Predictive Modeling of Diabetes Using Python Neural Networks

Advik Vatsyayan
6/28/24

Abstract

This work uses machine learning approaches to predict the occurrence of diabetes in individuals. Specifically, a neural network model is developed by utilizing the sklearn Python package. Since diabetes is a common chronic metabolic illness, preventing complications and managing the condition effectively depend on early detection. The goal of the research is to accurately classify people as either healthy or diabetic based on pertinent medical characteristics. The model's overall accuracy, balanced performance metrics across the healthy and diabetic classes, and precision, recall, and F1-score are used to evaluate its effectiveness. The outcomes show a respectable performance, suggesting a good potential for diabetes occurrence prediction. The findings emphasize how crucial it is to keep improving the model in order to increase its predictive power and, eventually, help those at risk of diabetes receive better healthcare.

1. Introduction

The creation of a prediction model to detect the presence of diabetes in patients is the main research topic this study attempts to answer. Diabetes is a chronic metabolic disease marked by elevated blood sugar levels. Because of its high prevalence and related consequences, diabetes poses considerable issues in the healthcare system. For diabetes to be well managed and to avoid negative consequences, detecting the disease early is essential. It is estimated that up to 40% of adults older than 30 years with type 1 diabetes might have been misdiagnosed with type 2 diabetes, further demonstrating how difficult it is to properly diagnose diabetes (Manov et al., 2023). Therefore, the goal of the research is to determine whether it is possible to forecast the incidence of diabetes based on pertinent patient variables by utilizing machine learning techniques, notably neural networks constructed using the sklearn Python package.

The significance of this research lies in its potential to aid healthcare professionals in early diagnosis and intervention, thereby improving patient outcomes and reducing the burden of diabetes-related complications. According to the CDC, effective blood sugar management can reduce the risk of eye disease, kidney disease, and nerve disease by 40%, along with reducing chances of other chronic conditions (CDC.gov). By accurately identifying individuals at risk of diabetes, preventative measures and targeted interventions can be implemented, leading to better disease management and overall public health.

The problem at hand is a supervised classification task, where the model is trained on a labeled dataset containing information about patients' demographic, clinical, and lifestyle factors, and their corresponding diabetes status (healthy or diabetic). The data includes a mix of numerical and categorical features, such as age, body mass index (BMI), blood pressure, and glucose levels. The output of the project is binary labels indicating whether a patient is classified as healthy or diabetic based on the input features after the algorithm runs its own analysis on the dataset.

2. Background

Neural networks look promising for diabetes prediction, as it is a multifactorial, chronic disease driven by different risk factors. However, developing predictive models that are accurate remains challenging because of the complicated interplay of genetic, lifestyle, and environmental factors contributing to diabetes onset and progression. Several researchers have used different approaches to try to deal with the problem. Kannadasan et al. proposed a DNN model using stacked autoencoders cascaded with a softmax classifier and got an accuracy of 86.26% in the classification of type 2 diabetes data. While the approach showed that deep learning techniques are viable in this area, further improvement can be achieved by diversifying and representing data. Zhou et al. used the CNN model to predict future glucose level data of diabetic patients. The CNN performs very well when a representation of a data set with spatial and temporal patterns is to be done, which is best in the analysis of time-series data. However, their research is in the prediction of glucose level data, not for the diagnosis of diabetes, which might take a whole different dataset and architecture. On the other hand, Bae et al. explored the prediction of risks for type 2 diabetes from both common and rare genetic variants. The addition of genetic data could potentially increase the power of prediction models, but it is hard to collect such data and raises ethical issues such as privacy and accessibility to data. The project lies in developing a Python-based Neural Network model that utilizes the flexibility and scalability of Python's ecosystem for data preprocessing, model training, and deployment. By selecting the correct dataset, engineering features related to the data, and trying different neural network architectures, the model's accuracy and generalizability of diabetes prediction models can be increased. The approach, therefore, is to attain a model that strikes a good balance between model complexity and interpretability, to ensure that the predictions obtained are accurate but also explainable to both healthcare professionals and patients, factoring in techniques for handling imbalanced data and addressing potential biases in the training dataset to ensure fair and equitable predictions for different demographic groups.

3. Dataset

The numerical data in this project constitutes the dataset made from 70,692 samples. The model considers the samples from the 2015 CDC survey about the health of responders and if they are diabetic or not. Most of the questions asked were yes or no questions, like "Have you smoked at least 100 cigarettes?" or "Do you have high blood pressure?" Responders were later asked if they are diabetic or not.

