Publication

Personalized Treatment Based On Patient History

Rohit Sharma
Computer Department
Thakur college of Engineering and Technology
Mumbai, India
rohit.tps123@gmail.com

Nitesh Tiwari
Computer Department
Thakur college of Engineering and Technology
Mumbai, India
niteshtiwari1971@gmail.com

Ganesh Wartale
Computer Department
Thakur college of Engineering and Technology
Mumbai, India
wartaleganesh@gmail.com

Dr. Megharani Patil
Computer Department
Thakur college of Engineering and Technology
Mumbai, India
megharani.patil@thakureducation.org

Abstract—**Health care practices involve collecting all kinds of patient data which would help the doctor correctly diagnose the condition the subject is likely suffering from. This data could be everything from the simple symptoms observed by the subject, initial diagnosis by a physician or a detailed test result from a lab. Thus far this data is only utilized for subjective analysis by a doctor who then ascertains the disease in play using his/her personal medical expertise. We posit that there is definite potential for application of data mining routines on this rich reserve of patient data. Employing apposite data mining techniques, useful patterns and conclusions could be drawn from the raw data at our disposal. These findings could in turn be utilized in a number ofproductive ways like to carry out automated diagnosis, equip doctors with a better understanding of the causes and factors in play behind a particular disease. We studied different data mining techniques like Logistic Regression, KNN, SVM, Naïve Bayes, Decision trees and Random Forest and found out that KNN was giving the highest accuracy of 78.57%. Automated diagnosis in particular could prove very useful for the determination of non-critical diseases or in cases where a doctor may not be available to carry out diagnosis such as in case of being in a remote location. We also aim to incorporate it in an application which will provide diagnosis, storing of patient information and provide visualization of that data.**

I. INTRODUCTION

With the advance in technology and methodology people have begun automating tasks in order to save time and energy. Most of these tasks were repetitive or simple, but nowadays we have begun automating processes which seemed impossible to do so before. Techniques in data mining has led to the ease of extracting not only information, but patterns and conclusions from data which previously had to be done manually. By using a technique of our own a web application will automate the process of medical diagnosis for users who are in need of it where it isn't possible to consult a doctor on an immediate or daily basis. Diabetes is one of the leading causes of death and most people even don't know that they have diabetes at an early stage. Type 1 diabetes can occur in childhood or adolescence age. Type 2 diabetes usually affects the adults who are obese this type, the body does not produce insulin or resistance to insulin was observed. Type 2 generally occurs in middle age or group. In addition, there are various other reasons which cause diabetes, such as diabetes, bacterial and viral infections, or toxic or chemical food content, obesity, poor diet, lifestyle changes, eating habits, environmental pollution etc. Diabetes leads to various diseases such as cardiovascular complications, renal issues, retinopathy, foot ulcers,etc.

A. Problem Definition

Going doctors to at regular intervals for a health checkup is an advise that not many people follow. There can be many reasons for not visiting a doctor at regular intervals

which might include not having time, money or any other reason. We posit that an application can be developed which has a potential to solve this problem of most of the people. An application which will be able to predict early signs of diabetes in patients based on various health parameters and also provide health advise to them.

II. LITERATURE SURVEY

Following are some of the search which has been reviewed for the proposed system: Deeraj Shetty ,Kishor Rit , Sohail Shaikh and Nikita Patil ." Diabetes Disease Prediction Using Data Mining ". *International Conference on Innovations in Information, IEEE(2017).[1]* This paper discusses a system which uses machine learning model to cluster the dataset of diabetes and find out the accuracy of the model. It mainly focuses on machine learning algorithm deployment of classification. The proposed system uses machine learning model as prediction and web technologies as user interface. In the model selection they got KNN as best model which works on diabetes datasets. Primarily they are maintaining the database for the patient access. Their examination concentrates on this part of Medical conclusion learning design through the gathered data of diabetes and to create smart therapeutic choice emotionally supportive network to help the physicians.

Samrat Kumar Dey, Ashraf Hossain and Md. Mahbubur Rahman. "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm". *21st International Conference of Computer and Information Technology (ICCIT) 2018. [2]*

This study concentrates on various diseases and their prediction with the help of intelligent system . In this paper they have collaborated various disease prediction with web technologies. They have made a web portal for patients to diagnose their disease. In this they proposed work was done on data collection on Diabetes from various reports then preprocessing is done using machine learning techniques. Their Our main aim of

this exploration is to build a web application based on the higher prediction accuracy of some powerful machine learning algorithm. We have used a benchmark dataset namely Pima Indian which is capable of predicting the onset of diabetes based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy which drives us to develop an nteractive Web Application for Diabetes Prediction.

