Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

A Study on Various Machine Learning Classification Algorithms for Diabetes Prediction

Jiby T C
Department of Computer Applications
Cochin University of Science and Technology
Cochin, India

***Abstract*-Diabetes is a condition that the blood glucose level is high than normal level in the blood. Which is a lifelong disease and causes lots of health problems. To predict the diabetes which helps for the treatment to preclude diabetes and its related health problems. Early prediction of diabetes using various efficient machine learning methods can helps people to prevent the diabetes and to get the treatment earlier. This paper does a study on different machine learning classification algorithms for predicting diabetes. The different classification algorithms are k-Nearest Neighbor, Naive Bayes, Random Forest, Support Vector Machine, Decision Tree, etc are done on diabetes data set.**

Keywords: Type 2 Diabetes; Machine Learning Algorithms; kNN,; SVM; NB; RF; DT;

I. INTRODUCTION

Nowadays, diabetes is one of the most life-threatening disease in the world. It is becoming more common and millions of peoples are diabetic in India. Diabetes mellitus is a major health care problem that potentially outbreak proportions in India and its enormous complications may cause many healthcare issues on patients. Alarmingly, diabetes is a major health problem which causes lots of health problems even in the younger age. Diabetes is highly visible across all areas in India and mostly affecting in India large scale and lots of research is carried at different levels to predict the diabetes in earlier to decline the increase rate in diabetes for the coming years.

Diabetes, a long-term disease of sugar glucose in the blood is abnormally high due to the deficiency of insulin which helps to control the blood sugar levels of the body. Mainly two types of diabetes: Type1 and Type2 diabetes. Type1 diabetes is the severe condition that the body does not have the ability to produce insulin. Type2 diabetes is most commonly seen in whole of the world , it is insulin resistant or it defies insulin. Too much sugar in the blood may cause serious, sometimes life threatening health problems as it can mutilate patients eyes, kidneys, nerves and even cause heart disease stroke. Type-2 diabetes is more common, almost 90-95% have type-2 diabetes.

Diabetes Mellitus Type-2 is marked by the body resisting insulin as the body cells react differently to insulin than they normal would and this type of diabetes is commonly found in people with high BMI or those who lead an inert lifestyle[1]. The diabetes type 2 presents 90% of all diabetes cases and is characterized by chronic hyperglycaemia, and the body's inability to regulate blood sugar levels, which causes a too high glucose (sugar) level in the blood[2]. There are different factors which effect the development of type 2 diabetes. One of the factor is lifestyle behaviors commonly associated with urbanization. Most of the studies indicates that type 2 diabetes could be prevented through regular physical exercises and healthy diet. It develops more slowly, usually over a period of months or even years as the symptoms may appear very gradually, which can make detecting the signs more difficult and the ascendancy of type 2 diabetes mellitus is less in rural population and more in urban population[3].

Predicting diabetes earlier will helpful to control such disease and save life. So this in concern by taking various risk factors related to diabetes, to predict it in early stage is a challenge and too much research work is on going in this area using various machine learning algorithms to build models for early prediction of diabetes. These algorithms are very efficient in result predicting by constructing predictive models from the data sets collected from various diagnostic centres.

A lot of distinct researches is done on diabetic datasets using Pima Indian Diabetes data set and confer a review on various kinds of prediction done for diabetes patients and the corresponding machine learning algorithms and techniques used[2].

II. RELATED WORKS

Using various classification algorithms, there have been performed many studies in the area of diabetic prediction by using the available medical data set. This paper aims to review on various classification algorithms for predicting diabetes by selecting the recent published papers from various electronic databases.

