Diabetes Prediction Using Ensemble Machine Learning Techniques

Nicole Dsouza A020

Kunjal Shah A065

Pranav Singh A075

*Problem Statement*

Diabetes is a critical disease. The timely prediction of this disease is essential in order to avoid drastic side effects. The current practice in the medical industry is to gather the required information for diabetes diagnosis through numerous tests and treatment is later provided on the basis of diagnosis. However, in a lot of cases, the early stages go undetected and it is also quite challenging for medical practitioners to diagnose it due to complex interdependence on various factors.  A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision-making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. This paper proposes the creation of an early diabetes detection model, this model not only will provide better accuracy then human beings but will also reduce the burden on medical professionals.

*Research Gaps*

Upon an extensive literature survey, we have found that there are various methods to tackle this situation but each of them comes with certain disadvantages or a set of problems.

The first challenge is related to data acquisition. One of the major problems while conducting a literature survey was finding articles and papers which were unrelated to the PIMA India dataset. This dataset is well established and provides fairly decent results but there are others that are either too small or inadequate or lack real-time data. Small datasets lead to overfitting on a model, which leads to higher accuracy, but these do not work that well on new testing data. Thus, it is not feasible for real-time implementation.

It is important to bring in real and recent patient data for continuous training and optimization of models. The quantity of the data-set should be large enough to train appropriately and predict with higher efficiency.

Another problem is feature selection. Some authors neglected a few of the features, while some grouped them for feasible training. It is necessary to select a model which uses features to optimize the performance. Debugging was also a problem because tools like Jupyter Notebooks which divide the code into cells cannot be used on automation batch processes.

Authors in some cases required a time-series dataset. It's hard to replicate such work because it's not available on any online resource. For both training and testing, such specialized models demand extensive tuning and large datasets.

Another challenge is the construction of an actual model. Many parameters must be adjusted to achieve perfect accuracy. While building a model, factors such as random states, kernel, number of trees, hyperparameter tuning, and others are taken into account. It is really essential to choose the correct algorithm with the appropriate hyperparameters. Some classification models only train on a single parameter, due to which the model’s real-time detection accuracy is reduced. The analysis of these schemes in all cases reveals that the majority of them have either a single data input parameter or a feature selection that is not optimal. Along with such parameter constraints, few classification-based schemes are solely dependent on specific types of hardware devices, making their availability and adaptability more difficult.

*Solutions*

K-means algorithm can be used for a variety of data types, but it is sensitive to the initial positions of cluster centers which determine the final result. These either provide an efficiently clustered dataset for the logistic regression model or give a lesser amount of data as a result of incorrect clustering of the original dataset, limiting the performance of the logistic regression model as a consequence.

Principal components analysis can be used to reduce the dataset to a lower dimension to ensure that the least information is lost, and provide a better centroid point for clustering.

*Future Scope*

On further reading we inferred that in order to improve the accuracy of the models and provide better performance, we must work towards feature selection techniques like LDA, SVD and PCA, this will enable us to better choose important features and will also consequently reduce training time of the model.

The implementation of Neural Networks like ANN, CNN and RNN in combination along with other algorithms will help achieve higher accuracy in diabetes detection. This is because hybrid schemes play an important role in improving the performance of the models. Patients can be treated more effectively owing to early detection which will help avoid further risks in cases of diabetes.

One limitation of this study is that a structured dataset has been utilised for training, however in the future, unstructured data must also be considered, this will paint a clearer picture on the accuracy of the model.

The learnings from diabetes detection models can be applied to other medical domains for prediction, such as for different types of cancer, sleep apnoea, patterns of mental health illnesses etc.

Additionally, attributes like family history, lifestyle habits etc can also be incorporated in the model

Extra resources referred in addition to the literature survey:

Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V et al (2020) Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. Diabetes Metab Res Rev 36(2):e3252. <https://doi.org/10.1002/dmrr.3252>

Thenappan S, Rajkumar MV, Manoharan PS (2020) Predicting diabetes mellitus using modified support vector machine with cloud security. IETE J Res. <https://doi.org/10.1080/03772063.2020.1782781>.

Bai BGM, Nalini BM, Majumdar J (2019) Analysis and detection of diabetes using data mining techniques-a big data application in health care. In: Shetty NR, Patnaik LM, Nagaraj HC, Hamsavath PN, Nalini N (eds) Emerging research in computing, information, communication and applications, vol 882. Springer, Singapore

Islam MT, Raihan M, Farzana F, Aktar N, Ghosh P, Kabiraj S (2020) Typical and non-typical diabetes disease prediction using random forest algorithm. In: Abstracts of the 11th international conference on computing, communication and networking technologies, IEEE, Kharagpur, 1-3 July 2020. <https://doi.org/10.1109/ICCCNT49239.2020.9225430>

Islam MMF, Ferdousi R, Rahman S, Bushra HY (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In: Gupta M, Konar D, Bhattacharyya S, Biswas S (eds) Computer vision and machine intelligence in medical image analysis. Advances in intelligent systems and computing, vol 992. Springer, Singapore, pp 113–125. <https://doi.org/10.1007/978-981-13-8798-2_12>

Daanouni O, Cherradi B, Tmiri A (2019) Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In: Abstracts of the 4th international conference on smart city applications, ACM, Casablanca, 2-4 October 2019. https://doi.org/10.1145/3368756.3369072

Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585. <https://doi.org/10.1016/j.procs.2018.05.122>