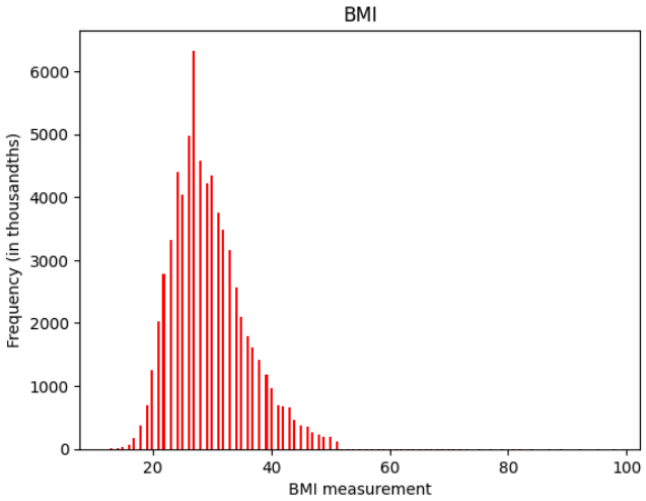

Figure 1: A graph showing the distribution of one of the variables in the dataset, showing frequency in the thousands for every BMI measurement. Most entries in this dataset had a BMI of 20-25, around the healthy range.

Data preprocessing procedures applied are minimal and mainly deal with the handling of null values. It contains 21 feature variables and a target variable named "Diabetes_binary," which is itself of two classes: 0 for those who have no diabetes and 1 for those who have either prediabetes or are diabetic. The dataset distribution is balanced, with a 50-50 split between the respondents without diabetes and those that have either prediabetes or diabetes.

The dataset contains a variety of demographic, clinical, and lifestyle features that are generally associated with diabetes risk. For instance, features could be age, body mass index, measurements of blood pressure, and levels of glucose or cholesterol, family history of diabetes, physical activity, and dietary habits. Each of these features forms the basis on which the predictive model draws to discern patterns and accurately predict the possibility of diabetes incidence among users.

4. Methodology / Models

The MLP, or Multi-Layer Perceptron, is a neural network model composed of multiple layers of interconnected neurons. These artificial nodes in a neural network are arranged in layers, just like neurons in the brain, with each layer being in charge of examining one particular aspect of the input data. They are very strong and can represent complex nonlinear relationships between features and target labels. This ability of a neural network model to interpret information and learn from examples is crucial in providing accurate results.

4.1. MLP Classifier Application to Dataset

In more detail, the model takes its input, then forward propagates it through the network; activation functions at each layer add non-linearity, while weights get adjusted through backpropagation to minimize a certain loss function. MLPs are flexible and have the ability to capture complex patterns in the data. The dataset was split using a 70:30 split for the train and test datasets, respectively—70% for training the models and 30% for testing the models. Some categorical data have been omitted from the analysis, since they might need other preprocessing techniques before they can be fed into some algorithms. In the case of the neural network, the MLP classifier, some adjustments of its hyperparameters were made to fine-tune its performance. This includes tuning parameters like alpha and c to prevent overfitting and hence improve generalization. Two models were trained: one with base parameters and another with tuned hyperparameters, hence allowing the comparison of their performance. The model learning procedure involves the preprocessing of the data, selecting relevant features, training multiple machine learning algorithms, and evaluating their performance using the right metrics. This iterative process helps identify the most suitable algorithm and parameters that yield the best results in terms of predicting diabetes occurrence in patients.

5. Results and Discussion

In the results section, the findings from the developed models to predict diabetes occurrence in patients are presented. In this regard, performance metrics such as accuracy, precision, recall, and F1-score of each model for classification were used to evaluate their effectiveness. Some models also tune hyperparameters or select them to optimize model performance. All models differed in performance regarding precision and recall; with the main MLP classifier model achieving an accuracy of 75%.

	precision	recall	f1-score	support
Healthy	0.76	0.73	0.75	10681
Diabetes	0.74	0.77	0.75	10527
accuracy			0.75	21208
macro avg	0.75	0.75	0.75	21208
weighted avg	0.75	0.75	0.75	21208

Table 1: Accuracy measurements of the model's performance after hyperparameter tuning. Precision and recall are both averaged for an F1-score and the two F1-scores are averaged for the accuracy.