A. SVM Classifier

SVM Classifier is a supervised learning algorithm for linear and non-linear data. Examine the given data and construct the function which can be used for depicting new data. SVM uses a hyperplane to separate data from two classes. [4] Framed a classifier using WEKA tool to predict diabetes based on classification algorithms such as Naive Bayes, Support Vector Machine and Random Forest and SVM gives 79.13 % accuracy in diabetic prediction than other methods. [5] Focused on performance analysis to predict diabetes using different algorithms such as SVM, KNN,J48 and Random Forest. Experiment results shows that SVM algorithm gives the high accuracy of 77.9% among other data mining techniques. In paper [6] “Prediction of Diabetes using Support Vector Machine” examine to muster a diabetes disease prediction system. This paper proposed to use SVM Classifier to predict diabetes and experimental result obtained 75.3 % accuracy. [7] build a model to predict type2 diabetes based on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

non-linear SVM in that the features were derived from the OGTT. It secured an accuracy of 96.80% for Type 2 diabetes prediction using SVM classifier. The proposed work [8] shows the study on predicting diabetes using classification techniques such as Decision tree, KNN and SVM using Pima Indian Diabetic Data Set. The experimental result proved 90.23% of accuracy in the prediction of diabetes and, it is better than other two methods.[9] Proposed a model to predict type2 diabetes, using different data mining techniques, and also performed a performance comparison of these algorithms. SVM algorithm secured an accuracy of 95% in prediction of Type -2 diabetes mellitus.

B. KNN Classifier

K-Nearest Neighbor is a supervised classification algorithm which used for, pattern recognition and statistical estimation and classifies the new data or case based on a similarity measure. It identifies the data points which separated in to different categories and it tries to predict the classification of a new sample. The proposed work [10] focused on diabetes prediction using K- Nearest Neighbor classifier and achieved 100% accuracy by feature reduction in D3. The paper [11] proposed a diabetes disease prediction system which uses Naive Bayes and K Nearest Neighbour to predict diabetes by taking v attributes such as age, pregnancy, BMI etc. Related to work [12] shows the study on predicting type2 diabetes using weighted KNN algorithm, and the experimental result is compared with other techniques and weighted KNN technique achieved with accuracy of 80.12%. [13] A comparative study on seven classification algorithms to analyze the performance of each algorithm in predicting diabetes mellitus and observed that KNN was given the accuracy of 99.0% than other algorithms.

C. Random Forest

Random Forest is a tree classifier consists of different decision trees of different sizes. It dispenses the class of dependent variable based on many trees. It is a random sampling of train data when building tree and finally selects the decision based on the outcome of majority of the decision trees. [14] Developed a model to predict diabetes using the following three, classifiers such as L R, S V M and R F, in these the Random Forest as the ideal algorithm in predicting diabetes with accuracy of 84%. In [15] used three classifiers are DT, RF and NN by using PCA and mRMR to predict diabetes mellitus and the all experiments shows accuracy of using mRMR gives good results than PCA and secured 80.80% accuracy in Random Forest classifier in predicting diabetes. The paper [16] exploits a random forest model for the diagnosis of diabetes using Pima Indian data set to predict the accuracy and result shows that Random forest classifier is more efficient in comparison to other method of machine learning techniques. Related work on Random Forest Classifier in diabetes prediction, the work [17] Developed an early prediction diabetes system for patients with a high accuracy by using R F algorithm. At work [2] conducted a analysis using machine learning and deep learning algorithms to predict diabetes, and overall results shows Random forest has secured 83.76 % accuracy in predicting diabetes. [19] Build a classifier model based on classification algorithms as KNN, S V M, R F, Decision tree and Logistic Regression using Pima Indian Data set to predict diabetes and Random Forest secured 75 % accuracy in diabetic prediction than other methods.

D. ANN Classifier

An Artificial Neural Network is a methodical information processing system. It is as similar as the human brain operates. ANN is an interconnection of the assembly of nodes with their structure using a directed link. Information about the input is associated to each connection link as weights. [20] Proposed a web-based model for predicting diabetes disease using various classification algorithms. From different algorithms ANN secured 82.35% accuracy using min-max scaling method on Pima Indian diabetes data set. [21] Proposed a model using Artificial Neural Network and this model is ideal for predicting the diabetes with 92% accuracy. Reference [22] proposed many popular machine learning techniques such as NN, RF, SVM etc and a comparative study had done on these methods, in that NN was given the best accuracy(80.4%) than any other techniques. [23] Proposed a model using three algorithms, in that the result have shown as ANN outperforms the other methods with a best accuracy of 75.7%. [24] Proposed a model for predicting diabetes disease using ANN and had done a comparison in accuracy of ANN model with other models on Pima Indian Diabetes data set and the experimental result secured 85.09% accuracy in ANN classifier to predict diabetes.

