Early detection of diabetes based on lifestyle

Manav Mandal
mmandal@iu.edu

Atharva Gurav
athgurav@iu.edu

DataMining-athgurav-mmandal

Abstract

Amid the global diabetes epidemic that afflicts over 420 million people and gives rise to critical health issues, early detection has emerged as a crucial intervention. This study harnesses the power of advanced machine learning techniques to predict diabetes risk among a population of US citizens, drawing exclusively from data within the United States. With a substantial dataset comprising more than 250,000 instances, the primary objective is to categorize individuals into one of three groups: Healthy, Diabetic, or Pre-Diabetic. Lifestyle modifications, including dietary adjustments and lifestyle changes, offer a promising avenue for reducing diabetes risk, particularly within high-risk demographics. The project seeks to uncover essential lifestyle factors that contribute to the onset of diabetes. Building upon prior research that compared prediction models and utilizing diverse data mining techniques, the study's outcomes will furnish valuable insights into diabetes risk predictors, facilitating early intervention and preventive strategies.

Keywords

Machine Learning, Diabetes, Naive Bayes, Decision Trees, Glmnet, RF, XGBoost, LightGBM

1 Introduction

Diabetes has become a global epidemic where over 420 million people suffer from diabetes worldwide. It can cause other health related concerns such as cardiovascular disease, blindness and kidney failure if not managed on time. Early detection of diabetes has become crucial for timely treatment and prevention. Advances in machine learning has opened up new avenues for predictive modelling that helps us identify high risk individuals based on their day to day lifestyle. Machine learning models such as logistic regression, random forest and even deep learning nwtworks such as CNN analyze huge amounts of data that help uncover latent relationships between patterns associated with diabetes.

We leverage machine learning approaches to detect diabetes using the Centers for Disease Control and Prevention database. This dataset contains information about the lifestyle, physical activity and other demographics. Our goal is to use machine learning approaches to classify an individual into one of three classes: Healthy, Diabeteic or Pre Diabeteic. Answers to questions like Which lifestyle modifications can be made such as diet changes, prohibiting smoking and alcohol consumption to significantly reduce the progression even among high-risk groups. This project provides insights into the most important lifestyle predictors for diabetes onset.

Previous work

The paper A Machine Learning-Based Intelligent System for Predicting Diabetes [1] uses Naive Bayes and an enhanced Naive Bayes with clustering. This enhanced version first clusters the training instances and then trains the model on the cluster containing test instances to make predictions. The paper concludes that clustering before Naive Bayes improves the accuracy and sensitivity significantly for diabetes prediction compared to standard Naive Bayes on the dataset.

Many screening tests for Type 2 Diabetes (T2DM) were initially developed using statistical methods and later simplified into scoring systems. The study [2] compared different types of prediction models for identifying undiagnosed type 2 diabetes using data collected from various sources. They tested traditional regression models and more advanced machine learning models like Glmnet, RF, XGBoost, and LightGBM. The performance was evaluated by predicting fasting plasma glucose levels over time in batches of 6 months, simulating the arrival of new data.

The study found that using more sophisticated prediction models did not significantly improve clinical relevance. It suggests that in developing clinical prediction models, factors like stability of variable selection over time, interpretability, and model calibration should also be taken into account.

The next study that we saw [4] aimed to create a predictive model for ketosis-prone Type 2 Diabetes (T2DM) based on patients' clinical characteristics. In conclusion, their study offers a practical way to predict the risk of ketosis-prone T2DM, which can assist in better classifying and managing this condition.

In the paper Performance Analysis of Data Mining Classification Techniques to Predict Diabetes [3], the authors use three algorithms to compare the performance on a dataset from the Canadian Primary Care Sentinel Surveillance Network. Out of the three methods Ada Boost with J48 as the base classifier gave the maximum accuracy where area under the curve (AROC) was the evaluation parameter.

2 Methods

The study [2] involved looking at the medical records of 27,050 adults who didn't have type 2 diabetes. These records came from 10 healthcare centers in Slovenia, and the data was made anonymous before being put together into one big database. The dataset initially had 111 different pieces of information, including questions about people's lifestyles and health history. First, they removed records with missing answers to certain questions to compare their results to a model developed for the Slovenian population. Then, they got rid of unusual values and any data with more than 50% missing information.The final dataset included information about cholesterol levels, lifestyle factors, blood pressure, and other health conditions. They split the data into five groups based on when it was collected (the first 6, 12, 18, 24, and 30 months) and filled in any missing information using a method called Multiple Imputation by Chained Equations. This helped make the data more complete and ready for analysis. With just 6 months of data, the simple regression model performed the best with the lowest error, followed by RF, LightGBM, Glmnet, and XGBoost. When more data was added, Glmnet improved the most. LightGBM models were the most consistent in selecting the important variables over time.

The authors [4] included 964 newly diagnosed T2DM patients in their study. They collected the patients' baseline clinical information and used statistical analysis to identify the key risk factors. With these factors, they developed a model and created a visual tool called a nomogram to predict the risk of ketosis-prone T2DM. The model's accuracy was tested and confirmed to be reliable, both internally and externally. Three different data mining techniques were used by [3] to predict diabetes in patients. These methods are a standalone decision tree called J48, and two ensemble methods called bagging and adaboost, which use J48 as their base. The goal was to generate useful knowledge for making decisions about diabetes patients by analyzing various patient data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) dataset. Decision trees are great for classifying and predicting things, and in this study, they built effective models to classify diabetic patients across different age groups in Canada. The results showed that adaboost worked better than bagging and the standalone J48 decision tree. In the future, these ensemble methods can be used to predict other diseases like hypertension, heart disease, and dementia. Additionally, other techniques like Naïve Bayes, SVM, and neural networks can also be used as base methods in these ensemble approaches. The research [1] aimed to improve the accuracy of predicting diabetes using a smarter system. They combined a method called Basic Sequential Clustering with the Naïve Bayes technique to create this system. They tested it with a dataset called Pima and compared it to the regular Naïve Bayes method.

The comparison showed that their new method was 10.34% more accurate than the regular one. However, it had a slight drop in one measurement (specificity) and a tiny increase in the chance of making a mistake. But it significantly improved another measurement (sensitivity) by 53.11%. This means the new method is a better choice for predicting diabetes. The study had some limitations, like not having enough data from female patients and only using five features to make predictions. But they believe the method can be adapted to work with more data and features related to diabetes, as long as some conditions are met. If that happens, they'll need to test how well the method performs. Our primary aim is to identify lifestyle factors in US citizens that are linked to diabetes using data exclusively gathered from people in the United States. With over 250,000 instances in the dataset, there is a substantial sample size available for building prediction models and conducting various analysis for feature extraction. By employing various machine learning methods we would categorize individuals into one of three groups: Healthy, Diabetic, or Pre-Diabetic.

References

[1] Nabila Shahnaz Khan, Mehedi Hasan Muaz, Anusha Kabir, and Muhammad Nazrul Islam. A machine learning-based intelligent system for predicting diabetes. *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, 4(2):1–20, 2019.

[2] Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1):11981, 2020.

[3] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, 82:115–121, 2016.

[4] Jia Zheng, Shiyi Shen, Hanwen Xu, Yu Zhao, Ye Hu, Yubo Xing, Yingxiang Song, and Xiaohong Wu. Development and validation of a multivariable risk prediction model for identifying ketosis-prone type 2 diabetes. *Journal of Diabetes*, 2023.