International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

DIABETES PREDICTION BASED ON FOOD CONSUMPTION

Sneha R Shetty[1], Ananya K S[2], Apeksha K J[3], Disha[4], Manisha[5]

1Assistant Professor Computer Science Department, SDMIT Ujire
2BE Computer Science Student, SDMIT Ujire
3 BE Computer Science Student, SDMIT Ujire
4BE Computer Science Student, SDMIT Ujire
5BE Computer Science Student, SDMIT Ujire

Abstract - Diabetes is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body's cells do not respond properly to insulin. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the eyes, kidneys, nerves, heart, and veins. The proposed method aims to focus on selecting the attributes that are in early detection of Diabetes using Predictive analysis.

Key Words: Diabetes, Machine learning, Healthcare

1.INTRODUCTION

Diabetes is a chronic disease that has no cure, where the body is unable to produce Insulin hormone as normal body do and which due to the blood glucose level is too high in the body. As we know, normal body obtained blood glucose level from the meal that human take daily and the insulin hormone is very important in providing energy to body. However, excessive or high glucose level in body can cause many serious problems such as it can damage eyes, kidney, and nerves. Usually, due to lack of knowledge about diabetes is the reason diabetic patient does no know to self-manage their illness. Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely KNN, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage based on the food consumption.

There are two main types of diabetes: Type 1 and Type 2. Gestational diabetes occurs during pregnancy and affects about 18 percent of all pregnancies. Diabetes is considered as one of the deadliest and chronic diseases which cause an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic centre and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely SVM and Random Forest are used in this experiment to detect diabetes at an early stage based on the food consumption.

This paper is used to predict diabetes based on daily monitoring of food consumption. In order to alleviate the burden of a doctor, a system that provides an early warning can be of help. Additionally, the user is able to view basic information regarding diabetes and advice for diabetes managed in future. Main aim of this project is to predict diabetes and prevent from being diabetic.

Arguably, health issues are one of the serious matters that directly affect the wellbeing of our community. One of the major health problems that faced by the community members are the diabetes Mellitus diseases. This project aims in developing system that helps to reduce the time between the patient and doctor in order to identify whether they have diabetes by symptomatic selection.

2. LITERATURE SURVEY

Literature Survey is an important activity, which we have to do while gathering information about a particular topic. It will help us to get required information or ideas to do work. The following paragraphs discuss the related work and issues in the area Prediction and Analysis of Energy Consumption in a Computer Laboratory using machine learning algorithm.

Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid and Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare" [1]. This paper discusses the predictive analytics in healthcare, six different machine learning algorithms are used in this research work. For experiment purpose, a dataset of patient's medical record is obtained and six different machine learning algorithms are applied on the dataset. Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes.

Deepti Sisodia and Dilip Singh Sisodia,"Prediction of Diabetes using Classification Algorithms" [2]. In this paper performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques" [3]. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project also aims to propose an effective technique for earlier detection of the diabetes disease. The technique may also help researchers to develop an accurate and effective tool that will help them make better decision about the disease status.

Debadri Dutta , Debpriyo Paul, Parthajeet Ghosh ," Analysing Feature Importances for Diabetes Prediction using Machine Learning" [4]. In this paper we will discover what are the critical elements for the reason of diabetes are. Likewise we will centre on the most essential features to predict whether a person will have chances to develop diabetes in the future. We can conclude that Random Forest is the most ideal algorithm for predicting Diabetes, which gives an accuracy of around 84%. And if people want to prevent Diabetes, they should really keep their glucose level down and with increase in age they should follow a proper diet.

Sushant Ramesh, H. Balaji, N.Ch.S.N Iyengar1 and Ronnie D. Caytiles, "Optimal Predictive analytics of Pima Diabetics using Deep Learning" [5]. This deep neural network, coded on python, will help to obtain numeric results on the severity and the risk factor of the diabetics in the data set. At the end, a comparative study is done between the implementation of this model on type 1 diabetes mellitus, Pima Indians diabetes and the Rough set theory model. The comparison shows that the deep learning models are definitely more effective in terms of precision than the rough set theory model.

