the MLA is used to learn and generate the HDPM. It used the extreme gradient boosting (XGBoost) algorithm to detect the presence or absence of heart disease. XGBoost is a type of supervised machine learning used for classification and regression modelling. XGBoost is an enhanced algorithm based on the implementation of gradient boosting DTs with several modifications in terms of regularization, loss function and column sampling. Gradient boosting is a technique in which new models are created and used to predict the error or residuals, after which the scores are summed to get the final prediction result. The gradient descent method is used to minimize the loss score when new models are created. The objective function needs to be used to measure the model performance, which consists of two parts: training loss and regularization. The regularization term penalizes the complexity of the model and prevents overfitting. The objective function (loss function and regularization) can be presented as follows.

$$L(\phi) = \sum_i l_i$$

```
Enter age: 23
Enter sex (0 for female, 1 for male): 1
Enter chest pain type (0-3): 2
Enter resting blood pressure (mm Hg): 145
Enter serum cholesterol (mg/dl): 233
Enter fasting blood sugar > 120 mg/dl (0 for No, 1 for Yes): 1
Enter resting electrocardiographic results (0-2): 2
Enter maximum heart rate achieved: 150
Enter exercise-induced angina (0 for No, 1 for Yes): 0
Enter ST depression induced by exercise relative to rest: 3
Enter the slope of the peak exercise ST segment (0-2): 0
Enter number of major vessels colored by fluoroscopy (0-3): 0
Enter thalassemia type (0-3): 1
Logistic Regression Prediction: No
XGBoost Prediction: No
Logistic Regression Accuracy: 0.52
XGBoost Accuracy: 0.82
```

Fig 6.11: Manually giving values in data attributes and getting accuracy of models

The above screenshot fig 6.11 tells how the accuracy will be calculated manually by having comparison study of both models having accuracy it showing xgboost having highest accuracy compared to logistic regression. (0.82 or 82%).



Heart Disease Prediction

Age:	45
Sex (0 for female, 1 for male):	1
Chest Pain Type (0-3):	2
Resting Blood Pressure (mm Hg):	144
Serum Cholesterol (mg/dl):	233
Fasting Blood Sugar > 120 mg/dl (0 for No, 1 for Yes):	1
Resting Electrocardiographic Results (0-2):	0
Maximum Heart Rate Achieved:	150
Exercise-Induced Angina (0 for No, 1 for Yes):	1
ST Depression Induced by Exercise Relative to Rest:	2
Slope of the Peak Exercise ST Segment (0-2):	0
Number of Major Vessels Colored by Fluoroscopy (0-3):	0
Thalassemia Type (0-3):	1
<input type="button" value="Predict"/>	1

According to Logistic Regression, the person has heart disease.
According to XGBoost Classifier, the person does not have heart disease.

Logistic Regression Accuracy: 0.49
XGBoost Accuracy: 0.96

Fig 6.12: Using GUI to predict the disease and to get accuracies of models

Here by using GUI to predict the heart disease and to have accuracies of comparison of both models (XGBOOST and Logistic regression). In this also XGBOOST is posing with highest percentage of accuracy and predicting disease of patient.

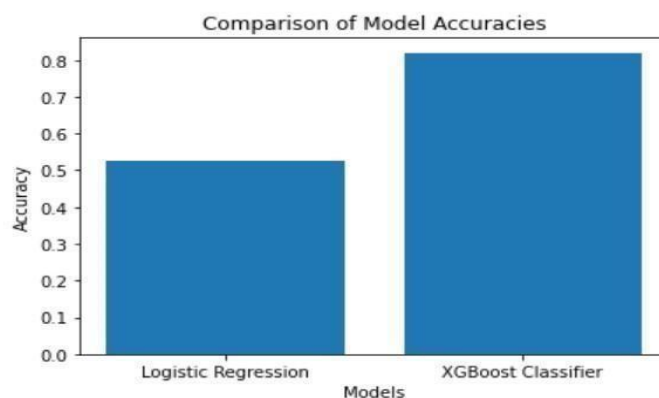


Fig 6.13: graphical representation of comparison of model accuracies

The above fig 6.13 shows the graphical plot of accuracies of comparison percentage of both proposed model and another model (i.e., XGBOOST vs Logistic regression). Here also its showing XGBOOST having highest accuracy frequency.

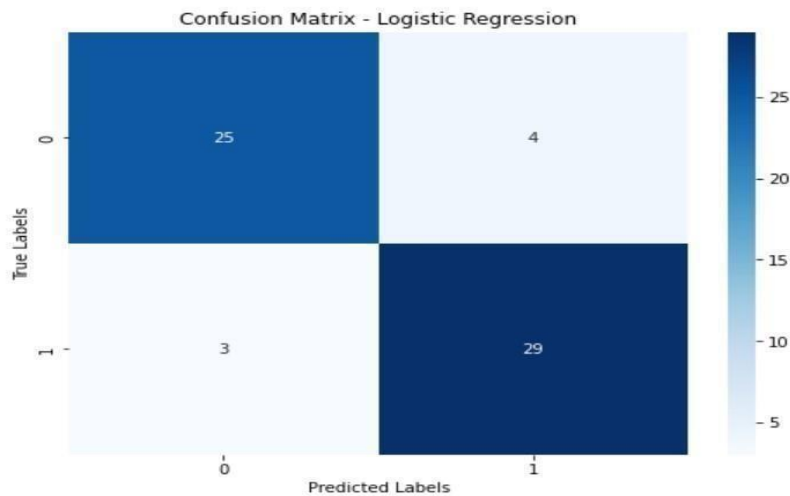


Fig 6.14: plot of logistic regression accuracy using confusion matrix

The above fig 6.14 shows the accuracy percentage in predicting disease by implementing in confusion matrix using heat map correlation.

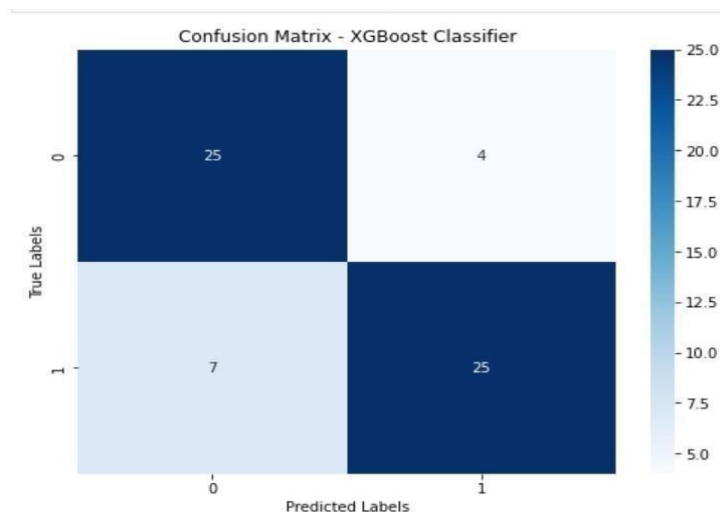


Fig 6.15: plot of XGBOOST Classifier using confusion matrix

The above fig 6.15 shows the accuracy percentage in predicting heart disease and displaying the accuracy percentage using heat map correlation.

