Heart Disease Predictor

Pooja Mudgil[1, #], Swarnim[2], Tushar Goyal[3]

[1,2,3]*Bhagwan Parshuram Institute of Technology, Delhi, India.*
[#]Corresponding Author, Email: engineer.pooja90@gmail.com

***Abstract*— Heart diseases are one of the most fatal diseases known in existence. Heart pumps blood to the entire body, if it fails due to any malfunctioning the body has to bear every consequence. The diseases related to heart are high risk due to the fragile nature of the organ and hence a signal to prevent it can be very useful. The purpose of this research paper is to put forth the idea as to how data that is managed by the hospitals around the world can be used to predict heart diseases in certain individuals with help of some parameters that are taken as input. This paper focuses on various machine learning algorithms that were applied on the given data to gain maximum accuracy in prediction, data analysis of the data is done and a percentage is predicted as to how much extent the predicted results are accurate.**

***Keywords*— Machine Learning, Data Mining, Support Vector Classifier, K-Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier.**

1. INTRODUCTION

Heart diseases are one of the most fatal diseases known in existence. Unhealthy lifestyle with less physical activities and extreme mental stress with improper routines and diet has led the world to suffer with many diseases known, yet heart diseases are the most prevalent and the most catastrophic ones [1]. Specialists conduct numerous surveys and research on heart diseases and collect information of affected persons and their symptoms to track progressions in the results.

The term 'heart disease' includes various number of diseases that affect heart and degrade the proper functioning of heart. People being affected by these heart diseases are increasing drastically. There are more and more cases rising over frequent years. The World Health Organization (WHO) has released a report that states that every year a large amount of people suffer and die due to heart related diseases in which Africa being the worst affected region [2], it stated an estimated 17.9 million people died from cardiovascular diseases in 2016, representing 31% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. In such a fast-paced life no one relies on testing every now on then, most people ignore the slightest of symptoms and prefer to medicate themselves rather than visiting a proper medical practitioner for advice which leads to hazardous consequences for some people. Hence, it is quite important to keep track of your health. With the time technologies have changed and hence with the concept of data analysis we can predict with quite an accuracy that if we are suffering from a disease or so. So, that one can seek medical advice when it is curable.

Data mining and machine learning have been very crucial in the field of medical science. It gives a meaningful base to analytical decisions. Data mining is useful in producing unbiased results as it provides treatments that are completely based on machines analysis without human interference. Data mining also helps to predict results from massive amounts of datasets which is physically not possible, in less amount of time. Medical industries collect large amounts of data which contain some hidden information that can be used to make effective decisions and retrieve precise results. [3]. Heart predictor system will use the datasets and analyse the hidden patterns in data to give a user-oriented approach for better understanding. The knowledge which is implemented can be used by the health care experts to get better quality of service and to reduce the risk of life at the least.

2. LITERATURE REVIEW

Numerous studies on diagnosis of heart disease have been done. Few of them have been listed below:

[4] In their model they predicted probability of heart diseases in an individual using multiple regression model. The work is done using thirteen attributes and over 3000 records. The dataset is divided in ratio 7:3 for training and testing.

[5] have proposed a system that can predict chronic heart diseases by data mining techniques using Artificial Neural Network, Support Vector Machine, Decision Tree and Naïve Bayes. They have compared various parameters to achieve better accuracy.

Another author [6] recommended different algorithms for better performance such as Artificial Neural Network, KNN, Support Vector Machine, Decision tree, Naïve Bayes.

[7] The authors suggested to use big data tools like Hadoop Distributed File System, Map reduce in addition to Support Vector Machine for predicting heart diseases, optimization of attributes was done to obtain better results.

Another author [8] proposed a heart disease prediction system with the help of data mining techniques and used naïve bayes, K-means algorithms for determining the accuracy. The system used historical heart database and used thirteen risk factors. The model predicts heart disease on the 13 listed attributes. Data integration and cleaning was also done.

[9] They provided reviews on types of works already present on heart disease topic and a comparative study to show how each paper had which technique and what accuracy obtained.

[10] They proposed a system which helps practitioners in decision making effectively. The system provides accuracy up to 87% in training and up to 86% in testing phase.

Another author [11] predicted multi-diseases using data mining techniques. The work mainly focused on heart diseases, diabetes and breast cancer and some of the other similar diseases.

[12] Authors proposed a system focusing on predicting disease with a smaller number of attributes and uses data mining techniques to extract information from a record and predict efficiently.

[13] In this system they pre-processed the data and used ID3 algorithm to build decision trees and also used k-means algorithm along with naïve bayes to predict the disease.

[14] Authors suggested a system which used naïve bayes and K-means algorithms to predict the heart disease. This system uses 13 attributes to predict the disease and uses historical dataset on which data mining techniques are performed.

