JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A review on diabetes prediction using machine learning techniques

Kajal[1] , Sunita Beniwal[2]
[1]M.Tech Student, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana
[2]Assistant Professor, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana

Abstract:-Machine learning is widely used nowadays on medical datasets. A lot of work has already been carried out on datasets of different disease. Diabetes is affecting a lot of people worldwide. International diabetes federation has reported that 382 million people are living with diabetes across the whole world and this number is going to double by 2035. This paper aims at reviewing some of the work done on diabetes dataset using machine learning techniques.

Keywords: Machine learning, Diabetes, classification

1. Introduction:

Machine learning can be known as a part of artificial intelligence that can be used to for discovering patterns in given data for classification and prediction on unknown data. Various statistical tools, data mining techniques and optimization techniques are put together to form ML models. Machine learning techniques are classified in three categories:-[1]

a. Supervised learning: It is used to construct predictive models which predict missing values using other values present in the dataset. Decision Tree, the Bayesian method, artificial neural networks and SVM etc. are examples of the supervised learning techniques.

b. Unsupervised learning: When inputs are known but the output is unknown, unsupervised techniques are used. Clustering algorithms like K-means clustering etc come in this category.

c. Semi-supervised learning: Labeled and unlabeled data both are used. Labeled dataset is used for making predictions, and unlabeled data is used for larger data distribution. [6]

Diabetes is a chronic disease that has affected many people in the world and number of diabetic patient are rapidly increasing. According to a report by the International Diabetes Federation (IDF) 463 million people have been affected by diabetes globally in 2019.It has also been reported that 578 million adults will have diabetes by 2030, and 700 million by 2045. Diabetes if left untreated and undetected can cause several other problems in long run like comma, renal and retinal failure etc.[2] Early detection and treatment of diabetes is important so that further complications arising due to diabetes can be avoided. Machine learning can be used to build prediction models for detecting diabetes and its type. It can also be used to detect risk factors causing diabetes so that precautionary measures can be taken. A huge amount of data of diabetic patients is available online in different data repositories like- KAGGLE, UCI, MNIST, etc. Machine learning models build on data available can be used by medical practitioners for early diagnosis and treatment.

2. Literature Review:

Different techniques have been applied to build prediction and classification models for medical datasets. An ensemble model has been proposed by Tafa et al. [3] using naïve Bayes and support vector machine for predicting diabetes on data of 402 patients collected from different locations and they reported an accuracy of 97.6% which is better than the techniques used individually. Sisodia and Sisodia [4] reported higher accuracy of 76.3% accuracy using naïve bayes after applying naïve bayes, decision tree, and support vector machine techniques on the Pima Indians Diabetes Dataset. Kumari et al. [5] used Pima Indians Diabetes Dataset, and applied a technique which is an ensemble of Random Forest, Logistic Regression, and Naïve Bayes and compared the performance of ensemble approach with other machine learning techniques and reported a better performance by ensemble approach with an accuracy of 79%. Gupta et al. [6] used support vector machines and Naïve Bayes algorithms on PIMA India diabetes dataset and reported better performance by support vector machine.

Perveen et al [7] used CPCSSN diabetes dataset and applied J48 decision tree, and two ensemble approaches i.e. bagging using J48 and adaboost using J48. They reported that performance of bagging was better for large sample size and adaboost ensemble method outperformed others in small sample size. Tigga and Garg [8] collected data through an online and offline questionnaire and Pima Indian Diabetes Dataset and applied Logistic regression, K-Nearest Neighbour, Support vector machine, Naïve Bayes and Decision tree and Random forest. They reported that best results were obtained by random forests for both datasets achieving an accuracy of 94.1% and 75% for their own dataset and PIMA dataset respectively.

