International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Heart Disease Risk Predictor using Support Vector Machine

Pooja Saharan, Rashmi Mishra, Charvee Garg, Aman Payal

Abstract: ***Nearly 17.5 million deaths occur due to cardiovascular diseases throughout the world. If we could create such a mechanism or system that could tell people about their heart condition based on their medical history and warn them of any risk than it could be of huge help. In our work, we will use machine learning algorithms to forecast the heart disease risk factor for a person depending upon some attributes in their medical history. The data mining technique Naive Bayes, Decision tree, Support Vector Machine, and Logistic Regression is analyzed on the Heart disease database. The accuracy of different algorithms is measured and then the algorithms are compared.***

IndexTerm: Data Mining, Heart Disease, Jupiter notebook, SVM (Support Vector Machine).

I. INTRODUCTION

A popular saying goes that we are living in an "information age". Large amounts of data are produced every day. Efficient tools are not much prevalent to extract knowledge from these databases for clinical detection of diseases or other purposes. A big challenge that healthcare organizations are dealing with (medical centers, hospitals) is the facility of better facilities at prices which are lower and most which most people can afford. Quality service means checking patients rightly and performing treatments that give better results. Bad clinical decisions can result in risky consequences which may not be appreciated and accepted by people. Hospitals can achieve these things by appointing proper decision support systems or computer-based information. "To make intelligent clinical decisions on how to transform information into information which can help medical people to make better clinical related decisions?"

Revised Manuscript Received on 30 July 2019.
* Correspondence Author
Pooja Saharan*, Computer Science & Engineering, ABESEC, Ghaziabad, India, pooja.saharan@abes.ac.in
Rashmi Mishra, Computer Science & Engineering, ABESEC, Ghaziabad, India,rashmi.mishra@abes.ac.in
Charvee Garg, Computer Science & Engineering, ABESEC, Ghaziabad, India,charvee.15bcs1028@abes.ac.in
Aman Payal, Computer & Engineering, ABESEC, Ghaziabad, India,aman.15bcs2037@abes.ac.in

We have planned a system to forecast heart disease risk by employing data mining techniques and machine learning algorithms. In the existing scenario many authors elaborated on the prediction of heart disease by employing data mining methods i.e. classification, regression, decision tree and many more. Sellappan.et.al found how the vast amount of data generated by the healthcare industry is not utilized for making the better decision making.

The finding of masked patterns and ideas and correlations often go undiscovered. In this situation, new improved data mining techniques can aid us [1]. This paper provided an Intelligent Heart Disease Risk Predictor System. Marija.et.al analyzed various data mining techniques for the prediction that major causes of death have been found out to be heart disease around the world. Its prediction is not easy as it requires high expertise and accuracy. This paper according to some input attributes of a personal medical history points out the problem of heart disease prediction.[2] The authors made the Heart Disease Risk Predictor through the Weka software using SMO, Multilayer Perceptron, KStar, J48, Bayes Net. Each algorithm performance is compared and measured by adding the results of predictive accuracy, AUC value and ROC curve using collected data and standard data. Based on results Bayes Net and SMO technique performed better than that of KStar, J 48, and Multilayer Perceptron. Soodeh.et.al put forward a machine-learning algorithm to forecast the risk of coronary artery atherosclerosis. The missing values in the database of atherosclerosis are estimated by a Ridge expectation Maximization Imputation Technique (REMI). A conditional likelihood Maximization Method is used for dimensionality reduction. The UCI and STULONG databases are used to extend the performance of prediction Systems. The results showed an improvement in the percentage of accuracy of risk prediction in their proposed method checked in relation to other works when they performed experiments. The effect of left out values in the algorithm for prediction is also evaluated and compared to any other traditional techniques the proposed REMI approach works better.[3] In world, Medical theorist facing a problem of predicting the heart related issues before they happen. Different Techniques are employed in reducing medical cost related problems and early prediction of disease.Naik.et.al provides shallow analysis and understanding of various prediction models based on data mining from last seven years [4]. The authors presented the comparison between each model given by various authors in terms of accuracy. Poornima.et.al explained that remarkable progress had already made in diagnosing the heart disease, but early prediction is still need to be explored.

The objective of authors were to develop a prediction system based upon apriority algorithm of association rule [5].The investigation with various risk factors such as sex, smoking glucose etc. are analyzed for the prediction and a total of 369 cases were collected from CHZ survey for the development of system. Authors discussed the previous prediction system in which. C4.5 algorithm applied for the performance analysis.
In Medical science data mining technique play an important role to extract the important information form the data base. As per the author, in many countries people are suffering from the cardiovascular disease. In current scenario many pharmaceutical companies' dealings with information constantly, some of the information is used for examine and produce the result for difficulties faced by the electrocardiography. For this data mining techniques gives proper results. Many scientists all over the world analyzing the factual of these disease by using the Data mining Apparatuses [6]. Authors found that the level of accuracy in finding the coronary illness using hybrid strategy are completely explored.

