Victoria De Cun
Asif Abbas
Moein Afrouzmehr
Fidel Gutierrez
Indronil Bhattacharjee

CS508 - Intro to Data Mining
Project Proposal

24 September 2023

Diabetes Predictive Modeling: Leveraging Machine Learning for Early Detection

Introduction

Diabetes is a widespread health issue affecting millions worldwide, originating from the body's incapacity to produce or utilize insulin effectively, thereby resulting in sustained high blood sugar levels. Its onset is influenced by a combination of genetic, lifestyle, and environmental factors. Uncontrolled, diabetes has the potential to cause serious complications such as heart ailments, strokes, renal damage, and visual impairments. Establishing a predictive model to discern individuals at an elevated risk of contracting diabetes is imperative. This would enable timely interventions and preventative measures, potentially saving lives and reducing healthcare costs.

Objective

The chief objective of this project is to employ a variety of machine learning classifiers on a dataset comprising individuals both with and without diabetes, with the aim of developing a sturdy predictive model. This model will utilize a diverse array of features, spanning clinical and non-clinical, to ascertain an individual's risk of developing diabetes. Through harnessing the predictive capabilities of machine learning algorithms, we aspire to improve existing screening methods and proactively pinpoint individuals at elevated risk.

Motivation

1. **Early Detection and Intervention**: A majority of complications arising from diabetes are a result of delayed diagnosis and treatment. By predicting susceptibility, one can instigate early interventions thereby reducing the risk of complications.
2. **Enhancing Existing Diagnostic Tools**: While traditional diagnostic tools focus on overt symptoms and clinical markers, a machine learning model can consider a wide array of features, potentially uncovering novel indicators of diabetes risk.
3. **Holistic Understanding of Contributing Factors**: By evaluating the significance of various features in predicting diabetes, we can gain insights into less-understood factors that contribute to its onset.

Methods

To construct an optimal diabetes prediction model, a pivotal step is securing a comprehensive dataset encompassing a variety of relevant features. Clinical features such as blood glucose levels, insulin levels, age, BMI, blood pressure, and lipid profiles are essential as they are direct indicators of diabetes. Additionally, incorporating non-clinical features like lifestyle, dietary habits, and family history can enhance the model's predictive capabilities. The dataset should also be sufficiently large to train robust models and validate their

performance.
A diverse array of algorithms should be explored to identify the most effective model. We propose starting with simpler models such as Decision Trees, kNNs and Naive Bayes before expanding to more advanced techniques. Support Vector Machines and Logistic Regression are also well-suited for binary classification problems like this. While neural Networks can uncover complex patterns, they necessitate large volumes of labelled data to uncover intricate patterns and relationships. In the realm of medical diagnostics, acquiring sufficient quality and quantity of patient data, whilst maintaining privacy and ethical standards, can pose a significant challenge. More over, they suffer from a lack of interpretability. In medical applications, understanding the rationale behind a prediction is crucial for clinician acceptance and trust. The inability to interpret model decisions can hinder the deployment of such models in a healthcare setting. Lastly, to ensure reliability and robustness of the classifier, we will evaluate performance using techniques like cross-validation and assessing metrics such as accuracy, precision, recall and F1 score.

Dataset

The chosen dataset and the associated statistical analysis is presented in detail in a separate document.

Proposed Timeline

Statistical Analysis and Data Pre-Processing: Currently, we have completed the statistical analysis and the pre-processing of the dataset.
Application of different classification Models: Since the classification models can be applied independently from each other, we propose splitting the tasks among members to ensure prompt completion. We envision this will take 3-4 weeks.
Final Report and Presentation: Upon obtaining a robust classification model, we will be utilising the rest of the allotted time to write the final manuscript and prepare for the presentation.