Pahulpreet Singh Kohli and Shriya Arora. "Application of Machine Learning in Disease Prediction" . *4th International Conference on Computing Communication and Automation (ICCCA) 2018. [3]*

This paper focuses on prediction of various deadly diseases such as Heart, Breast and Diabetes. They have collected datasets from Kaggle and used Data Mining for removing missing values from it. They have proposed all types of supervised machine algorithms for classification. As they have predicted various deadly disease therefore the accuracy was checked in all algorithms. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases.

V.Swathi Lakshmi, V.Nithya, K.Sripriya , C. Preethi and K. Logeshwari. "Prediction of Diabetes Patient Stage Using Ontology Based Machine Learning System". *Proceeding of International Conference on Systems Computation Automation and Networking 2019. [4]*

This paper proposed the work methodology of automated system working on the intelligence system for predicting diabetes disease. l related problems. In this paper a proposed method to identify patient with diabetes disease risk level is indentified. In this work diabetes patient risk level is been detected by using ontology and machine learning technique. Ontology holds disease symptoms, causes and treatments. In machine learning, naïve base algorithm is used to make decision on patient record also it defines possibilities of risk level. The proposed algorithm will be evaluated against the following metrics namely confusion matrix, precision level, mean and this proposed work is found to have better prediction level when compared with existing work.

Santosh Rani and Dr. Sandeep Kautish. "Association clustering and Time Series based data mining in Continuous data for Diabetes Prediction" . *Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018) IEEE Xplore*. [5]

The project proposed that large amount of health related data is being produced in various levels of health system. Due to the size of the data it will be difficult to process the data and then extract the analysis. But with the use of

different machine learning based approaches data can be processed. Machine learning based approach will provide the efficient data for curing of the patients. Even helps in forecast the future disease. Patients past history for different parameters can contribute to his probability of various health related issues. But using Association clustering and Time Series based data mining in Continuous data early warning system can be developed. This prediction based system can define the disease while analyzing his existing parameters.

III. EASE OF USE

This application can be used by anybody who is having basic sense of understanding of web based applications. The users can access and track their Health condition using the application from anywhere provided with an internet connection and a computing device like Laptop or smart phone.

IV. PROPOSED ARCHITECTURE

In this project we have divided the entire project into four major modules. The Modules are namely Registration, Diagnosis, Diagnosis History and Visualization modules

Modules

1. Registration Module

The Registration module is used for registration of new users. The user needs to login after the registration to use the system.

2.Diagnosis Module

This module is the brains of the system which is responsible for identifying the diseases based on the symptoms. It takes input from the user with the help of a simple form and uses the knowledgebase containing information about the diseases and their symptoms and finally provides the result using the K-nearest neighbor algorithm.

3. Exercise Module

In this module the exercise plan is generated whenever the user fills the analysis form. The exercise plan generated is stored in the database with date and time so that the user can refer to that in the future as well.

4. Diet Module

In this module the diet plan is generated whenever the user fills the analysis form. The diet plan generated is stored in the database with date and time so that the user can refer to that in the future as well.

5. Visualization module

The module provides visualizations based on the diagnosis data collected. These visualizations will be in the form of bar graphs. We provide the visualization of both the input parameters as well as the visualization of the results generated.

6. Appointment Module

This module deals with the appointments. The patients will be able to book appointments with the doctors and the doctors will be able to view the booked appointments.

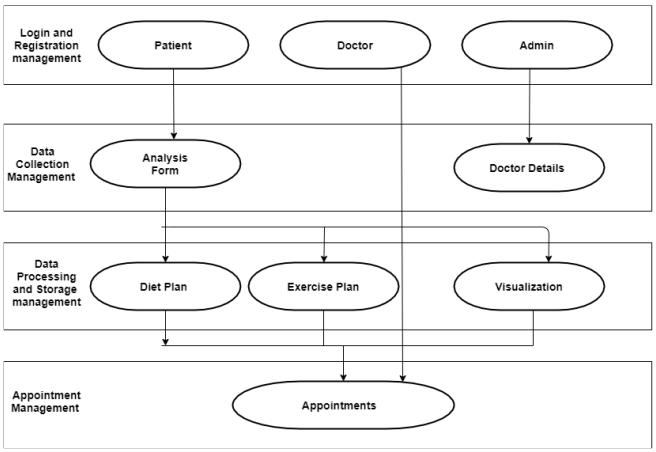

Fig. 1. Flowchart of the proposed model.

