Diabetes Analysis and Prediction Report

Ben Weisenburger

**Executive Summary**

In this project, I decided to work with a dataset that concerned the results of a diabetes test amongst a group of individuals. My goal was to create a model that could predict the outcome—whether or not an individual would be diagnosed with diabetes, based on 8 variables in the dataset that serve as predictors. I used kNN to create this model, and logistic regression to analyze the impact weights of each variable. In conclusion, I created a model that was over 75% accurate and found that the biggest predictors of diabetes are Diabetes Pedigree Function test score, BMI, and number of pregnancies.

**The Problem**

The problem that I wanted to solve in my project was being able to create an accurate predictive model for the outcome of having diabetes based on the provided variables in the dataset, as well as to be able to identify which specific variables impacted the outcome. I wanted to create a predictive model that concerned individuals or patients could use to test the odds of diabetes based on their inputs for the specific variables, as well as inform these individuals about the variables that had the most positive increase on the odds of diabetes, so that these individuals in turn could work on improving their health in respect to these specific variables.

**Technique**

To create the predictive model for this project, I used the k-Nearest Neighbors supervised machine learning method. I used this method because it is very strong and viable option for classification, in this case predicting 0 or 1, the outcome for diabetes. This method takes all variables into effect and results can be visualized through a confusion matrix that is easy to explain and understand for the average individual or executive in the form of a presentation. To find out which variables were impacting the outcome the most, I ran a logistic regression test via Analytics App, an app created by my Marketing Analytics professor, to determine the impact on the odds of the outcome based on each variable. I used logistic regression because it can easily be expressed in terms of odds and is very easy to understand for the typical patient or user who would be completing this diabetes analysis/prediction, if they were curious as to which variables are affecting their odds of diabetes the heaviest, which would in-turn allow them to alter their behavior/health based on these specific variables.

**Conclusions**

To conclude the results yielded from the kNN model that I created in R, I can describe the model as having an accuracy of slightly over 75% (75.16%). These means that my predictive model can guess whether or not an individual will have diabetes based on the 8 predictors with an accuracy of 75.16%. My model predicted 115 of 153 instances correctly. After my presentation in class, you suggested that I add a couple additional metrics from the confusion matrix in addition to accuracy. I have calculated precision, or TP/TP+FP, and my model is 67.44% precise. The negative predictive value of my model, or TN/TN+FN, is 78.18%.

To conclude the results of my logistic regression output, it is clear to see that the top 3 predictors, in order, are Diabetes Pedigree Function test (one unit increase in this variable increases odds of being diagnosed with diabetes 157%), Pregnancies (one unit increase in this variable increases odds of being diagnosed with diabetes 13%), and lastly BMI (one unit increase in this variable increases odds of being diagnosed with diabetes 9%). It is interesting to note that none of the variables had a significant decrease on the outcome of diabetes, however there were several (Glucose, Blood Pressure, Skin Thickness, Insulin, and Age), that did not affect the outcome in a significant statistical way.

Those who are concerned about possibly having diabetes can simply plug their values of these 8 predictors into my model (which can all be acquired with one simple doctors visit/blood test to acquire glucose levels, insulin levels, etc., as well as to be scored on the Diabetes Pedigree Function test). This predictive model, while requiring a little work, is significantly less stressful and invasive than actual tests for diabetes, not to mention it is more time efficient. Based on the results of this project, I strongly recommend concerned individuals to check their family genetic history of diabetes through receiving a Diabetes Pedigree Function test, as well as strongly recommend these individuals to keep a healthy BMI (weight/height proportion).

**Appendix**