An Application on Diabetic Prediction System Using Machine Learning Algorithm

Siji Joju[1] Dr. G. Kiruthiga[2]
[1]Student [2]Head of Department
[1,2]Department of Computer Science and Engineering
[1,2]IES College of Engineering, Thrissur, India

Abstract— Diabetes is a fast-growing disease among the people nowadays. Diabetes is caused due to the excessive amount of sugar condensed into the blood. Diabetes can be classified into two categories such as type 1 and type 2. Type 1 diabetes is an autoimmune disease. The body destroys the cells essential to produce insulin to absorb the sugar to produce energy, cause obesity. The obesity is the increase in body mass index (BMI) than the normal level of BMI of an individual. Type 2 diabetes usually affects the adults who are obese. The body resists observing insulin or fails to produce insulin. Diabetes leads to various diseases such as cardiovascular complications, renal issues, retinopathy, foot ulcers, etc. We can early diagnosis by relatively inexpensive calculations without going to a diagnostic center. Depends on subjective data from informative survey more than the observed clinical value. Entering the values for the attributes and enrolling in a web based application it can able to predict whether a person is diabetic.

Keywords: Diabetes, Machine Learning, Body Mass Index (BMI), KNN

I. INTRODUCTION

Diabetes is caused due to the excessive amount of sugar condensed into the blood. Currently, it is considered as one of the lethal diseases in the world. People all around the globe are affected by this severe disease knowingly or unknowingly. Other diseases like heart attack, paralyzed, kidney disease, blindness etc. are also caused by diabetes. Numerous computer-based detection systems were designed and outlined for anticipating and analyzing diabetes.

Usual identifying process for diabetic patients needs more time and money. But with the rise of machine learning, we have that ability to develop a solution to this intense issue. Therefore we have developed an architecture which has the capability to predict where the patient has diabetes or not.

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.

- Can use in many fields
- Personalized Medicine
- Autonomous Robotic Surgery
- Automatic Treatment or Recommendations.
- Continuing genetic research to prevent disease and reduce its complications (gestational diabetes).

Control of gestational diabetes to avoid complications of fetal uterine growth

II. MOTIVATION

It has significant potential in the field of medical science for the detection of various medical data accurately. It is used for computerizing the existing biomedical System. The use of automated learning, artificial intelligence branch is to analyze and predict a predictive diabetes model. It can depend subjective data from informative survey more than the observed clinical value. It will increase the self awareness within second, apart from Building a decision support system is to predict and diagnose a particular disease. Developing a web based application on its higher prediction accuracy of some powerful machine learning algorithm.

III. LITERATURE SURVEY

Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm proposed by Mohamed El Koudri et al. [1] Presents automatic classification of Arabic web documents. Such a classification is very useful for affording directory search functionality, which has been used by many web portals and search engines to cope with an everincreasing number of documents on the web. In this paper, Naive Bayes (NB) which is a statistical machine learning algorithm is used to classify non-vocalized Arabic web documents (after their words have been transformed to the corresponding canonical form, i.e., roots) to one of five predefined categories. Cross validation experiments are used to evaluate the NB categorizer. The data set used during these experiments consists of 300 web documents per category. The results of cross validation in the leave-one-out experiment show that, using 2,000 terms/roots, the categorization accuracy varies from one category to another with an average accuracy over all categories of 68.78 %. Furthermore, the best categorization performance by category during cross validation experiments goes up to 92.8%.

A Data Mining Approach For The Diagnosis of Diabetes Mellitus proposed by Sonu Kumari et al. [2] presents data mining techniques and tools are widely used in almost every field like marketing, E -business, Retails, Healthcare Systems etc. Healthcare System is one of the new emerging research areas where can apply data mining techniques and tools. Diagnosis of a disease is done mostly by expertise and experienced doctors, but still there are cases of wrong diagnosis and treatment. Patient have to undergo various test which are very costly and sometimes all of them are not required so in this way it will hugely increase the bill of a patient unnecessarily.

In such cases proposed work will be very helpful. There have been many computerized methods to diagnose Diabetes Mellitus but the drawback of all these methods is

that patient still has to undergo various medical tests to provide the input values to the computerized diagnostic system and the overall cost for diagnosis will remain almost same for the patient but in proposed model user itself sitting from home can diagnose whether he/she is suffering from Diabetes Mellitus or not.