[15] The authors suggested that the classifiers alone are not good enough to predict any discrepancy so ensemble technique should be used which combines the weak learners to implement hybrid model that provides better result than individual classifier themselves.

[16] The paper focuses on machine learning tree classifiers, such as, Random Forest, Decision tree, Logistic regression, Support vector, K-nearest neighbours. They were breaking down on their precision and AUC ROC scores. The random woodland machine learning algorithm accomplished higher than others mentioned.

[17] The authors have used supervised machine learning models to predict the diseases. They have used Cleveland database that has 76 attributes in total but only 14 are taken into consideration for predicting. The KNN algorithms has given the maximum accuracy among KNN, Decision tree, Random Forest, Naïve Bayes.

3. PROPOSED WORK

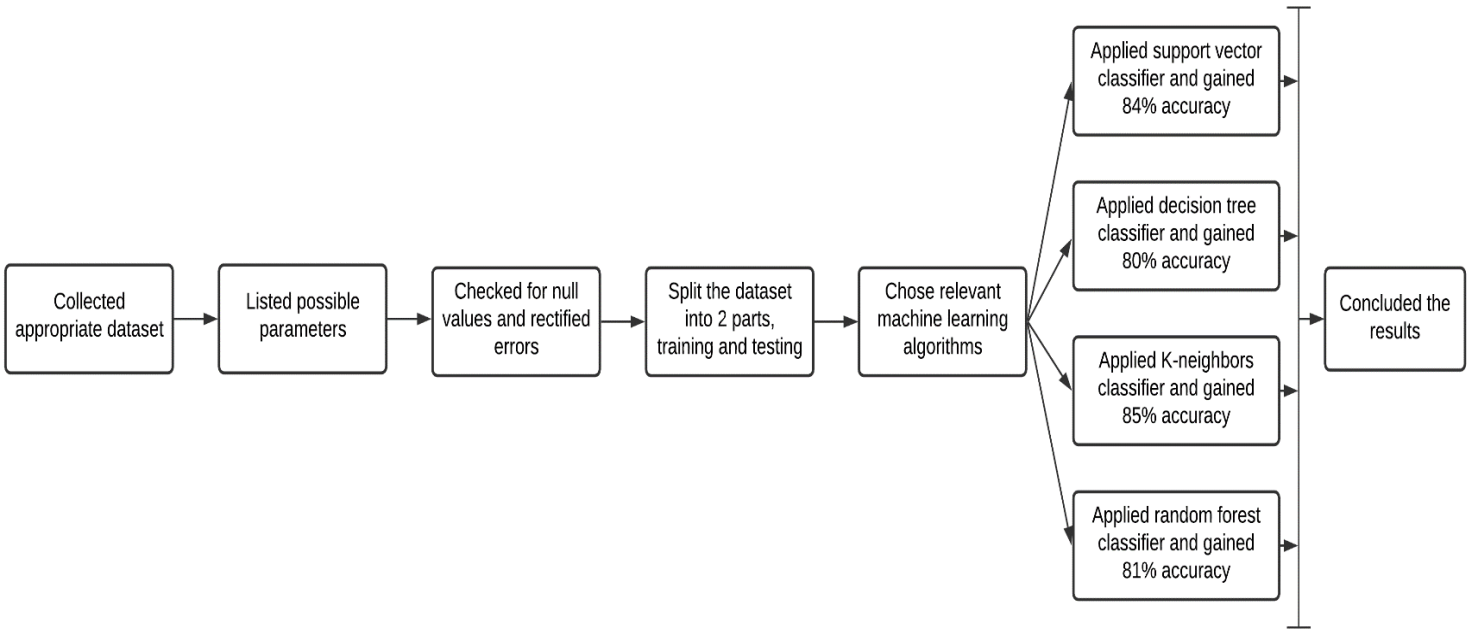

Fig. 1. Depiction of proposed work in a flowchart

In this system, the dataset was collected from Kaggle which contains 303 entries and has 13 attributes and one target value. The dataset was checked for null values and as a result no null values were found so the experiment was proceeded with division of dataset into 2 parts with 65% training dataset and 35% testing dataset.

Four machine learning algorithms namely K-Neighbours, Support Vector, Random Forest and Decision Tree Classifiers were used and accuracy for each algorithm was calculated on changing parameters according to the algorithm's demand. Each algorithm performed well with accuracy of 85%, 84%, 81%, 80%. Among these K-Neighbours gave 85% accuracy with 8 neighbours which was the maximum among all.

