EARLY STAGE DIABETES RISK PREDICTION IN BANGLADESH

Final Report

Herman Autore

Introduction to Statistical Consulting | STA 5939

Abstract

Epidemiological data from a Bangladeshi study of diabetes prevalence was analyzed to find what features are important in predicting a patient's likelihood of developing diabetes. Through the course of the analysis the author learned ways of analyzing data, including entirely binary data. Although machine learning software packages allow for the production of interpretable decision tree graphics, which are useful for a patient's self-assessment of diabetes risk, logistic regression proved to be the more reliable analytical tool in terms of accuracy and model generalization.

Motivation

Diabetes mellitus is a disease where a human cannot properly process sugar in the body. There are two types of diabetes, type 1 and type 2. Type 2 can be acquired in adulthood and occurs when insulin receptors on cells don't respond to insulin, and don't allow glucose into cells. Type 1 diabetes occurs when the pancreas fails to produce insulin, so cells don't receive insulin, the signal that commands them to let glucose in.

Diabetes is a widespread problem with an average prevalence rate of 8.39% across the world's countries. The highest national rate is 30%. Diabetes by region is higher in pacific island small states, and smallest in central European and the Baltics. Bangladesh's national prevalence rate is 9.2%, which is within the world's 50% most common rates, from 5.6% to 10.4%. The United States is in the 50% most uncommon rates, at 10.8%.

Several complications arise as a result from diabetes. These are eye problems, chronic kidney diseases, cardiovascular diseases, and depression. In Bangladesh the per capita income is $2,100, and a typical household in that country can expect to pay $160 per year in diabetes treatment. So, for a single person household that amounts to 7.6% of annual gross income.

Prevention and early intervention is difficult. Up to 50% of diabetes patients go undiagnosed. For this reason, we propose the development of an easy tool for Bangladeshis to use so they can self-assess their risk of developing diabetes so they may avoid all of its consequences.

Exploratory Analysis

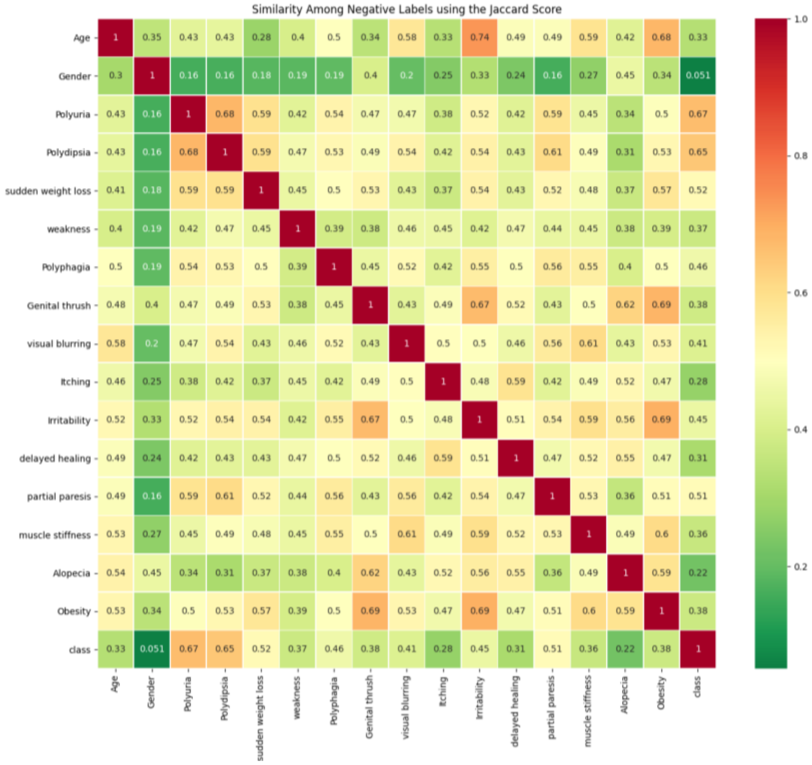
The data in this report consisted of 520 patients surveyed in Bangladesh for 16 traits and diabetes diagnosis for a total of 17 variables. We used diabetes diagnosis as our response variable in a binomial classification problem. The entire list of variables are age, alopecia, gender, genital, irritability, itching, obesity, polydipsia, polyphagia, polyuria, blurring, delayed, healing, loss, muscle, paresis, partial, stiffness, sudden, thrush, visual, weakness, and weight.

