Machine Learning Based Model for Detecting Progression of Diabetes on Multiple Daily Readings

Ibrahim Usmani, S.M.Ashar Rizvi, Sania Uzair, Mustafa Haider, Dr.Saad Qasim

---

Abstract - *Diabetes is among one of the most common and critical diseases and lots of people suffer from this disease yearly. Various factors such as weight, inherited diabetes, lack of exercise and eating habits results in Diabetes. If diabetes is undiagnosed and its progression is not maintained it can result in fatal compromise of vital organs of human body. Therefore timely detection of diabetes is very important for diabetes patients. Current practice in hospitals is to collect required information through a series of test and then diagnose the disease and proposed a treatment. Machine learning can be used to study the hospitals dataset volumes and can play a significant part in healthcare industry. Machine learning is widely used for diabetes detection and diagnosis however it can also be used for monitoring the progression of diabetes. In this paper we have proposed a machine learning model for detecting diabetes and predicting the progression of diabetes in patients through multiple daily glucose level reading of patients .The parameter of datasets include age, gender, fasting reading, pre meal reading, and post meal reading. A parameter of HBA1C reading is included for prediction of prediabetes. Medication, Exercise and Carbohydrates parameters assist in predicting the progression of diabetes patients.*

Key Words: Diabetes, HBA1C, Diabetes progression, Machine learning, Random Forest, Decision Tree

1. INTRODUCTION

Diabetes is a medical disorder in which blood sugar (glucose) is not metabolized by the body. There are different type of Diabetes. In type 1 diabetes glucose level in the blood is raised to alarming level in this condition body is not able to produce insulin hence injection of insulin is required in it. In type 2 diabetes human body cells are not able to use insulin properly. Type 3 diabetes usually occur because of increase blood level in pregnant women. Diabetes is an incurable disorder. A diabetic person can develop severe complications like heart attack, kidney failure, nerve damage and stroking. Therefore it's very important to control the diabetes in order to avoid severe complications. Maintaining daily readings of glucose level is very crucial for patients. Progression of diabetes in patients helps doctors in treatment and assist them in recommending the medication.

Machine learning is an important application applied in the healthcare system. Machine learning is getting greater attention in disease detection and prediction. In disease diagnosis it can help the physician to build a system that helps in detection of diseases and predicting the progression. Different models are proposed by the researchers for training that helps in diabetes diagnosis but progression of diabetes through multiple daily readings is unexplored.

In this study we used different machine learning algorithms to train models that not only diagnose diabetes but also projects the progression of diabetes. The paper focus on utilizing different classification algorithms for model building of diabetes diagnosis and predicting the progression of diabetes through multiple daily readings. All the models are trained on datasets that include parameters a physician requires to conduct the test.

Section 2 of this paper has brief discussion on experiment and step take for model building and dataset collection. Section 3 contains the result and section 4 comprises of related work. Finally in section 5 discussion is done on the research and models built and concluding note is presented in section 6 of the paper.

2 EXPERIMENT

The main goal is to implement different classification algorithms and train them on data repository. The input data to the model and steps taken are discussed below in the subsection respectively

2. I Dataset collection:

We have utilized two datasets during our series of experiments .The first dataset repository was used for diabetes diagnosis. This dataset was acquired from UCI Machine Learning Repository. The data available was not adequate and dataset was append manually. The dataset was studied and helped in understanding the pattern and trends which helped in predicting the results and disease diagnosis. For prediction of progression of diabetes required dataset was not available on online forums therefore dataset was generated manually under the guidance of doctors. Multiple glucose level reading (fasting) was

monitored and appended in data repository. The data was approved by the doctors.

Dataset parameters used in diabetes diagnosis model are listed in table 1.1

Age
Gender
Weight
BMI
HBA1c
Glucose level Fasting
Glucose level Pre-meal
Glucose level Post-meal

Table No. 1.1

Dataset parameters used in predicting the progression of diabetes are listen in table 1.2

Age
Gender
Glucose level Fasting(5 days)
Daily Exercise
Medication
Carbohydrates Intake

Table No. 1.2

HBA1C reading assess the last three month glucose level of patients. This reading is integral for detecting the phase of diabetes in patient. Standard reading of HBA1C test are as follows.