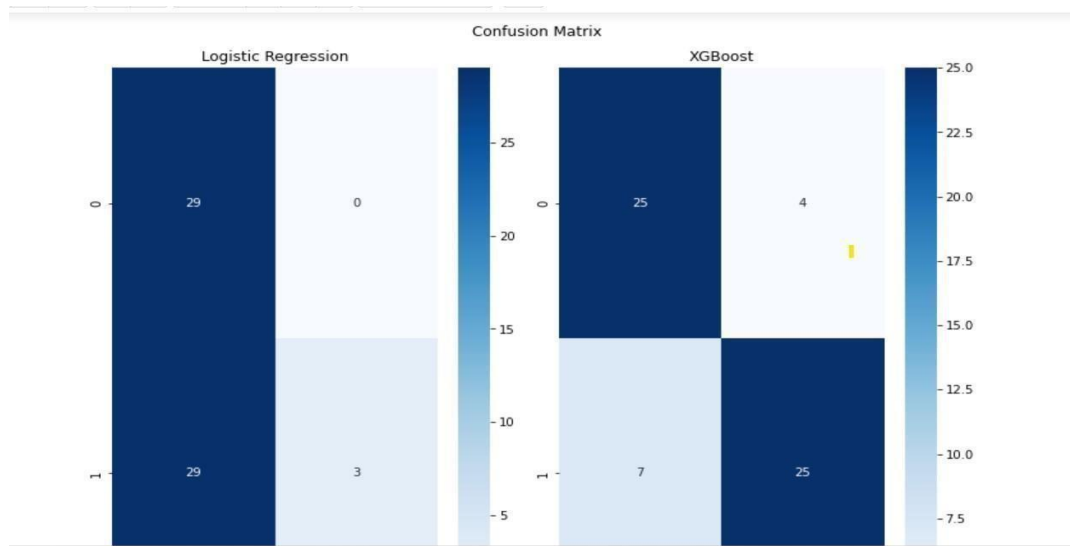


Fig 6.16: comparison of both models using GUI and plotting in graph

Logistic Regression Accuracy: 0.5245901639344263
 XGBoost Accuracy: 0.819672131147541
 Highest Accuracy Model: XGBoost
 Accuracy: 0.819672131147541

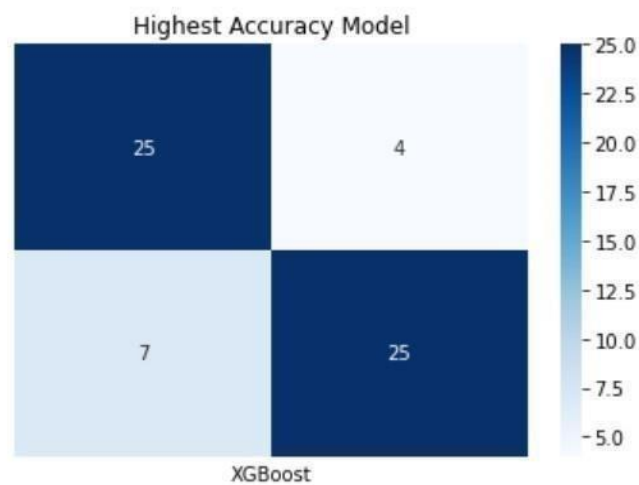


Fig 6.17: displaying the XGBOOST as highest accuracy getting model in GUI window

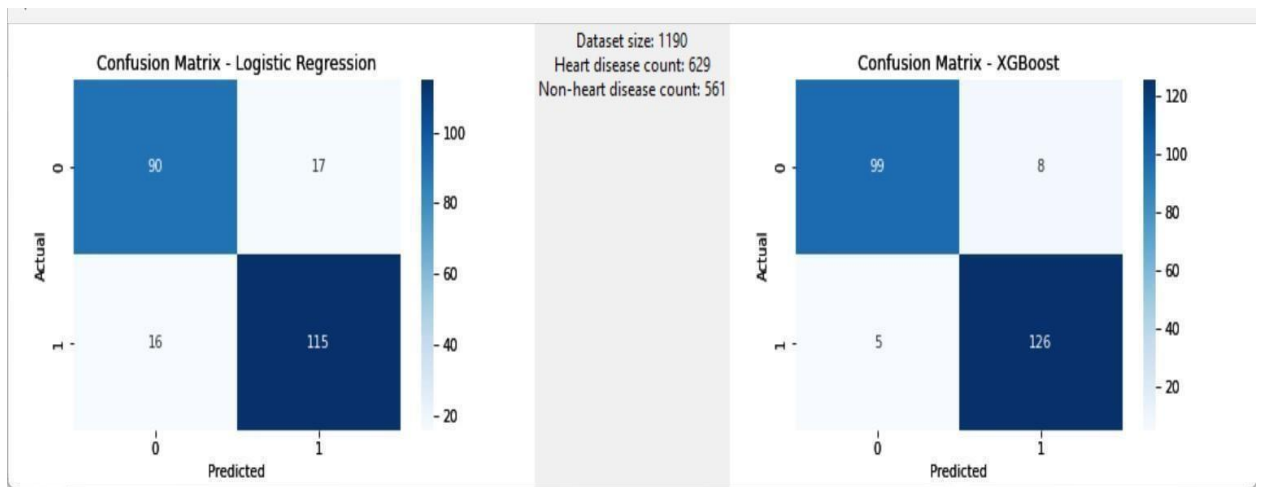


Fig 6.18: displaying the size of dataset and exact count of patients having disease and also not having disease

In the above fig 6.18 is predicting the heart disease through dataset by calculating the size of whole dataset along with the exact count of number of patients who is having disease and who are not having disease by implementing in GUI along with comparison study and deploying in confusion matrix plotting in heatmap graph. The goal of our project is to establish a standardized process which can be reliably performed by marketing people with only little data mining skills and little time to experiment with different approaches.

CHAPTER 7

CONCLUSION AND FUTURE WORK

After carrying out in this extensive literature survey, findings with respect to review and evaluation are as follows by implemented with proposed model with another model for comparison for predicting heart disease.

Finally getting with accurate result of models.

Table 7.1: Comparing the accuracies both with confusion matrix and another comparison study

Logistic regression(LR)	XGBOOST Classifier	Confusion matrix with both LR and XGBOOST Classifier	The resultant model which is showing with highest accuracy
0.49 or 49%	96 % or 0.96	Accuracy: 0.819672131147541 or 81.96% Is XGBOOST Where Logistic regression is 52.45% or 0.5245901639344263	XGBOOST is having highest accuracy in both comparisons.