Diabetes Mellitus Forecast Using Different Data Mining Techniques proposed by Rakesh Motka et al. [3] presents with the improvements in expert systems and ML tools, the effects of these innovations are entering to more application domains day-by-day and medical field is one of them. Decision-making in medical field can sometimes be a trouble. Classification systems that are used in medical decision-making provide medical data to be examined in shorter time and in a more detailed manner. The work four different approaches have been proposed for the classification of subjects into two classes namely: Diabetic & Non-diabetic. The techniques undertaken are ANFIS, PCA + ANFIS, Neural Networks & PCA + Neural Networks. The results obtained are very interesting and show improvement from the previous works. There is enough scope for improvement in this field and with the advent of faster and more accurate learning techniques the results can surely be improved considerably. Again the application on live subjects rather than relying on stored datasets can lead to breakthrough research in the field of diabetes. Various available traditional methods for diagnosing diabetes are based on physical and chemical tests.

Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus proposed by Veena Vijayan V et al. [4] presents diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. A number of Data mining algorithms were designed to overcome these uncertainties. Among these algorithms, amalgam KNN and ANFIS provides higher classification accuracy than the existing approaches. The main data mining algorithms discussed are EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm and ANFIS algorithm.

EM algorithm is the expectation-maximization algorithm used for sampling, to determine and maximize the expectation in successive iteration cycles. KNN algorithm is used for classifying the objects and used to predict the labels based on some closest training examples in the feature space. K means algorithm follows partitioning methods based on some input parameters on the datasets of n objects. Amalgam combines both the features of KNN and K means with some additional processing. ANFIS is the Adaptive Neuro Fuzzy Inference System which combines the features of adaptive neural network and Fuzzy Inference System. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.

Application of K- Nearest Neighbor Classification in Medical Data Mining proposed by Hassan Shee Khamis et al. [5] presents medical data is an ever-growing source of information from hospitals in form of patient records. When mined, the information hidden in these records is a huge resource bank for medical research. The data contains hidden patterns and relationships, which can lead to better diagnosis. Unfortunately, discovery of these patterns and relationships often goes unexploited.

Studies have been carried out in medical diagnosis to predict heart diseases, lungs diseases, and various tumors based on the past data collected from patients. However, they are mostly limited to domain-specific systems that predict diseases restricted to their area of operations. In retrospect, the performance of the k-nearest neighborhoods (k-NN) classifier is highly dependent on the distance metric used to identify the k nearest neighbors of the query points. The standard Euclidean distance is commonly used in practice. This study uses vast storage of information so that diagnosis based on historical data can be made. It focuses on computing the probability of occurrence of a particular ailment by using a unique algorithm.

A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using Anfis proposed by Hassan C. Kalaiselvi et al. [6] presents the multi factorial, chronic, severe diseases like diabetes and cancer have complex relationship. When the glucose level of the body goes to abnormal level, it will lead to Blindness, Heart disease, Kidney failure and also Cancer. Epidemiological studies have proved that several cancer types are possible in patients having diabetes. Many researchers proposed methods to diagnose diabetes and cancer.

To improve the accuracy and to achieve better efficiency a new approach like Adaptive Neuro Fuzzy Inference System (ANFIS) is proposed. The Pima Indian diabetic dataset is used as data set for classification. . The main data mining algorithms discussed are EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm and ANFIS algorithm. The algorithm performs well and classifies the dataset well compared to traditional methods. The proposed work reduces the cost for different medical tests and helps the patients to take precautionary measures well in advance. In future the same method can also be applied in diagnosing other diseases like liver cancer etc.

Artificial Neural Networks In Medical Diagnosis proposed by Filippo Amato1 et al. [7] presents an extensive amount of information is currently available to clinical specialists, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. Each type of data provides information that must be evaluated and assigned to a particular pathology during the diagnostic process.

To streamline the diagnostic process in daily routine and avoid misdiagnosis, artificial intelligence methods (especially computer aided diagnosis and artificial neural networks) can be employed. These adaptive learning algorithms can handle diverse types of medical data and integrate them into categorized outputs. The philosophy, capabilities, and limitations of artificial neural networks in medical diagnosis through selected examples. In addition, their use makes the diagnosis more reliable and therefore increases patient satisfaction.

However, despite their wide application in modern diagnosis, they must be considered only as a tool to facilitate the final decision of a clinician, who is ultimately responsible for critical evaluation of the ANN output.

Methods of summarizing and elaborating on informative and intelligent data are continuously improving and can contribute greatly to effective, precise, and swift medical diagnosis.

Comparing Performances of Back Propagation and Genetic Algorithms in the Data Classification Diagnosis proposed by H. Hasan Örkcü et al. [8] presents artificial neural networks (ANN) have a wide ranging usage area in the data classification problems. Back propagation algorithm is classical technique used in the training of the artificial neural networks. Since this algorithm has many disadvantages, the training of the neural networks has been implemented with the binary and real-coded genetic algorithms.

