ISSN 2395-1621

Diabetes Detection and Prediction System using Machine Learning Algorithms

Prof. D. S. Zingade, Prajakta Apte, Vedika Chandel, Vijayalaxmi Bhambure, Aditi Sutar

deeplakshmisach@gmail.com,
apteprajakta@gmail.com,
aditisutar1398@gmail.com,
vedikachandel1398@gmail.com,
bhamburevijayalaxmi@gmail.com

Computer Engineering Department, AISSMS Institute of Information Technology, Pune, India

ABSTRACT

Over 250 million people worldwide with a majority of them being women are being affected due to diabetes. Diabetes is one of the deadliest diseases recorded by United States. The condition is defined by metabolic abnormalities and by long-run issues affecting nerves, eyes, blood vessels and kidneys. With the enhancement of knowledge and technologies, desktop application shall be evolved for patient self-management and determine the chance of being affected by diabetes. Diabetes could be a chronic and manner sickness and immeasurable individuals from everywhere the globe fall victim to the disease. Although there are some mobile apps keeping track of calories, lifestyle, sugar, pressure level medicine doses, blood sugar, weight of people and giving suggestion about exercises and food to manage their health. No such application exist that can predict the risk level of diabetic patient. Thus, the aim of this project is to develop a system which detect the early stage of diabetes which is supported by machine learning to check his/her probability of being diabetic, prediabetic or nondiabetic without the help of any doctor.

Keywords— CSV Dataset, Support Vector Machine, Decision Tree.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

Blood glucose is main input to body and it comes from the food we eat in day to day. Insulin is one of the main hormones of the body which is made by the pancreas that helps in making the glucose from food. The glucose is absorbed by cells to be used for energy. When the level of glucose is too high in the blood it leads to many health problems. And diabetes is one of the serious health issue cause due to high glucose level. Diabetes is not only one single disease it leads to many other diseases like heart disease, kidney disease and blindness etc. High level of diabetes may lead to death. There are different types of diabetes based on the reason of their cause. The diabetes can be categorized into

1. Diabetes Mellites
    Type 1 Diabetes
    Type 2 diabetes
2. Gestational Diabetes

Type 1 diabetes is autoimmune condition. This means that the body's immune system is permanently destroys the cells in the pancreas which produces the insulin hormones [7]. The damage cause by type 1 diabetes is permanent. This is also called as juvenile diabetes and occurs due to genetic disorders [2].

Type 2 diabetes is most common form of diabetes which occurs due to high sugar level in the body for longer duration. Mostly adults over the age of 40 years are being diagnosed to this type of diabetes but today children are also suffering from type 2 diabetes. Type 2 diabetes is caused due to obesity and lack of exercise [8].

Gestational diabetes -It is type of diabetes which occurs in pregnant women. It is high blood sugar (glucose) that develops during pregnancy and usually gets vanishes after giving birth. It can happen at any stage of pregnancy; it occurs because the body cannot make enough insulin that can meet extra need of the body. Even women having gestational diabetes can leads to type 2 diabetes in later stage of life.

II. ALGORITHMS

A. SVM - It is supervised learning technique, where classifier is build based on training dataset which contains the target attribute and the output of testing dataset is predicted by classifier. Each individual in dataset is represented as a point in space for graph plotting. If the dataset is linearly separable SVC (Support Vector Classifier) plots the hyperplane and divides the dataset into two classes. Many hyperplanes can be plotted but the one with maximum margin is selected for effective classification. For non-linear dataset, kernel function is used. There are 3 types of kernels- linear, polynomial and gaussian [3]. Kernel function is a function where non-linear separable data of lower dimension is given as input which converts it into linear separable data of higher dimension and returns linear separable data as output.

B. Decision Tree - Decision tree algorithm falls under the category of supervised learning. DT uses the tree representation to solve the problem in which the leaf node represents a class label and attributes are represented on the internal node of the tree. In this algorithm the major challenge is to identification of the attribute for the root. To solve this problem, there are two popular attribute selection-Information gain and Gini index.

This paper deals with the early detection of diabetes with help of machine learning algorithm such support vector machine and decision tree. We are using pima Indian dataset for the prediction of the diabetes and on that dataset, we perform various steps to get accuracy of the model. Then we compare the accuracy level of both the algorithm SVM and DT to get the best results.