II. PROPOSED SYSTEM

Figure 1 shows the working of our system wherein there is a model that is trained using the test data and then that trained model is used to classify the data on the set of values provided by the user. We have planned a system in a medical analysis that would improve the medical debate and it can similarly lower the costs that can effectively find the doctors to predict the risk level of patients on the basis of some parameters about their health.

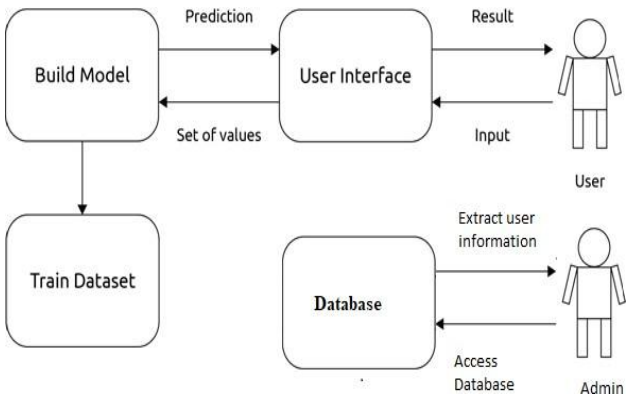

Figure 1: Prediction system

The main aim of our study is to help the not so specialized doctors to make a proper decision about the risk of heart disease for a patient. It aims at a combination of medicinal decision support with computer-based patient data which will reduce medical faults, increase patient safety, and improve the overall outcome. This would help to lower down the treatment price and also enhances how we see it and the ease of meaning with large knowledge and better data in this field. Large companies can invest heavily in this study to help divert attention on possible activities and risks that may be involved. Such type of work can bring together all the available data, as a basis on which we can develop rational assumptions about the future.

III. SYSTEM INTERFACE

System interface shows the capability of the model that it is works on all types of browsers and servers.

Table 1: Interface

Login page	Log into the system as a User
Admin Interface	Whether to check up or find your medical history
User interface	Health report will be displayed
Decide	Decision based on 1 and 0.

Default page is the login page from where a user can login or navigate to about us, sign up or admin login page. Every user after logging in will be redirected to the predict page where he can enter the details and predict the result. Here the user will have an option to redirect to the profile page or to logout. Profile page will show the details and profile picture of the user and the past predictions he has made if any. Hardware interface: Minimum requirement of system is 2 GB RAM, Pentium processor, Server to connect to the database. Software interface: Operating System: Windows 10, Ubuntu 17.04,Web Browser: Google Chrome, Mozilla Firefox, MySQL Database (v5.7.19): For connectivity with background databases, Anaconda- to run the machine learning code, Django and Python- for web application. Communications Interfaces: Web Browser with good Internet connection, Minimum Internet Speed for uploading, processing, details are 1Mbps. Memory Constraints: Minimum 4 GB RAM to run the project. Product Functions: The product will have the function of user login, signup and admin login, Admin will have the function to add new entry into the dataset or access the records, User can enter the values of various parameters on the basis of which his risk factor will be calculated. User Characteristics: Any user that wants to predict the risk needs to know the attributes required on the prediction page. Any wrong value of attribute can mislead the user, the values of each attribute should be in correct format and range. Constraints: The dataset used is small, The database used will be regular database which can store a limited amount of data, This product uses the data of users for prediction purposes only, as is evident as there is no" Name" input field. Thus, the privacy of the user is preserved. But the program uses open source libraries and any developer having doubts about safety can directly check the source code. The system needs to be reliable. If unable to process the request then appropriate error message will be displayed. Web pages are loaded within few seconds. The algorithm with the highest accuracy needs to be chosen for the web application

Retrieval Number: B3423078219/19©BEIESP
DOI: 10.35940/ijrte.B3423.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A. Use Case Diagram

