Collins Kariuki
Kenny Gonzalez
Brisa Salazar
11/14/23

Project Proposal

Summary

Diabetes is one of the most preventable diseases in the United States, almost entirely determined by an individual's diet and lifestyle. Given this, for this project, we aim to determine which health risk factors have the biggest impact on determining the likelihood of a person to develop prediabetes or diabetes. We will use a probabilistic model to determine which factors have the most weight in determining diabetes. We are still deciding on the classifier, and we are debating other options like a neural network or a linear classifier (feedback greatly appreciated here!). The main metric we will use for our evaluation is accuracy rates, depending on in the different features.

Resources

The dataset we will use can be found here:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/
This data set is a collection of around 70k surveys. The survey has a couple of features and a variety of values for the features ranging from binary to non-binary. In terms of the labels, it has two, either 0 or 1 (0 for diabetes and 1 for prediabetes or diabetes). It is important to note that this data is split 50-50, meaning 50% of these responses have no diabetes and the other 50% prediabetes or diabetes.

In terms of the code that we will be using, we are still deciding the classifier, but there is some code that we want to largely reuse and make sure we include. First, we would like to reuse some of the data readers and data parsers that were given to us as part of most assignments, and we want to work through these and make the necessary changes so that they are applicable and usable with the format of your data. Some things that come to mind are, for instance, making sure that functions and classes that use the data work. For example, we want to make sure that functions like “get data, ” “get label,” “get feature set” and more still work for out data set and code. We can do this two ways, by changing the format of the data or changing the code itself.

Second, we want to use the general code structure (either the code we wrote or pseudo code) from the classifier that we decide to use. If we want to do a probabilistic model, then we want to use this code and adapt so that we can use it for our new purposes.

In the case that we change the code quite a bit, we will reuse some of the smaller data sets to make sure that it is working properly.