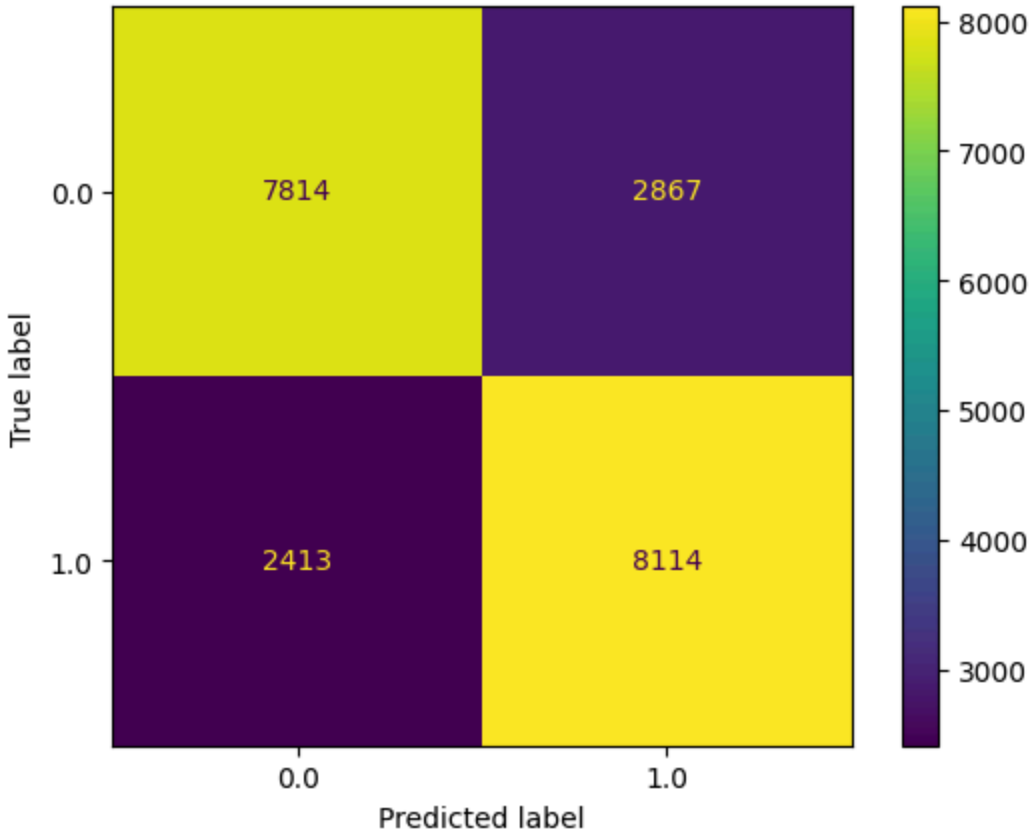

Table 2: The confusion matrix for the model. The table shows how many test entries were classified correctly and incorrectly, with the 0.0-0.0 and 1.0-1.0 grids showing correct classifications and 0.0-1.0 and 1.0-0.0 showing incorrect classifications.

Hyperparameters for the MLP classifier had been tuned to optimize its performance. Alpha and c parameters had been adjusted to avoid overfitting of the model and supported in generalization. The fine-tuning of these hyperparameters resulted in higher accuracy and better performance of the neural network model compared to its base configuration.

Despite the promising results obtained, it is important to note that there could be potential errors and drawbacks to the approach. In certain cases, the models might not give very good performance due to several reasons. Some possibilities might arise from noisy or irrelevant features included in the dataset, which possibly have negative effects on the model's ability to generalize to unseen data. In addition, class imbalance may lead to partial predictions where the model might not identify the minority classes correctly. Lastly, too little data or poor representation of some demographics in the dataset may contribute to the inability of the model to learn diverse patterns and trends in diabetes occurrence. The data did come from a 2015 CDC survey, which could have played a part in the model accuracy due to data that did not properly fit in some categories. Moreover, a high level of complexity may involve some challenges of interpretability and computational resources required for training and inference. In fact, neural

networks are flexible enough to model really complicated patterns but suffer from different effects like vanishing gradients or overfitting if not properly regularized or tuned.

The developed models promise satisfactory performance in forecasting diabetes onset among patients. However, these potential errors or limitations must be kept in mind, and a careful choice of hyperparameters, features, and dataset characteristics is required. Future research may look into how to alleviate these challenges, along with advanced techniques to improve model robustness and generalization capabilities for real-world healthcare applications.

6. Conclusions

The task of predicting diabetes becomes like trying to put together a difficult puzzle with odd-shaped pieces. Unlike some medical conditions where certain characteristics distinctly point to the presence of an ailment, diabetes takes many shapes among diverse populations. This can only be treated as an important problem to address, like trying to fit a square peg into a round hole. There's no one universal type of person with diabetes. These multiple facets in diabetes are largely caused by genetics, lifestyle, and environmental influences that make this a problem both for the patients and those who treat them. Innovative approaches must be developed to navigate through these challenges. In using predictive Python AI modeling, a new promising step forward is presented in combating this problem. By developing the power of machine learning algorithms, this research effort enables a competent tool to help the decisions of healthcare providers. However, it must be emphasized that this first version of the model is not intended for direct communication with patients but will function as a behind-the-scenes application with the purpose of empowering healthcare practitioners with more insights and predictive capabilities. The differentiation is critical because it specifies the role of technology as a supporting partner and not a replacement for human expertise and judgment. Only through such a partnership between technology and health professionals can the accuracy and efficiency in diabetes diagnosis and management be improved.