Our exploratory analysis began by trying to see the correlation between the variables. However, since all but one of the 17 variables were binary, a typical correlation matrix would not suffice. Calculating the Jaccard score for the positive and negative labels of all 17 variables allowed us to produce Jaccard Score similarity matrices. When predicting negative cases (Figure 1), polyuria and polydipsia have strong association with each other and class. Many other variables have moderately strong associations with each other. E.g., Lack of obesity is strongly associated with lack of genital thrush and irritability. When predicting positive cases (Figure 2), polyuria and polydipsia have moderate positive association with each other and class. Genital thrush and obesity have moderate to strong negative association with almost all the other variables. Interestingly, genital thrush and partial paresis are strongly anticorrelated.

To include age, our only continuous variable, in the similarity matrices, we decided to categorize it by choosing a threshold value. The decision to choose a threshold value was based on the age which maximizes the ROC area under the curve for all variables. First we plotted the ROC curve for Age in predicting each of the other 16 variables (Figure 3). Then we identified the optimum threshold for each curve (Table 1), and from those values picked the most common Age value (Figure 4).

One benefit of using Jaccard scores is that differences in weighted and unweighted Jaccard scores imply label imbalances. There are notable imbalances in the labels for genital thrush, irritability, and obesity. All three of these have noticeably different weighted and unweighted means for the Jaccard score (Figure 5). Indeed, for “Obesity” there were 432 negative labels, but only 88 positive labels. We don’t expect the imbalance in labels to have affected our analyses, however, since we stratified all training and testing sets.