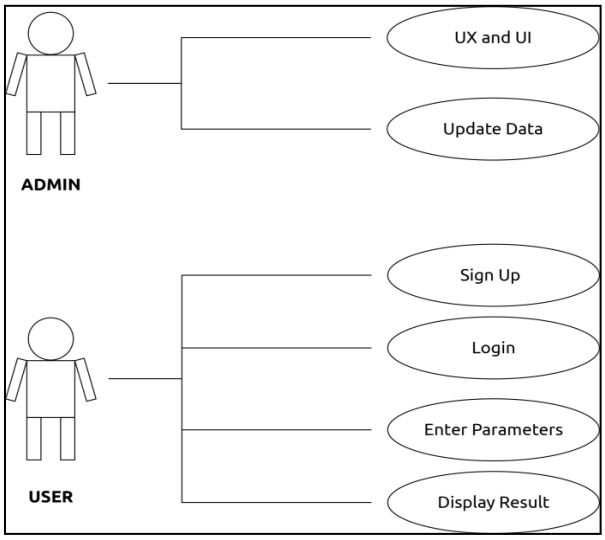

Figure 2: Use Case Diagram

Figure 2 above represents the use case diagram of our project. We have two kinds of user one is admin which will have access to the database and other is the user or the doctor which can login or signup and predict risk by filling the details.

A. Sequence Diagram

Figure 3 represents the sequence diagram of how activities are performed and how response is sent back to the user. It shows how data is sent to the model and how it classifies the data

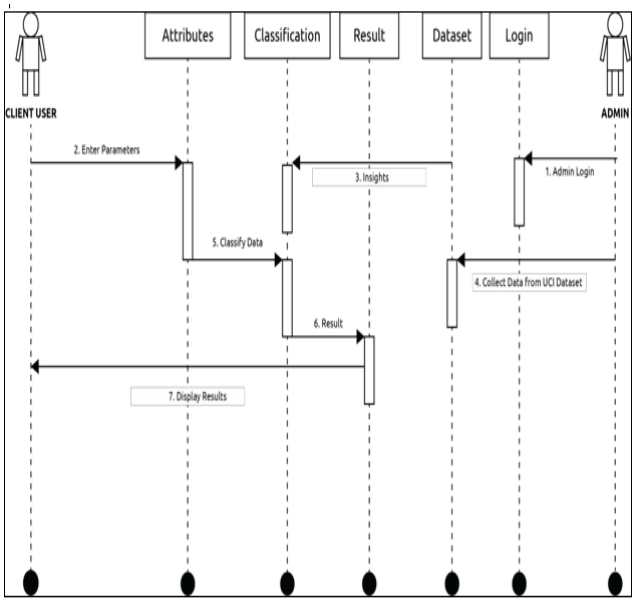

Figure 3 : Sequence Diagram

B. Activity Diagram

The prediction system allows the user to predict heart disease by creating the account login. Figure 2 represents the dynamic aspect of how the risk predicted using supervised learning. Initially, users and doctor's login their accounts with the credentials, if they are registered the user. The Collected attributes of patients are entered into the system and then data proceeded for the prediction.

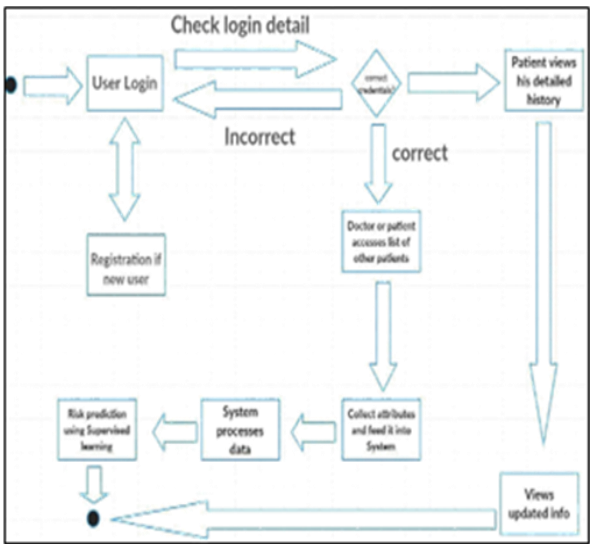

Figure 4: Activity Flow of Prediction System

IV. SYSTEM DESIGN

The proposed system is divided into 3 modules:
Module 1: User Registration and data set creation.
Module 2: Finding the chance of heart disease.
Module 3: Producing the remedy by the doctor to the patient.
In the first module user needs to register him/her self by entering their personal information into the login page and data set is created using the prediction engine has been implemented using various libraries of python like scikit-learn, panda and numpy. To predict if there is a risk of heart disease or not, they use various attributes like cholesterol, age, blood sugar. It allows users to browse through datasets, download datasets and donate datasets. The Cleveland dataset The UCI Data Repository contains 351 datasets maintained by the University of California; Irvine contains 303 records. The dataset contains 76 attributes but, we have only considered a subset of 14 attributes. Figure 5 shows major visualizations of the increment, how the different attributes in the UCI dataset are distributed. By seeing the entries entered by the user's doctor take the decision then another set of questions will be generated. Then remedies will be generated by the doctors.