These algorithms can be used for the solutions of the classification problems. The real-coded genetic algorithm has been compared with other training methods in the few works. It is known that the comparison of the approaches is as important as proposing a new classification approach. For this reason, in this study, a large-scale comparison of performances of the neural network training methods is examined on the data classification datasets. This technique uses the values of a set of variables associated with each observation or unit to classify observation or unit of unknown group or class membership. Generally, the major classification methodologies include statistical methods, neural networks and other machine learning approaches.

Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques proposed by C Kalaiselvi [9] presents the heterogeneous, chronic diseases like heart diseases and cancer are commonly occur and increased nowadays in diabetic patients. Most of the people do not know the symptoms of these diseases and its chronic complications. The aim of this paper is to predict the diseases such as heart diseases and cancer in diabetic patients. The association between these diseases can be analyzed based on the factors that cause these diseases which include obesity, age, associated diabetic duration, and some other life style factors. Methods: This work consists of two stages. In the first stage, the attributes are identified and extracted using Particle Swarm Optimization (PSO) algorithm.

Comparison of Neat and Back propagation Neural Network on Breast Cancer Diagnosis proposed by Hamza Turabieh [10] presents a comparison between Neuro Evolution of Augmenting Typologies (NEAT) algorithms with Back propagation Neural Network for the prediction of breast cancer. Machine learning algorithms could be used to enhance the performance of medical practitioners in the diagnosis of breast cancer. NEAT is a promising machine learning algorithm, which combines genetic algorithms and neural network. Results demonstrate that NEAT outperforms Back propagation Neural Network, and show that experimentally that NEAT has better generalization and much lower computational cost.

Compare the performance between NEAT algorithm and standard neural network to predict breast cancer. NEAT is attempting to evolve optimal or nearoptimal neural network topologies to solve problems. Thus, a very good approach to evolving a neural network to solve a given problem seems to be to start with relatively minimal architectures while including both complexifying and simplifying mutations, their rates balanced in such a way to produce an even search in both directions through the range of possible architectures.

This approach should yield both more efficient searches and more efficient solutions than either complexifying or simplifying algorithms alone. However, work remains to be done to verify these conclusions across a wide range of domains of varying difficulty

IV. PROPOSED SYSTEM

Supervised Learning algorithms learns the pattern from pre-existing data and try to predict new result based on the previous learning. ML algorithms are used to identify existing data like probability-based, function-based, rule-based, treebased, instance-based, etc. Here indicates the ample architecture of my proposed Model. According to the model patients are required to provide their medical data for successful diagnosis of their diabetes test.

A. Data Preparation

As a Data Scientist the most tedious task which encounter is the acquiring and the preparation of a data set. Even though there is an abundance of data in this era, it is still hard to find a suitable data set which suits the problem you are trying to tackle. If there aren't any suitable data sets to be found, you might have to create your own. In this tutorial aren't going to create own data set, instead, be using an existing data set called the "Pima Indians Diabetes Database" provided by the UCI Machine Learning Repository (famous repository for machine learning data sets). Will be performing the machine learning workflow with the Diabetes Data set provided above.

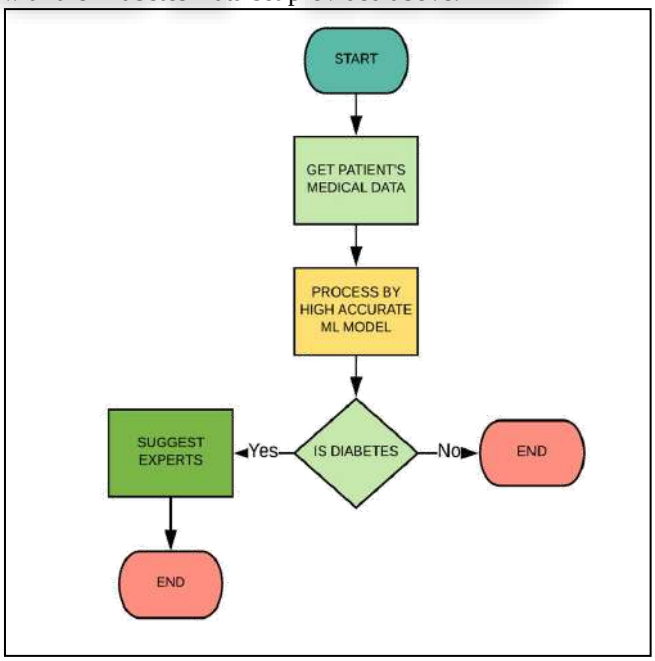

Fig 1: Flow Chart of Diabetes Prediction System.