Figure 1



positive association

no association

negative association

Figure 2

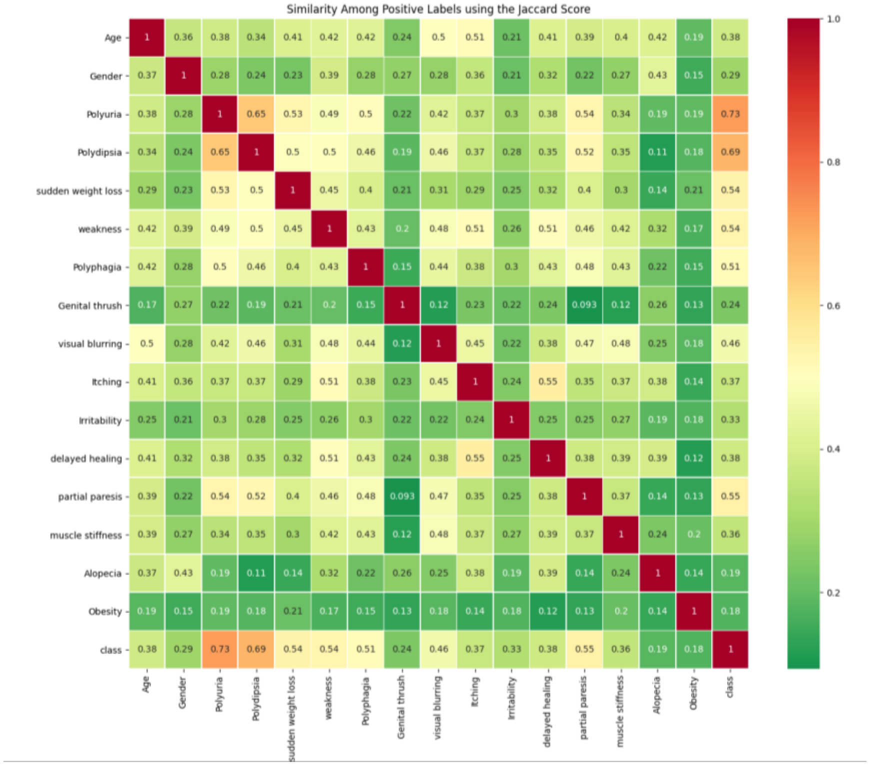


Figure 3

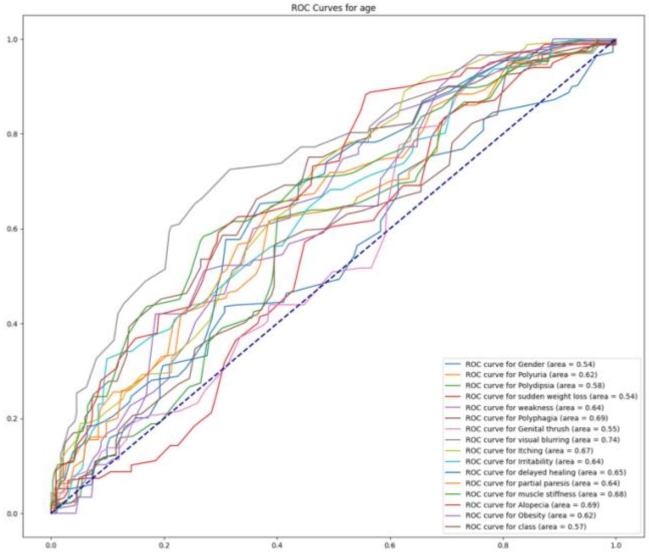


Figure 4

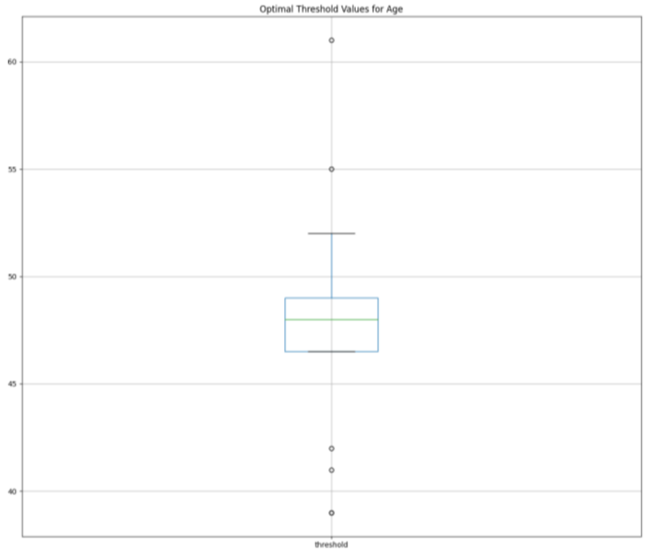
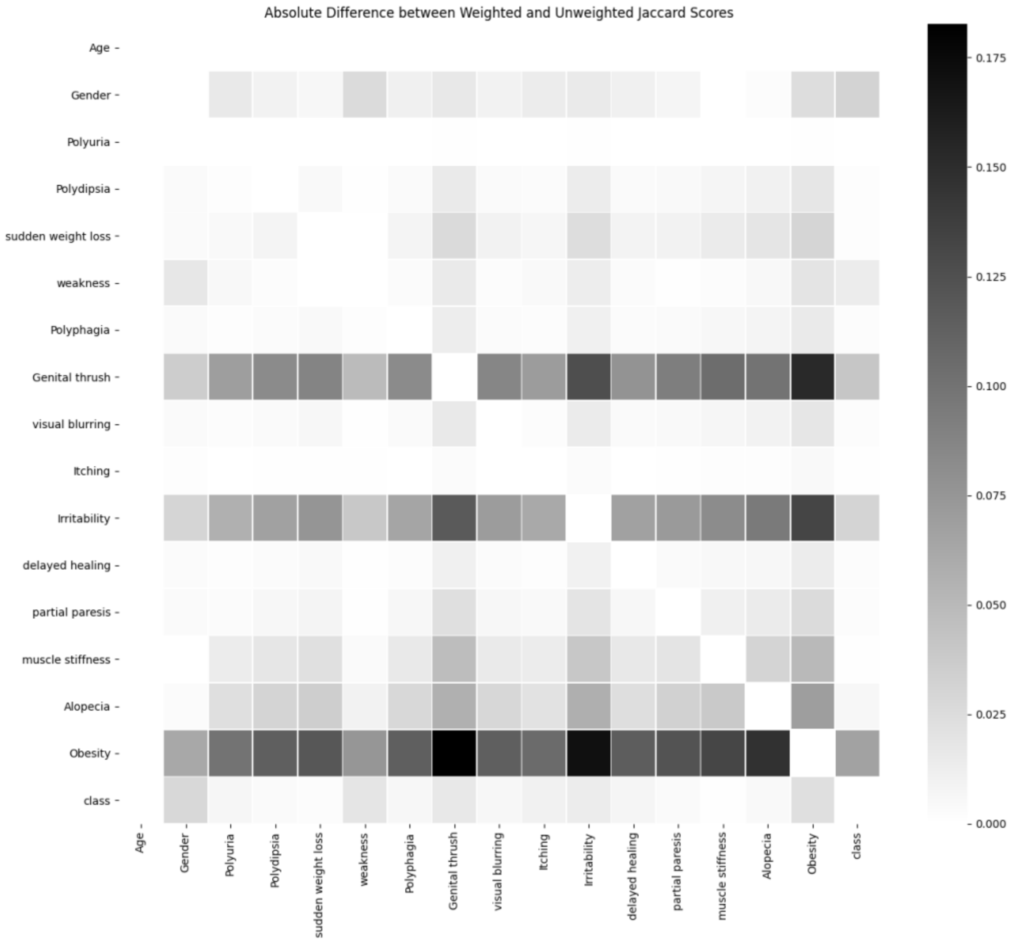