V. METHODOLOGIES

A. K NEAREST NEIGHBOUR

This algorithm is a supervised learning algorithm, K- means a number of a vector. The working methodology of KNN is pretty simple, it's predict based on the value of K parameter. Graphical representation of K Number Nearest Neighbours are depicted in following Fig. 3

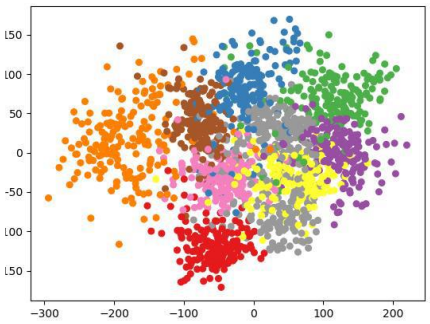

Fig. 2. Graphical representation of KNN working Procedure

B. Support Vector Machine (SVM)

Support vector machine is a supervised machine learning algorithm mainly used for classification problem. Overfitting derives an ML model to misclassify the data from a given dataset. In that case SVM can prevent overfitting nature from samples data and produce better accuracy [8][9]. SVM possess a linear hyperplane with a margin which divides the dataset into positive and negative samples [8][9].

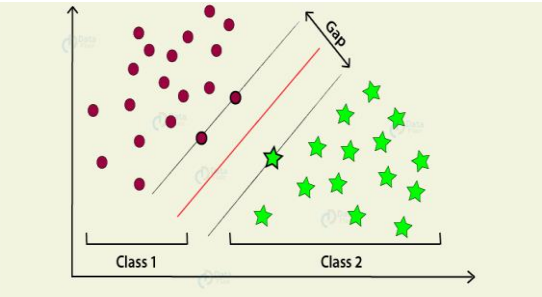

Fig. 3. Graphical working procedure of SVM

C. Artificial Neural Network (ANN)

ANN is considered to represent a numerical method by mirroring human neurons their learning and speculation techniques. Neural network model can be extended a highly nonlinear systems, the relationship between the variables which are unknown or very complex[11]. Neural networks consists of a variety of neurons and then followed by some layers. Structure represents human neuronal axons and dendrites as nodes, which represent links between the nodes and the weighted connections axons.The total structure of the neural network is shaped by input layer, one or more hidden layers, and the output layer where ith neuron represents a connection with jth neuron of total structure and Wij represents as the strength of the link between neurons.

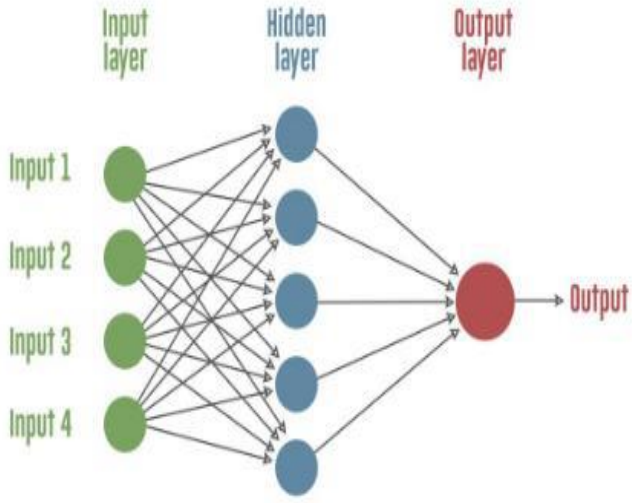

Fig. 4. ANN with double hidden layer

The nodes of the structure of an ANN take inputs (features) then send them to next hidden layer through some sort of weighted link where ith nodes send data to jth nodes for processing and calculating the weighting sum and totaling a bias term .

D. Logistic Regression

Logistic regression is an algorithm mainly used for classifying observations into different sets of classes. Linear regression computes exact output for a given input in form of numerical values where predicted value was a continuous quantitative variable. What if we want to predict a discrete value Eg Male or Female based on height.