B. Data Exploration

When encountered with a data set, first should analyse and "get to know" the data set. This step is necessary to

familiarize with the data, to gain some understanding about the potential features and to see if data cleaning is needed. First will import the necessary libraries and import data set to the Jupyter notebook. It can observe the mentioned columns in the data set. It should be noted that the above data set contains only limited features, where as in reality numerous features come into play. It can examine the data set using the pandas' head () method.

It can find the dimensions of the data set using the panda Dataframes' 'shape' attribute. Diabetes data set dimensions. Can observe that the data set contain 768 rows and 9 columns. 'Outcome' is the column which are going to predict, which says if the scaling of the image and in-plane rotation of the face. patient is diabetic or not. 1 means the person is diabetic and 0 means person is not. It can identify that out of the 768 persons, 500 are labeled as 0 (non-diabetic) and 268 as 1 (diabetic).

Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn etc.

C. Data Cleaning

The phase of the machine learning work flow is the data cleaning considered to be one of the crucial steps of the work flow, because it can make or break the model. There is a saying in machine learning "Better data beats fancier algorithms", which suggests better data gives you better resulting models. Ignore/remove these cases: This is not actually possible in most cases because that would mean losing valuable information. And in this case "skin thickness" and "insulin" columns means have a lot of invalid points. But it might work for "BMI", "glucose "and "blood pressure" data points. Put average/mean values:

This might work for some data sets, but in case putting a mean value to the blood pressure column would send a wrong signal to the model. Avoid using features: It is possible to not use the features with a lot of invalid values for the model. This may work for "skin thickness" but it's hard to predict that. By the end of the data cleaning process have come to the conclusion that this given data set is incomplete.

D. Feature Engineering

Feature engineering is the process of transforming the gathered data into features that better represent the problem that are trying to solve to the model, to improve its performance and accuracy. Feature engineering create more input features from the existing features and also combine several features to produce more intuitive features to feed to the model.

"Feature engineering enables to highlight the important features and facilitate to bring domain expertise on the problem. It also allows to avoid over fitting the model despite providing many input features". 'Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes Pedigree Function', 'Age' By a crude observation can say that the 'Skin Thickness' is not an indicator of diabetes. But can't deny the fact that it is unusable at this point. Therefore will use all the features available.

It separate the data set into features and the response that are going to predict. Assign the features to the X variable and the response to the y variable. Generally feature engineering is performed before selecting the model. However for this tutorial follow a different approach. Initially will be utilizing all the features provided in the data set to the model, will revisit features engineering to discuss feature importance on the selected model.

E. Model Selection

Model selection or algorithm selection phase is the most exciting and the heart of machine learning. It is the phase where select the model which performs best for the data set at hand. First will be calculating the "Classification Accuracy (Testing Accuracy)" of a given set of classification models with their default parameters to determine which model performs better with the diabetes data set. Will import the necessary libraries to the notebook. Import 7 classifiers namely K-Nearest Neighbors, Support Vector Classifier, Logistic Regression, Gaussian Naïve Bayes, Random Forest and Gradient Boost to be contenders for the best classifier. Initialize the classifier models with their default parameters and add them to a model list.

F. Model Parameter Tuning

Scikit Learn provides the model with sensible default parameters which gives decent accuracy scores. It also provides the user with the option to tweak the parameters to further increase the accuracy. Out of the classifiers will select Logistic Regression for fine tuning, where change the model parameters in order to possibly increase the accuracy of the model for the particular data set. Instead of having to manually search for optimum parameters, can easily perform an exhaustive search using the GridSearchCV, which does an "exhaustive search over specified parameter values for an estimator". It is clearly a very handy tool, but comes with the price of computational cost when parameters to be searched are high.

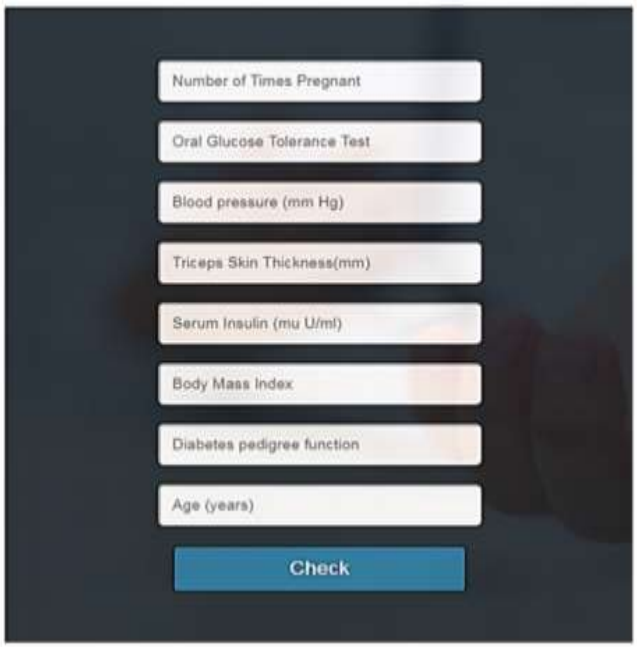

Fig. 2: Attributes Using For Prediction.