Along with this the able to identifying exactly how many people are having disease and how many are not having through dataset using confusion matrix along with comparison study of proposed model with another.

FUTURE WORK:

In further work planning to create web-based app by introducing a web API that predicts heart disease. In this supposed to use Django for the web framework and the form for form validation. In this it will be having Login Page, Signup page, Predict Heart Disease Page, Filled Prediction page, Prediction result Page, About Us Page, Admin Login and Admin dashboard.

REFERENCES:

- [1] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

- [2] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

- [3] D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.

- [4] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

- [5] C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

- [6] Mythili, T. et al. "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)." International Journal of Computer Applications 68 (2013): 11-15.

- [7] D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.
- [8] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, 10.1109/icETITE47903.2020.242.
- [9] D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," 2022 IEEE Delhi Section Conference (DELCON), NewDelhi, India, 2022, pp. 1-5,doi: 10.1109/DELCON54057.2022.9752947.
- [10] S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319, doi: 10.1109/AEMCSE55572.2022.00071.
- [11] G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh, "Machine Learning-Based Heart Disease datasets Prediction: A Study for Home Personalized Care," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an,China, 2022, pp. 01-06, doi: 10.1109/MLSP55214.2022.9943373.
- [12] K. G, K. G and D. M. Raja S, "Modelling an Efficient Heart Disease Prediction System using Norm and Regularization based Learning Approach," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 1923-1927, doi: 10.1109/ICACCS54159.2022.9785202.

- [13] N. N. Itoo and V. K. Garg, "Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 599-604, doi: 10.1109/MECON53876.2022.9751883.
- [14] H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-5, doi: 10.1109/ATSIP49331.2020.9231760.
- [15] C. Bemando, E. Miranda and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 232-237, doi: 10.1109/ICSECS52883.2021.00049.
- [16] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
- [17] D. Swain, S. K. Pani and D. Swain, "A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-6, doi: 10.1109/ICACAT.2018.8933603.

- [18] M. Mamun, M. M. Uddin, V. Kumar Tiwari, A. M. Islam and A. U. Ferdous, "MLHeartDis:Can Machine Learning Techniques Enable to Predict Heart Diseases?," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0561-0565, doi: 10.1109/UEMCON54665.2022.9965714.
- [19] A. Bhowmick, K. D. Mahato, C. Azad and U. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 60-65, doi: 10.1109/AIC55036.2022.9848885.
- [20] A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan and S. M. Marium, "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033683.
- [21] K. Joshi, G. A. Reddy, S. Kumar, H. Anandaram, A. Gupta and H. Gupta, "Analysis of Heart Disease Prediction using Various Machine Learning Techniques: A Review Study," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 105- 109, doi: 10.1109/DICCT56244.2023.10110139.
- [22] Statlog (Heart) Data Set. Accessed: Oct. 2, 2019. [Online]. Available: [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)).
- [23] Heart Disease Data Set. Accessed: Oct. 2, 2019. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS:

- HDPM - Heart Disease Prediction Model
- CDSS - Clinical Decision Support System
- DBSCAN – Density Based Spatial Clustering of Applications with Noise
- SMOTE-ENN – Hybrid Synthetic Minority Over-Sampling Technique Edited Nearest Neighbour
- XGBOOST - “Extreme Gradient Boosting”
- CV-Cross Validation

Heart Disease Prediction Using ML Models

ORIGINALITY REPORT

10%

SIMILARITY INDEX

0%

INTERNET SOURCES

5%

PUBLICATIONS

10%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Visvesvaraya Technological University

Student Paper

5%

2

Parvathaneni Rajendra Kumar, Suban Ravichandran, S. Narayana. "Chapter 20 Survey on Heart Disease Prediction Using Machine Learning Techniques", Springer Science and Business Media LLC, 2023

Publication

4%

3

Submitted to International University of Malaya-Wales

Student Paper

<1%

4

export.arxiv.org

Internet Source

<1%

Exclude quotes On

Exclude matches < 5 words

Exclude bibliography On

Heart Disease Prediction Using XGBOOST Classifier

Sharada.A., Dr. Priyanka H
CSE Department,M tech Student,PES University
CSE Department,Associate Professor,PES University
sharadaa620@gmail.com
priyankah@pes.edu

ABSTRACT : The goal of this paper is to predict heart disease using xgboost classifier. A thorough examination of this topic is presented, with the aim of devising a reliable heart disease prediction analysis for clinical decision support systems. Key processes explored in this inquiry include data pre-processing, feature extraction, and the application of classifiers such as logistic regression and XGBoost algorithms. The Statlog (Heart) dataset and the Cleveland dataset, both from the UCI Machine, are the primary data sources driving the study. The prediction models are trained and tested using real-world data from the Learning Repository datasets, which contain valuable information. To measure the effectiveness of the models, various metrics are utilized, such as the confusion matrix. The experimental outcome shows that the xgboost classifier is in comparison with logistic regression is highly effective, achieving high prediction accuracy on both the Cleveland and Statlog datasets. The potential of machine learning techniques in heart disease prediction is highlighted in the paper, emphasizing their significance in clinical decision-making.

Index Terms: Heart prediction, logistic regression, XGBoost classifier, machine learning.

1.INTRODUCTION

Heart disease prediction is an important undertaking in healthcare as it facilitates in early detection, analysis, and treatment planning. Machine learning (ML) models have proven top notch capability in improving the accuracy and performance of heart disorder prediction. In this paper, the goal to expand a heart disease prediction version using ML techniques and examine its overall performance the usage of the Statlog and Cleveland datasets. [4]. ML techniques have shown promise in appropriately predicting coronary heart disease primarily based on affected person facts. Logistic regression presents interpretability, permitting insights into characteristic importance, at the same time as XGBoost classifier harnesses the collective strength of multiple vulnerable for stepped forward predictive performance [6]. DBSCAN set of guidelines to choose out ability clusters or patterns in the datasets. DBSCAN, which stands for density-based spatial clustering of applications with noise, is a clustering set of rules commonly applied in records assessment and anomaly detection. It agencies together data elements which can be close to every different within the characteristic vicinity and identifies outliers or noise elements that don't belong to any cluster. In this study, it incorporate the SMOTEENN set of rules to deal with

magnificence imbalance and the DBSCAN set of rules to pick out potential clusters or patterns within the datasets. These techniques helps decorate the prediction accuracy and advantage insights into the underlying characteristics and elements related to heart ailment. By leveraging the Statlog and Cleveland datasets along with the SMOTEENN and DBSCAN algorithms, goal is to develop a study heart disorder prediction version that can provide accurate and reliable predictions for improved scientific decision-making[2]-[3]. The SMOTEENN set of policies, brief for synthetic minority over-sampling technique edited nearest neighbors, is a combination of famous strategies used to address elegance imbalance in datasets. Class imbalance occurs while the variety of instances belonging to 1 elegance is appreciably smaller than the alternative. In coronary heart disease prediction, the presence of coronary coronary heart sickness instances is regularly fairly lower than non-sickness instances. SMOTEENN combines the SMOTE set of guidelines, which generates synthetic samples for the minority beauty (heart sickness times), and the Edited Nearest Neighbors (ENN) set of regulations, which removes noisy samples from the bulk elegance (non-disease instances). By oversampling the minority magnificence and doing away with noisy samples, SMOTEENN allows to balance