4. EXPERIMENTATION AND RESULT

In this project, heart disease predicting system is implemented with use of various machine learning algorithms such as Support Vector, Random Forest, K-Neighbours and Decision Tree Classifiers. We have used 13 attributes and 1 target value for implementation.

Attributes are listed below:

Attributes	Names and meaning
age	Age (in years)
sex	Gender (1- male and 0 - female)
cp	Chest Pain (Level of pain (0,1,2,3))
trestbps	Resting Blood Pressure
chol	Cholesterol in serum (in mg/dl)
fbs	Fasting blood Sugar (>120 mg/dl)
restecg	Resting Electrocardiographic Results (0,1,2)
thalach	Maximum heart rate achieved
exang	Exercise induced angina (0,1)
oldpeak	Old peak
slope	Slope of peak exercise
ca	Number of major vessels fluoroscopy (0-3)
thal	Defect-7, Normal-3, Fixed Defect -6
target	Target value (0 or 1)

The dataset that is being used by default contains 303 entries and can be replaced with any other data set. It was ensured that there are no null values in the dataset.

To choose dataset that was optimal is key to obtain accuracy. Suppose, if we chose a dataset with 100 records in which only 1 was suffering from the disease, the accuracy would obviously be 99% as without any training or learning, the model will produce accurate results. So, to choose a dataset that has equal number of patients and non-patients is important for proper research purposes. Hence, bar plot for almost balanced target classes was obtained from the dataset with 0 for no disease and 1 for disease. Yellow colour is indicated for individual with no disease and orange with an individual suffering from disease. The data is processed, some of the variables were converted into dummy variables and scaling was done before training the machine learning models. The dataset was split into 65% training dataset and 35% testing dataset.

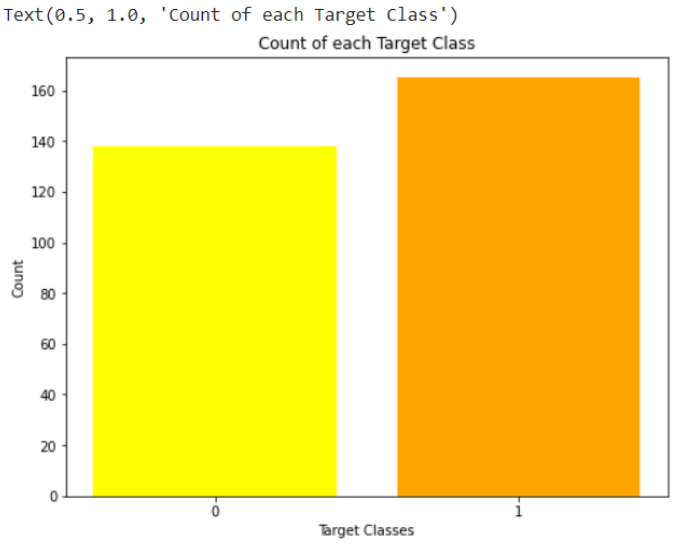

Fig. 2. Count of each target class

The various machine learning algorithms were implemented as explained:

4.1 Support Vector Classifier

Support vector classifier is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used for classification problems. The algorithm uses a set of mathematical functions that are defined as kernel. The function of kernel is to take data as input and transform it into the required form. Different support vector algorithms use different types of kernel functions. These functions can be different types. For example, linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. Here in this model, the graph consists of polynomial, linear, radial basis function and sigmoid. Among all the kernels linear provided best performance with 84% accuracy.

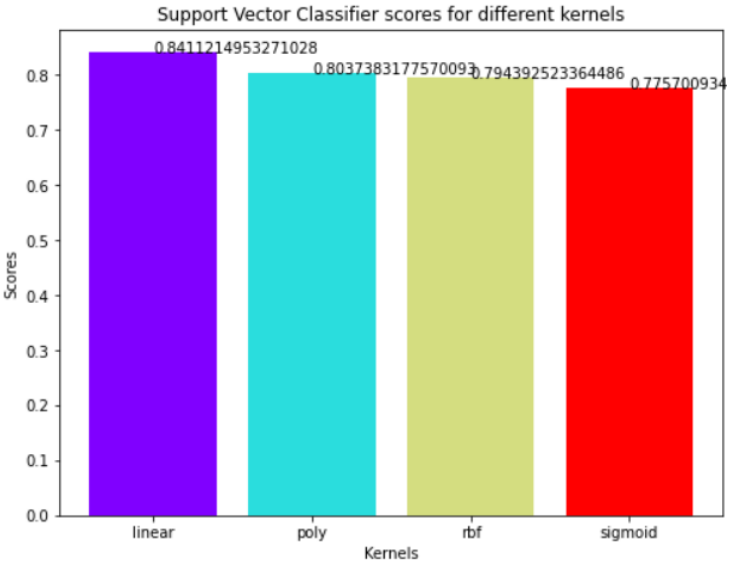

Fig. 3. Support Vector Classifier graph representation showing different kernels

4.2 Decision Tree Classifier

Decision tree classifier is a supervised machine learning algorithm that can be used for both classification and regression problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches. Here in this model graph is plotted where the maximum number of features were split in a range from 1 to 30 and the scores were determined. The maximum score obtained for the classifier for different number of maximum features. The model obtained best accuracy of maximum features at 18 of 80.37%.

