Diabetes Prediction

Alexander Wen, Raymond Wong, Michael Eirikson

Table of contents

Summary

In this project we attempt to build a model to predict diabetes disease. We compared a decision tree model and naive bayes model and found the decision tree is stronger in this context. We used f2-score as our scoring function because detecting diabetes is the priority: a false negative could be much worse then a false positive.

In the test dataset: the decision tree model correctly detected 8283 of 10604 positive cases (recall rate is about 78%). This result does come at a fairly significant cost in terms of false positives (precision rate is about 30%) with 19650 false positives. Depending on the actual cost of false positive this may need significant improvement to be a viable screening model.

Introduction

In Canada and the USA approximately 10% of people are living with diabetes. In Canada in 2023 approximately 3.7 million people were living with diabetes and in the USA in 2021 approzimately 38.4 million people were living with diabetes. ("Snapshot of Diabetes in Canada, 2023" (2023)) In the USA it is the 8th leading cause of death. (Rios et al. (2017)) Globally an estimated 44% of people living with diabetes are undiagnosed. (Stafford et al. (2025))

In this project we try to predicted diabetes disease based on common health factors. A reliable model could help to prescreen people and recommend following up with a physician for people who are at risk. Given the large number of people living with undiagnosed diabetes this could potentially have a significant positive impact of world health.

The analysis uses the American CDC Behavioural Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators dataset (UCI ID 891), containing 253,680 survey responses with 21 health-related features and a binary diabetes outcome (0 = no diabetes/pre-diabetes, 1 = diabetes). (Dane and Teboul (2021))

No missing values were present and all features were already encoded numerically. The target classes is imbalanced (86% non-diabetic, 14% diabetic).

Methods

Modeling Approach

Results

EDA

Data Summary

Visualizations

Modeling

Classification Analysis

Result Visualizations

Discussion

References

Dane, Sohier, and Alex Teboul. 2021. "Diabetes Health Indicators Dataset." https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data; Center of Disease Control.

Rios, Nilka Burrows, Isreal Hora, Linda S. Geiss, Edward W. Gregg, and Ann Albright. 2017. "Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico." *Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico*, no. 66(43): 1165–70.

"Snapshot of Diabetes in Canada, 2023." 2023. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/snapshot-diabetes-canada-2023.html; Public Health Agency of Canada.

Stafford, Lauryn K, Anna Gage, Yvonne Yiru Xu, Madeleine Conrad, Ismael Barreras Beltran, and Edward J Boyko. 2025. "Global, Regional, and National Cascades of Diabetes Care, 2000–23: A Systematic Review and Modelling Analysis Using Findings from the Global Burden of Disease Study." *The Lancet Diabetes & Endocrinology*, no. 13(11): 924–34.