Prediction of Diabetes using Support Vector Machine

Ajay Kumar Tiwari, Avadhesh Kumar Dixit, Piyush Rai
IET, Dr. RML Avadh University, Ayodhya

Keyword: Diabetes, SVM, KNN, Random Forest

Abstract: Diabetes is a very common disease in the world. If diabetes is detected in the early stage, it can be cured easily. Several machine learning techniques are available to predict diabetes in an earlier stage using a data set. This paper proposes support vector machine based methods to predict diabetes. This paper also provides the comparative analysis of Naive Bayes, SVM, KNN, Random Forest, Logistic Regression and Decision Tree to predict diabetes. In this paper the proposed SVM based approach achieved the accuracy 77.08% that is better in compare to other machine learning based approaches.

1 INTRODUCTION

Diabetes is a coarse disease in our society. In this type of disease, a person has high blood sugar, either insulin production inefficient or the body cell do not return correctly to insulin, or by both reason. In human body blood sugar level is controlled by insulin hormone released by pancreas. When due to any reason secretion of insulin hormone becomes irregular, blood sugar level also affected. In this way a person may be affected from diabetes. The patients affected from diabetes can be cured by regular exercise, and by adopting healthy lifestyle. To control blood sugar level some medicine may be given or insulin may be given explicitly. To know whether a person is affected from diabetes, some diagnosis is required. If we came to know about the disease in early stage, we may prevent this harmful disease.

2 RELATED WORKS

In literature there are various machine learning based approaches used for diabetes prediction. He, B., Shu, K., & Zhang, H. (2019, August) authors has been proposed a deep learning based method for the prediction of diabetes. Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). authors has been proposed a disease prediction model to provide an early prediction for T-2 diabetes based on individual's risk factors data.Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). authors has been proposedk-nearest Neighbour, Decision Trees, Random Forest, Naive Bayes, and AdaBoost and Multilayer Perceptron (MLP) method for the prediction of diabetes.Woldemichael, F. G., & Menaria, S. (2018, May). authors has been proposed to predict diabetes using data mining techniques.Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018) authors has been proposed a predictive analysis using different machine learning methods for the prediction of diabetes. Rout, M., & Kaur, A. (2020) authors has been proposed to predict of diabetes diseaseusing various researchof machine learning Techniques. Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017) authors has been proposed K-Nearest Neighbour and the Naïve Bayes for the prediction of diabetes. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., & Abbas, Z. (2019). authors has been proposed to predict diabetes using various machine learning algorithms to determine significant attribute selection for clustering, prediction, andassociation rule mining for diabetes. Ahmed, T. M. (2016) authors has been proposed a new predicted model base on patients HbA1c. Aljumah et.al, (2013) authors has been proposed SVM based approach by using the Dataset of disease in Saudi Arabia to observe obesity and predicts chances of diabetes in a person. Chen, Peihua, and Chuandi Pan (2018) have been proposed diabetes prediction model based on boosting algorithms. They performed non parametric testing using two algorithms, Adaboost and Logitboost on test data of 35669 individuals and got

Tiwari, A., Dixit, A. and Rai, P.
Prediction of Diabetes using Support Vector Machine.
DOI: 10.5220/0010563800003161
In *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering (ICACSE 2021)*, pages 119-122
ISBN: 978-989-758-544-9

area under characteristics curve 0.99. The authors Mercaldo, Francesco (2017) worked on the concept of classification. They used Pima Indians dataset and obtained precision value 0.770 and recall value 0.775.The authors Patil, Bankat M (2010) proposed a prediction Model which used simple K-means algorithm and C4.5 algorithm using Pima Indians diabetes data and achieved an accuracy of 93.5%.The authors Kavakiotis, Ioannis, et al (2017) have been proposed SVM based model and got the accuracy of 85%. The authors Kohli, Pahulpreet Singh (2018) have been proposed logistic regression on separate datasets of heart, breast cancer and diabetes and got accuracy of 80.77%.The authors Perveen, Sajida, et al.(2016) performed classification technique, decision tree J48 and achieved Area under ROC is 0.98.The authors Sisodia, Deepti (2018)used DT, SVM and NB classification methods on Pima Indians Diabetes datasets. NB gave accuracy of 76.30%.

3 MATERIAL AND METHODS

Here, in this paper the description of diabetes dataset, methodology to compare the performance of various machine learning models has been provided.

3.1 Data Set Description

In this paper the PIMA Indian diabetes dataset Thomas, Jensia, et al. (2019) was used which was taken from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database). It is made to predict diabetes in women more than 21 years of age. It contains eight attributes or input variables and one output variable. The attributes are as follows:

Pregnancies: It represents number of pregnancies of a woman. During pregnancy the glucose level of women may increase which is called gestational diabetes. If women got pregnant number of times the gestational diabetes may leads to diabetes mellitus.

Glucose: It represents glucose concentration in blood. If glucose concentration in blood increases than a certain value then it may cause diabetes.

Blood Pressure: It represents BP(diastolic in mm Hg). Higher diastolic blood pressure increases the risk of diabetes.

BMI: It represents body mass index (weight (kg)/height (m)^2). It determines the obesity of the patient. Hence it is an important metric to predict diabetes.

