Diabetes Mellitus Prediction and Diet Recommendation

Aishwarya Jadhav
Student, Pimpri Chinchwad College of Engineering, Pune 411044

Kajal Kadam
Student, Pimpri Chinchwad College of Engineering, Pune 411044

Vishal Dusane
Student, Pimpri Chinchwad College of Engineering, Pune 411044

Gopal Kabra
Student, Pimpri Chinchwad College of Engineering, Pune 411044

Ganesh Deshmukh
Assistant Professor, Pimpri Chinchwad College of Engineering, Pune 411044

ABSTRACT
Type 2 diabetes is a lifelong disease that keeps your body from using insulin the way it should. People with type 2 diabetes are said to have insulin resistance. Insulin resistance, a condition in which fat, muscle, and liver cells do not consume insulin properly leading to this form of diabetes. In this paper, the system is designed to predict diabetes and recommend the diet to the user. Classification algorithms were studied and implemented to calculate accuracy. Based on the results obtained, the highest accuracy was given by the Random Forest classifier. Thus, the same is used in the system for prediction purposes. Diet is recommended to diabetes-affected patients using a diet recommendation module to control his/her blood sugar levels. The recommended diet is based on an individual's requirement of calories calculated by the Harris-Benedict equation. A unified system that predicts whether an individual is diabetic and helps to manage diet according to his/her caloric needs is implemented.

Keywords
Diabetes mellitus, prediabetes, machine learning, diet recommendation, random forest classifier, disease prediction

1. INTRODUCTION
Diabetes of all types can increase the risk of dying prematurely and affect many parts of the body [11]. The number of individuals with prediabetes is predicted to grow fairly and is estimated to globally affect 482 million people by 2040 [1]. Type 2 diabetes is caused because the body uses insulin ineffectively. Most people with diabetes around the world are diagnosed with type 2 diabetes. Excessive body weight and physical inactivity result in type 2 diabetes. Risk factors are increased weight, abnormal cholesterol, high triglycerides, sleep apnea,family history, high blood pressure etc. The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014[1]. Diabetes is a major cause of kidney failure, heart attacks, blindness, lower limb and stroke amputation. To stop the increase of diabetic patients, they should be made aware about prediabetes and diabetes both. They should take care right from the prediabetic stage. Early diagnosis of diabetes mellitus can save many lives. Treatment of type 2 diabetes primarily involves lifestyle changes, monitoring of your blood sugar, along with diabetes medications, insulin, or both. So as to do this a system should be implemented which consists of significant predictors of diabetes and will predict whether a person is prone to prediabetes as of now and diabetes in the long run. This system will help in early diagnosis of diabetes and will help to manage diabetes through its diet recommendation system.

2. RELATED WORKS
In this section we have learned different attributes required to detect and predict diabetes and use the best attributes along with a classifier which provides greater accuracy. We studied various databases and selected the one which provides non-clinical assistance.

In [2] the authors have used the mentioned algorithms - Neural Networks. The features are: Gender, Age, Weight, Height, Loss in weight, Thirst increase, Hunger increase, appetite increase, Nausea, Fatigue, Skin infections, Blurred vision[2]. The dataset of 100 entries was collected randomly, out of which 70 were used to train the model and the remaining were used to test it. This study aims to provide an affordable home efficient way to detect diabetes. It consists of 13 features which don't require to be clinically tested and saves money. The neural network has 28 nodes out of which 13 nodes are input nodes, 14 nodes are hidden nodes and one output node. If the output is 0 then the patient is not affected or else he is considered as affected. But the test cases considered were only 20 cases. The dataset is not mentioned and some of the features like thirst increase and hunger increase are ambiguous and can't certainly be due to diabetes. Artificial Neural Network with back propagation algorithm the accuracy achieved was 92.8 %. It is cost efficient and there is no need to visit any clinics.