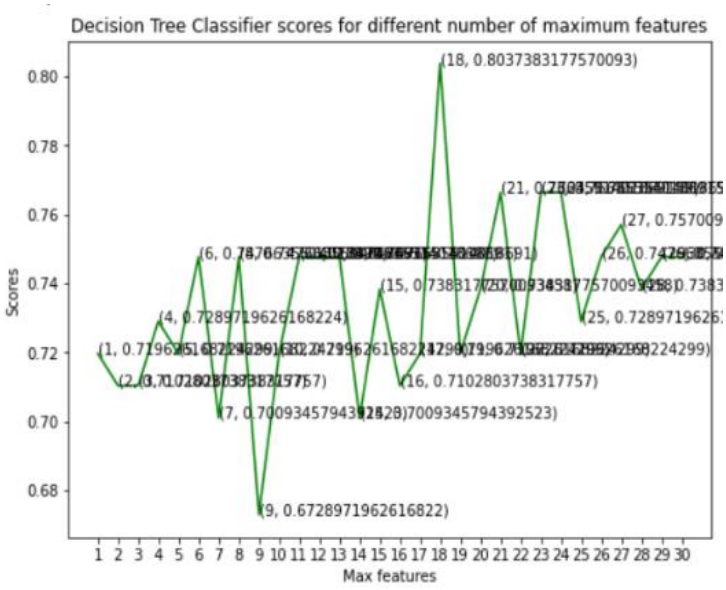

Fig. 4. Graph representing decision tree classifier for different maximum features

4.3 Random Forest Classifier

Random forest is a supervised machine learning algorithm used for classification, regression, and other tasks using decision trees. The random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a

randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction. Here in this model, the number of estimators is varied to obtain different accuracy with accuracy of 81% with 100,200 estimators.

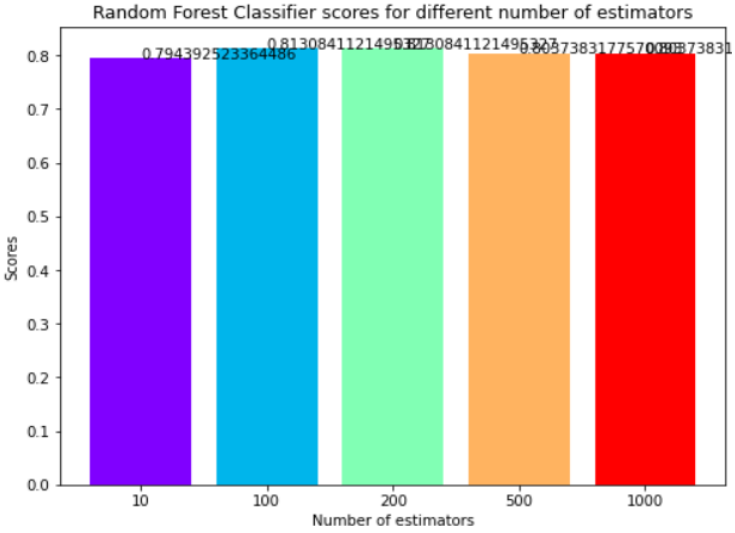

Fig. 1. Graph representing random forest classifier for different number of estimators.

4.4 K-Neighbors Classifier

K-Neighbour is one of the simplest machine learning algorithms based on supervised learning technique. It assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. Here in this model, the graph represents different values of neighbors varied with classification scores. The best accuracy was achieved with 8 neighbors with accuracy of 85.

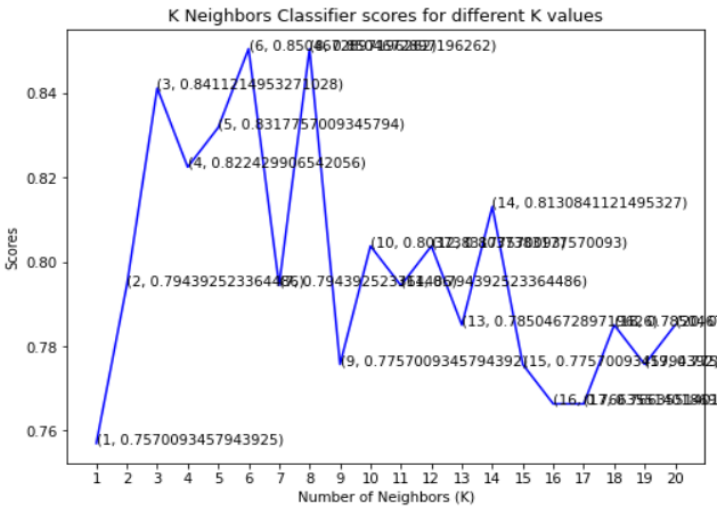

Fig. 2. Graph representing K-Neighbors classifier for different K values.