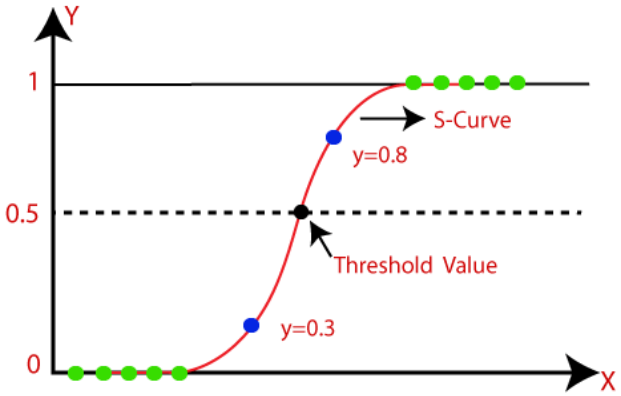

Fig. 5. Logistic Regression

If the prediction has to be an output label and not to be a value then this is termed as classification problem and the difference over here is 1 for probability of happening of an event say p and 0 for not happening of an event. The sum of both the probabilities should be equal to 1 . Major uses of Logistic Regression are Detection of spam mails, Credit card fraud detection, Diseases Prediction , Weather Forecast, etc. The Formula below is the mathematical way to represent Logistic Regression.

1 / (1 + e^-value)

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

E. Decision Tree

Decision tree learning is one of the prescient displaying approaches utilized in insights and information mining. It makes inductive inferences based on if then rules .Decision trees is a recursive approach that follows divide and conquer method. Decision trees have good interpretability as

compared to other algorithms. The leaf node in the tree is either an output label for classification or a numerical value for regression. Decision trees are more robust to error thus may work on data with missing value. Tree models where the target variable can take a discrete course of action of characteristics are called gathering trees; in these tree structures, leaves address class names and branches address conjunctions of features that lead to those class marks. Decision trees have a natural "if … then … else …" construction that makes it fit easily into a programmatic structure. They likewise are appropriate to arrangement issues where qualities or highlights are efficiently checked to decide a last class. For example, a decision tree could be used effectively to determine the species of an animal.

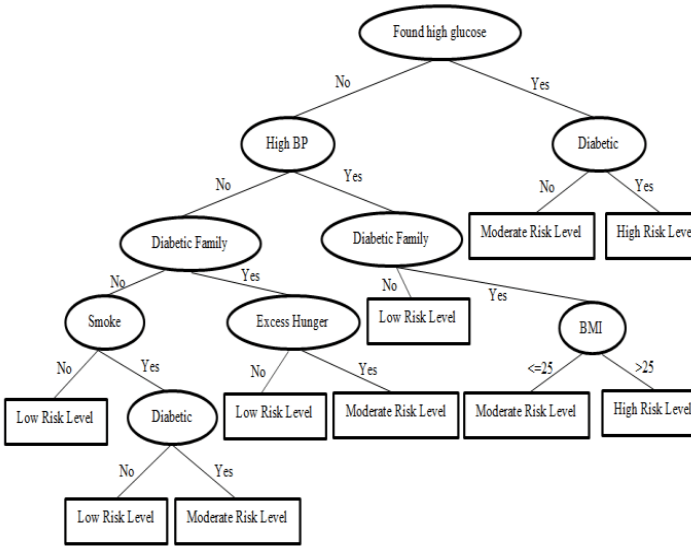

Fig. 6. Decision Tree

VI. RESULTS

In this section we will discuss regarding our results which we have achieved after experimental design. Following TABLE I. represents an insight description of our Pima Indian Dataset. This dataset is mainly based on the females those were living at Pima Indian heritage. Following 8 features (a-h) of Pima Indian dataset helps us to predict the diabetes of any Individuals with the help of our proposed methodologies.

a) Numbers of time Pregnant
b) Glucose Test
c) Blood Pressure
d) Triceps skinfold thickness
e) 2-Hour Serum Insulin
f) Body Mass Index
g) Diabetes Pedigree function
h) Age

Class	Attribute Number
Pregnancy Count	1
Glucose concentration in plasma	2
Blood pressure (diastolic, mm Hg)	3
Thickness of triceps skin fold (mm)	4
2-Hour serum insulin (μ U/ml)	5
Body mass index	6
Pedigree function of diabetes	7
Years of age	8

Table I. List of Attributes in Dataset

Following TABLE II depicts the accuracy rate of various algorithms used during the experimental design.