In [3] the authors have used the mentioned algorithms - Logistic Regression, Adaptive Boosting, Random Forest, and Support Vector Machine. The Pima Indians Diabetes Dataset contains a total of 768 instances, with 8 attributes including following: Blood pressure, Skin thickness, Insulin, BMI, Diabetes, pedigree function, Age, No of times pregnant, Glucose concentration found in oral glucose tolerance test (glucose level) [3]. The outcome attribute holds either 0 value (not diabetic) or value 1 (diabetic) for the instances. The missing values for the attributes glucose level, blood pressure, skin thickness, insulin, BMI found in the dataset were filled by their mean in the dataset respectively. Support Vector Machine (SVM) with a linear kernel was found to perform better than others.

In [4] the authors have used the mentioned algorithms - Random Forest, Naive Bayes, Support Vector Machine, Simple CART algorithm. Pre-process the input dataset for diabetes disease Perform percentage split of 70% to divide dataset as Training set and Test set. Select the machine learning algorithm i.e. Random Forest and Simple CART algorithm, Naive Bayes, Support Vector Machine. Build the classifier model for the mentioned machine learning algorithm based on the training set. Test the Classifier model for the mentioned machine learning algorithm based on the test set. Perform Comparison Evaluation of the experimental performance results obtained for each classifier. After

analysing based on various measures, conclude the best performing algorithm. The Accuracy of SVM is the highest, which is 0.7913. The Accuracy of Naive Bayes is 0.77 better than Random Forest and Simple CART. The accuracy of Random Forest and Simple CART is almost equal with value 0.765.

In [5] the authors have used the mentioned algorithms - Decision Tree, RandomForest, and Artificial Neural Network. The features used are: Left systolic pressure (LSP), Right systolic pressure (RSP), Left diastolic pressure (LDP), Right diastolic pressure (RDP), Age, Pulse rate , Breathe, Height, Weight, Physique index, Fasting glucose, Low density, lipoprotein (LDL), High density lipoprotein (HDL), Waistline [5].

This study uses two different datasets with different feature sets. For the Pima Indians dataset hold and wait technique was applied for the distribution of dataset. This dataset contains 768 entries. In the Luzhou dataset five-fold validation was used, randomly patients were selected from the Luzhou dataset, which consisted of both healthy and diabetic patients. It was extracted five times and provided to the algorithms and the average of the all 5 results were considered for the output.

PCA and mRMR was also used to reduce the dimensionality of the dataset. The index for detecting diabetes is limited to fasting glucose for the Luzhou dataset and oral glucose tolerance test for the Pima Indian dataset. Random Forest is the best algorithm with accuracy of 80 % in the Luzhou dataset, and 77% accuracy in the Pima Indians dataset.

In [6] the authors have used the mentioned algorithms - Logistic regression, K-Nearest Neighbour, Support vector Machine, Naive Bayes, Decision Tree, Random Forest. The features used are : age, Gender ,Family history of diabetes, High Blood Pressure, Physical activity, BMI, Smoking, Alcohol, Sleep hours, Sound sleep hours, Regular Medicine, Junk Food, Stress, Pregnancies, Urination frequency, Blood Pressure level, Prediabetes[6]. A dataset containing 952 rows collected by survey. It uses non clinical parameters to predict diabetes. No diet recommender model is suggested for diabetes affected patients. Random Forest provides greater accuracy of 94.1%. It is cost efficient and there is no need to visit any clinics.

In [7] the author have mentioned following modules -

(1) Diabetes Intelligent Meal Recommender Module, (2) Educational Module for Diabetics with Food Recognition Engine, (3) Activity Tracking Module, and (4) Medication Reminder Module. Diabetes Intelligent Meal Recommender Module schedules a diet for diabetes management by generating whole meals for breakfast, lunch, and supper to meet the nutritional requirements of the diabetic patient. The diet type of the patient is determined from the obtained data, and calorie needs calculated using the Harris-Benedict's equation. The food is recommended using KNN Algorithm. The present system tracks the user's walk and saves the route but does not relate the saved route to the calories burned. We have studied all these papers along with their pros and cons and have tried to implement a system with higher accuracy in prediction and a diet recommender system especially designed for diabetic patients.