Table 1: Best Thresholds in ROC

	Age	False Positive Rate	True Positive Rate
Gender	52	0.31	0.44
Polyuria	48	0.39	0.62
Polydipsia	48	0.40	0.62
sudden weight loss	39	0.70	0.83
weakness	48	0.36	0.60
Polyphagia	48	0.36	0.66
Genital thrush	41	0.64	0.81
visual blurring	48	0.32	0.73
Itching	39	0.63	0.89
Irritability	61	0.10	0.33
delayed healing	48	0.37	0.65
partial paresis	48	0.39	0.65
muscle stiffness	52	0.27	0.58
Alopecia	42	0.56	0.88
Obesity	55	0.27	0.51
class	48	0.40	0.57

Figure 5



Analysis

Our analysis involved several rounds. Our first round involved determining the accuracies and ROC areas under the curve (AUC). To do this we performed 10,000 stratified samples from our data where 30% was allocated to the training set and 70% was allocated to the test set. For our random forest classifier, we used 100 trees and a max depth of 2. The 30-70 split was used to assess how good the models would be in generalizing classification. In this aspect logistic regression came out on top, performing better on the train and test sets (Table 2). This performance also translated to superior ROC AUC results (Table 3). The accuracy and ROC AUC values are reported with 95% confidence intervals based on the 10,000 samples.

The second round of analysis was to develop a decision tree graphic that patients could use to self-assess diabetes risk. To do this we changed the train-test split to 50:50. This resulted in accuracies and ROC AUC values (Table 4) almost approximating those of the first-round logistic regression model. The training and testing sets were sampled 1,000 times with stratification from our data to construct 95% confidence intervals for the accuracies and ROC AUC values, and the random forest classifier was set to one tree with a max depth of 8.

For feature selection we computed the feature importances from the random forest models. These feature importances are based on a GINI calculation, a concept related to entropy, which measures how separable the classes are when discretized by a certain feature. For example, a GINI score of 1 means perfectly inseparable, and a GINI score of 0 means perfectly separable. The feature importances and their 95% confidence intervals based on the 1,000 samples are reported in Table 5. Note that feature importances range from 0 to 1, like GINI values, but with 1 being most important, and 0 being least important. The reader will notice that many of the most important features in the random forest classifier were also those which had the most telling Jaccard scores.