Diwani and Sam. [9] used J48 decision tree and naïve bayes on the Pima Indian Diabetes Dataset and compared their performance with rule based techniques and random tree and reported best performance for naïve bayes i.e. 76.3% followed by J48 with an accuracy of 73.8%. Linear discriminant analysis (LDA) was used for Feature selection by Ahuja et al. [10] on Pima Indian Diabetes dataset to select the important features and conducted a comparative analysis on the preprocessed dataset of several machine learning algorithms namely Support Vector Classifier, Decision Tree, Random Forest, Logistic Regression and the Multilayer Perceptron and reported that MLP performed better than other classifiers. Malik et al. [11] performed a comparative analysis 10 machine learning algorithms on a dataset taken from a Frankfurt hospital in Germany and reported that K-NN, random forest and decision tree performed better than other techniques. Soltani and Jafarian[12] used probabilistic neural network for the prediction of diabetes on Pima Indians Diabetes Dataset and reported an accuracy of 89.56% and 81.49% on training set and test set respectively.

Hasan et al [13] selected features using Principle Component Analysis (PCA), Independent Component Analysis, and Correlation-based techniques on Pima Indian diabetes dataset.and then applied k-NN, Naïve Bayes, Decision Tree, Random Forest, Multilayer perceptron, Adaboost and gradient boost and reported that the ensemble of adaboost and gradient boost classifiers performed better than other techniques. Alehegn et al [14] reported that in order to achieve maximum accuracy, an ensemble of different machine learning algorithms should be used. Rathore et. al[15] applied Support Vector Machines and Decision Tree and reported an accuracy of 82% using SVM.

Shetty et al. [16] developed an intelligent diabetic prediction system using KNN and Bayesian algorithms to assist doctors in forecasting diabetes. Nai-Arun & Moungmai [22] designed a web application for forecasting the risk of diabetes. For thisthey first constructed a dataset of data collected after preprocessing, from 26 primary care units and applied various data mining techniques like bagging, decision tree, random forest, naïve bayes, logistic regression etc. reporting best performance by random forest. Majji, Ramachandra, and co-authors [23] created a website that gives the risk score of diabetes of the individual using different parameters like age, Smoking Status, Steroid usage, BMI (body mass index), Ethnicity, etc.

3. Conclusion:

Machine learning techniques applied to diabetes dataset have been reviewed in this paper. It can be seen that machine learning techniques improvers the prediction accuracy. It can also help medical practitioners to detect diabetes at an early stage or to prevent diabetes.

References:-

1. Saravanan, R., & Sujatha, P. (2018, June). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)* (pp. 945-949). IEEE.
2. IDF Diabetes Atlas. Key Figures from the IDF Diabetes Atlas 9th Edition. 2019.Available online: **https://web.archive.org/web/20211208190021/https://diabetesatlas.org/**
3. Tafa, Z., Pervetica, N., & Karahoda, B. (2015, June). An intelligent system for diabetes prediction. In *2015 4th Mediterranean Conference on Embedded Computing (MECO)* (pp. 378-382). IEEE.
4. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.
5. Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.
6. Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In *Operations Management and Systems Engineering: Select Proceedings of CPIE 2019* (pp. 365-376). Springer Singapore.
7. Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.
8. Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.
9. Diwani, S. A., & Sam, A. (2014). Diabetes forecasting using supervised learning techniques. *Adv Comput Sci an Int J*, *3*, 10-18.
10. Ahuja, R., Sharma, S. C., & Ali, M. (2019). A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, 2516-0281.
11. Malik, S., Harous, S., & El-Sayed, H. (2021). Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. In *Modelling and Implementation of Complex Systems: Proceedings of the 6th International Symposium, MISC 2020, Batna, Algeria, October 24-26, 2020 6* (pp. 95-106). Springer International Publishing.
12. Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications*, *7*(6).
13. 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.
14. Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426-436.
15. Rathore, A., Chauhan, S., & Gujral, S. (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women. *International Journal of Advanced Research in Computer Science*, *8*(5).
16. 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.
17. Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132-142.
18. Majji, R & Ravi, B. (2018). Type 2 Diabetes Classification and Prediction Using Risk Score. International Journal of Pure and Applied Mathematics. 119. 1099-1111.