3. PROPOSED SYSTEM

PART 1: Diabetes Prediction System

With a good prediction model and an accurate detection technique, diagnosis can be made more efficient. To accurately predict the disorder a good model which can represent the presence of diabetes through input characteristics is required. The random forest classifier is used for training the model.

The dataset used for training the model is collected by a survey by the following authors: Neha Prerna Tiggaa, Shruti Garga, from the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India [6]. The dataset used for training the model consists of 17 attributes and 1 target variable. The total number of instances in the dataset is 952. The 17 features consist of age, gender, family history, high blood pressure, physical activity, BMI, smoking, alcohol, sleep hours, sound sleep hours, regular medicine, junk food, pregnancies, urination frequency, blood pressure level, prediabetes.

In the pre-processing step, out of the above 17 attributes, authors have removed 3 attributes namely 'high blood pressure, 'sound sleep hours', and 'prediabetes'. Here 'high blood pressure' and 'sound sleep hours' are redundant attributes as 'blood pressure' and 'sleep hours' are already included. If the 'prediabetes' attribute is included then that will be the most relevant attribute as if a person is non-prediabetic then he is not diabetic and this attribute may also require clinical assistance.

Random Forest Classifier:

In the random forest algorithm, decision trees are created on sample data and the result from each tree is taken and then the best solution is selected using majority voting. It is an ensemble learning method that is better than a single decision tree as it reduces the over-fitting by averaging the result.

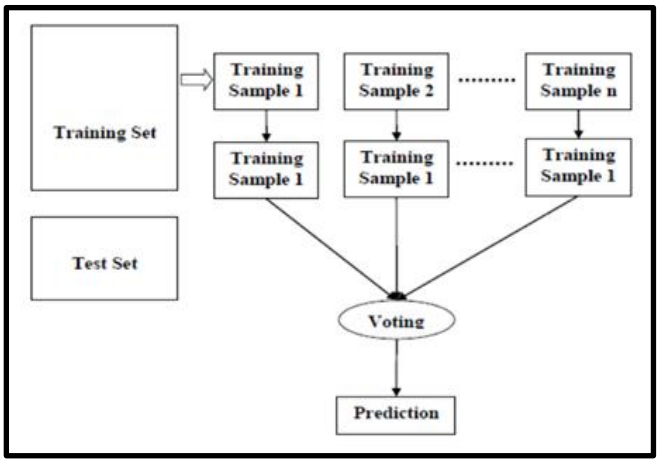

Fig. 1: Working of Random Forest

PART 2: Diet recommendation module

In this module, diet is recommended to the diabetic patient by calculating required calories according to the patient's age, height, weight, and activity.

The algorithm consists of 3 steps which are as follows:

Step 1: Calculate Basal Metabolic Rate (BMR) using Muffin - St. Joer Equation [7]

- $\boldsymbol{Male}$: $(10 \times weightinkg) + (6.25 \times heightincm) - (5 \times ageinyears) + 5$
- $\boldsymbol{Female}$: $(10 \times weightinkg) + (6.25 \times heightincm) - (5 \times ageinyears) - 161$

Step 2: Calculate total daily calorie needs: Harris-Benedict Formula [7]

To determine a person's total daily calorie needs, multiply basal calories by an appropriate activity factor, as follows:

- $Calorie - Calculation = BMR \times 1.2$, If the person is sedentary (little or no exercise)
- Calorie − Calculation = BMR × 1.375, If person is lightly active (light exercise/sports 1-3 days/week)
- Calorie − Calculation = BMR × *1.725*, If person is moderately active (moderate exercise/sports 3-5 days/week) : If person is very active (hard exercise/sports 6-7 days a week)
- Calorie − Calculation = BMR × *1.9*, If the person is extra active (very hard exercise/sports & physical job or 2x training)

Step 3: According to the total energy requirement select a diet chart for a week. Each meal is decided using the plate method. The plate method uses the image of a **9-inch standard dinner** plate to help people visualise nutritional balance as they plan their meals. The **Centres for Disease Control and Prevention (CDC)** [8] recommend imagining that a plate full of food includes:

- 50% non-starchy vegetables
- 25% lean protein, such as lentils, tofu, fish, or skinless and fatless chicken.
- 25% of high-fibre carbohydrates, such as whole grains or legumes [9].