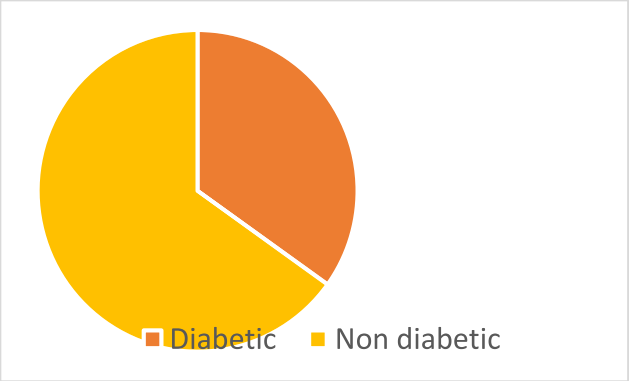

Fig1. Proportion of diabetic patients

III. LITERATURE SURVEY

1. **Lejla Alic et.al.**[1] In this paper CGMS (Continuous Glucose Monitoring System) is used for measuring the level of glucose in patient and with the help of machine learning techniques like Non-linear auto regressive Neural Network prediction of diabetes is done.
2. **Alessandro Aliberti et.al.**[2] Paper describes the use of ML technologies like LSTM (long short-term memory networks) and NAR (non-linear autoregressive neural networks) to find the solution for glucose level and these results were compared three literature approaches those are FNN (feed-forward neural networks), AR (autoregressive models). NAR gave good results for short-term predictions i.e. prediction horizon is 30 minutes while LSTM gave good performance for both short-term and long-term predictions.
3. **Debadri Dutta et.al.**[3] Paper compares three algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest and from results it concludes that Random forest predicts with highest accuracy i.e. 84%, for that detail study of dataset is done and important features for Random Forest are selected.
4. **Deepti Sisodia et.al.**[4] WEKA tool is used for diabetes prediction and precision, recall, f-measure, accuracy, ROC area for multiple machine learning algorithms like Naïve Bayes, SVM, Decision Tree are compared, among which Naïve Bayes shows better results.
5. **Nabila Shahnaz Khan et.al.**[5] The objective of this paper is to develop an intelligent mHealth application based on machine learning to assess his/her possibility of being diabetic, prediabetic or nondiabetic without the assistance of any doctor or medical tests.
6. **M. Alghamdi et.al.**[6] The henry ford exercise testing project,2017 is used. In this study, the relative performance of various machine learning methods such as Decision Tree, Naive Bayes, Logistic Regression, Logistic Model Tree and Random Forests for predicting incident diabetes using medical records of cardio respiratory fitness are described. The metrics like kappa, recall, precision, specificity were used to compare the algorithms in which Random forest and NB tree model performed better.
7. **Preeti Verma et.al.**[7] This paper gives detailed information about various papers for diabetes detection using machine learning and data mining techniques. According to the results modified spline SSVM gives promising results.

IV. METHODOLOGY

This paper proposes methods to predict and monitor the effects of diabetes, a polygenic disease in users based on their recent pathological reports and past medical history. The proposed system uses advanced machine learning models SVM and Decision Trees to predict diabetes based on inputs submitted by the user. The input data includes blood sugar levels, blood pressure, levels of insulin, age and BMI. To promote usability, a web-based application will be developed that can accept user inputs and further provide them to the model which are based on SVM and decision trees. The dataset used to train the model consists of multiple records with some values being null and duplicate records which can be removed to avoid errors or poor accuracy. The system will monitor sugar levels in blood on monthly basis and based upon prediction the application would provide suggestions. To interact with the model, a web-based application would be developed using python programming language and for storage of data we would be using MySQL. The application would be used to monitor all 3 types of diabetes by monitoring users blood sugar levels and would display precautions and tips with respect to diabetes, overall health and diet restrictions. The application would have in-built security measures to protect user data, also based on users selected options the health analytics report can further be forwarded directly to registered doctors saved in the application by the users.

SVM uses

$$p.q = p_1q_1 + p_2q_2 = \sum_{n=1}^{2}(p_nq_n)$$

Decision tree formulae:

$$Information\ Gain\ I(a,b) = \frac{a}{a+b}\log_2\frac{a}{a+b} - \frac{b}{a+b}\log_2\frac{b}{a+b}$$

$$\mathrm{Entropy\ E(A)} = \sum_{m=1}^{n}\frac{a_m + b_m}{a+b}I(a_m, b_m)$$

$$Gain = I_t(a,b) - E(A)$$

V. SYSTEM ARCHITECTURE

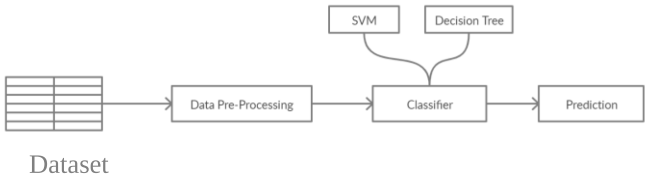

Fig2. General overview of system

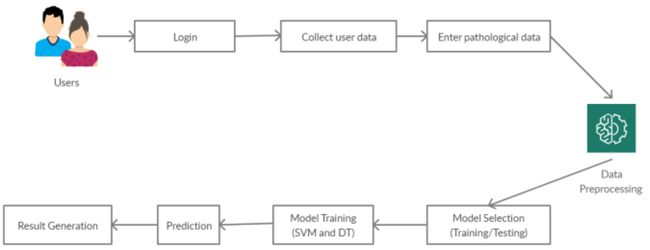

Fig3. Block Diagram

In the proposed system, PIMA Indian data set is being used containing 8 attributes [3] having 768 entries of women in Arizona in which there are positive (1) and negative (0) entries. The dataset emerges from National Institute of Diabetes and Digestive and Kidney diseases and all the records have attribute type as real and integer. The application provides a secure login feature utilizing customer credentials as well as collect basic information such as name, age, gender, date of birth etc. Upon the receiving the pathological reports user feeds the data such as age, insulin, blood sugar levels, diabetic pedigree function, etc into the application.