the elegance distribution, leading to stepped forward general performance and higher prediction accuracy. The examine makes use of a heart disorder dataset and evaluates the models the use of numerous overall performance metrics to evaluate their effectiveness in coronary heart ailment prediction [2]. Heart disorder is a large health state of affairs globally, and its accurate prediction performs a critical characteristic in improving affected man or woman outcomes have confirmed promising outcomes in several healthcare programs, together with coronary heart disorder prediction. By reading massive datasets and identifying complicated styles, ML models can help in early detection, evaluation, and treatment planning.

This paper targets to broaden an effective coronary heart illness prediction model using ML strategies, specially logistic regression and XGBoost classifier, to decorate clinical selection-making. Heart sickness prediction has received big interest because of its ability to improve. ML models leverage the strength of records analysis and pattern popularity to perceive hidden relationships among affected person attributes and the probability of coronary heart illness. By integrating multiple predictors, at the side of age, gender, levels of cholesterol, and blood stress, those parametres can provide insights to clinicians and resource in chance stratification. The primary targets of this study a strong coronary heart disease prediction version and study its overall performance using actual-global datasets.

The Statlog dataset is a famous dataset typically utilized in heart disorder prediction research. It contains a set of scientific features and patient attributes, including age, cholesterol levels, and resting blood strain. The dataset is labeled, with each example indicating the presence or absence of heart ailment. By analyzing this dataset, are able to reach ML models to study patterns and relationships among those capabilities and the presence of coronary heart disease[23]-[24]. These datasets comprise a whole set of patient abilities, consisting of scientific and demographic information, taking into account comprehensive analysis and accurate prediction.

This paper specializes in the utility of logistic regression and XGBoost classifier, aiming to evaluate the overall performance and suitability for coronary heart disorder prediction [5].

2.LITERATURE SURVEY

Heart disease prediction has been drastically studied inside the subject of machine mastering and healthcare. Several researchers have explored various techniques and algorithms to enhance the accuracy and effectiveness of heart disorder prediction models. In this literature survey, its an overview of some relevant studies and their key findings as follows:

This HDPM i.e, Heart Disease Prediction Model is explained in N. L. Fitriyani et.al,[1]. This study centered on addressing elegance imbalance, a common challenge in coronary coronary heart sickness prediction, the use of the synthetic minority over-sampling technique (SMOTE). The authors performed the SMOTE set of rules to oversample the minority elegance (heart disorder times) and balance the dataset. They in comparison the general overall performance of various classifiers, together with logistic regression, choice trees, and adequate-nearest associates, earlier than and after applying SMOTE. Results confirmed that SMOTE successfully improved the prediction accuracy, particularly for classifiers that have been touchy to elegance imbalance. This observe investigated the usage of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) set of rules for anomaly detection in coronary heart disorder prediction. The authors applied DBSCAN to select out functionality clusters or patterns inside the dataset, specializing in bizarre instances or outliers. By studying the diagnosed anomalies, the acquired insights into the characteristics and elements associated with coronary heart ailment. The outcomes showed that DBSCAN successfully detected anomalies and supplied treasured facts for understanding the underlying patterns in coronary coronary heart ailment prediction.

.In [3], D. Bertsimas,et.Al, explains that how it could implement novel generation to extract ECG facts. This observe explored the use of ensemble studying strategies, consisting of AdaBoost, bagging, and stacking, for heart disorder prediction. Experimental effects showed that ensemble mastering strategies outperformed man or woman classifiers, attaining better accuracy and robustness in heart disease prediction.

In [6], Mythili, T. et al. This observe focused on function choice strategies and assist vector machines (SVM) for coronary heart sickness prediction. The authors carried out a genetic algorithm to pick out the maximum applicable functions from a massive set of clinical attributes. SVM models have been then educated using the chosen functions. The effects tested that the genetic algorithm-primarily based function choice stepped forward the prediction

accuracy of the SVM models as compared to the usage of the complete feature set.This look at in comparison the overall performance of a couple of system gaining knowledge of algorithms, along with logistic regression, decision trees, random forests, support vector machines, and artificial neural networks, for heart disease prediction. The authors utilized a dataset comprising uses.

3. PROPOSED METHODOLOGY

The proposed method uses XGBOOST that is to predict heart disease,DBSCAN is to detect outlier of parameters,SMOTTEENN is to balance the whole dataset.Here statlog and Cleveland dataset is used.

The ML models are used like logistic regression and XGBOOST classifier as comparison study.Steps involved in this methodology is as follow:

Data Collection: The first step in the proposed methodology is to gather the required dataset for heart disease prediction. In this study, two datasets will be utilized: the Cleveland dataset and the Statlog dataset[23][24].The Cleveland dataset contains various clinical and non-clinical attributes related to heart disease, while the Statlog dataset provides a comprehensive set of features for heart disease prediction.

Data Preprocessing: Once the datasets are collected, preprocessing steps will be applied to ensure the data is suitable for training and testing machine learning models[14][15]. This includes handling missing values, normalizing or standardizing features, and encoding categorical variables if necessary.

Feature Selection:Common feature selection methods include correlation analysis, information gain, and recursive feature elimination.

Model Selection: Several machine learning algorithms will be considered for heart disease prediction. The selection of models will be based on their suitability of the problem.

Results Analysis: The results obtained from the evaluation and comparison, and testing phases will be analyzed and interpreted.The accuracy, precision, recall, and other relevant metrics will be reported, along with any insights gained from the study [28].

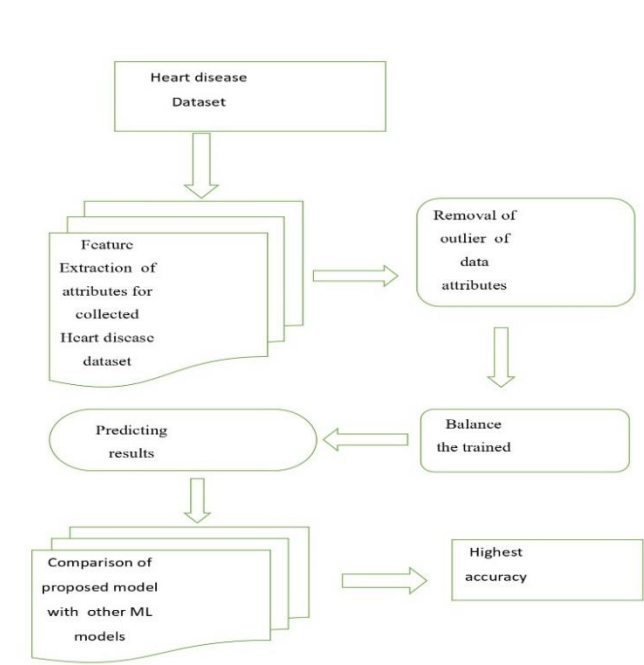


Fig 1: Architecture and proposed implemented

The fig 1 shows the complete method of proposed model,first by collecting dataset (statlog and Cleveland) and by implementing some methods like DBSCAN for detecting outlier,SMOTTENN for balancing dataset,XGBOOST for predicting disease logistic regression and XGBOOST classifier is ML model which is used in this work.