A final result from our second-round analysis was the production of an example decision tree. We call this an example, because it is only one of the 1,000 such trees that were made from the samples. However, there is a 95% chance its model metrics are similar to those of its peers. Of note in the decision tree is that all final nodes have a GINI score of 0, indicating perfect separation. The tree also includes only up to 7 nodes, meaning that not all 16 features are needed to achieve perfect separation, and even less than the 8 allowed.

Table 2: First round accuracy with 95% confidence intervals

	Train Lower Bound	Train mean	Train Upper Bound	Test Lower Bound	Test mean	Test Upper Bound
Logistic Regression	0.943	0.944	0.944	0.908	0.908	0.908
BernoulliNB	0.885	0.885	0.886	0.871	0.872	0.872
Random Forest	0.907	0.907	0.908	0.888	0.888	0.889

Table 3: First round ROC AUC with 95% confidence intervals.

	Train Lower Bound	Train mean	Train Upper Bound	Test Lower Bound	Test mean	Test Upper Bound
Logistic Regression	0.943	0.944	0.944	0.907	0.908	0.908
BernoulliNB	0.890	0.890	0.891	0.875	0.876	0.876
Random Forest	0.905	0.905	0.906	0.886	0.886	0.886

Table 4: Second round random forest model evaluation with 95% confidence intervals

	Train Lower Bound	Train mean	Train Upper Bound	Test Lower Bound	Test mean	Test Upper Bound
Accuracy	0.958	0.959	0.960	0.903	0.905	0.907
ROC AUC	0.959	0.960	0.961	0.904	0.905	0.907

Table 5: Feature importance based on random forest GINI values with 95% confidence intervals

	Lower Bound	Mean	Upper Bound
Polyuria	0.217	0.223	0.228
Polydipsia	0.188	0.193	0.198
Gender	0.094	0.096	0.098
Age	0.091	0.092	0.094
partial paresis	0.056	0.058	0.061
sudden weight loss	0.052	0.054	0.056
Alopecia	0.038	0.039	0.040
Irritability	0.037	0.038	0.040
Polyphagia	0.031	0.032	0.034
delayed healing	0.030	0.031	0.032
visual blurring	0.027	0.028	0.029
Itching	0.027	0.028	0.028
muscle stiffness	0.023	0.024	0.024
weakness	0.022	0.023	0.024
Genital thrush	0.021	0.022	0.023
Obesity	0.018	0.019	0.020

Additional Analysis

Our initial analyses did not include feature selection as an element in our model construction strategy, so a post hoc feature selection analysis was performed using coefficient significance and logistic regression on our entire data's observations and features. The first round of regression showed that sudden weight loss, weakness, visual blurring, delayed healing, muscle stiffness, alopecia, and obesity were not significant (Table 6). The second regression, with the latter variables removed, showed that age was not significant (Table 7), and the final model is shown in Table 8.

Discussion

The random forest decision tree can serve as an easy and reliable tool for patient self-assessment for diabetes risk. Based on the model evaluation, it can be up to 90.5% accurate. The provided figure can be made even easier to use by using plain English. Note that the figure's "<= 0.5" really means "is the patient a negative for this trait?", since all traits are binary; 0 or 1. However, before a final product can be made, we should construct the model on our entire data using only the features deemed significant by our post hoc feature selection. The feature importances, like the features shown in Figure 6, are based on random samples from our data. So it's possible that some samples result in certain features appearing to be more important than they actually are.

Also, the 95% confidence intervals were also not correctly calculated since they were not done using the bootstrap method. Samples were picked without replacement with a size smaller than that of the data. The bootstrap method requires that samples be picked with replacement and of a size equal to that of the data.