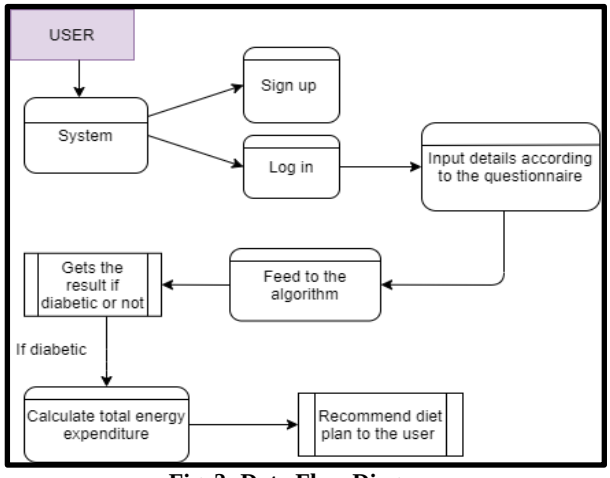

Fig. 2: Data Flow Diagram

Figure.2 shows a data flow diagram that elaborates the flow of the system. The user has to log in or sign up to the system. Then he/she has to answer the questionnaire. The data obtained will be given to the prediction module and the result will be obtained. If the patient is diabetic then total energy expenditure will be calculated and diet will be recommended.

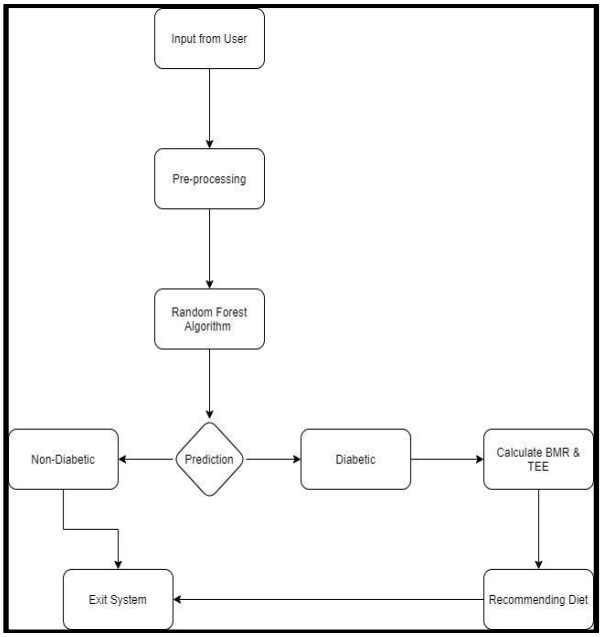

Fig.3: System Flowchart

Figure.3 shows the flowchart of the proposed system where it accepts the input from the user, predicts the result and then recommends diet accordingly.

4. RESULTS & DISCUSSION

Algorithms applied on Diabetes Dataset and their accuracies:

Here authors have used 5 different classification algorithms based on the research papers studied. Then the accuracy of these 5 algorithms were compared. Based on these accuracies, the final algorithm for the system was decided.

Table no. 1: Accuracy of different algorithms.

SR NO.	ALGORITHM	ACCURACY
1.	Logistic Regression	87.08%
2.	K-nearest Neighbour	91.59%
3.	Support Vector Machine	86.78%
4.	Decision Tree	87.98%
5.	Random Forest	96.99%

Random Forest Algorithm provides comparatively greater accuracy (~97%) and is simpler to implement. The algorithm is implemented on the survey Dataset. The test cases are 30% of the actual dataset. Here the correct predictions are = 277. The total predictions are = 286. Thus the accuracy obtained during implementation is 96.99 % Authors are predicting diabetes which will save the patient's time. Because of this, the system is capable of predicting

whether the person is diabetic or is prone to diabetes. If the training set size is increased then the accuracy also improves but on a minute scale.

