Diabetes Mellitus Prediction Using Machine Learning Algorithms

Project Category: Life Sciences

Project Mentor: TBD

Name: Xinxie Wu

SUNet ID: xinxiewu

Department of Computer Science

Stanford University

xinxiewu@stanford.edu

Motivation

Diabetes Mellitus (DM) is a chronic (long-lasting) disease that affects how your body turns food into energy, the 7th leading cause of death in the United States. There are three main types of diabetes: type 1, type 2, and gestational diabetes.

According to Centers for Disease Control and Prevention (CDC), in the last 20 years, the number of adults diagnosed with diabetes has more than doubled. Currently more than 37 million US adults have diabetes, and even 1 in 5 of them don't know they have it.

Working in the healthcare industry for 5 years, with 3 years in the area of diabetes mellitus, I have witnessed how the prevention and the early detection help pts out of diseases. However, early diagnosis/detection of DM is quite challenging for medical practitioners, since DM has a complex interdependence on various factors from human's different organs. As a data scientist, I believe machine learning models, based on pts' medical data, would help the early identification/prediction of DM. Therefore, this application research will explore how the machine learning models would help in DM early prediction, and discuss the prediction accuracy among several models.

Method

1. Classification models would be the top ones to be used and compared in this research, such as logistic regression;
2. Features' selection will be discussed, since the important/high-related features could be suggested to medical practitioners and so pts being taken care specifically;
3. Also, if time and resources allowed, deep learning models, such as SVM & ANN, will be introduced and discussed.

Intended Experiments

1. **Experiment**: This research is to develop several machine learning models (classification & deep learning), based on pts' medical data. Also, features' selection will be conducted, important features/variables will be discussed. Finally, we will compare models' prediction accuracy;
2. **Evaluation**: Prediction accuracy is the top metric in this research. To achieve this, the dataset will be divided into training and testing sets, randomly.

Dataset

Due to PHI concern, this research is going to use public dataset, 3 **potentials** listed as below. Taking data size and team resources (work alone on this project) into consideration, "**2 only**" is preferred for now, but further detailed check is still required. If needed, additional datasets will be included during the research.

1. **Pima Indians Diabetes Database**: This dataset is widely used for diabetes prediction and is available from the UCI Machine Learning Repository. It contains data on 768 female Pima Indians, including features such as age, BMI, blood pressure, and glucose levels.
2. **Diabetes 130-US hospitals for years 1999-2008 Data Set**: This dataset contains data on over 100,000 diabetic patients from 130 hospitals in the United States. It includes features such as patient demographics, lab test results, and hospital admission details.
3. **Diabetes Data from the National Institute of Diabetes and Digestive and Kidney Diseases**: This dataset contains data on 443 patients with diabetes, including features such as age, sex, BMI, and blood pressure.

Reference

Reviewed papers which discussed DM prediction methodology and used logistic regression for prediction, more literature will be reviewed and included as reference during the research, ~10 reference in the final report. For now, reference as below:

1. Toshita Sharma, and Manan Shah. A comprehensive review of machine learning techniques on diabetes detection, 2021.
2. Jobeda Jamal Khanam, and Simon Y. Foo. A comparison of machine learning algorithms for diabetes prediction, 2021.
3. Jingyu Xue et al 2020 J. Phys.: Conf. Ser. 1684 012062. Research on Diabetes Prediction Method Based on Machine Learning.
4. Shahabeddin Abhari, Sharareh R. Niakan Kalhori, Mehdi Ebrahimi, Hajar Hasannejadasl, and Ali Garavand4. Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care: Focus on Machine Learning Methods, 2019.
5. Aishwarya Mujumdar, and Dr. Vaidehi V. Diabetes Prediction using Machine Learning Algorithms, 2019.