paper-title

author-list

March 4, 2025

Abstract

Diabetes is an endocrine disease that results in presence of excess blood sugar, which leads to a multitude of complications. The past decade has seen an alarming increase in diabetes prevalence, with the number of people with diabetes expected to exceed 570 million by this year. In parallel, recent years have seen many leaps in development and performance of machine learning models. This study is an attempt at harnessing these models to predict diabetes, using a well-established benchmark dataset. Existing models suffer from being unable to differentiate between different types of diabetes due to lack of data points in a single dataset. Our model hopes to alleviate this weakness by compounding two datasets with adequate data, such that one predicts the presence of diabetes, and the other, the different types of diabetes. Prediabetes, Type 1 Diabetes, Type 2 Diabetes, and Pancreatogenic Diabetes were chosen for detection. Naive Bayes Classifier, K-Nearest Neighbours, Logistic Regression, Random Forest and XGBoost algorithms were used, with XGBoost showing the most promise. The XGBoost model was thus selected, and its ¡¿ was used to obtain a feature importance graph in order to validate the outcome against existing data, and further interpret the patterns among different diabetes types. A model like this could potentially be used for diabetes detection using readily available parameters.