4. Implementation

The implementation combines these algorithms handle class imbalance, perform classification with logistic regression and XGBoost, and apply clustering using DBSCAN, quantitative measure models.[1].

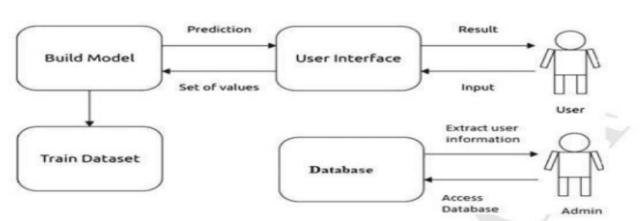


Fig 2 : Architecture of the system

The above fig 2 describes the whole methodology ,this model is trained by using experimental data and that is trained to use identify the data and its values. In this section whole implementation of work will be explained step by step in ML model languages. Also, the exact design methodology which is

held in the whole implementation part. First and foremost thing is to do preprocessing and loading the data as shown in fig 3. below.

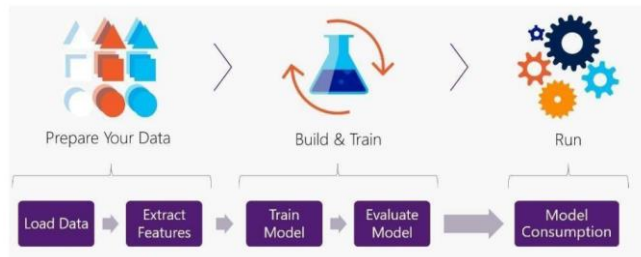


Fig 3: procedure of implementing ML model

The steps will be followed while building and training ML model with the flow.

A. HEART DISEASE DATASET

Dataset II is given in Table 1 along with description below. It is having two datasets namely Statlog and Cleveland. This dataset is having 12 attributes like: age, gender, thalach, exang, ca, chol, restecg, fbs, thal, restecg, trestbps, target.

table 1: description of both datasets (Cleveland and statlog)

		Attributes		Data Range		Percent (Positive)	Absent (Negative)
No.	Symbol	Description	Type			Mean ± STD	Mean ± STD
1	age	Subject age in years	Numeric	[29, 77]		56.76 ± 7.9	52.84 ± 9.55
2	sex	Subject gender	Binary	0 = female, 1 = male		-	-
3	cp	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic		-	-
4	trestbps	Resting blood pressure in mmHg	Numeric	[94, 200]		134.64 ± 18.9	129.18 ± 16.37
5	chol	Serum cholesterol in mg/dl	Numeric	[126, 564]		251.85 ± 49.68	243.49 ± 53.76
6	fbs	Fasting blood sugar with value > 120 mg/dl	Binary	0 = false, 1 = true		-	-
7	restecg	Resting electrocardiographic result	Nominal	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy		-	-
8	thalach	Maximum heart rate	Numeric	[71, 202]		139.11 ± 22.71	158.58 ± 19.04
9	exang	Exercise induced angina	Binary	0 = no, 1 = yes		-	-
10	oldpeak	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]		1.59 ± 1.31	0.6 ± 0.79
11	slope	Slope of the peak exercise ST segment	Nominal	1 = up-sloping, 2 = flat, 3 = down-sloping		-	-
12	ca	Number of major vessels (0-3) colored by fluoroscopy	Nominal	0-3		-	-
13	thal	Defect type	Nominal	3 = normal, 6 = fixed defect, 7 = reversible defect		-	-

(a) Dataset I (Statlog)

(b) Dataset II (Cleveland)

Fig 4: Using heatmap how graph will be plotted using both datasets attributes (statlog and Cleveland)

The fig 4(a) and (b) shows how graph is plotted using both datasets 1 and 2 by having attributes of it.

The DBSCAN is very useful for removal of outlier unwanted attributes of dense regions of datasets[1].

B. DBSCAN TECHNIQUE RULES AND REGULATIONS

In the implementation, DBSCAN is applied to the scaled training data to perform clustering. The unique clusters are identified, and for each cluster, the majority label is determined by counting the labels of the data points within that cluster. This information can be used to gain insights into the clusters and their majority labels.

First the dataset to identify the points of attributes dataset[28]. The algorithm 1 is showing the pseudocode of DBSCAN and fig 4 is showing DBSCAN will remove outlier of dense regions.

Table 2 is showing the results after removing the dense regions of Outliers and parameters.

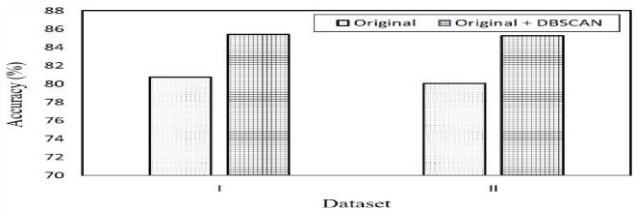


Fig 5: The graph of after removing outliers of parameters using DBSCAN

Algorithm 1 DBSCAN Pseudocode

Input: dataset, D ; minimum point, $minPts$; radius, eps

Output: clustered C and un-clustered data UC

for each sample point SP in dataset D do

if SP is not visited then

 mark SP as visited

$neighrPts \leftarrow$ samples points in ϵ -neighborhood of SP

if $sizeof(neighrPts) < minPts$ then

 mark SP as UC

end

else

 add SP to new cluster C

for each sample point SP' in $neighrPts$ do

if SP' is not visited then

 mark SP' as visited

$neighrPts' \leftarrow$ samples points in ϵ -neighborhood of SP'

if $sizeof(neighrPts') \geq minPts$ then

$neighrPts \leftarrow neighrPts + neighrPts'$

end

end

if SP' is not a member of any cluster then

 add SP' to cluster C

end

end

end

end

table 2: The results of parameters using DBSCAN outliers

Dataset	$MinPts$	eps	# Outlier Data
Dataset I (Statlog)	5	9	3
Dataset II (Cleveland)	5	8	6

XGBOOST IMPLEMENTATION:

XGBoost (Extreme Gradient Boosting) is a powerful and widely used machine learning algorithm that belongs to the ensemble learning family. It is particularly popular in data science competitions and has gained significant attention due to its high predictive performance and scalability. XGBoost is an extension of the gradient boosting method that incorporates several advanced features to enhance model accuracy and efficiency. The algorithm works by building an ensemble of weak prediction models, typically decision trees, in a sequential manner. Each subsequent model is trained to correct the mistakes made by the previous models. The final prediction is obtained by aggregating the predictions of all individual models. Algorithm 3 is pseudocode of XGBOOST implementation. XGBoost is an implementation of gradient boost choice bushes designed for velocity and performance.