V. METHODOLOGY

The proposed system is using SVM Data mining technique. We have use training dataset from Cleveland Heart Disease database with medical attributes. SVM attributes which we have taken for the prediction of heart diseases is listed below:

Table 2: Attribute

Serial No.	Attribute	Description
1	Sex	value 1: Male value 0: Female
2	Chest Pain Type	value 1: typical type 1 angina value 2: typical type angina value 3: non-angina pain value 4: asymptomatic
3	Fasting Blood Sugar	value 1: > 120 mg/dl value 0:< 120 mg/dl
4	RestECG	resting electrographic results value 0: normal value 1: 1 having ST-T wave abnormality value 2: showing probable/definite left ventricular hypertrophy
5	Exang	exercise induced angina value 1: yes value 0: no
6	Slope	the slope of the peak exercise ST segment value 1: unsloping value 2: flat value 3: downsloping
7	CA	number of major vessels colored by flourosopy (value 0 – 3)
8	Thal	value 3: normal value 6: fixed defect value 7: reversible defect
9	Trest Blood Pressure	(mm Hg on admission to the hospital)
10	Serum Cholesterol	(mg/dl)
11	Thalach	maximum heart rate achieved
12	Oldpeak	ST depression induced by exercise
13	Age	In year
14	Height	In cms
15	Weight	In kgs

A. DFD Level 1

Proposed system is based on Jupiter notebook. In the proposed system user need to register using their credential, after the successful of registration users record is stored into the database. After the successful login users need to answers the predefined questions. Then the answers are evaluated by the doctors to predict the heart disease. Figure 6 represents the data flow diagram at level 1. It shows a more detailed view of how data flows and the different entities and features of the system.

B. DFD Level 2

Figure 7 represents data flow diagram at level 2. It show how request and response is transmitted from system to user and from the user to the system for accessing all the modules of system.