Skin Thickness: It represents skin thickness (mm). In case of varying ratio of muscle mass and fat mass BMI is not adequate parameter to assess obecity which may lead to diabetes. Hence triceps skinfold thickness plays an impotant role to predict the patient may be diabetic or not.

Insulin: It represents serum insulin(mu U/ml). It is 2 hour serum insulin which indicates that how the body of a person respond on taking food.

Diabetes Pedigree Function: It is a function which determines the probability of diabetes on the basis of diabetic family history of a person.

Age: It is generally observed that person having age greater than 60 years are m ore prone to diabetes

3.2 Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised learning techniques, which is used for classification as well as regression analysis. It is used to find a hyper plane in an N-dimensional space. SVM is based on discrimination. Support vectors represent data points that are closet to hyper plane.

4 RESULT AND COMPARATIVE ANALYSIS

Here, in this paper the result of proposed approach and comparative analysis of various Machine Learning Techniques is presented. In this paper the different machine learning techniques are used and results for all techniques are different because the working principle of each technique is different. The results are evaluated on the basis of accuracy. Here, by using the K-Nearest Neighbourachieved an accuracy of 71.48%,the Decision Table obtain accuracy of 73.04%, Random Forest achieved an accuracy of 75.13%, Naïve Bayes technique obtained accuracy of 75.78% ,Logistic Regression obtained an accuracy of 76.82% andsupport vector machine obtained better an accuracy of 77.08%. See (Table-1)and Figure-1.

Table-1. Accuracy for the prediction diabetes using Machine learning Techniques

Sl. No.	Techniques	Accuracy
1	K-Nearest Neighbour(KNN)	71.48%
2	Decision Table(DT)	73.04%
3	Random Forest(RF)	75.13%
4	Naïve Bayes	75.78%
5	Logistic Regression(LR)	76.82%
6	Support Vector Machine (SVM)	77.08%

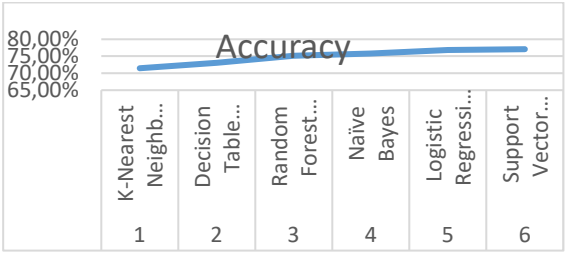

Figure 1: Comparison of accuracy of machine learning Techniques.

5 CONCLUSION

To predict the early stage of diabetes is one of the most challenging and important task. If diabetes is detected in an early stage, it can be cured easily. Machine learning methods have different power in different data set. Several machine learning techniques are available to predict diabetes in an earlier stage using data set. This paper proposed a support vector machine based methods to predict diabetes. This paper also provided the comparative analysis of Naive Bayes, SVM, KNN, Random Forest, Logistic Regression and Decision Tree to predict diabetes. In this paper the proposed SVM based approach achieved the accuracy 77.08% that is better in compare to other machine learning based approaches.

REFERENCES

He, B., Shu, K., & Zhang, H. (2019, August). Diabetes Diagnosis and Treatment Research Based on Machine Learning. In 2019 IEEE SmartWorld, Ubiquitous Intelligence &Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 675-679). IEEE.Hammoudeh, A., Al-Naymat, G., Ghannam, I., & Obied, N. (2018). Predicting hospital readmission among diabetics using deep learning. Procedia Computer Science, 141, 484-489.

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. IEEE Access, 7, 144777-144789.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

Woldemichael, F. G., & Menaria, S. (2018, May). Prediction of diabetes using data mining techniques. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 414-418). IEEE.

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning Algorithms in healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

Rout, M., & Kaur, A. (2020, June). Prediction of Diabetes Risk based on Machine Learning Techniques. In 2020 International Conference on Intelligent Engineering and Management (ICIEM) (pp. 246-251). IEEE.

Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017, March). Diabetes disease prediction using data mining. In 2017 international conference on innovations in information, embedded and communication systems (ICIIECS) (pp. 1-5). IEEE.

Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.

Ahmed, T. M. (2016). Using data mining to develop model for classifying diabetic patient control level based on historical medical records. Journal of Theoretical and Applied Information Technology, 87(2), 316.

Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad KhubebSiddiqui. "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences 25.2 (2013): 127-136.

Chen, Peihua, and Chuandi Pan. "Diabetes classification model based on boosting algorithms."BMC bioinformatics 19.1 (2018): 109.

Mercaldo, Francesco, Vittoria Nardone, and Antonella Santone. "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques." Procedia computer science 112 (2017): 2519-2528.

Patil, Bankat M., Ramesh Chandra Joshi, and DurgaToshniwal. "Hybrid prediction model for type-2 diabetic patients." Expert systems with applications 37.12 (2010): 8102-8108.

Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal 15 (2017): 104-116.

Kohli, Pahulpreet Singh, and ShriyaArora. "Application of Machine Learning in Disease Prediction."2018 4th International Conference on Computing Communication and Automation (ICCCA).IEEE, 2018.

Perveen, Sajida, et al. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms."Procedia computer science 132 (2018): 1578-1585.

Thomas, Jensia, et al. "Machine Learning Approach For Diabetes Prediction." International Journal of Information 8.2 (2019).
https://www.kaggle.com/uciml/pima-indians-diabetes-database.