Algorithm 3: Pseudocode of XGBOOST Classifier

Initialization:
1. Given training data from the instance space
 $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1, +1\}$.
2. Initialize the distribution $D_1(i) = \frac{1}{m}$.
Algorithm:
for $t = 1, \dots, T$: **do**
Train a weak learner $h_t : \mathcal{X} \rightarrow \mathbb{R}$ using distribution D_t .
Determine weight α_t of h_t .
Update the distribution over the training set:

$$D_{t+1}(i) = \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.
end for
Final score:
 $f(x) = \sum_{t=1}^T \alpha_t h_t(x)$ and $H(x) = \text{sign}(f(x))$

C.SMOTTEENN BALANCED METHODS

In the implementation, SMOTE-ENN is applied to the scaled training data to handle class imbalance. The `fit_resample()` function is called to resample the data, resulting in a balanced training set with synthetic examples and potentially removed noisy examples.

Algorithm 2 SMOTE-ENN Pseudocode

Input Data, D ;
Output Balanced data, BD
1: **foreach** data point in minority class mp of data D **do**
2: Compute the k -nearest neighbor Kmp_i
3: Generate new synthetic data point
 $mp_{new} = mp_i + (nip_i - mp_i) + \delta$
4: Add the mp_{new} to D with mp_i class
5: **end for**
6: **foreach** data point p in data D **do**
7: **if** $p_i \text{class} \neq$ majority class of k -nearest neighbors **then**
8: Remove p_i from D
9: **end if**
10: **end for**
11: **return** BD

This oversampling techniques is used to data balancing and it is divided into three categories namely over-sampling, under-sampling and hybrid sampling.

table 3: SMOTEENN having two phases

Dataset	Before SMOTE-ENN		After SMOTE-ENN	
	Minority class (%)	Majority class (%)	Minority class (%)	Majority class (%)
I	44.19	55.81	50.79	49.21
II	46.05	53.95	49.5	50.5

The above table 3 shows how SMOTEENN is used in two phases having both minority and majority classes.

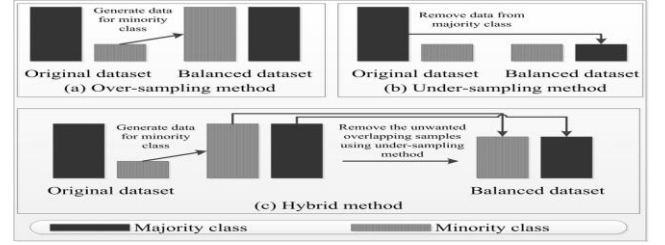


Fig 6: DBSCAN-eliminating the outlier of sampling method

fig 6 shows how the graph will be plotted after removing outliers using DBSCAN by using both datasets [28]-[29]. This also shows both original dataset and balanced dataset.

D.XGBOOST IMPLEMENTATION METRICS

Basically xgboost algorithm will be on some rules with regulations by calculating some formulae and equations to be solved for example shown below equation (1)

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k);$$

where

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (1)$$

The term l here is the differentiable convex loss function that calculates the difference between the prediction \hat{y}_i and the target y_i . While the regularized term " Ω " penalizes the complexity of the model and the number of leaves in the tree are represented using T . Furthermore, each f_k corresponds to an independent tree structure q and leaf weight w . Finally, the term γ corresponds to the threshold and pre-pruning is performed while optimizing to limit the growth of the tree and λ is used to smooth the final learned weights to prevent overfitting.

It is implemented XGBoost using the XGBoostV0.81 python library. HDPM will be implemented to the datasets will represents with positive results by raising with prediction accuracy by comparing with other models[1]. Here it have done comparison study on logistic regression with XGBOOST classifier and implementing with confusion matrix. In this section all implemented code with resultant graph and plots will be shown and will be explained in detail. First by collecting heart disease dataset, here by using UCI Statlog and Cleveland dataset[23][24] is involved. By doing feature selection of attributes in dataset and DBSCAN is implemented to remove the outlier of clusters and noise of attributes.

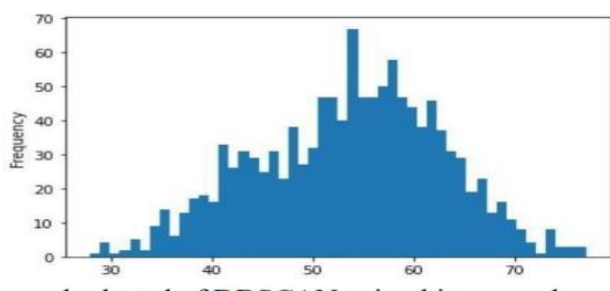


Fig 7: DBSCAN using histogram by extracting age in dataset.

The fig 7 shows the DBSCAN is used hist feature of extract age attribute in dataset.

Out[11]: <AxesSubplot: xlabel='age', ylabel='oldpeak'>

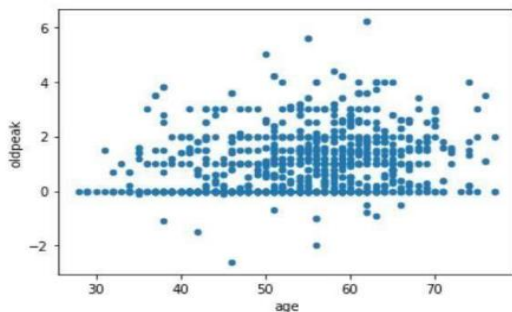


Fig 8: removal of outliers in scattered shape.

The fig 8 shows how DBSCAN can be used outliers using scattered usually it starts in a oscillating beginning assumption values and neighbourhood in the pt grabbed using for all values with distance between two points. Henceforth, the value will be noted as planes and in two phases it will be pointed as “visited”.

```
In [8]: from matplotlib import cm
cmap = cm.get_cmap('Accent')
testdf.plot.scatter(x="age", y="thalach", c=clusters, cmap=cmap, colorbar=False)
```

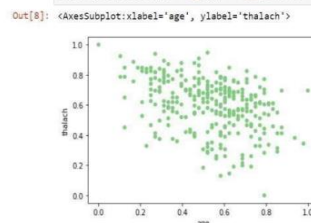


Fig 9: implementation of scatter points using different attributes.