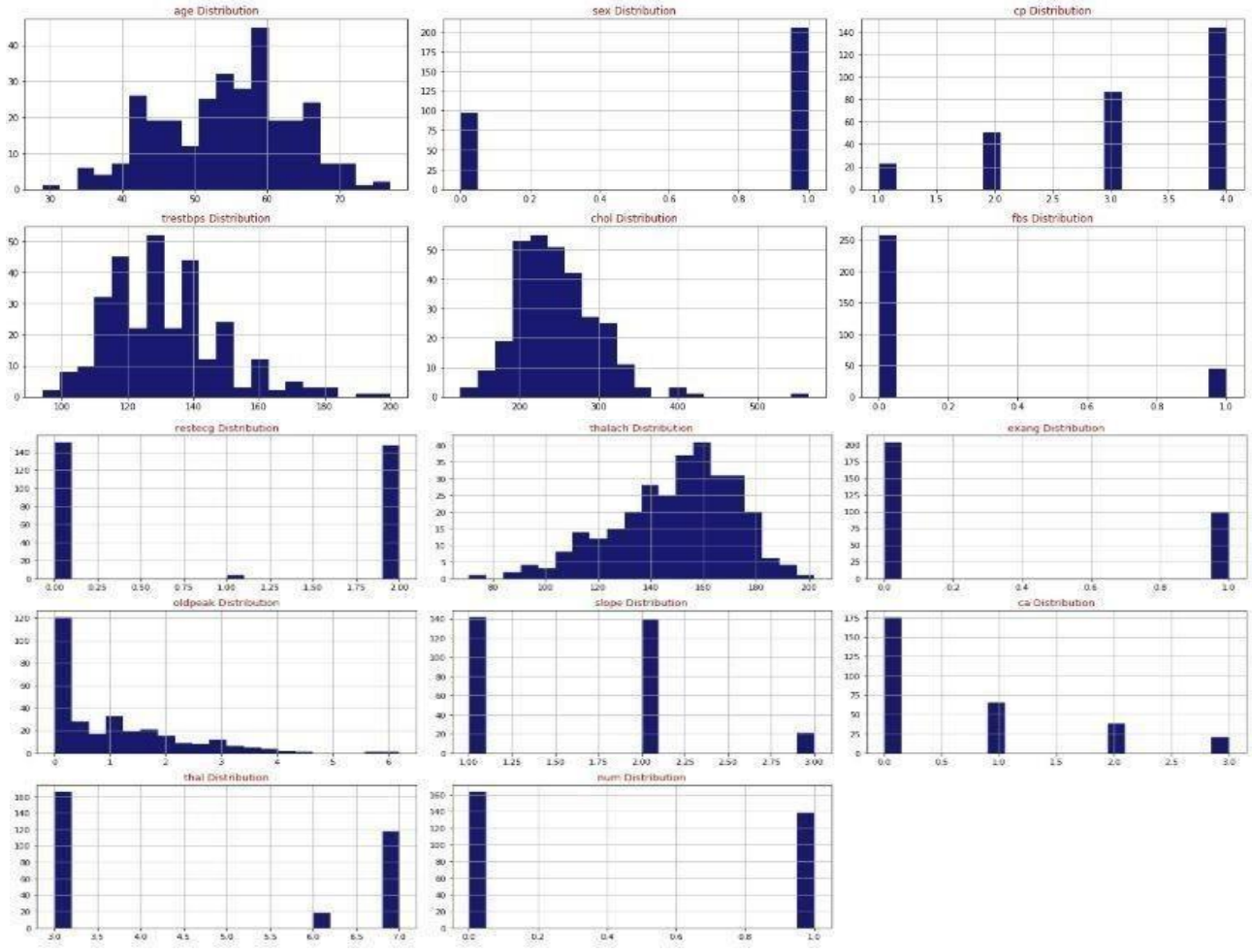

Figure 5: Histograms for the Attributes of Heart Disease

Retrieval Number: B3423078219/19©BEIESP
DOI: 10.35940/ijrte.B3423.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

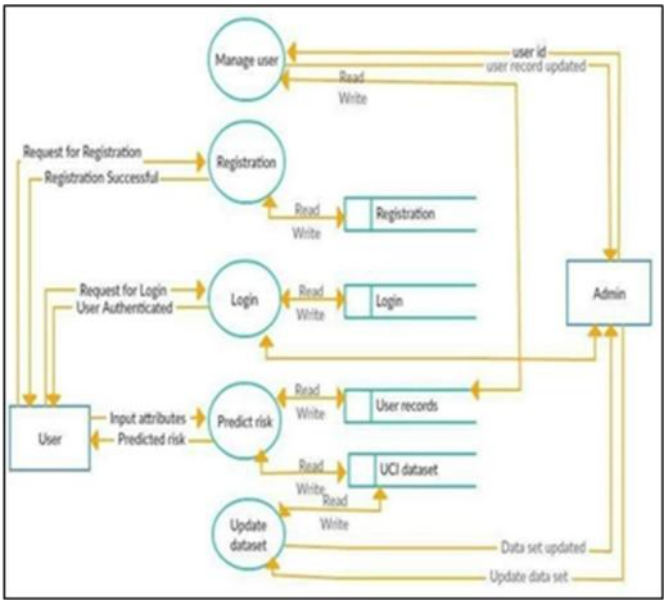

Figure 6: Data Flow Diagram (Level 1)

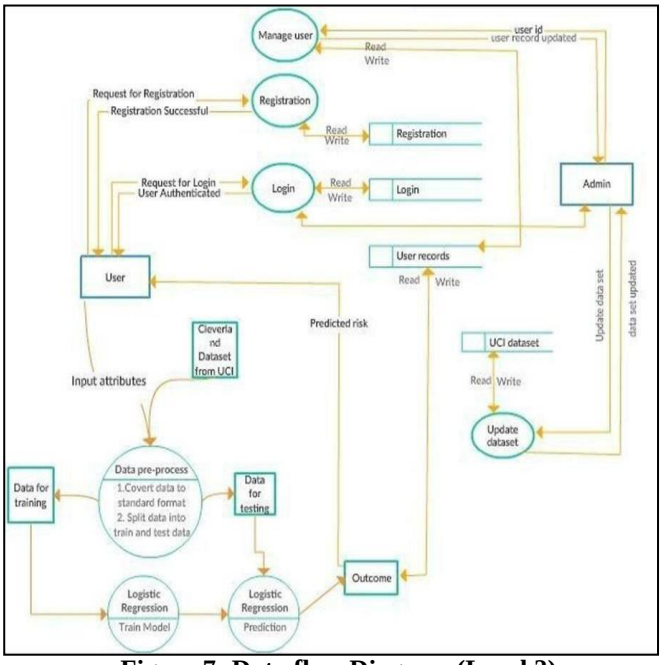

Figure 7: Data flow Diagram (Level 2)

VI. IMPLEMENTATION

A.Login Page: User need to register for the login. In login phase user enter his/her user id and password. Then authentication process is done through database.

Figure 8: login

B.Prediction page: User need to answers the pre-defined questions. By seeing the answers filled by the user, doctor generate the another set of questions which will help doctors to predict the heart disease.