In this research paper, the important task of predicting diabetes occurrence in patients with the help of machine learning techniques was taken into account. This was aimed at formulating accurate predictive models in order to give aid and intervention in early diagnosis and proper diabetes management. This study is therefore of value since it will help bring about identification of those people who are at risk of contracting diabetes, leading to health outcomes through timely intervention.

The predictive model was developed using the python sklearn’s MLP classifier—a neural network algorithm. This paper covered a sequence of methodologies that comprised data preprocessing, feature selection, model training, and model evaluation to produce good results in predicting diabetes occurrence.

The models performed competitively, where the MLP classifier (neural network) had a good accuracy and good overall performance metrics, showing room for improvement. Hyperparameter tuning further optimized the performance of the neural network model to perform better; this has shown the importance of fine-tuning model parameters for better predictive accuracy.

While the models' performance was overall good, there are some factors that must have contributed to their performances: either negatively or positively. The inclusion of relevant

features and the careful preprocessing of the dataset probably contributed to the models' predictive abilities. However, the challenges that probably negatively influenced the performance of the models include imbalanced class distributions, noisy features, and model complexity.

There are several next steps that can be considered for further improvement of the research. Trying out different machine learning models or ensemble methods would provide insight into alternative ways of predicting diabetes occurrence.

While the models' performance was overall good, there are some factors that must have contributed to their performances: either negatively or positively. The inclusion of relevant features and the careful preprocessing of the dataset probably contributed to the models' predictive abilities. However, the challenges that probably negatively influenced the performance of the models include imbalanced class distributions, noisy features, and model complexity.

There are several next steps that can be considered for further improvement of the research. Trying out different machine learning models or ensemble methods would provide insight into alternative ways of predicting diabetes occurrence. Further, obtaining more diverse and comprehensive datasets, especially with larger sample sizes and better representations of demographic groups, could improve model robustness and generalization capabilities. Moreover, the application of advanced preprocessing techniques and feature engineering methods can help get rid of data quality issues and enhance model performance.

This research has paved the way for further research and improvement of predictive models used in the prediction of diabetes. Continuing to innovate and refine methodologies aims to contribute to the advancement of healthcare technologies and ultimately improve patient outcomes in diabetes management.

Acknowledgments

I would like to extend sincere gratitude to Ronil Synghal, my research mentor, for their invaluable support and contributions to this research endeavor. Ronil's dedication and expertise in the field have been instrumental in guiding and shaping the development of the project. Their insights and encouragement have significantly enriched the research process, and their commitment to excellence has been truly inspiring. The author is deeply appreciative of Ronil's unwavering support and acknowledges their significant role in the success of this endeavor.

References

Zhou, H., Myrzashova, R. and Zheng, R. (2020) *Diabetes prediction model based on an Enhanced Deep Neural Network - EURASIP Journal on Wireless Communications and networking*, *SpringerOpen*. Available at: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-020-01765-7 (Accessed: 28 June 2024).

E P, P. *et al.* (2022) *Implementation of artificial neural network to predict diabetes with high-quality health system*, *Computational intelligence and neuroscience*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9170457/ (Accessed: 28 June 2024).

Bae, S. and Park, T. (2017) *Risk prediction using common and rare genetic variants: Application to type 2 diabetes | IEEE conference publication | IEEE xplore*, *IEEE*. Available at: https://ieeexplore.ieee.org/document/8217926/ (Accessed: 28 June 2024).

Health and economic benefits of diabetes interventions (no date) *Centers for Disease Control and Prevention*. Available at: https://www.cdc.gov/nccdphp/priorities/diabetes-interventions.html (Accessed: 28 June 2024).

Kannadasan, K., Edla, D.R. and Kuppili, V. (2018) *Type 2 diabetes data classification using stacked autoencoders in deep neural networks*, *Clinical Epidemiology and Global Health*. Available at: https://cegh.net/article/S2213-3984(18)30277-X/fulltext (Accessed: 28 June 2024).

Manov, A.E. *et al.* (2023) *Unmasking type 1 diabetes in adults: Insights from two cases revealing misdiagnosis as type 2 diabetes, with emphasis on autoimmunity and continuous glucose monitoring*, *Cureus*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10450099/ (Accessed: 28 June 2024).