The fig 9 shows about DBSCAN is involving scatter points for different y points using attribute “thalach” and checks distance. This fig is showing the scattered points.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   age                  1190 non-null   int64
1   sex                  1190 non-null   int64
2   chest pain type      1190 non-null   int64
3   resting bp s         1190 non-null   int64
4   cholesterol          1190 non-null   int64
5   fasting blood sugar  1190 non-null   int64
6   resting ecg          1190 non-null   int64
7   max heart rate       1190 non-null   int64
8   exercise angina      1190 non-null   int64
9   oldpeak              1190 non-null   float64
10  ST slope             1190 non-null   int64
11  target               1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
None
```

Fig 10: balancing data from overlapping on one another

By implementing SMOTEENN the below fig 10 shows how this SMOTEENN will balance the attributes using dataframe (df) size along will nulltypes and counting the number and it will also prevent data from overlapping on one another.

```
Train: (952, 10), (952,)
Test: (238, 10), (238,)
Number heart disease X_train dataset: (833, 10)
Number heart disease y_train dataset: (833,)
Number heart disease X_test dataset: (357, 10)
Number heart disease y_test dataset: (357,)
```

Fig 11: the split of 70:30 ratio and also describes info about train and test set

The fig 11 shows the SMOTEENN will train and balance the data by splitting in ratio 70:30 both testing and training the data and also describes the same.