Rahul Joshi, Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach" [6]. In this system the most known predictive algorithms apply KNN, Naïve Bayes, Random forest, and J48. Single algorithm provided less accuracy than ensemble one.in most study decision tree provided high accuracy.in this study hybrid system Weka and java are the tools to predict diabetes dataset.

Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes using Classification Mining Techniques" [7]. This

paper aims at finding solutions to diagnose the disease by analysing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. Experimental results show the effectiveness of the proposed model. The performance of the techniques was investigated for the diabetes diagnosis problem.

Aishwarya. R, Gayathri. P and N. Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes" [8]. One of the promising techniques in machine learning is Support Vector Machine (SVM). SVM is used for classification of system. Upshot of SVM has provided with classification of system. The accuracy of the proposed system is good 4 while pre-processing, when compare with previous work which has been done without preprocessing the system. Pre-processing has played a key role in classification of diabetes.

Uswa Ali Zia, Dr. Naeem Khan," Predicting Diabetes in Medical Datasets Using Machine Learning Techniques"[9]. In this study a medical bioinformatics analyses has been accomplished to predict the diabetes. The WEKA software was employed as mining tool for diagnosing diabetes. In this study we aim to apply the bootstrapping resampling technique to enhance the accuracy and then applying Naïve Bayes, Decision Trees and k Nearest Neighbors (kNN) and compare their performance. The accuracy can be increase by improving the performance of the data, the algorithms or even by algorithm tuning.

Panigrahi Srikanth and Dharmaiah Deverapalli, "A Critical Study of Classification Algorithms Using Diabetes Diagnosis"[10]. In this paper Classification Algorithm Examine of the Decision Tree Algorithm, Byes Algorithm and Rule based Algorithm. Popular Classification Algorithms were considered for evaluating their classification algorithm and performance measurements can apply to calculate accurate results in classifying Diabetes pregnant patient's Pima dataset.

There are various existing solution for finding relations between the diseases, symptoms and medications, but these algorithms have their own limitations; numerous iterations, high computational time and the continuous arguments etc. Naïve Bayes overcomes various limitations and can be applied on a large dataset in real time. Several challenges are also highlighted which includes governance issues including ownership, security, privacy have however to be addressed. By overcoming the existing limitation as defined above will help in more fast progress in analyzing.

2.1 Problem Formulation

Before attempting to solve a problem, we need to first formulate or define the problem. It is important to precisely define the problem you intend to solve. Problem formulation is the act of a problem, determining the cause of the problem and, identifying the solution. People nowadays are having difficulty to seek the doctor or undergoes any medical checkup in order to get knows their body health due to increasing workload which lead to insufficient of time. This is the reason we develop diabetes detection system using mobile. This application can help reduce time between doctor and patient. Nowadays, health problems in our country are increasing rapidly especially diseases that related to blood disorders. There are many types of blood disorder diseases, such as diabetes, anemia, blood cholesterol, hemophilia, HIV/AIDS, leukemia, cancer and so on. Diabetes Mellitus affects nearly 400 million in worldwide. Hundreds of thousands of people are afflicted with this chronic disease. Thus, in order to identify their health condition, these systems have been developed.

The objectives of the proposed project are as follows:

- The objective is to raise awareness about the importance of diabetes as a global public health problem.
- To act as an advocate for the prevention and control of diabetes in vulnerable populations.
- To diagnose people with diabetes at an early stage based on food consumption.
- To diagnose patients the importance of lifestyle in the management of diabetes and the prevention of complications, especially the role of exercise, nutrition.

Serious action need to be taken by each individual in order to reduce the number of people that suffer with diabetes disease from the early stage. Besides that, if someone already knows that they have diabetes, their focus should be on preventing the complications, which can cause serious disabilities such as blindness, kidney failure requiring dialysis, amputation, or even death. Furthermore, having a healthy diet also required in order to prevent diabetes.

