Machine Learning to Identify Diabetes - Report

Ryan McGuinness

12/12/2021

## Introduction and Problem Statement

It is hard to overstate the impact that diabetes has had on many Americans, especially in recent years. According to the CDC, the disease is the seventh largest killer in the US, and the number of cases of diabetes has doubled in the last 20 years. But perhaps the most disturbing statistic about diabetes is that a suspected 1 in 5 cases go completely undiagnosed. We attempt to combat this crisis by developing a classification algorithm to help identify cases of diabetes (type II) and prediabetes in adults.

The data used in this analysis comes from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) and is cleaned up and made available for use by Alex Teboul on Kaggle. The dataset includes 253,680 responses from adults in the United States and includes a three-valued target variable (no diabetes or diabetes only during pregnancy, prediabetes, and diabetes) and 21 feature variables.

## Methodology

Several approaches were attempted in the course of this project. The most successful approaches were to use the k-NN technique directly on the data, and to use the Naïve-Bayes classifier on this data. The k-NN technique produced the highest accuracy model. However, this is likely a product of the relatively low number of prediabetic cases in the data set used, as the model performed extremely poorly on those cases. While the Naïve-Bayes classification technique produces a model with worse accuracy, it is able to handle the prediabetic cases fairly well. As trying to predict these cases is part of the goal of the project, we will chiefly be using this approach in this analysis.

One of the unfruitful approaches that were tried was to split the problem into two by developing a model to first identify cases of either prediabetes or diabetes (collectively “some diabetes” in this report), and then to test records where some diabetes has been detected with a separate model to classify these records as either prediabetic or diabetic with k-NN. Perhaps this method is worth exploring more in depth in the future, but this method ended up producing somewhat worse results than the direct k-NN model for this project. For that reason, we will not include it further in this analysis.

Another method that was tried in this project was to use the data where the target variable is binary (no diabetes or diabetes only during pregnancy, and prediabetes or diabetes), and the proportion of cases positive for some diabetes is increased by removing some of the negative cases from the data set. While this did produce a more accurate model with both k-NN and the Naïve-Bayes classifier, the results were not significant enough to justify losing the distinction between prediabetes and diabetes.

## Analysis

The confusion matrix for the Naïve-Bayes classification approach is shown below. For the purposes of this analysis, class 0 refers to having no diabetes or diabetes only in pregnancy, class 1 refers to having prediabetes, and class 1 refers to having diabetes.

Text

Description automatically generated

Figure 1: Naïve-Bayes algorithm confusion matrix and overall statistics

The 75.1% accuracy may be pleasing at first, but it perhaps isn’t as high as one would like. Considering the use case of detecting diabetes, it may be more telling to look at statistics relating to each class anyway.

Text

Description automatically generated

Figure 2: Naïve-Bayes algorithm statistics by class

And here we can see the main flaws in this approach. The model is decent on class 0 and even on class 2, but struggles greatly with class 1. This may be a product of there being relatively few prediabetic responders to the BRFSS 2015 survey, but as described above, shortening the data set by removing some negative cases to increase the proportion of diabetic and prediabetic responders in the data pool did not produce better results.

The alternate model is the k-NN model applied directly. Below is the confusion matrix for that model.

Text, letter

Description automatically generated

Figure 3: k-NN confusion matrix and overall statistics

Right away, the problem with this model is obvious: The model does not predict any cases to be prediabetic. Despite the higher accuracy of this model, this is not considered the main model as attempting to identify prediabetic cases is part of my problem statement. Additionally, its true positive rate is much worse than the Naïve-Bayes model. This is better seen in the detailed class statistics, as shown below.

Text

Description automatically generated

Figure 4: k-NN statistics by class

This alternate model is proof that the accuracy measure is not the defining measure in this data – the imbalance in the data set allows the model to ignore class 1 and predict the vast majority as class 0 in order to obtain a high accuracy.

## Conclusions

With an accuracy of 75.1%, the model is not as effective as one would hope for. Considering the current rate of diabetes detection is close to 80% according to the CDC, this model looks even worse by comparison, as the true positive rates that this model for detecting prediabetes and diabetes are much, much worse at 57%. While this model probably falls short of helping medical professionals in identifying diabetes to a significant degree, this model may provide some use for preliminary “at-home” assessments where suspecting people may test their likelihood of having diabetes or prediabetes before seeing a physician. Still, it is likely no help for the doctors in a hospital in assessing the patient’s condition.