Hence, we achieved maximum accuracy of 85% with K-Neighbours classifier.

5. CONCLUSION AND FUTURE WORK

The dataset is modelled into 65% training and 35% testing. Then scaling was done on some of the features while generation of dummy variables for some. Four machine learning algorithms were applied and parameters were varied according to the models and at last, K-Neighbours Classifier achieved highest score of 85% with 8 nearest neighbours.

Future work for the project will be to make a web-based system for users to get their information in hand after giving in information of the required 13 attributes. This system will also take any other datasets to produce accuracy also. Apart from that, one can also try to implement systems that can predict various types of heart diseases specifically and provide some instant solution or some particular doctor or medical facility could be recommended to the user.

Some of the work has been implemented, a web-based system is built that gives user the access to the application “Heartline”, this web application has login and register portal followed by a dashboard where users can request the admin to display the blogs that have adequate information about the cardiovascular diseases and recent blogs, news or any information can be put up on dashboard via admin. There is another page that predicts the disease, it takes 13 attributes as input and predicts the severity of the any heart related disease via K-Neighbour’s prediction model, this model was chosen as it acquired maximum accuracy in the research done. The application also stores any information that user has entered before for further access. This system can further be improvised with recommendations and even at the end detailed report could be provided.

Few screens are shown below that are part of web application “Heartline” based on this research.

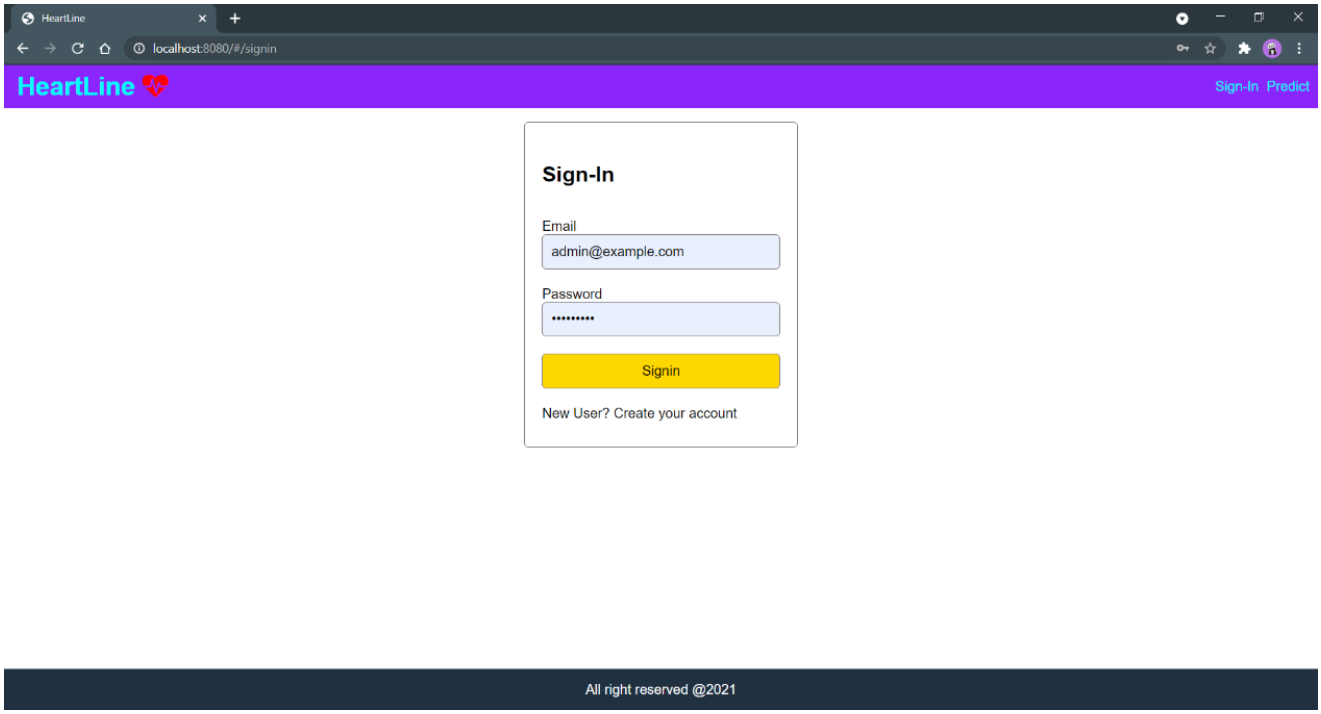

Fig. 7. Login screen where user can enter their email and password to login.