2.2 Requirements and Methodology

The hardware requirements for the proposed project are depicted in the table below:

Table 1: Hardware requirements

Sl. No	Hardware / Equipment	Specification
1	RAM	5GB (Minimum)
2	Processor	I5 processor or above
3	Windows	7 or above

The software requirements for the proposed project are depicted in the table below:

Table 2: Software requirements

Sl. No	Software	Specification
1	Operating system	Windows 10
2	Server	Jupyter Notebook

Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. It shows the flow of the research conducted in constructing a model.

2.2.1 Working Procedure

Step 1: Input

Load the file/dataset into the tool for pre-process step.

Step 2: Data Pre-processing

Select the classifier, in that choose algorithms.

Step 3: Segmentation

In test options, click percentage split options and also give the percentage split option and also give the percentage for splitting the data set into training set and test set. Example: Training set=80% and Test set=20%

Step 4: Output

After completing all the above steps, click start. The results are displayed in the screen. It shows the correctly classified instances, incorrectly classified instances, prediction on test set, total number of instances and confusion matrix.

The training set used is the dataset which is a pre-loaded dataset on the tool. An existing training dataset can be accessed through the tool. Various limitations were placed while choosing the instances from a much larger dataset. The device is utilized for information pre-handling. As a first step the dataset must be pre-processed as the data obtained might be incomplete, noisy and dirty. The data might lack attributes, have errors and outliers and could be inconsistent. The process of automatically selecting only those features of the given data that contributes the most to the prediction variable or output in which we are interested in, is known as Feature Selection. This step applies the various supervised classification algorithms namely, Naïve Bayes, and Support Vector Machine (SVM), on the training dataset that is obtained after following the first to third steps. This is the final step of comparing and analysing the accuracy measures and performance on the training dataset by the machine learning algorithms.

2.3 System Design

System design is a one important phase in software or system development. System design can be defined as method of defining different modules required for software or system to fulfil all requirements.

System design shows the overall design of the proposed diabetes prediction model, which is depicted in Figure 1 shown below.

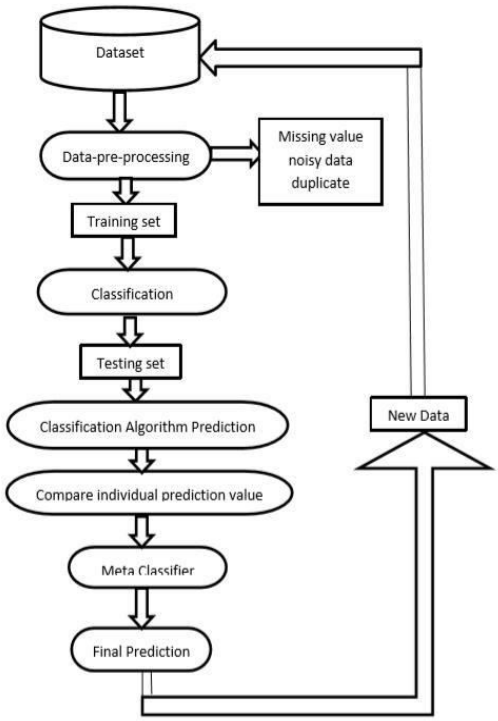

Figure 1: System Design

2.3.1 Phases in Diabetes Prediction

The diabetes prediction system is trained using unsupervised learning approach in which it takes dataset of different food consumption. The system includes the training and testing phase followed by preprocessing data, classification algorithm prediction, comparing individual prediction values, and Meta classifier. Phases for implementing the Proposed Project are:

1. ML model comparison for PIMA dataset.
2. Django backend with a bootstrap form in the front end to test input data and to see the performance of the model
3. Daily Calorie/ Sugar recommender for people based on their Age, BMI and giving a management of their daily calorie/ sugar needs based on the food that they eat

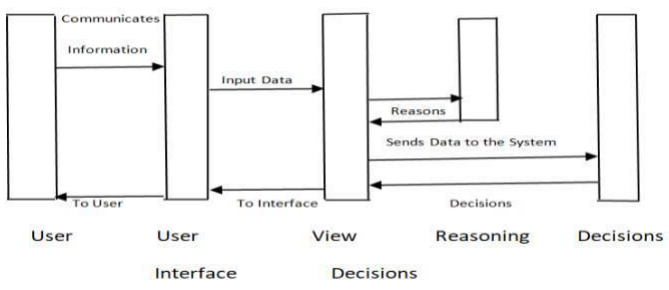

Figure 2: Sequence diagram

2.4 System Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner.

Software testing is the process of checking whether the developed system is working according to the original objectives and requirements. Software testing process commences once the program is created and the documentation and related data structures are designed. Software testing is essential for correcting errors. Otherwise the project is not aid to be complete.