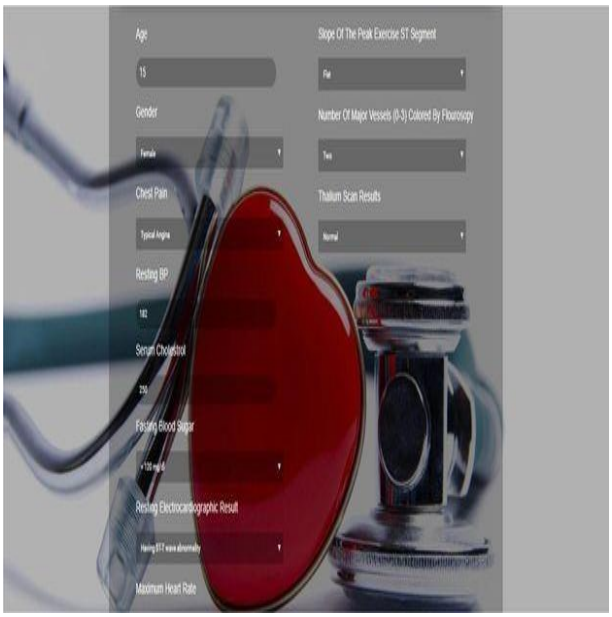

Figure 9: Prediction Page

C. Predict heart disease page: Medical information is filled by the users which is generated by the doctors in the last module, based on the estimated result measured by the test is filled by the doctor in the predict your heart disease page.

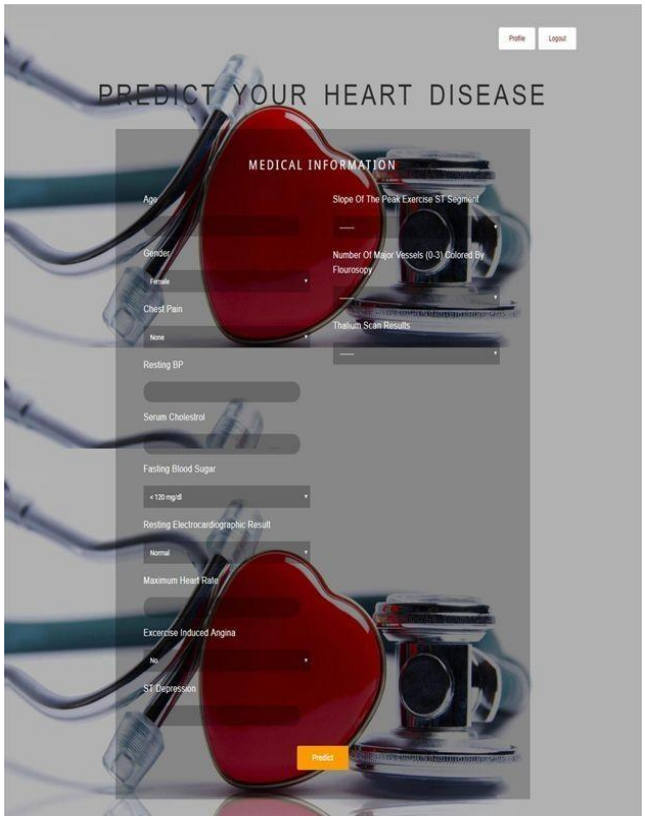

Figure 10: Predict Heart Disease

D. Prediction results Page: The Prediction System is user friendly and provide the faster diagnosis of heart disease. By reducing the medical errors, system showing whether there is a risk of heart diseases. The attribute "Risk of Heart Disease" having two discrete values (0,1). The attribute Value "1" showing the risk of heart Disease whereas "0" showing that there is no risk of heart-diseases. To make system convenient for the user, Django provide the API for the automatically generated database.