Normal < 5.7%
Pre diabetes 5.7% -- 6.4%
Diabetes > 6.5%

Glucose levels are maintained on daily bases and they helped in detecting the progression of diabetes. Standard reading of glucose levels are as follows

Fasting 100 normal
Pre meal 110 normal
Post meal 130 normal

The dataset has record of multiple reading within the standard range and deviated values and the models are trained on the dataset. All the models are trained to get accurate results and later the accuracy of models is tested and compared.

2. II Dataset preprocessing:

In this step we deal with handling of inconsistent data to get more accurate results. The dataset acquired from UCI repository doesn't have any missing value however data cleaning was required as some reading of glucose level, BMI had zero value entry. Data cleaning was performed by deleting the rows of inconsistent data and then refined data was trained.

2. III Model building:

This part of experiment includes model building and training for detection and predicting the progression of diabetes. In the first step different machine learning models are implemented for detecting diabetes, pre-diabetes and normal conditions in patients. In the UCI data repository there are 167, 202 and 399 entries for diabetes, prediabetes and no diabetes. The 67 % of data repository is used for training and 33% of data repository is used for testing. Nearly 500 sample is used for training of diabetes, prediabetes and no diabetes and 268 sample is used for testing.

In the later stage dataset generated manually was trained for prediction of diabetes progression. In the dataset there were 35 entries. The 65% of dataset is used for training and the remaining 35% was used for testing. Python language is utilized for machine learning model building.

Mostly machine learning study done by the researchers for disease diagnosis such as diabetes revolves around utilizing different classification algorithms therefore in our experiment for training a machine learning model that diagnose and predicts the progression of diabetes we have opted a similar approach and used the following algorithms. All the models for diabetes diagnosis were trained using these algorithms and then comparative study was performed to find out the best accurate model. For predicting progression of diabetes we have opted decision tree algorithm. Following is the precise description about each algorithm working used in this research.

K-nearest neighbors (KNN)

KNN is a simple machine learning algorithms based on supervised learning. This algorithms divides the dataset into classes and new cases are classified on a similarity measure. If new data wants to be classified neighbor element is located which is based on the majority number of votes for the label. The data is initialize and distance between classes and neighbor locating and voting for labels is calculated.

Support vector machine (SVM):

SVM is a machine learning algorithm that is used for both classification and regression problem. The goal of this algorithm is to create decision boundary that can segregate n-dimensional space into classes so new data points are put in the correct category. In SVM there are different hyper planes that divides the data and we have to find the hype planes that divides the class better. For this we calculate the distance between planes and data which is called Margin. Classes with high margin are selected as they proposed lower chance of misconceptions.

Decision Tree:

Decision tree is a machine learning algorithm. It's a tree structured classifier in which all the data is represented

inside the shape of tree. The internal nodes represents the attributes of dataset, branches represents the decision rules and leaf nodes represents the output. Decision nodes are used to make decisions. The fundamental goal is to identify the foundation node in every step.

Random Forest:
Random forest is a popular machine learning algorithm that belongs to the supervised learning technique. It's applicable to classification and regression problem solving. Random forest contains number of decision trees on various subsets of dataset and consider the average to improve the predictive accuracy of the dataset. Output is predicted on majority votes which is calculated from the prediction of each tree. Number of trees structured present in data is directly proportional to accuracy of the result.

3. RESULT

In this section of experiment the results of the experiments are evaluated. The performance of classification algorithms KNN, SVC, Random forest and Decision Tree for diabetes diagnosis is analyzed using performance metric such as accuracy and their results are compared.
Table 3.1 shows the algorithms which gives the best accuracy on dataset used for diabetes diagnosis model

Algorithms	Accuracy (%)
K-nearest neighbors (KNN)	72.83
Support vector machine (SVM)	62.43
Decision Tree:	97
Random Forest	98.09

Table No. 3.1 Accuracy Result

Random forest gives the best accuracy for diabetes diagnosis model. Confusion matrix of random forest is given in table 3.2

	Diabetes	Normal	Pre-Diabetes
Diabetes	48	0	6
Normal	0	123	0
Pre-Diabetes	0	0	65

Table No. 3.2 Confusion Matrix of Random Forest

The model built for predicting progression of diabetes was trained on decision tree which gives the accuracy of 87%.The confusion matrix for decision tree is given in table 3.3