2.4.1 Unit Testing

Unit testing is a level of software testing that involves individually testing unit of code to ensure that it works on its own, independent of the other units. The key purpose is to validate that every single unit of the software performs as perfectly designed. System will get input from user through user interface. After getting input the system will process dataset, then the output of preprocess module will give as input for segmentation process, after segmentation the system will do feature extraction process, the output of feature extraction process is fed as input for classification and meta classifier where classification of dataset using Classification Algorithm will happen and finally accurate result will be given as output in the editable form through UI for user.

2.4.2 Component Testing

Component testing is defined as a software testing type, in which the testing is performed on each individual component separately without integrating with other components. System will get input from user through user interface. Predicting accuracy is the main evaluation parameter that we used in this work. Accuracy can be defied using equation. Accuracy is the overall success rate of the algorithm.

$$\text{Accuracy} = (TP+TN) / (P + N) \quad (1)$$

All predicted true positive and true negative divided by all positive and negative. True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) predicted by all algorithms.

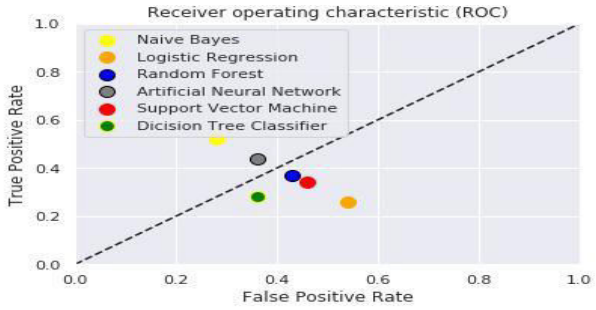

Figure 3: Comparison graph

2.4.3 Integrated Testing

Component testing is defined as a software testing type, in which the testing is performed on each individual component separately without integrating with other components. System will get input from user through user interface.

In this testing we have combined the loss functions used during finding out the performance of each machine learning model by giving it variety of data sets with distinct ranging values. We have also tested the endpoints for the Django web app and written unit tests specific to the web app. We have also written unit test for the APIs used in the sugar/ calorie management web application.

2.5 Result Analysis

The aim of this project is to detect the calories consumed based on their food record. In early stage system was trained using random algorithms based on the dataset by various classification algorithms. Data set as partitioned into training and testing. Dataset consists of different samples which are selected randomly from the research analysis. Based on our personal information number of calories intake per day is recommended. We can add the food that is already intake and can also include food that e are interested to consume. This information calculates the total calories and represents the result in the form of pie chart. The most ideal algorithm is SVM and Random forest with Accuracy of 82% and 85% respectively.

Home page where the project application is started. Once we click on the application it proceeds to next page where we need to enter the details. This application is not user specific. Anybody can use this application and check the result. Since the data is not stored there is no need for the user to login

Figure 4: Input Data

Since the system is trained with more number of dataset the accuracy will be more. In this page we are entering the information for further analysis.

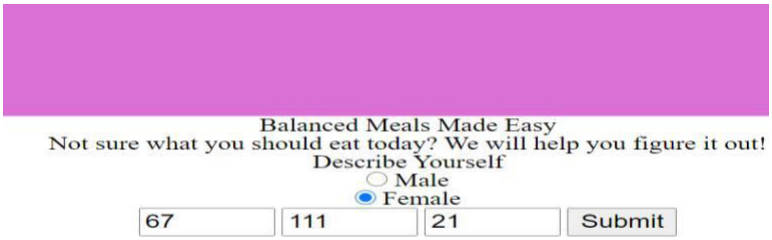

Figure 5: Example Input

In this application the height and weight is entered in terms of inches and pounds respectively. The procedure used to calculate the total calories depends on height and weight in inches and pounds only.