Sr.No	Algorithm	Accuracy (in%)
1	Logistic Regression	71.42
2	KNN	78.57
3	SVM	73.37
4	Naïve Bayes	71.42
5	Decision Trees	68.18
6	Random Forest	75.97

Table II: Accuracy of Various Algorithms

We have experimented with different machine learning algorithms like Logistic Regression, KNN, SVM, Decision trees, Naïve Bayes .

Implementation of Algorithm

```
In [31]: # Evaluating using accuracy_score metric
         from sklearn.metrics import accuracy_score
         accuracy_logreg = accuracy_score(Y_test, Y_pred_logreg)
         accuracy_knn = accuracy_score(Y_test, Y_pred_knn)
         accuracy_svc = accuracy_score(Y_test, Y_pred_svc)
         accuracy_nb = accuracy_score(Y_test, Y_pred_nb)
         accuracy_dectree = accuracy_score(Y_test, Y_pred_dectree)
         accuracy_ranfor = accuracy_score(Y_test, Y_pred_ranfor)

In [32]: # Accuracy on test set
         print("Logistic Regression: " + str(accuracy_logreg * 100))
         print("K Nearest neighbors: " + str(accuracy_knn * 100))
         print("Support Vector Classifier: " + str(accuracy_svc * 100))
         print("Naive Bayes: " + str(accuracy_nb * 100))
         print("Decision tree: " + str(accuracy_dectree * 100))
         print("Random Forest: " + str(accuracy_ranfor * 100))

         Logistic Regression: 71.42857142857143
         K Nearest neighbors: 78.57142857142857
         Support Vector Classifier: 73.37662337662337
         Naive Bayes: 71.42857142857143
         Decision tree: 68.18181818181817
         Random Forest: 75.97402597402598
```

Fig. 7. Implementation of different algorithms on the Pima Indian Diabetes dataset .

Here we have preprocessed diabetes dataset and applied different machine leaning algorithms .Then we found that KNN gives highest accuracy with 78.56% .

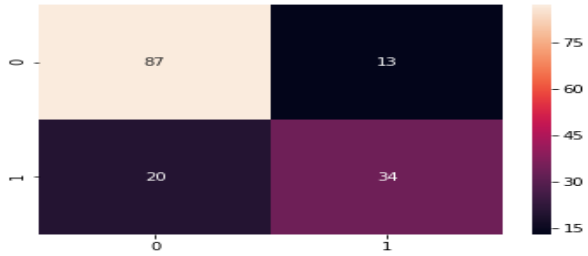

Fig. 8. Confusion Matrix for accuracy of KNN

The actual implementation is done and the screenshots of the main components of the system are attached herein.

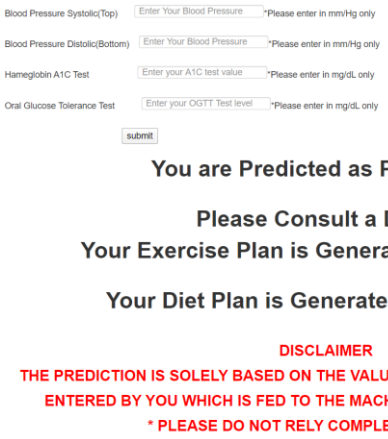

Fig. 9. The system predicts the results based on the input provided and the machine learning algorithms are applied on the input and results are generated.

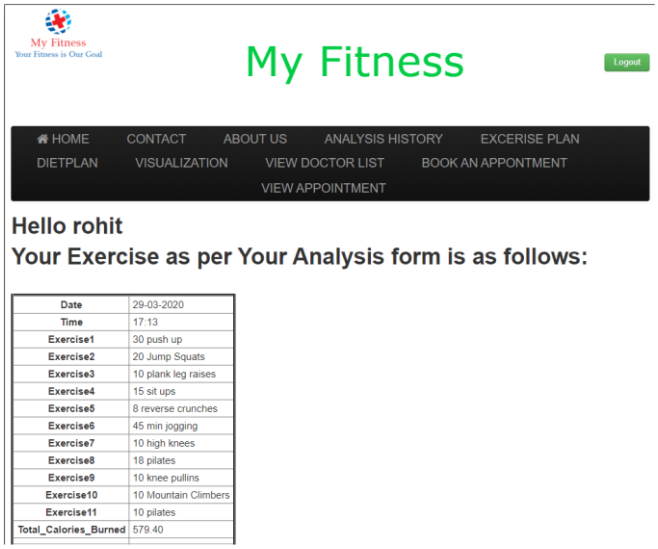

Fig. 10. Exercise plan generated every time the user fills the analysis form. The users need to fill the analysis form to get the exercise plan and diet plan generated everytime.