First import GridSearchCV: Logistic Regression model has some hyperparameters that doesn't work with each other. Therefore provide a list of grids with compatible parameters to fine tune the model attributes to find the best parameters and the best estimator. Can feed the best parameters to the Logistic Regression model and observe whether its accuracy has increased. Could conclude that the hyper parameter tuning didn't increase its accuracy.

V. IMPLEMENTATION

My goal is to develop a Web-based application which can predict diabetes based on patients health data. In order to find most suitable ML algorithm that is capable of predicting diabetes more precisely, I have tested some sort of powerful machine learning models like SVM, KNN, Naive Bayes and ANN. For the successful evaluation of these models, we have used some machine learning libraries such as Scikit-learn, numpy, matplotlib, pandas, and tensorflow.js. In order to get rid of overfitting problem we split up the dataset into two subparts: one is for testing and another is for training. Based on different training dataset size we have achieved the accuracy rate for every defined Model. However, preprocessing of Indian Pima Dataset can produce higher prediction accuracy. Therefore, we also calculate the accuracy for our defined model in order to improve the prediction accuracy.

Data normalization is an effective way to increase the accuracy of certain machine learning models and some machine learning model do not perform well without data Normalization. Numpy provides a high-performance multidimensional array object, and tools for working with these arrays

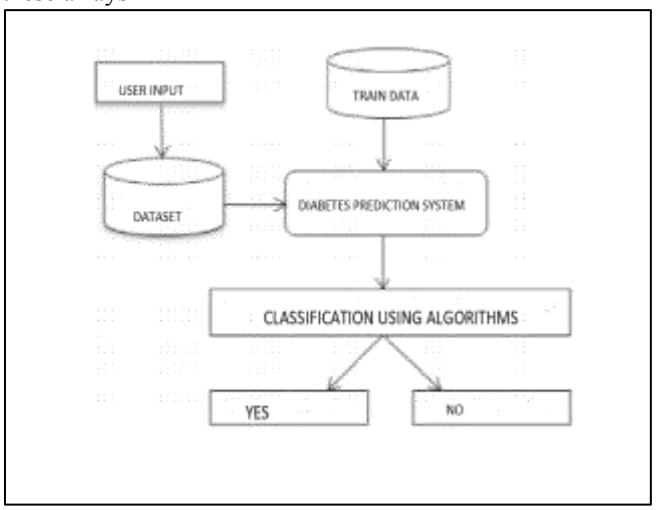

Fig. 3: Architecture diagram

matplotlib is a 2D plotting library which produces publication quality figures in a variety of hardcopy formats. Imutlils modules which allow developers to write applications or services without having to handle such low-level details as protocols. Scipy is an All-in-one library-based ecosystem for scientific and technical computing and collection of algorithms for image processing. Flask is a web frame work. Provide tools libraries and technologies for build web application. Glob module finds all the pathnames matching a specified pattern.