The system is designed in such a way that people can easily access it through a website that has a user-friendly interface, designed in HTML, CSS. The framework used is Django and the database used is SQLite. The prediction system is implemented using the python sklearn library. The prediction results are stored in a pickle file to save time. The diet recommender module is implemented using python. This web interface is accessible on various platforms (Windows, Linux, Mac, etc.) and various browsers (chrome, safari, Firefox, etc.)

System User Interface

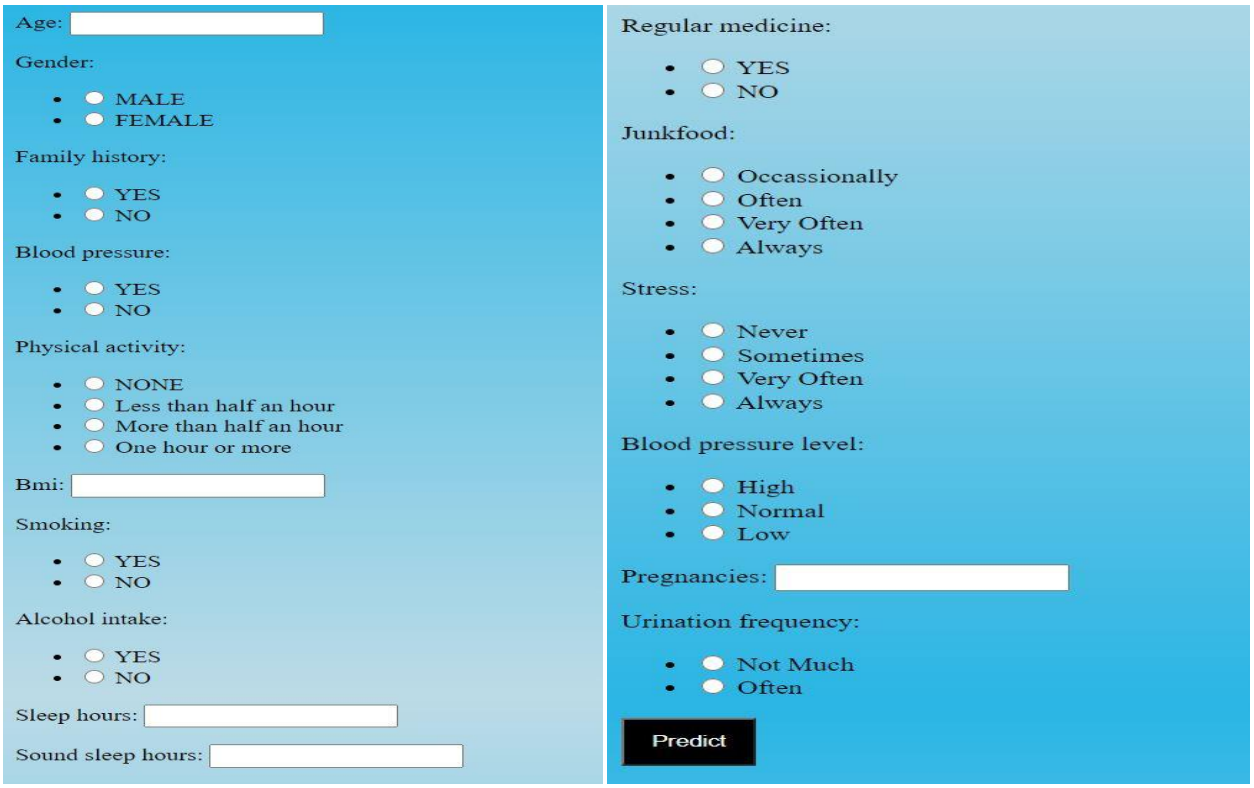

Fig 4. Diabetes Prediction Form

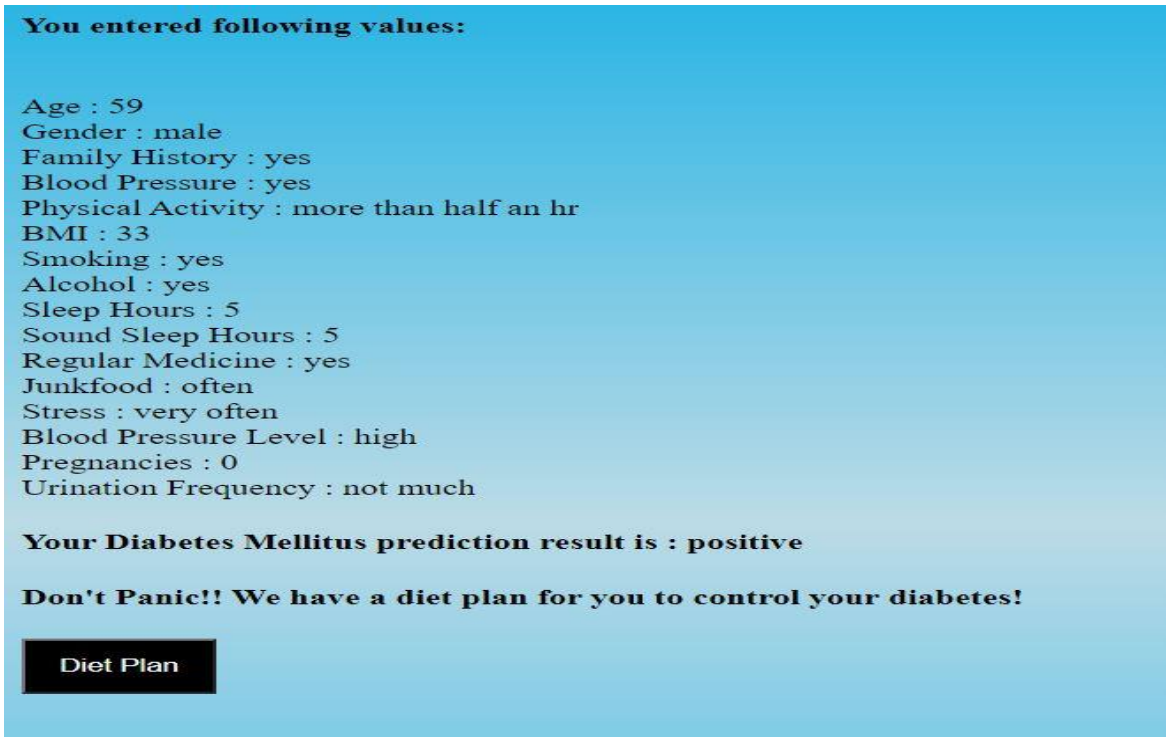

Fig 5. Prediction Result

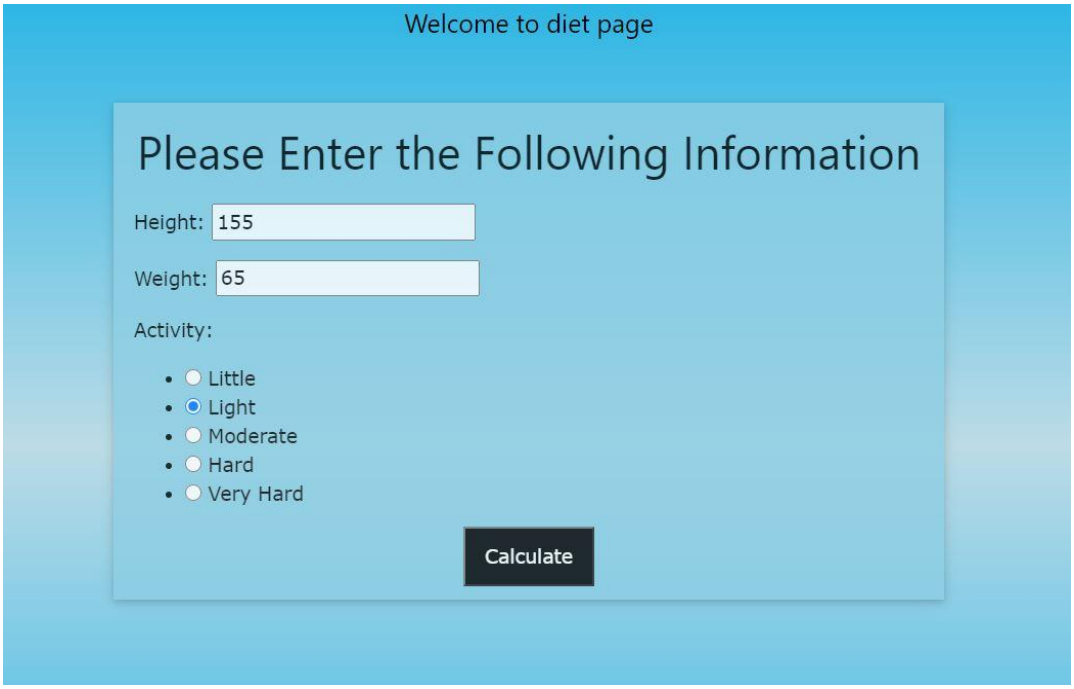

Fig 6. Diet Plan Requirements Form

Fig. 7: Diabetes Result Prediction

	BREAKFAST	LUNCH	DINNER	SNACK
MONDAY	Cereal (cooked) ½ cup	Peas (cooked) ½ cup Paneer/Chicken (white, no skin) 1 cup Whole milk 1 cup	Corn (cooked) ½ cup Eggs 1 Beets (A serving is ½ cup of cooked vegetables or 1 cup of raw vegetables) Butter 1 tsp	Apple 1
TUESDAY	Beans (cooked or canned) 1/3 cup	Corn (cooked) ½ cup Buttermilk	Corn on the cob(sweetcorn) 1 Chicken/fish/paneer Broccoli Nuts or seeds 1 Tbsp	Banana (medium) ½

