Enhancing Diabetes Risk Classification Using Machine Learning

Alain Areeba Siddiqui

Faculty Mentor: Dr. Masoud Rostami

Division of Data Science, College of Science, The University of Texas At Arlington

Introduction

Diabetes is a global health concern, affecting over 38.4 million people in the United States and more than 830 million people worldwide. Early detection and risk assessment are crucial for improving patient outcomes and reducing the burden on healthcare systems. Machine learning (ML) models can enhance diabetes prediction by identifying patterns in patient data that are not immediately evident through traditional diagnostic methods. This study uses ensemble learning algorithms to predict diabetes risk level based on various health indicators, including BMI, patient history, and socioeconomic factors. Using the CDC's Diabetes Health Indicators dataset, multiple ML models were trained and combined to improve predictive accuracy. Such models can support healthcare providers in early diagnosis and treatment planning. By using AI-driven insights, these models can identify key risk factors and contribute to more personalized and efficient diabetes management strategies, ultimately improving patient care.

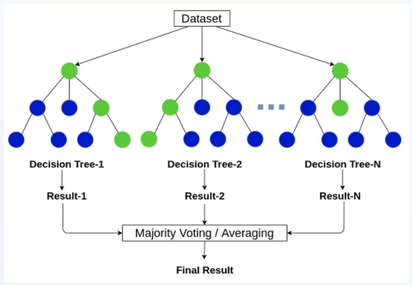
Background and Objective

Traditional risk assessment of diabetes relies on the clinical judgment of healthcare providers and standard diagnostic criteria, which can sometimes lead to an oversight and not take into consideration other factors. Machine learning offers an alternative by analyzing complex relationships in patient data to improve risk prediction accuracy.

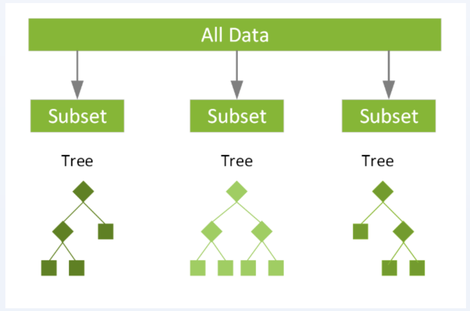
Developing robust ML models for diabetes risk classification could assist healthcare providers in early diagnosis and treatment planning and increase efficiency.

Methods

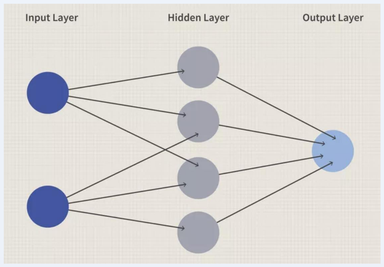
This study employs ensemble learning algorithms in Python, such as Random Forest, XGBoost, Neural Network, and Stacking to create different models Evaluation metrics such as accuracy, precision, recall, and F1-score were compared, with recall being the most valued. These models are trained to predict diabetes risk based on various health indicators, including BMI, patient history, and socioeconomic factors. The dataset used for training and evaluation is the CDC's Diabetes Health Indicators dataset, outlining different factors such as cardiac history, fruit intake, mental health, medical costs, and education. Multiple machine learning models were trained and combined to enhance accuracy. The ensemble learning approach integrates the strengths of different models to improve overall performance.



Decision Tree and Random Forest Visualization



XGBoost Visualization

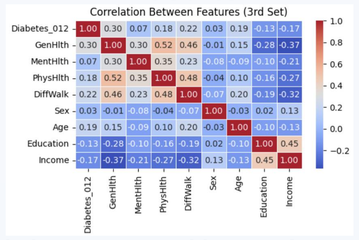


Neural Network Visualization

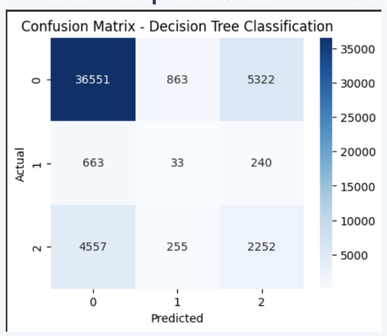
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Equation for Recall Metric

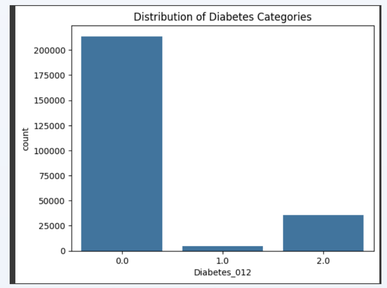
Results



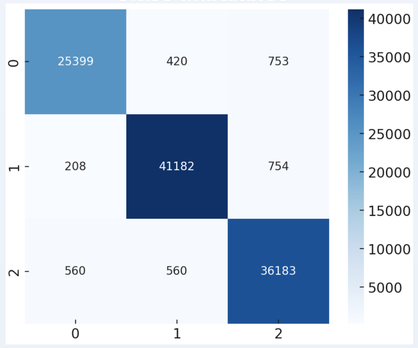
Correlation Heatmap with some factors in dataset



Baseline Model Confusion Matrix



Class Imbalance

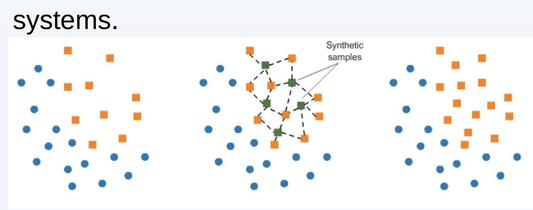


Final Model Confusion Matrix

Using the confusion matrices as a visual representation to compare the predicted class of data point versus the actual class of the data point, the machine learning model classifies each data point based on the values for the different indicators. The target variable is ordinal encoded; 0 is no diabetes, 1 is prediabetes, 2 is diabetes. There is a difference in our baseline and final models due to the model having errors in misclassification, which is why the color spread differs.

Conclusions

The biggest obstacle in this study was using imbalanced class techniques such as SMOTENN to solve the issue where there were too many data points of one classes. The final ensemble model achieved strong performance, with an accuracy of 0.97, a precision of 0.96, recall of 0.97, and an average F1-score of 0.96, demonstrating its effectiveness in diabetes risk classification. By using AI-driven insights, these models can support healthcare providers in early diagnosis and treatment planning. Additionally, they can help identify key risk factors, enabling more efficient diabetes management strategies. Implementing such ML-based solutions could significantly enhance patient care and reduce the long-term burden of diabetes on healthcare systems.



How SMOTE works

Models	Accuracy	Precision	F1 Score	Recall
Baseline	0.77	0.78	0.77	0.77
Random Forest	0.84	0.79	0.81	0.84
XGBoost	0.85	0.81	0.81	0.85
XGBoost with SMOTEEN	0.85	0.85	0.85	0.85
Neural Network	0.51	0.52	0.51	0.51
Stacking with SMOTEEN	0.97	0.97	0.97	0.97

Comparing All Models

Future Works

The findings highlight the potential for integrating machine learning models into patient software systems such as Epic Systems or Oracle Health. This model could be implemented into the electronic medical record in healthcare facilities and as well as automatically update as the patient's diagnostics change and see if their risk has changed based on changes they've made to certain factors to provide more holistic care.