VI. RESULT

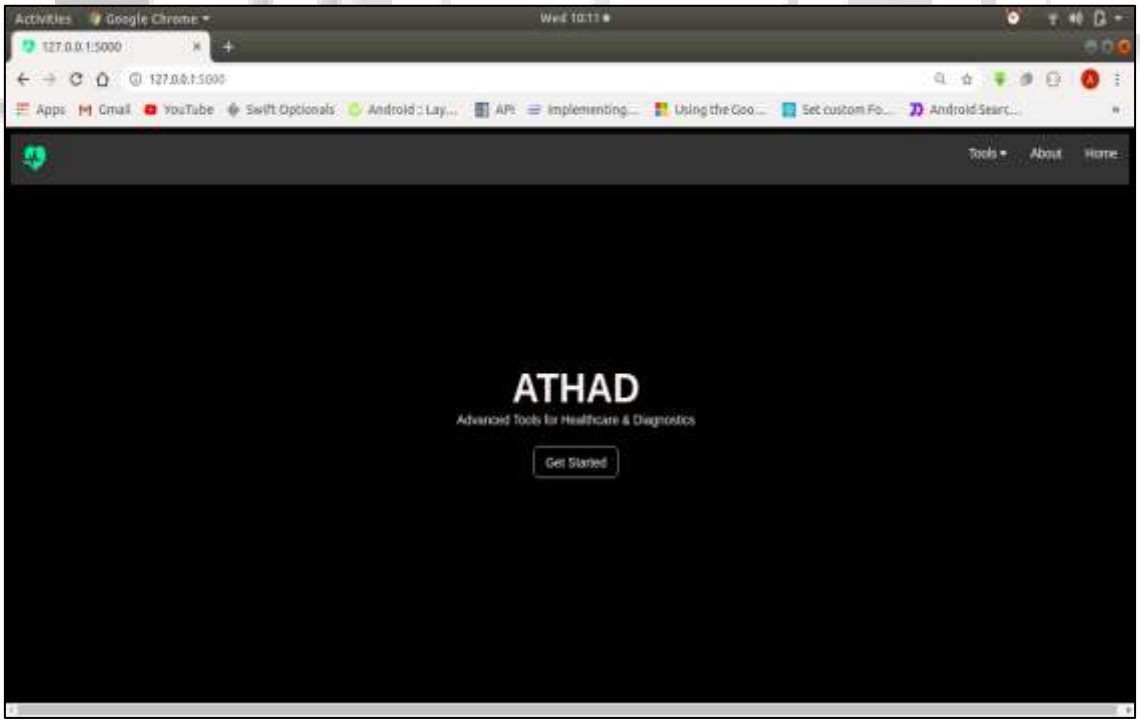

Fig. 4: Advanced Tool for Healthcare Diagnosis.

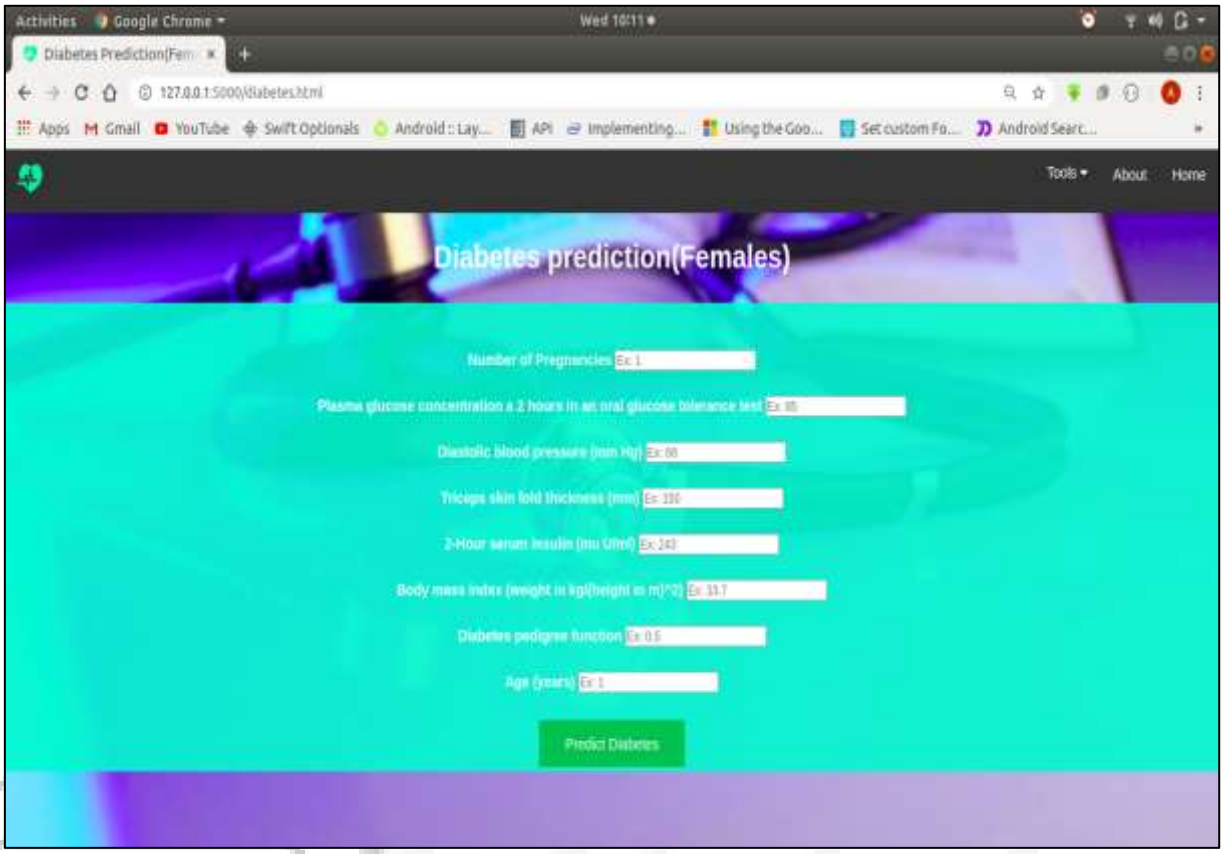

Fig. 5: Patients details entering page.

