STATS/CSE 780
Project Proposal

Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-10-23

Introduction

Diabetes is a chronic disease that occurs when the body cannot effectively produce or use insulin to regulate sugar levels in the blood. According to the World Health Organization, 422 million people have diabetes worldwide and 1.5 million deaths that occur every year are linked to the disease (World Health Organization, n.d.). In this paper, we propose a study that leverages machine learning techniques to predict and prevent diabetes.

The data in this proposal was originally collected by Islam et al. from patients in Sylhet Diabetes Hospital in Sylhet, Bangladesh (2020) and was later openly published on Kaggle (Larxel, 2023).

Islam et al.'s study also used machine learning techniques to predict diabetes. In particular, they used naive Bayes, logistic regression, and random forest techniques and found that their random forest model had the best accuracy (Larxel, 2023). This study will expand on their research by validating their results through another random forest model and assessing whether a neural network would better predict results.

Methods

1) decision tree with boosting/bagging/random forest
2) clustering or GAM or neural network

Preliminary Analysis

Timelines

The presentation slides are aimed to be completed by November 23rd, 2023. The oral presentation will be on November 30th, 2023. The final written project will be completed by December 12, 2023.

References

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In M. Gupta, D. Konar, S. Bhattacharyya, & S. Biswas (Eds.), *Computer vision and machine intelligence in medical image analysis* (pp. 113–125). Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Larxel. (2023). *Early classification of diabetes*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

World Health Organization. (n.d.). *Diabetes*. https://www.who.int/health-topics/diabetes#tab=tab_1