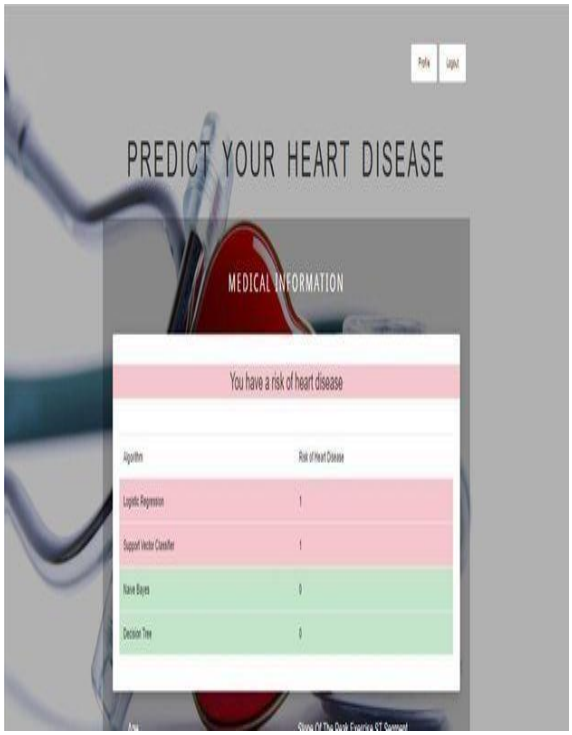

Figure 11: Result

VII. CONCLUSION

Currently, in India, the number of heart disease patients is more than 30 million. The unaware attitude of people towards their health status further leads to various diseases and can be a threat to their lives. Huge amounts of data are generated by the health care industry. However, mostly it is not effectively used. The data generated has some hidden patterns and relationships between them. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes are not much prevalent. The system uses input features like sex, cholesterol, blood pressure like more features to forecast the likeliness of patients getting a Heart disease. The result of this would be a result of either 0 or 1 which would represent no risk or risk for any person. Django has been used for the implementation of algorithms. Data mining can be of very knowledge form such a suitable dataset. It has been concluded that in the making of the predictive model only some success is achieved and hence there is a need for the combination of different models and more complex models to increase the accuracy of the predicting the risk of heart disease. With the more amount of data being fed into the database, the system will be very intelligent. To improve the scalability and accuracy there can be many improvements possible. In many countries, a doctor is liable for any negligence in patient care and might be held accountable. Thus, this project should only be considered as an assistive tool and must be completely relied upon only after it has been validated in the future by integrating more practical data.

Performance Evaluation

The evaluation of the machine learning algorithms and comparison between different algorithms based on their accuracy is given in below table. By doing the comparison in between the four types of algorithms, Support Vector Machine gives the best result which is not done by any author in literature survey. The implementation of algorithm is performed using "Jupyter Notebook" to find the accuracy of the prediction with different algorithms and to find the correlation between the variables.

TABLE 3: Performance Evaluation of ML algorithms.

ALGORITHMS	ACCURACY
Logistic regression	83.38%
Naïve Bayes	80.73%
Decision tree	80.54%
Support vector classification	87.89%

