CSCE 585: Machine Learning Systems: Project Proposal: Implementation and Evaluation of Different Classification Algorithms to Predict Diabetes

Shima Oruji and Zach Thomas

September 15, 2022

Contents

1 Project Repository

This is the link to the github repository that we created for the course project.

2 Introduction

Diabetes Mellitus is among critical diseases and lots of people are suffering from it in recent years. According to the recent studies, age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases such as heart disease, kidney disease, stroke, eye problem, and nerve damage. With the recent advancements in the field of machine learning (ML), several researchers have tried to apply ML models to perform Diabetes prediction in patients based on various factors. However, there is no rigorous and comprehensive study on the evaluation of different ML models to determine the best practices in this specific problem.

In this project, we aim at implementing and evaluating different classification methods (e.g., decision tree, random forest, support vector machine, and neural network) on the given dataset and determine which methods perform better and under which conditions. We will use the Pima Indians onset of diabetes dataset. This is a standard machine learning dataset from the UCI Machine Learning repository. It describes patient medical record data for Pima Indians and whether they had an onset of diabetes within five years.

3 Problem Statement, Our Approach, and Evaluation

This is a binary classification problem (onset of diabetes as 1 or not as 0). As mentioned in the previous section, we will use the Pima Indians onset of diabetes dataset, which is a standard benchmark dataset for these studies. For the evaluation, we will use the accuracy, precision, and recall as general indicators of the performance of different classifiers. We are planning to tune each classifier with the best hyperparameters using the ROC curve and use it for comparison with other classifiers. We are not sure at this point, but we are speculating that neural network based model might reveal the best results.

4 Related Work

This problem has been extensively studied in the literature. However, none of the existing works has done a comprehensive evaluation on different classifiers. This is a short list of references: reference #1, reference #2. We will write a comprehensive related work in our final report.

5 Revisions According to the Feedback

Based on the discussion we had with Professor Jamshidi and to address his comments, we add the following two items to the scope of the project:

- In addition to the current dataset, we will try to find additional datasets and evaluate the performance of the trained models with them. This can indicate the performance dependency of the studied models on the new datasets.
- We will design a website which can be used by users to directly interact with the trained models for seeing the diabetes prediction of a given input. The website will use an API to do the predictions. We will release the API publicly so that interested folks can connect their applications (e.g., mobile apps) to it or interact with the models programatically.