	Decrease	Increase	Maintain
Decrease	1	0	0
Increase	0	7	1
Maintain	0	0	1

Table 3.3 Confusion Matrix of Decision Tree

4. RELATED WORK

Disease diagnosis using machine learning technique is done by the researchers to help the health professionals A lot of different machine learning algorithms using classification study are being used for diabetes prediction and diagnosis. A paper written by Naveen Kishore in international journal of scientific and technology research compares accuracy of different classification algorithms for diabetes prediction.[1] Large data repository of health care industry is used by researchers to build different machine learning models for diabetes prediction and diagnosis .Various classification and regression models were trained on huge hospitals dataset. In International Conference on Recent Trends in Advanced Computing published a research paper on diabetes prediction using machine learning algorithms [2]. Lot of work has recently been published mainly in anomaly detection in the area of healthcare. Related to diabetes detection [3]. A random forest algorithm is employed for prediction of diabetes. The author used the UCI diabetes data repository. An early prediction of diabetes is important for increasing the survival rate and providing a timely treatment to the diabetes patient [4].Based on these learning we develop a machine learning model for diabetes prediction and detection. We had tried to explore the techniques of machine learning and developed a model for progression of diabetes using multiple reading of patients and predict their progression.

5. DISCUSSION

We have shown that using different machine learning classification algorithms we cannot only perform disease diagnosis such as diabetes but by exploring techniques of machine learning we can also build a model that can perform the prediction of diabetes. Different parameters of dataset plays integral part in

diabetes diagnosis such as HBA1C which shows the last three months of glucose level and greatly help in model training for detecting prediabetes. Similarly glucose level reading of different timespan also helps in predicting the results. If the glucose levels are greatly deviated from standard ranges that can be an indication of diabetes. Random forest showed the highest accuracy shown in table 3.1. We have also tried to build a basic machine learning model for projecting the progression of diabetes using multiple daily glucose level reading (fasting) of diabetes patients and predict the trend based on increase, decrease or maintained output. For this model we have used decision tree Confusion matrix of decision tree is shown in table 3.3. The parameters of medication, exercise and carb intake also helped in training of the model.

6. CONCLUSION

In this paper two machine learning models are proposed using classification algorithms for diabetes diagnosis and progression monitoring. The first model that was for diabetes diagnosis showed highest accuracy on random forest 98.05%. The other model for predicting the diabetes progression was trained on very minimal amount of data however the model was trained successfully and showed the progression of diabetes as increase, decrease or maintained output. The advantage of the models is that dataset can be explored and new parameters can be added for training. Moreover the dataset that was generated manually can also be further append easily by adding the readings. All the parameters of dataset can be monitored and it's practical to generate the dataset without much hustle. The model was trained on decision tree with an accuracy of 87%. More rigorous studies are needed to investigate techniques for the betterment the model thereby enhancing the accuracy

7. REFERENCES

[1] INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 01, JANUARY 2020 ISSN 2277-8616 1805 IJSTR©2020 www.ijstr.org Prediction Of Diabetes Using Machine Learning Classification Algorithms Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy

[2] INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019 Diabetes Prediction using Machine Learning Algorithms Aishwarya Mujumdara , Dr. Vaidehi Vb a Vellore Institute of Technology, Chennai, India b Mother Teresa Women's University, Kodaikanal, India

[3] G. Swapna, U. Rajendra Acharya, S. VinithaSree, J. S. Suri, Automated detection of diabetes using higher order spectral features extracted from heart rate signals, Intelligent Data Analysis 17 (2) (2013) 309–326.

[4] Talha Mahboob Alama,, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb , Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc , Muhammad Awais Malikb, Muhammad Mehdi Razab , Salman Ibrarb, Zunish Abbas, A model for early prediction of diabetes, Informatics in Medicine Unlocked, 2019

[5] Kumari, V.A., Chitra, R., 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA) www.ijera.com 3, 1797–1801.

[6] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., 2013. Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology Vol.3, 334– 337. doi:JUNE 2013, arXiv:ISSN 2277 – 4106

[7] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014

[8] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109

[9] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus, International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, IEEE, 2019.

[10] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.