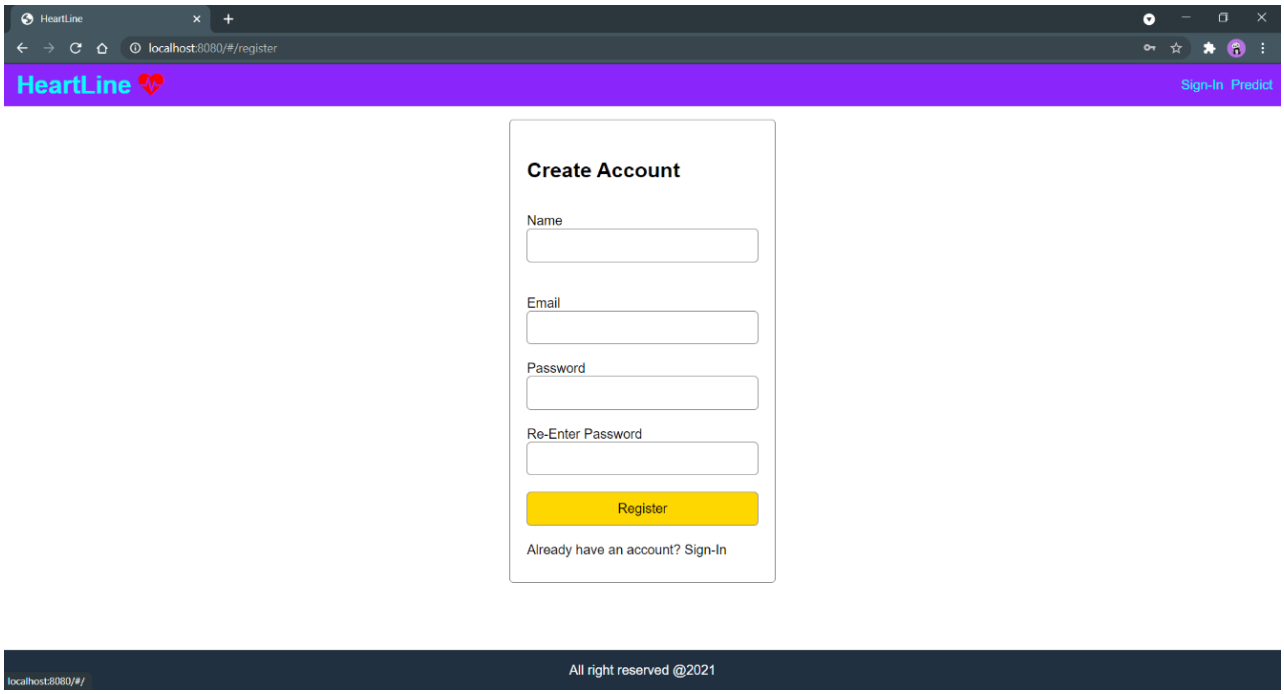

Fig.8. Create Account screen for user to create account on Heartline

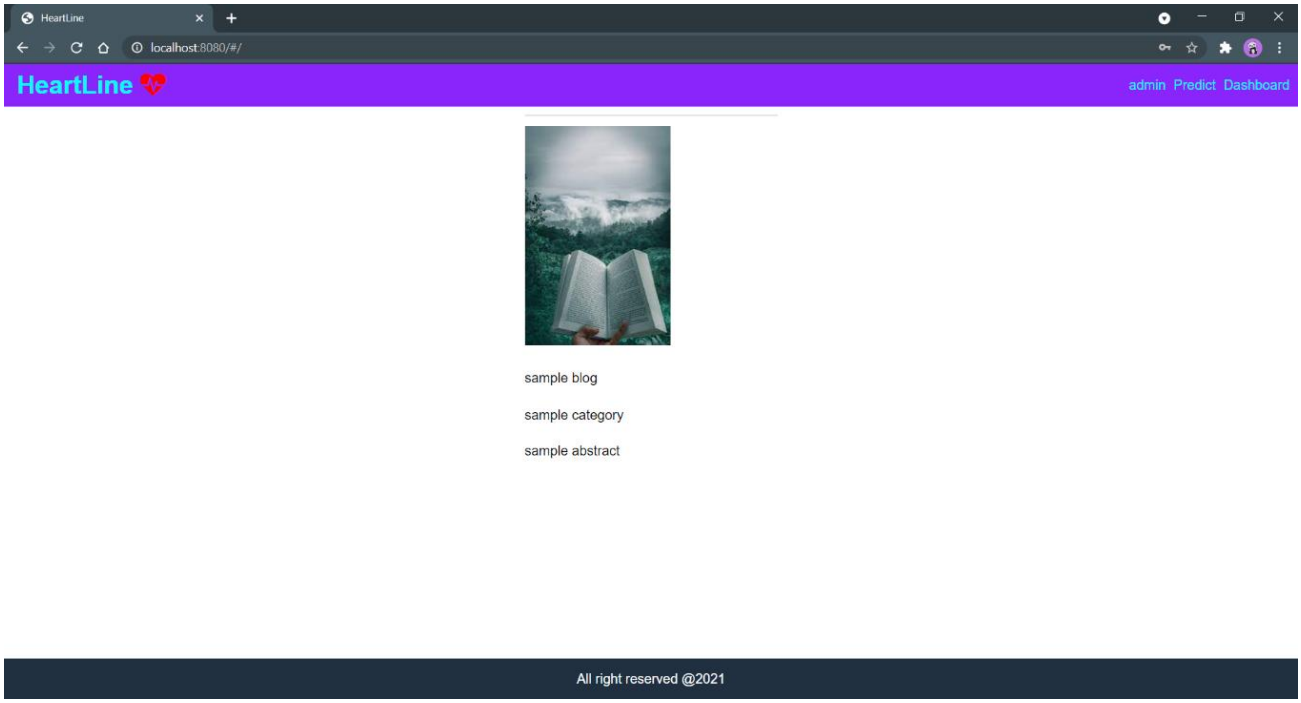

Fig.9. Dashboard that displays blogs, multiple users can post blogs here

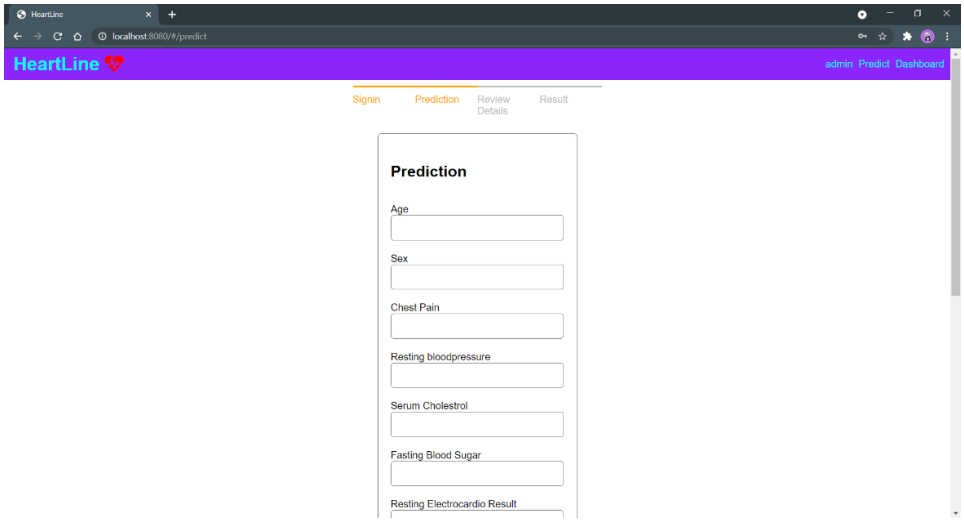

Fig.10. Prediction page where user can enter their details

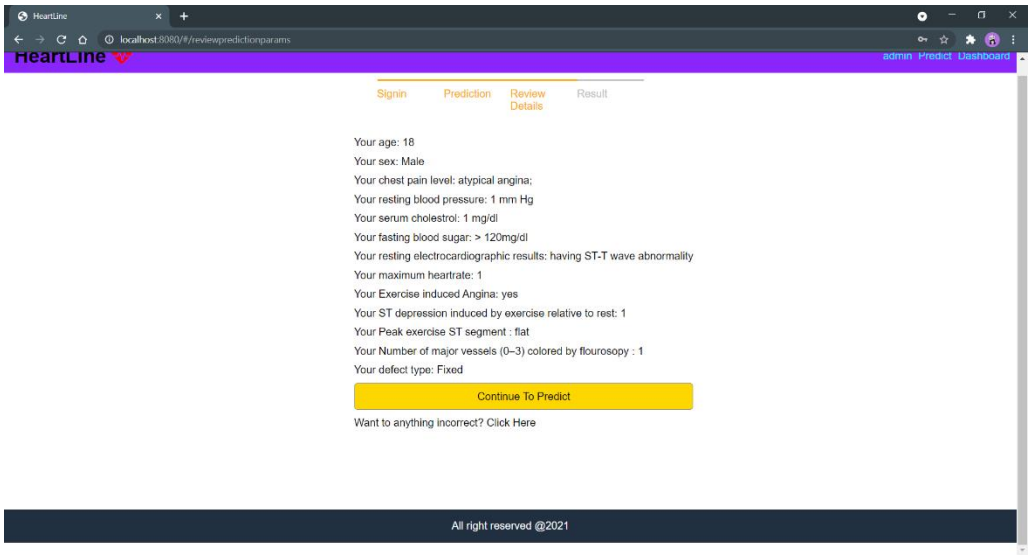

Fig.11. Prediction page where user can review their details

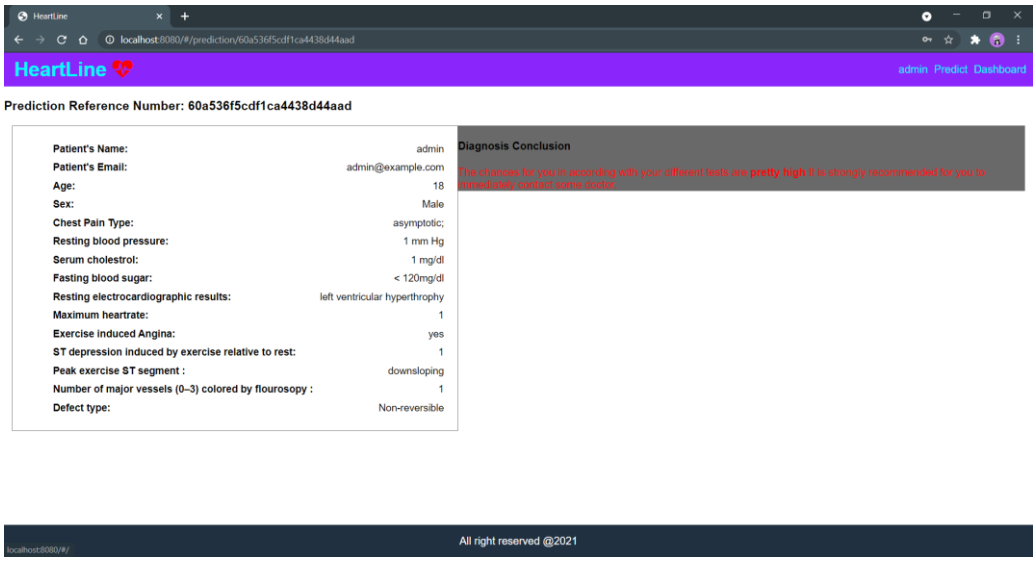

Fig.12. The predicted result page where patient's diagnosis is concluded

6. REFERENCES

[1] Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137-2159.

[2] https://www.researchgate.net/publication/331589020_Heart_Disease_Prediction_System

[3] http://www.researchgate.net/publication/327722009_A_Review_on_Heart_Disease_Prediction_using_Machine_Learning_and_Data_Analytics_Approach

[4] K. Polaraju, D. Durga Prasad: Prediction of Heart Disease using Multiple Linear Regression Model; International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.