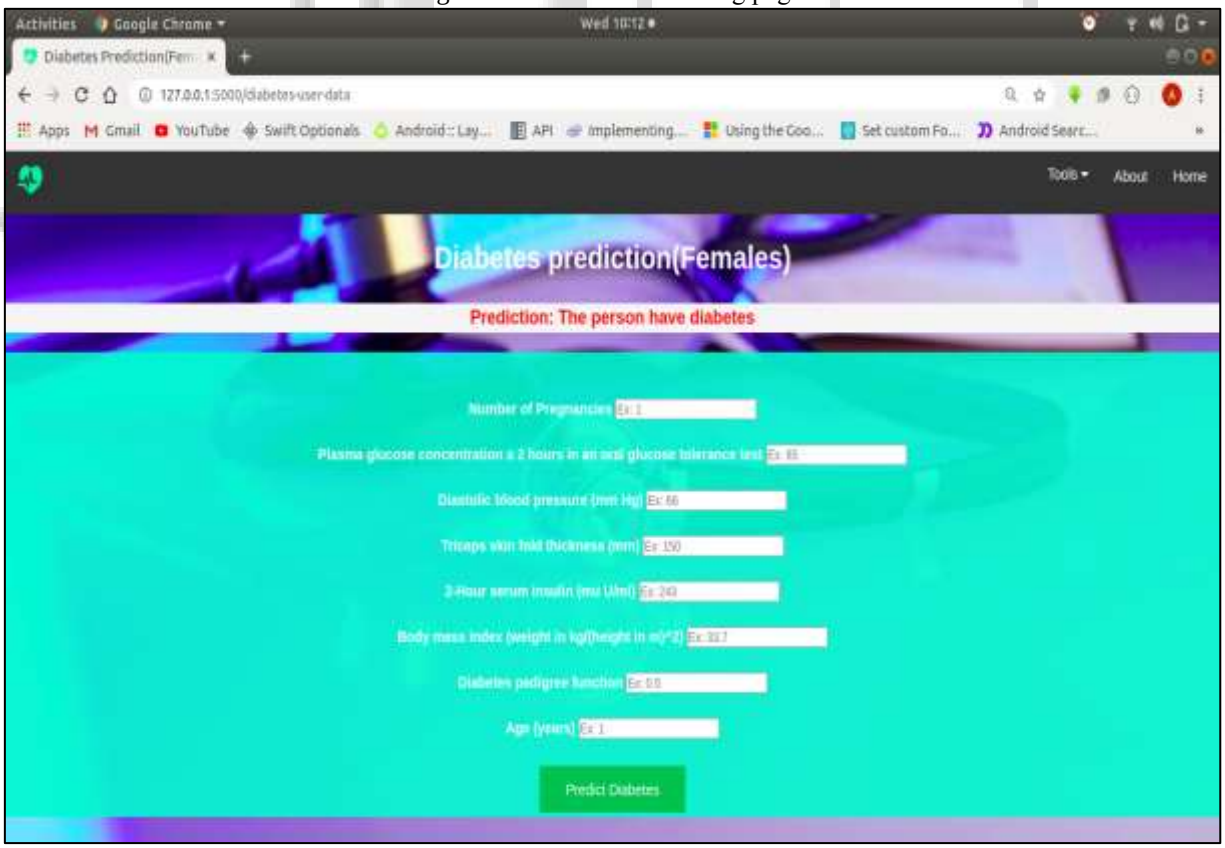

Fig. 6: The person having Diabetes.