Feature name	Importance
No. of times pregnant	0.06
Glucose	0.65
Skin thickness	0.00
Insulin	0.02
BMI	0.12
Diabetes pedigree function	0.08
Age	0.11
Blood Pressure	0.00

Table1: Feature Importance (Decision Tree)

After data is being loaded, pre-processing steps are applied where in null valued attributes are removed; null values present in attributes can be eliminated with the help of many methods like null value removal, replacing null values with the mean of all attributes. In the system we are using method of removing null values. Also, least important attributes are removed. Dataset is divided into 70% training set and 30% test set to train the classifier for classification. Classifier used are SVM and DT. In the proposed model Pima Indian dataset is given as input to SVC and decision tree. SVC internally plots the points on space and draws the hyperplane and separates the data into diabetic and non-diabetic with the help of python libraries. Decision tree calculates the entropy and information gain of each attribute and selects an attribute having maximum gain as a root node in first step. This process repeats and internally plots the tree and separates data as diabetic and non-diabetic. Lastly, the report is generated which is mailed to user.

VI. EXPERIMENTAL RESULTS

As Support vector machine algorithm is performed on pima Indian diabetes dataset, following results are achieved-

	Precision	Recall	F1-score	Support
Non-diabetic (0)	0.76	0.92	0.83	99
Diabetic (1)	0.77	0.49	0.60	55

Table 2: SVM Results

Accuracy of Support vector machine algorithm is 76.62%

For Decision tree algorithm probabilistic results are,

	Precision	Recall	F1-score	Support
Non-diabetic (0)	0.70	0.85	0.76	89
Diabetic (1)	0.69	0.52	0.65	45

Table 3: Decision Tree Probabilistic Results

VII.CONCLUSION

In the projected work, a non-intrusive system has been developed to check the level of diabetic. The order of diabetic sick is detected with the help of SVM classifier and DT classifier and accuracy of both the algorithms are compared, this comparison will tell which of the algorithm is better in terms of accuracy. The above mentioned two algorithms compare the classifier accuracy on the basis of correctly classified instances and find out time taken to build the model.

REFERENCES

[1] Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad Abdul Ghani, Khalid Qaraqe, "Predicating Diabetes in Healthy Population through Machine Learning", IEEE Computer based Medical Systems (CBMS), 2019.

[2] Alessandro Aliberti, Irene Pupillo, Stefano Terna, Enrico Macii, Santa Di Cataldo, "A Multi-Patient Data-Driven Approach to Blood Glucose Prediction", IEEE Access, Volume 7, May 2019.

[3] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analysing Feature Importance for Diabetes Prediction using Machine Learning", 2018.

[4] Deepti Sisodia, Dilip Singh Sisodia, "Prediction of diabetes using classification algorithms", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 2018.

[5] Nabila Shahnaz Khan, Mehedi Hasan Muaz, Anusha Kabir, Muhammad Nazrul Islam, "Diabetes Predicting mHealth Application Using Machine Learning", IEEE 2017.

[6] M. Alghamdi, M. Al Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using smote and ensemble machine learning approach the henry ford exercise testing project", 2017

[7] Preeti Verma, Inderpreet Kaur, Jaspreet Kaur, "Review of diabetes detection by Machine Learning and Data Mining", IJARIIT, 2016.

[8] Bum Ju Lee, Jong Yeol Kim, "Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning", IEEE Journal of Biomedical and Health Informatics, vol.20, No.1, January 2016.

[9] C. Lorenzo, K. Williams, K. J. Hunt, and S. M. Haffner, "Trend in the Prevalence of the Metabolic Syndrome and Its Impact on Cardiovascular Disease", Incidence: The San Antonio Heart Study, Diabetes Care, vol. 29, no. 3, pp. 625{630}, Mar. 2006.

[10] O. Tschritter, A. Fritsche, F. Shirkavand, F. Machicao, H. Haring, and M. Stumvoll, "Assessing the Shape of the Glucose Curve During an Oral Glucose Tolerance Test", Diabetes Care, vol. 26, no. 4, pp. 1026 1033, Apr. 2003.

[11] N. Unwin, J. Shaw, P. Zimmet, and K. G. M. M. Alberti, "Impaired glucose tolerance and impaired fasting glycaemia: the current status on definition and intervention", Diabetic Medicine, vol. 19, no. 9, pp. 708 723, Sep. 2002.