Fig. 8. Suggested Diet Plan

5. CONCLUSION & FUTURE WORK

The system is trained for different algorithms such as random forest algorithm, logistic regression, K-nearest neighbour, decision tree, support vector machine. Here RandomForest was selected as it provided the highest accuracy. This software model will help a lot of people in India to control diabetes. Diet will be recommended to maintain the blood glucose levels which will contain foods with a low glycemic index. The developed system provided results with an accuracy of 96.9%.

Future work includes faster implementation with higher accuracy. The system can be integrated within a Chatbot.

6. REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes - This is the official website of the World Health Organisation (WHO).

[2] Sonu Kumari, Archana Singh "A Data Mining Approach for the Diagnosis of Diabetes Mellitus". International Conference on Intelligent Systems and Control (ISCO 2013)

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

[4] Ayman Mir, Sudhir N. Dhage "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare",2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[5] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang "Predicting Diabetes Mellitus With Machine Learning Techniques" Frontiers in Genetics November 2018 | Volume 9 | Article 515

[6] Neha Prerna Tiggaa, Shruti Garga, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods", International Conference on Computational Intelligence and Data Science (ICCIDS 2019)the

[7] Robert A. Sowah, Adelaide A. Bampoe-Addo, Stephen K. Armoo, Firibu K. Saalia, Francis Gatsi, and Baffour Sarkodie-Mensah "Design and Development of Diabetes Management System Using Machine Learning", Hindawi International Journal of Telemedicine and Applications 2020.

[8] Francesco Mercaldoa, Vittoria Nardone, Antonella Santone "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques" International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017.

[9] Sonu Kumari, Archana Singh "A Data Mining Approach for the Diagnosis of Diabetes Mellitus". International Conference on Intelligent Systems and Control (ISCO 2013)

[10] https://www.cdc.gov/diabetes/managing/eat-well/meal-plan-method.html - This is the official website of the Centre for Disease Control (CDC).