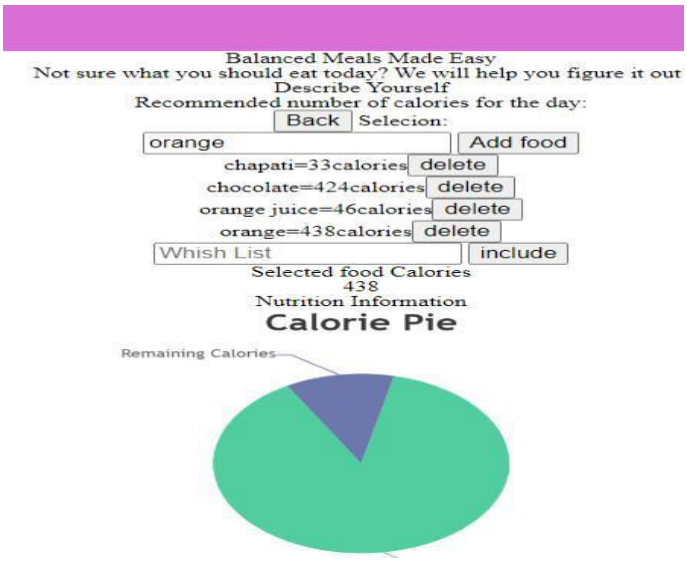

Figure 6: Result

This is the result of above example. As we can see the recommended number of calories per day is 1309.64485. We can add and delete food and even add food that we like to eat. Including food is optional. And the result is displayed in the form of pie chart. Communication is the key factor of living in this world. Diabetes is an uprising illness, particularly because of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes is fundamentally caused because of obesity or high glucose level, and so forth. So our project aims on controlling diabetes in terms of food and also creates awareness about it. Likewise we will determine whether the person intake calories are safe or not.

3. CONCLUSIONS

3.1 Conclusion

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. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status. This project focus on developing an application that detects diabetes in its early stage based on the food consumption and prevent from being diabetic

3.2 Scope for Future Work

The dataset analysed in this study was based on some main food. This study can be conducted with a larger dataset sample in rural and urban community settings in multiple states across India. This research study has only targeted on avoiding diabetes based on food and is applicable to all the users.

Various other key features in the medical records can also be analyzed. It will be interesting to perform a more exhaustive exploration of additional features in the dataset and study their relevance.

Living with diabetes is challenging and distressful. Diabetic patient's condition cannot be understood only from consumption of food chart. There is a need to collect and analyse both subjective and objective patient information I order to fully understand. Subjective data can be captured by interviewing patients or by conducting surveys which will enrich the depth of patient information. The conversation between doctor and patient can also be collected and analysed which could help to extract important features.

REFERENCES

1. 1. Analyzing Feature Importance for Diabetes Prediction Using Machine learning.
2. A Critical Study of Classification Algorithms using Diabetes Diagnosis.
3. Prediction and diagnosis of diabetes-A machine learning Approach.
4. A Critical study of Classification algorithms.
5. Optimal Predictive analytics of Pima Diabetics using Deep Learning.
6. Analysis and prediction of diabetes diseases using machine learning algorithm Ensemble approach.
7. Diagnosis of Diabetes using Classification Mining Techniques.
8. A Method for Classification Using Machine Learning Technique for Diabetes.
9. Predicting Diabetes in Medical Datasets Using Machine Learning Techniques

10. A Critical Study of Classification Algorithms Using Diabetes Diagnosis
11. Diabetes Prediction using Machine learning: https://youtu.be/HTN6rccMu1k
12. Machine learning with diabetes: https://youtu.be/QAUIzP4OR6Y
13. Prediction Analysis of Diabetes patients: https://youtu.be/H76_jbeCHPQ
14. "Canopy | Scientific Python Packages & Analysis Environment | Enthought." [Online]. Available: https://www.enthought.com/product/canopy
15. National Diabetes Information Clearinghouse (NDIC), http://diabetes.niddk.nih.gov/dm/pubs/type1and2/#signs