Table 6: First round of feature selection

	Coefficient	Std. Error	$P(Z > \|z\|)$
Age	0.0077	0.011	0.494
Gender	-3.9342	0.58	0
Polyuria	4.1756	0.685	0
Polydipsia	4.7945	0.791	0
sudden weight loss	0.2796	0.534	0.601
weakness	1.0201	0.522	0.051
Polyphagia	0.9188	0.488	0.06
Genital thrush	1.9635	0.559	0
visual blurring	0.7881	0.605	0.193
Itching	-2.6443	0.635	0
Irritability	2.0675	0.577	0
delayed healing	-0.5594	0.559	0.317
partial paresis	1.1224	0.486	0.021
muscle stiffness	-0.9027	0.537	0.093
Alopecia	-0.3914	0.562	0.486
Obesity	-0.2002	0.547	0.714

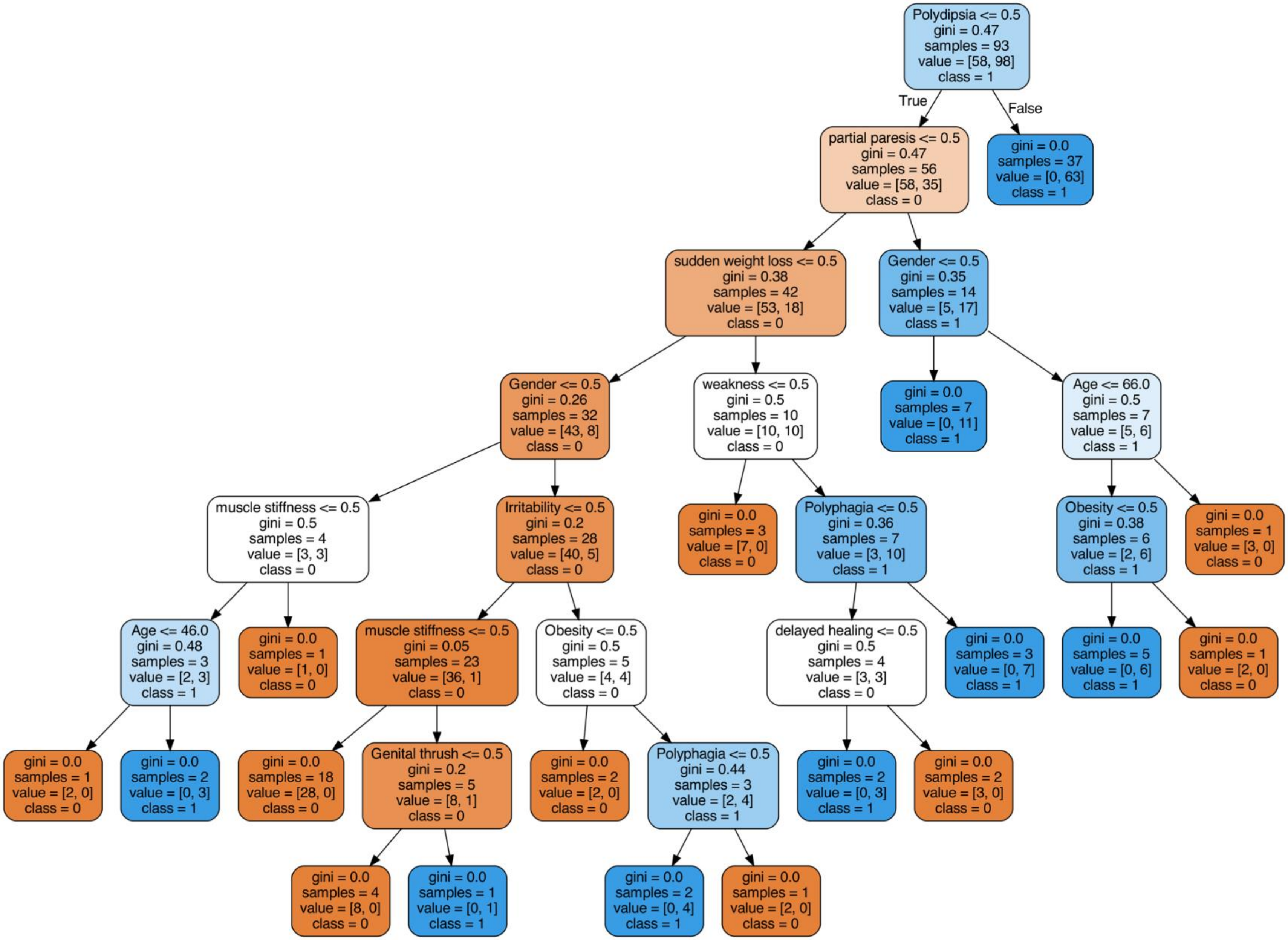
Table 7: Second round of feature selection