D. Naive Bayes Classifier

NB Classifier uses probability for classification process. It uses Bayes theorem, it predicts the occurrence of any events. [25] Proposed a model for predicting diabetes mellitus using three classification techniques such as NB, RF and NB- Tree. It is observed that Naive Bayes had shown with high accuracy 76.3% as compared to other techniques. [26] Proposed a model using Naive Bayes and Bayesian Network for predicting diabetes and concluded that this model shown accuracy of 99% over existing model. Work [27] designed a model for predicting diabetes at an early stage using three classification methods as Support Vector Machine, Decision Tree and Naive Bayes and the Naive Bayes classification was given 76.3% accuracy as compared to other algorithms. In work [28] used two Naive Bayes classifiers, MNB Classifier and GNB Classifier for predicting diabetes, in that Gaussian Naïve Bayes Classifier secured accuracy of 80.8% to predict the diabetes disease.

The table below shows the summarized results of the various classifies.

Table 1. Comparison of Classification Techniques

Reference	Techniques Used	Best Technique	Accuracy %
[2]	ANN	ANN	92%
[4]	SVM, NB, RF	SVM	73.5%
[5]	SVM, KNN, J48, RF	SVM	77.9%
[6]	SVM,	SVM	75.3%
[7]	SVM	SVM	96.80%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

[8]	KNN, DT, SVM	SVM	90.23%
[9]	KNN, SVM, LR, ANN	SVM	95%
[10]	KNN	KNN	100%
[11]	KNN, NB	KNN	
[12]	LR, NB, Weighted KNN, Bayes Network	Weighted KNN	80.12%
[13]	KNN, DT, SVM, RF, ANN, NB, LR	KNN	99.0%
[14]	LR, SVM, RF	RF	84%
[15]	DT, RF, NN	RF	80.80%
[16]	RF	RF	75%
[17]	DL, SVM, RF	RF	83.67%
[18]	KNN, LR, RF, DT, SVM	RF	75%
[19]	SVM, KNN, GNB, ANN	ANN	82.35%
[20]	NN, RF, SVM	NN	80.4%
[21]	ANN, RF, K-Means	ANN	75.7%
[22]	KNN, DB, NB, SVM, K-Means	ANN	85.09%
[23]	NB, RF, NB-Tree	NB	76.3%
[24]	NB, Bayesian Network		99%
[25]	SVM, DT, NB	NB	76.3%
[26]	Multinomial Naive Bayes(MNB), Gaussian Naive Bayes(GNB)	GNB	80.8%

III. CONCLUSION

Diabetes is a common and life threatening disease. This study performed a review on various machine learning classification techniques for predicting diabetes disease. SVM classifier is used by many researchers for predicting diabetes when compared with other classifiers. By summing up the study, it could be noted that in future more research works can be done on descriptive analysis of how to control glucose level and patient's mental, and physical problems caused due to diabetic and diabetic medicines using the other machine learning techniques.

REFERENCES

[1] Neha Prerna Tiggaa, and Shruti Garga, "Prediction of type2 diabetes using machine learning classification methods," Procedia Computer Science.167:706–716, 2020.

[2] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, Mirsat Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," 978-1-7281-3992, 2019.

[3] https://www.diabetes.co.uk/

[4] Ayman Mir, Sudhir N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare,"ICCUBEA, 2018.

[5] D. Jeevanandhini , E. Gokul Raj ,V. Dinesh Kumar, N. Sasipriyaa, "Prediction of type2 diabetes mellitus based on data mining," IJERT:2278-0181,Volume6,Issue 04, 2018.

[6] Harwinder Kaur, and Gurleen Kaur, "Prediction of diabetes using support vector machine," IJREAM :2454-9150 Vol-05, Issue-02, May 2019.