Fig. 11. Diet plan is generated everyt ime the user fills the analysis form.

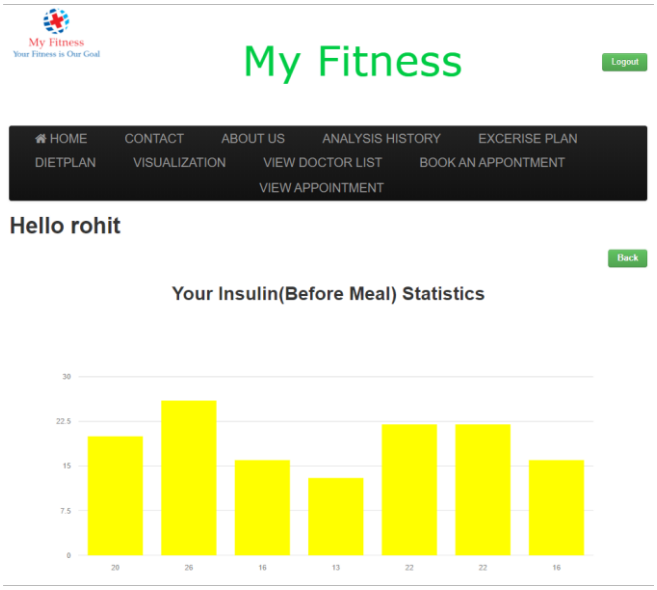

Fig. 11. The system also provides visualization of the input and the output parameters.

Fig. 12 The user can also check the history of the Analysis form.

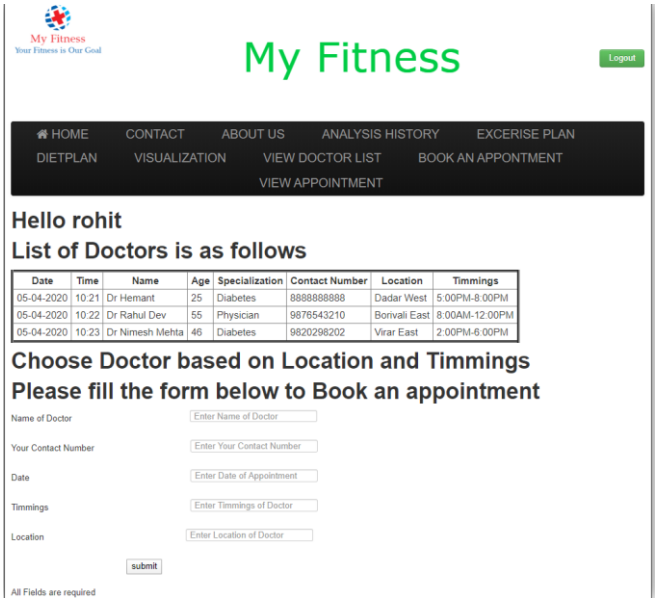

Fig. 13. The user can also book the appointment with the doctor.

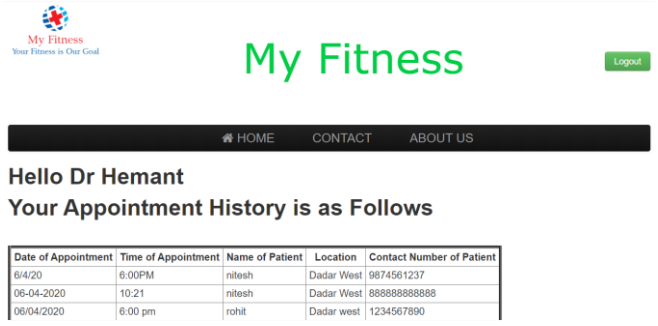

Fig. 14. The Doctors can login and view the appointments booked by the patients.