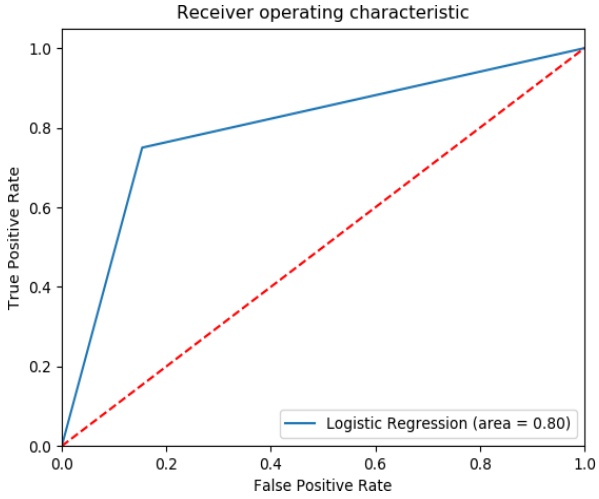

Figure 12: ROC curves for Decision Tree

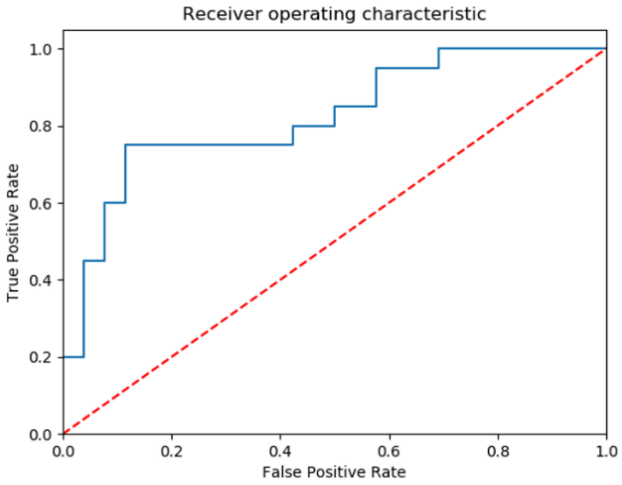

Figure 13: ROC curves for Logistic Regression

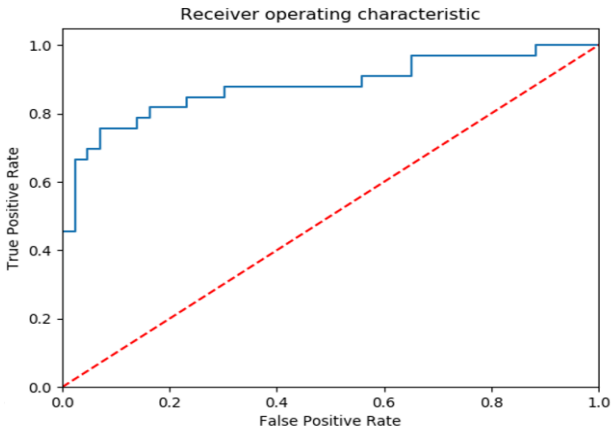

Figure 14: ROC curves for Naïve Bayes

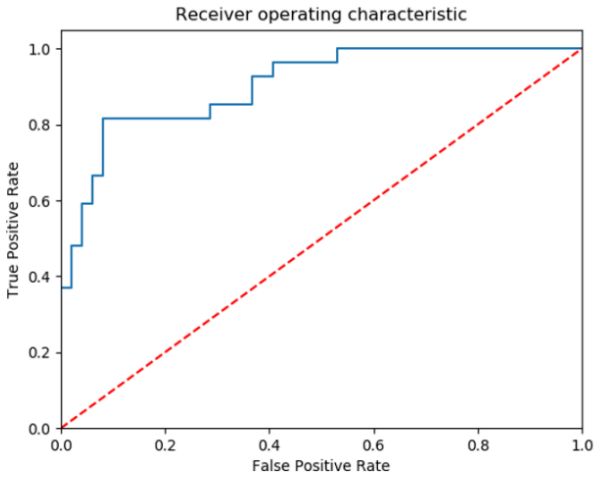

Figure 15: ROC curves for SVM

The ROC curve shows the difference between the two diagnostic groups using well a parameter. Since from the figures we can see that the AUC is highest for SVM followed by the Logistic Regression algorithm so we can say that SVM performs better than the other algorithm.

REFERENCES

1. S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2 IEEE/ACS International Conference on Computer Systems and Applications, Doha, pp. 108-115, 2008.
2. M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, pp. 1-5, 2016.
3. S. Nikan, F. Gwadry-Sridhar and M. Bauer, "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis," 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, pp. 34-39, 2016.
4. A. Naik and N. Naik, "Prognosis of Heart Disease using Data Mining Techniques: A Comprehensive Survey", International Journal of Computer Applications, vol. 181, no. 17, pp. 14-18, 2018.
5. V. Poornima and D. Gladis, "Analysis and Prediction of Heart Disease Aid of Various Data Mining Techniques: A Survey", International Journal of Business Intelligence and Data Mining, vol. 1, no. 1, p. 1, 2018.
6. A. Makwana and J. Patel, "Decision Support System for Heart Disease Prediction using Data Mining Techniques", *International Journal of Computer Applications*, vol. 117, no. 22, pp. 1-5, 2015.

I. AUTHORS PROFILE

Pooja Saharan received the B.Tech degree in Computer Science and Engineering Information from University Institute of Engineering and Technology, Rohtak, Haryana (M.D.U Rohtak) India, in 2009 and received M.Tech in Computer Science and Engineering from University Institute of Engineering and Technology, Rohtak, Haryana (M.D.U Rohtak) India, in 2011 and pursuing Ph.D from Ambedkar Institute of Technology, Govt. of NCT Delhi, Geeta Colony, New Delhi (Guru Govind Singh Indraprastha University, New Delhi), India. Her current research interest includes Data mining, clustering, recommender system, blockchain, .She has more than 8 years teaching experience. She is working as Assistant Professor in ABES Engineering college, Ghaziabad, Uttar Pradesh, India. She is the author/co-author of more than 6 publication in in International/National journals and conferences.

Rashmi Mishra received the B.Tech degree in Information Technology from Institute of Technology and Management, Gida, Gorakhpur (U.P. Technical Univ., Lucknow) India, in 2009 and received M.Tech in Information Security from Ambedkar Institute of Technology, Govt. of NCT Delhi, Geeta Colony, New Delhi (Guru Govind Singh Indraprastha University, New Delhi), India and perusing Ph.D from Delhi Technological University, Delhi, India. Her current research interest includes Wireless Sensor Network, MANET, Secure Semantic Web Services, Cryptography and Network Security, Cyber Security, Blockchain. She has more than 8 years teaching experience. She is working as Assistant Professor in ABES Engineering college, Ghaziabad, Uttar Pradesh, India. She is the author/co-author of more than 16 publication in in International/National journals and conferences.

Charvee Garg received the B.Tech degree in computer science from ABES-EC, Ghaziabad (U.P. Technical Univ., Lucknow) India, in 2019

Aman Payal received the B.Tech degree in computer science from ABES-EC, Ghaziabad (U.P. Technical Univ., Lucknow) India, in 2019