[7] Hasan T Abbas, Lejla Alic, Madhav Erraguntla, Jim X Ji, Muhammad Abdul-Ghani, Qammer H Abbasi, Marwa K Qaraqe, "Predicting Long-Term type 2 diabetes with Support Vector Machine using oral glucose tolerance test,"https://doi.org/10.1371/journal.pone.0219636,2019.

[8] Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema, "Diabetes Mellitus prediction using classificationtechniques," IJITEE:2278-3075, Volume-9 Issue-5, March 2020.

[9] Faranak Kazerouni, Azadeh Bayani,Farkhondeh Asadi, Leyla Saeidi, Nasrin Parvizi , and Zahra Mansoori, "Type2 Diabetes Mellitus prediction using data mining Algorithms based on the Long-noncoding RNAs Expression: A comparison of four data mining approaches," https://doi.org/10.1186/s12859-020-03719-8.

[10] Madhuri Panwar, Amit Acharyya, Rishad A. Shafik, Dwaipayan Biswas, "K-Nearest Neighbor based methodology for accurate diagnosis of diabetes mellitus," ISED:6, 978-1-5090-2541 2016.

[11] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil,"Diabetes disease prediction using data mining,," International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[12] Sreeja Vishaly Ma, Umamaheswari k, "Type 2 diabetic prediction using machine learning algorithm,"ASRJETS - Volume 45, No 1, pp 299-307,2017.

[13] Sara A. Aboalnaser, Hanan R. Almohammadi, "Comprehensive study of diabetes mellitus prediction using different classification algorithms," 978-1-7281-3021-2019.

[14] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh,"Analysing feature importances for diabetes prediction using machine earning.",978-1-5386-7266, 2018.

[15] Quan Zou, Kaiyang Qu , Yamei Luo , Dehui Yin , Ying Ju, and Hua Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques," doi:10.3389/fgene.2018.00515.

[16] Sofia Benbelkacem, Baghdad Atmani,"Random Forest for diabetes diagnosis". 978-1-5386-8125-1/19, 2019 .

[17] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.SofiaCaroline, "Random Forest Algorithm for the prediction of diabetes," 978-1-7281-1524-5, 2019.

[18] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, "Prediction Of Diabetes using machine learning classification algorithms," International Journal of Scientific & Technology Research Volume 9, Issue 01, 2020.

[19] Samrat Kumar Dey, Ashraf Hossain., Md. Mahbubur Rahman, "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm," ICCIT,978-1-5386-9242-2018.

[20] Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, Hemant Darbari, "Prediction of Diabetes using Artificial Neural Network approach," 10.1007/978-981-13-1642-5_59, ICoEVCI 2018.

[21] Protab Kumar Saha, Nazums Sakib Patway and Ifthakhar Ahmed, " A Widespread study of diabetes prediction using several machine learning techniques.", ICCIT -22 December 2019.

[22] Talha Mahboob Alama, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas,"A model for early prediction of diabetes.", https://doi.org/10.1016/j.imu.2019.100204.

[23] Nitesh Pradhan, Geeta Rania., Vijaypal Singh Dhaka., Ramesh Chandra Poonia,"Diabetes prediction using artificial neural network,"https://doi.org/10.1016/B978-0-12-819061-6.00014-8, 2020.

[24] B. Tamilvanan, Dr.V.Murali Bhaskaran,"An Experimental study of diabetes disease prediction system using classification techniques,"IOSR-Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278- 0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017).

[25] K.Priyadarshini, and Dr.I.Lakshmi, "Predictive analysis of diabetes using bayesian network and naive bayes techniques," ICACT 2018 ISSN: 2454-4248 Volume: 4 Issue: 2 71 – 74.

[26] Deepti Sisodia ,and Dilip Singh Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science 132 1578–1585.

[27] Krish Shah, Rajiv Punjabi, Priyanshi Shah, Dr Madhuri Rao, "Real time diabetes prediction using Naïve Bayes classifier on big data of healthcare," IRJET, e-ISSN: 2395-0056p-ISSN: 2395-0072, Volume: 07 Issue: May 2020.