[5] https://www.researchgate.net/publication/316530782_Predictive_analytics_to_prevent_and_control_chronic_diseases

[6] Beyene, C., & Kamat, P.: Survey on prediction and analysis the occurrence of heart disease using data mining techniques; International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173.

[7] Soni, J., Ansari, U., & Sharma, D; Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

[8] Ashwini Shetty A, Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281.

[9] Muthuvel, Marimuthu & Abinaya, M & Madhankumar, K & Pavithra, V. (2018). A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. International Journal of Computer Applications. 181. 975-8887. 10.5120/ijca2018917863.

[10] Purushottam, Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", 2016, pp.962-969.

[11] Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528.

[12] Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.

[13] Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna, "Prediction of Heart Disease using Machine Learning Algorithms", International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366.

[14] Sairabi H.Mujawar, P.R.Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Bayes", International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.

[15] Meenu Bhatia, Dilip Motwani, "Use of Ensemblers Learning for Prediction of Heart Disease", International Conference on Trends in Electronics and Informatics, **ISBN:**978-1-7281-5519-7, June 2020.

[16] N. Komal Kumar, G Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana , "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", International Conference on Advanced Computing and Communication Systems(ICACCS), ISSN-2575-7288, ISBN- 978-1-7281-5198-4, March 2020.

[17] https://www.researchgate.net/publication/346484346_Heart_Disease_Prediction_using_Machine_Learning_Techniques