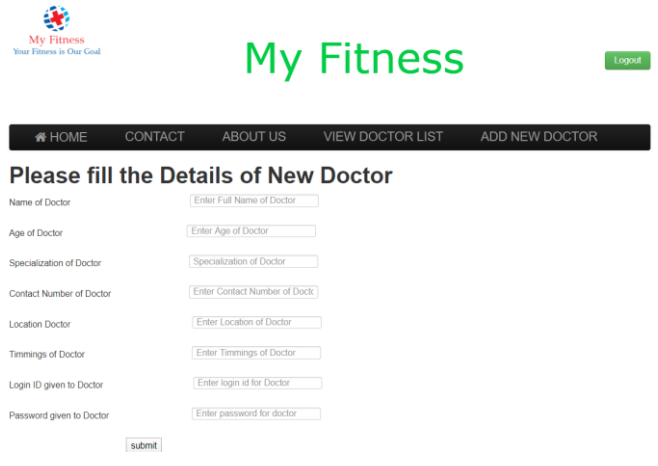

Fig. 15 The Admin can add the details of new Doctor in the system.

VII. CONCLUSIONS

In this project the team was able to develop a web application which will be able to predict diabetes disease or early signs of the disease with the help of machine learning algorithms. The users will be able to get personalized recommendation in the form of diet plan and exercise plan which the users can follow to maintain their health. The system also provides visualization of the various input parameters and the results generated in the form of bar graphs. The appointment feature is also included in the system where the users will be able to book appointment with the doctor. The Doctors will also be able to see the details of the booked appointments. Although many features are included in our system but many more features can be included and a more advanced system can be built in future which will be able to cover a greater number of diseases and provide more personalized recommendation to its users. These types of system play a crucial role to provide people with a stable or reliable recommendation of their general health condition. By providing a well-equipped system to a promising cause there will a significant amount of impact on the society's health and direct it towards a new direction of health and fitness.

VIII. ACKNOWLEDGEMENT

With deep sense of gratitude we would like to thank all the people who have lit our path with their king guidance. We are grateful to these intellectuals who did their best to help during our project. It is our proud privilege to express deep sense of gratitude to our Project Guide Dr. Megharani Patil for her continuous guidance and support throughout the project. It would never have been possible for us to complete this project successfully without her guidance and support.We remain indebted to our HOD Dr. Sheetal Rathi COMP Department, Dean Academic Dr. R.R Sedamkar and Principal Dr. B.K.Mishra for their comments and kind permission to complete this project. We would also like to thank them for their timely suggestions and valuable guidance. And lastly we would also like to thank our friends and the people who are directly or indirectly related to our project.

REFERENCES

[1] Samrat Kumar Dey , Ashraf Hossain "*Implementation of a Web Application to Predict Diabetes Disease: An Approach UsingMachine Learning Algorithm*" , 2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23

[2] Deeraj Shetty, KishorRit, Sohail Shaikh, Nikita Patil,"*Diabetes Disease Diseases Prediction Using Data*

Mining", 2017 InternationalConference on Innovations inInformation, Embedded andCommunication Systems (ICIIECS)

[3] Pahulpreet Singh Kohli, Shriya Arora, "*Application ofMachine Learning in Disease Prediction*", 2018 4th International Conference on Computing Communication andAutomation (ICCCA)

[4]Y. Cai, D. Ji,D. Cai, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor", Proceedings of NTCIR-8 Workshop Meeting, 2010.

[5] I. Rish, "An empirical study of the naive Bayes classifier", T.J. Watson Research Center, 2001.

[6]M.Elkourdi, A.Bensaid, T.Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", Alakhawayn University, 2001

[7]L.Wang, L.Khan and B.Thuraisingham,"An Effective Evidence Theory based on nearest Neighbor (KNN) classification", IEEE International Conference, 2008.

[8]M. Elkourdi, A. Bensaid, and T. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," Alakhawayn University, 2001.

[9]M. Seera, and C.P. Lim, "A hybrid intelligent system for medical data classification," Expert Systems with Applications, Vol. 41 pp. 2239– 2249

[10]W. Bergerud, "Introduction to logistic regression models with worked forestry examples: biometrics information handbook no. 7," no. 7, p. 147, 1996.

[11]J. R. Quinlan, "Induction of Decision Trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[12]V. Jakkula, "Tutorial on Support Vector Machine (SVM)," Sch. EECS, Washingt. State Univ., pp. 1–13, 2006.

[13]https://ieeexplore.ieee.org/document/7154738

[14] https://www.kaggle.com/uciml/pima-indians-diabetes-database

[15]https://developer.mozilla.org/enUS/docs/Learn/HTML/Howto/Use_Java