Proposal for the Reproduction of Diabetes Prediction Methods

Jeffrey Bird
jeffreyab@vt.edu

Joshua Uy
joshua88@vt.edu

Katie Geibel
kmgeibel@vt.edu

Abstract

Diabetes can be prevented if it is caught early enough through simple steps such as exercise & weight reduction. "Before developing type 2 diabetes, most people have prediabetes; their blood sugar is higher than normal but not high enough yet for a diabetes diagnosis. Prediabetes is really common – 96 million US adults have it, though more than 80% of them don't know they do. The good news is that prediabetes can be reversed."[1] Untreated and uncontrolled diabetes can lead to UTIs, Chronic Kidney Disease, eye damage, blindness, stroke, and even death. [2]

1 Motivation

We will reproduce the results from the following two papers: "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques" and "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective".

The two papers use three different diabetes datasets to study the efficacy of using machine learning to predict the likelihood of diabetes. The datasets contain different attributes that could be predictive of diabetes. Furthermore, the demographics of the patients are significantly different and the data is collected from multiple countries.

All of the features in the Sylhet dataset are questionnaire-type indicators that could be collected without the use of any medical testing equipment or exams. This is valuable because we could detect diabetes in individuals without a patient visit to a doctor. With the rise of telemedicine, an individual could answer a questionnaire remotely and the machine learning model would return a score that indicates the individual's likelihood of diabetes.

1.1 Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques - Islam, M.M.F., Ferdousi, R., Rahman, S., Bushra, H.Y. (2020)

https://link.springer.com/chapter/10.1007/978-981-13-8798-2_12

The dataset referenced in this paper contains 520 observations and 17 attributes that are collected using direct questionnaires and diagnosis results from the patients in the Sylhet Diabetes Hospital in Sylhet, Bangladesh. There are 2 demographic features, Age & Gender; the other 15 attributes are binary features that indicate whether the patient experienced a symptom of pre-diabetes such as: excessive urination, excessive thirst, episodes of sudden weight loss, etc.

[1] https://www.cdc.gov/diabetes/prevent-type-2/index.html
[2] https://www.webmd.com/diabetes/risks-complications-uncontrolled-diabetes

1.2 Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective - Chollette C.Olisah, Lyndon Smith, Melvyn Smith (2022)

https://doi.org/10.1016/j.cmpb.2022.106773

This paper compares the efficacy of using machine learning methods to predict diabetes with two different datasets that they name the "Pima Indian dataset" and the "Laboratory of the Medical City Hospital (LMCH) Diabetes Dataset". Both of these datasets contain clinical data such as blood-test results.

"The Pima Indian diabetes dataset contains information of 768 women from a population near Phoenix, Arizona, in the USA. The dataset can be assumed to yield gestational diabetes information because there are pregnant women represented." The Pima datasets contains the following attributes: "[number of] pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age."

"The [LMCH dataset] consists of data from 1000 patients of Iraqi nationals collected from [Al-Kindi teach hospital in Iraq]. In all, about 103, 53, and 844 patients belong to the normal, prediabetes, and diabetes class, respectively. Each patient is described using the following attributes: the number of patients, sugar level blood, age, gender, creatinine ratio (Cr), BMI, urea, cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides (TG) and HDL Cholesterol, HBA1C."

2 Methodology

We chose these two papers because the three datasets utilized contain different attributes that may be predictive of diabetes. Likewise, the datasets contain data on disparate populations.

We will compare the efficacy of predicting diabetes using questionnaire results to the efficacy of predicting diabetes from clinical test data.

3 Evaluation

The paper "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques" uses 4 different machine learning methods to predict the likelihood of diabetes. 1) Naive Bayes 2) Decision Tree 3) Logistic Regression 4) Random Forest. This paper uses tenfold cross-validation and an 80:20 percentage split to train and validate the model. We should be able to reproduce the accuracy metrics of this paper using these algorithms and validation methods.

The second paper "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective" also uses the Random Forest model in addition to support vector machine and DNN. This paper performs data preprocessing to improve the accuracy of the machine learning algorithm

4 Timeline

By the due date of milestone report on 10/23, we should be able to reproduce the results from the first paper. For the milestone report, we will each tackle an algorithm and attempt to reproduce the results from (Islam et al. 2020). With the exception of Decision Tree/Random Forest, this corresponds with the material taught on the syllabus.

After the milestone report, we replicate the second paper's implementation of RF and SVM to the diabetes prediction problem. We contrast this with the accuracy and methods in the first paper.

Finally, (Olisa et al. 2022) applies additional steps of data preprocessing & hyperparameter tuning to improve the accuracy of the algorithms. If we have time, we can circle back to the algorithms from the first paper and apply data preprocessing & hyperparameter tuning to the Sylhet dataset.