	Coefficient	Std. Error	$P(Z > \|z\|)$
Age	0.016	0.009	0.077
Gender	-3.9259	0.535	0
Polyuria	3.7213	0.528	0
Polydipsia	4.839	0.684	0
Genital thrush	1.7418	0.493	0
Itching	-2.417	0.497	0
Irritability	2.0679	0.507	0
partial paresis	1.2625	0.429	0.003

Table 8: Third and final round of feature selection

	Coefficient	Std. Error	$P(Z > \lvert z \rvert)$
Gender	-3.3991	0.43	0
Polyuria	3.5924	0.503	0
Polydipsia	4.6265	0.647	0
Genital thrush	1.788	0.488	0
Itching	-1.9727	0.415	0
Irritability	2.1138	0.503	0
partial paresis	1.5345	0.398	0

Figure 6: Example decision tree for patient self-assessment with 95% chance of having reported accuracy.


Polydipsia <= 0.5
gini = 0.47
samples = 93
value = [58, 98]
class = 1
True
False
partial paresis <= 0.5
gini = 0.47
samples = 56
value = [58, 35]
class = 0
gini = 0.0
samples = 37
value = [0, 63]
class = 1
sudden weight loss <= 0.5
gini = 0.38
samples = 42
value = [53, 18]
class = 0
Gender <= 0.5
gini = 0.35
samples = 14
value = [5, 17]
class = 1
Gender <= 0.5
gini = 0.26
samples = 32
value = [43, 8]
class = 0
weakness <= 0.5
gini = 0.5
samples = 10
value = [10, 10]
class = 0
gini = 0.0
samples = 7
value = [0, 11]
class = 1
Age <= 66.0
gini = 0.5
samples = 7
value = [5, 6]
class = 1
muscle stiffness <= 0.5
gini = 0.5
samples = 4
value = [3, 3]
class = 0
Irritability <= 0.5
gini = 0.2
samples = 28
value = [40, 5]
class = 0
gini = 0.0
samples = 3
value = [7, 0]
class = 0
Polyphagia <= 0.5
gini = 0.36
samples = 7
value = [3, 10]
class = 1
Obesity <= 0.5
gini = 0.38
samples = 6
value = [2, 6]
class = 1
gini = 0.0
samples = 1
value = [3, 0]
class = 0
Age <= 46.0
gini = 0.48
samples = 3
value = [2, 3]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
muscle stiffness <= 0.5
gini = 0.05
samples = 23
value = [36, 1]
class = 0
Obesity <= 0.5
gini = 0.5
samples = 5
value = [4, 4]
class = 0
delayed healing <= 0.5
gini = 0.5
samples = 4
value = [3, 3]
class = 0
gini = 0.0
samples = 3
value = [0, 7]
class = 1
gini = 0.0
samples = 5
value = [0, 6]
class = 1
gini = 0.0
samples = 1
value = [2, 0]
class = 0
gini = 0.0
samples = 1
value = [2, 0]
class = 0
gini = 0.0
samples = 2
value = [0, 3]
class = 1
gini = 0.0
samples = 18
value = [28, 0]
class = 0
Genital thrush <= 0.5
gini = 0.2
samples = 5
value = [8, 1]
class = 0
gini = 0.0
samples = 2
value = [2, 0]
class = 0
Polyphagia <= 0.5
gini = 0.44
samples = 3
value = [2, 4]
class = 1
gini = 0.0
samples = 2
value = [0, 3]
class = 1
gini = 0.0
samples = 2
value = [3, 0]
class = 0
gini = 0.0
samples = 4
value = [8, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 2
value = [0, 4]
class = 1
gini = 0.0
samples = 1
value = [2, 0]
class = 0