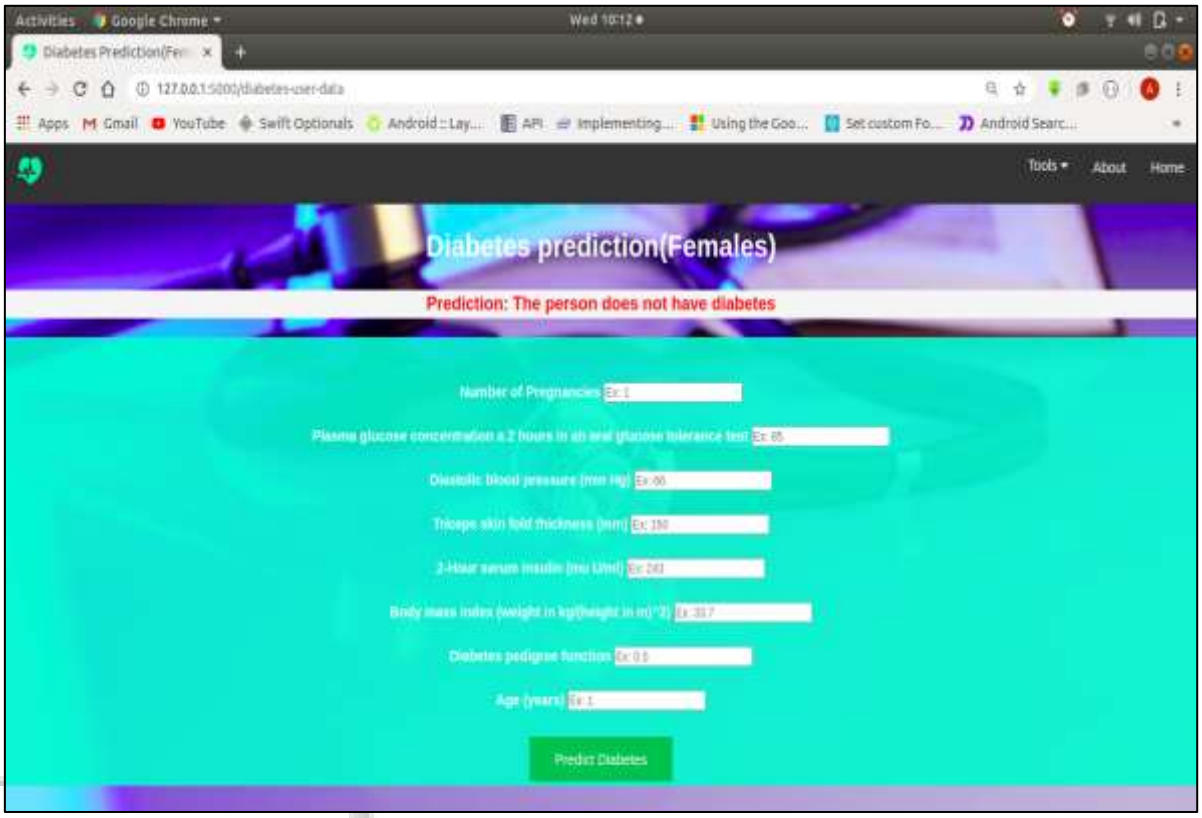

Fig. 7: The Person Not Having Diabetes.

VII. Conclusion

A web based application for the successful prediction of Diabetes Diseases from different machine learning algorithms. As I have proposed and developed an approach for diabetes disease prediction, Using ML techniques trained the system in such a way that it can classify the given input whether diabetic or not based on the given features. Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Computerizing the existing biomedical System, So that we can increase the self awareness within second apart from many doctor visits. Detection of diabetes in its early stages is the key for treatment.

Acknowledgment

I would like to convey my heartfelt gratitude towards my guide Dr. G. Kiruthiga for her constant guidance, encouraging help and inspiring words. I am thankful to the department of computer engineering for their support.

References

[1] M. Elkourdi, A. Bensaid, and T. Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm,” Alakhawayn University.
[2] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," 2013 7th International Conference on Intelligent Systems and Control (ISCO), Coimbatore.
[3] P. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining,” Pearson Education.
[4] R. Motka, V. Parmarl, B. Kumar and A. R. Verma, "Diabetes mellitus forecast using different data mining techniques," 2013 4th International Conference on Computer and Communication Technology (ICCCT), Allahabad.
[5] Vijayan, and A. Ravikumar, “Study of Data Mining algorithms for Prediction and Diagnosis of Diabetes Mellitus,” International Journal of Computer Application.
[6] H. Shee, K. W Cheruiyot and S. Kimani, “Application of k-Nearest Neighbour Classification in Medical Data Mining,” International Journal of Information and Communication Technology Research.
[7] C. Kalaiselvi and G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS," 2014 World Congress on Computing and Communication Technologies, Trichirappalli.
[8] Aleksander, and H. Morton, “An introduction to neural computing,” Int Thomson Computer Press, London.
[9] Chandra S.S, and A. Hareendran S, “Artificial Intelligence and machine learning,” PHI learning Private Limited, Delhi.
[10] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl and J. Havel, “Artificial neural networks in medical diagnosis”.
[11] Siji Joju, Dr. G. Kiruthiga, “A Survey on Diabetic Prediction System using Machine Learning Algorithm”. The IJSRD - International Journal for Scientific Research & Development| Vol. 8, Issue 1, 2020 | ISSN (online): 2321-0613