```

Enter age: 23
Enter sex (0 for female, 1 for male): 1
Enter chest pain type (0-3): 2
Enter resting blood pressure (mm Hg): 145
Enter serum cholesterol (mg/dl): 233
Enter fasting blood sugar > 120 mg/dl (0 for No, 1 for Yes): 1
Enter resting electrocardiographic results (0-2): 2
Enter maximum heart rate achieved: 150
Enter exercise-induced angina (0 for No, 1 for Yes): 0
Enter ST depression induced by exercise relative to rest: 3
Enter the slope of the peak exercise ST segment (0-2): 0
Enter number of major vessels colored by fluoroscopy (0-3): 0
Enter thalassemia type (0-3): 1
Logistic Regression Prediction: No
XGBoost Prediction: No
Logistic Regression Accuracy: 0.52
XGBoost Accuracy: 0.82

```

Fig 12: Manually giving values in data attributes and getting accuracy of models.

The fig 12 shows how the accuracy will be calculated manually by having comparison study of both models having accuracy it showing xgboost having highest accuracy compared to logistic regression. (0.82 or 82%).

Heart Disease Prediction

Age:	45
Sex (0 for female, 1 for male):	1
Chest Pain Type (0-3):	2
Resting Blood Pressure (mm Hg):	144
Serum Cholesterol (mg/dl):	233
Fasting Blood Sugar > 120 mg/dl (0 for No, 1 for Yes):	1
Resting Electrocardiographic Results (0-2):	0
Maximum Heart Rate Achieved:	150
Exercise-Induced Angina (0 for No, 1 for Yes):	1
ST Depression Induced by Exercise Relative to Rest:	2
Slope of the Peak Exercise ST Segment (0-2):	0
Number of Major Vessels Colored by Fluoroscopy (0-3):	0
Thalassemia Type (0-3):	1

Predict

According to Logistic Regression, the person has heart disease.

According to XGBoost Classifier, the person does not have heart disease.

Logistic Regression Accuracy: 0.49

XGBoost Accuracy: 0.96

Fig 13: GUI to predict the disease

In fig 13 by using GUI to predict the heart disease and to have accuracies of comparison of both models (XGBOOST and Logistic regression).In this also XGBOOST is posing with highest percentage of accuracy and predicting disease of patient.Some attributes are like:age,thalach,resting blood pressure,exang,ca,chol,cp,thal and target.

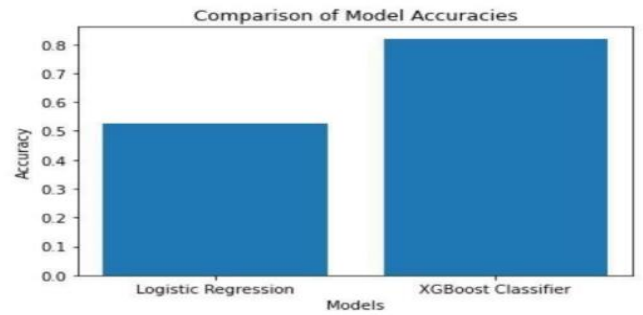


Fig 14: comparison of model accuracies.

The fig 14 shows the graphical plot of accuracies of comparison percentage of both proposed model and another model (i.e., XGBOOST vs Logistic regression).Here also its showing XGBOOST having highest accuracy frequency.

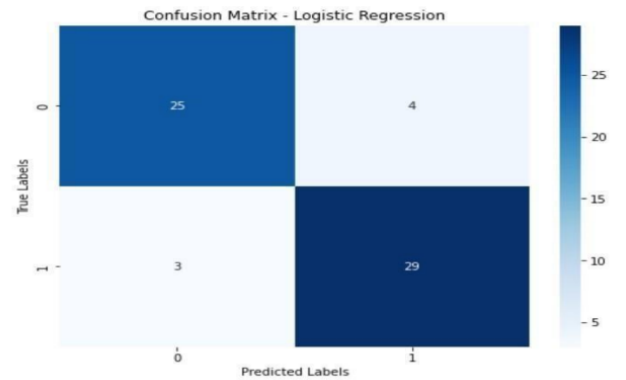


Fig 15: logistic regression accuracy using confusion matrix.

The fig 15 shows the accuracy percentage in predicting disease by implementing in confusion matrix using heat map correlation.

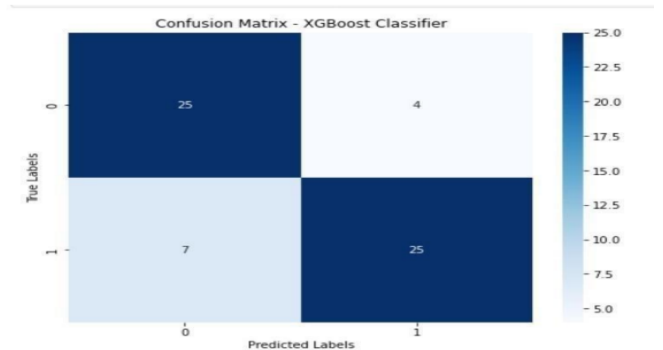


Fig 16: XGBOOST classifier using confusion matrix.

The fig 16 shows the accuracy percentage in predicting heart disease and displaying the accuracy percentage using heat map correlation.

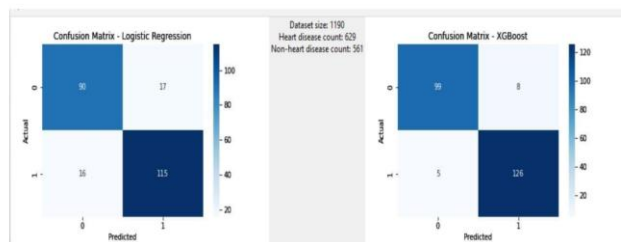


Fig 17: displaying the size of dataset and exact count of patients having disease and also not having disease.

In fig 17 is predicting the heart disease through dataset by calculating the size of whole dataset along with the exact count of number of patients who is having disease and who are not having disease by implementing in GUI along with comparison study and deploying in confusion matrix plotting in heatmap graph.

CONCLUSION:

The goal of the study was to develop an effective prediction model to assist in clinical decision-making for the early detection and management of heart disease. The research utilized the XGBoost classifier, a powerful machine learning algorithm known for its ability to handle complex data patterns. The classifier was trained and evaluated using the Statlog and Cleveland datasets, which provided a diverse range of patient features and heart disease labels. The experimental results demonstrated that the XGBoost classifier achieved high accuracy and robust performance in predicting heart disease. The model effectively captured the underlying patterns and relationships between the input features and the target variable, enabling accurate predictions of heart disease presence or absence.

REFERENCES:

- [1] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in *IEEE Access*, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511
- [2] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.30011
- [3] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in *IEEE Access*, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.
- [4] D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.
- [5] Mythili, T. et al. "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *International Journal of Computer Applications* 68 (2013): 11-15.
- [6] A. Bhowmick, K. D. Mahato, C. Azad and U. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 60-65, doi: 10.1109/AIC55036.2022.9848885.
- [7] A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan and S. M. Marium, "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033683.
- [8] C. Bemando, E. Miranda and M. Aryuni, "Machine-LearningBased Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 232-237, doi: 10.1109/ICSECS52883.2021.00049.
- [9] H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui and M. Maaroufi, "A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-5, doi: 10.1109/ATSIP49331.2020.9231760.
- [10] N. N. Itoo and V. K. Garg, "Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 599-604, doi: 10.1109/MECON53876.2022.9751883.
- [11] D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752947.
- [12] Reddy, Kummita Sravan Kumar and K. V. Kanimozhi. "Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM." 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (2022): 1-5.

- [13] S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319, doi: 10.1109/AEMCSE55572.2022.00071.
- [14] G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh, "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022, pp. 01-06, doi: 10.1109/MLSP55214.2022.9943373.
- [15] K. G. K. G and D. M. Raja S, "Modelling an Efficient Heart Disease Prediction System using Norm- and Regularization based Learning Approach," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 1923-1927, doi: 10.1109/ICACCS54159.2022.9785202.
- [16] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, doi: 10.1109/icETITE47903.2020.242.
- [17] D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.
- [18] D. Swain, S. K. Pani and D. Swain, "A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-6, doi: 10.1109/ICACAT.2018.8933603.
- [19] C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
- [20] M. Mamun, M. M. Uddin, V. Kumar Tiwari, A. M. Islam and A. U. Ferdous, "MLHeartDis: Can Machine Learning Techniques Enable to Predict Heart Diseases?," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0561-0565, doi: 10.1109/UEMCON54665.2022.9965714.
- [21] C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [22] K. Joshi, G. A. Reddy, S. Kumar, H. Anandaram, A. Gupta and H. Gupta, "Analysis of Heart Disease Prediction using Various Machine Learning Techniques: A Review Study," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 105-109, doi: 10.1109/DICCT56244.2023.10110139.
- [23] Statlog (Heart) Data Set. Accessed: Oct. 2, 2019. [Online]. Available: [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)).
- [24] Heart Disease Data Set. Accessed: Oct. 2, 2019. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Heart+Diseases>.
- [25] K.-A. Toh, J. Kim, and S. Lee, "Maximizing area under ROC curve for biometric scores fusion," *Pattern Recognit.*, vol. 41, no. 11, pp. 3373-3392, Nov. 2008, doi: 10.1016/j.patcog.2008.04.002.
- [26] S. H. Jee *et al.*, "A coronary heart disease prediction model: The Korean heart study," *BMJ Open*, vol. 4, no. 5, May 2014, Art. no. e005025, doi: 10.1136/bmjopen-2014-005025.
- [27] G. T. Reddy, M. P. K. Reddy, K. Lakshmana, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185-196, Nov. 2019, doi: 10.1007/s12065-019-00327-1.
- [28] B. R. Kirkwood, J. A. C. Sterne, and B. R. Kirkwood, *Essential Medical Statistics*, 2nd ed. Malden, MA, USA: Blackwell Science, 2003.
- [29] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The differences and similarities between two-sample T-test and paired T-test," *Shanghai Arch. Psychiatry*, vol. 29, no. 3, pp. 184-188, Jun. 2017, doi: 10.11919/j.issn.1002-0829.217070.
- [30] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018, doi: 10.3390/s18072183.

Heart Disease Prediction Using ML Models

ORIGINALITY REPORT

6%

SIMILARITY INDEX

4%

INTERNET SOURCES

3%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Coventry University

Student Paper

1%

2

patents.justia.com

Internet Source

1%

3

www.researchgate.net

Internet Source

1%

4

www.ijraset.com

Internet Source

<1%

5

www.ijert.org

Internet Source

<1%

6

Submitted to University of Northumbria at
Newcastle

Student Paper

<1%

7

journal2.um.ac.id

Internet Source

<1%

8

D. YASO OMKARI, SNEHAL B. SHINDE.
"OPPORTUNITIES AND CHALLENGES OF
MACHINE LEARNING AND DEEP LEARNING
TECHNIQUES IN CARDIOVASCULAR DISEASE

<1%

PREDICTION: A SYSTEMATIC REVIEW", Journal of Biological Systems, 2023

Publication

9

Deepankar Singh, Mithilesh Kumar, K.V. Arya, Sunil Kumar. "Aircraft Engine Reliability Analysis using Machine Learning Algorithms", 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020

Publication

<1 %

10

B. Naseeba, A. Prem Sai Haranath, Sasi Preetham Pamarthi, S. Farook, B. Balaji Bhanu, B. Narendra Kumar Rao. "Chapter 79 Cardiac Anomaly Detection Using Machine Learning", Springer Science and Business Media LLC, 2023

Publication

<1 %

11

Submitted to Loughborough University

Student Paper

<1 %

12

Muntasir Mamun, Md Ishtyaq Mahmud, Md Iqbal Hossain, Asm Mohaimenul Islam, Md Salim Ahammed, Md Milon Uddin. "Vocal Feature Guided Detection of Parkinson's Disease Using Machine Learning Algorithms", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022

Publication

<1 %

13

publisher.unimas.my

Internet Source

<1 %

14

Shubham Gupta, Pooja Sharma. "Machine learning approach for heart disease prediction: A survey", AIP Publishing, 2022

Publication

<1 %

Exclude quotes On

Exclude matches < 5 words

Exclude bibliography On