Running head: THE DEMOGRAPHICS OF DIABETES

The Demographics of Diabetes

Jessica Everett

Bellevue University

November 14, 2020

Abstract

The Pima Indians of Arizona are a tight knit community that is genetically isolated. They also have the highest diagnosed cases of type 2 diabetes than any other population group in the world (Baier, 2004). I use a dataset of medical information about the Pima Indians to determine if demographics and health statistics can identify people at higher risk of diabetes. This algorithm can help medical professionals identify at-risk individuals for early intervention and prevent possible development into type 2 diabetes.

The Demographics of Diabetes

Type 2 diabetes has symptoms that can mimic other diseases (fatigue, blurred vision, numbness in the extremities, etc.) (NIH, n.d., Diabetes). Genetics can play a part in developing the disease, but so can lifestyle factors like obesity and lack of exercise. 21% of adults with diabetes are undiagnosed (American Diabetes Association, n.d.). This costs the United States economy $18 billion a year (Radcliffe, 2018). I would like to answer this question: can demographics and medical data find type 2 diabetes and pre-diabetes cases for early intervention?

Data Source

I used a dataset that includes health and demographic information on the Pima Indians (Kaggle, 2016). The dataset only includes data for females over 21 years old. The dataset includes 9 columns and 768 rows. The columns are number of pregnancies, blood glucose level, blood pressure, skin thickness (triceps skin fold thickness in millimeters), blood insulin level, BMI (body mass index – weight in kg/(height in m)^2), diabetes pedigree function (family history/genetics), age, and diabetic status.

Methodology

I replaced zeroes with the median of each category because they didn't make sense in the context (skin thickness can't be 0, among others). I started with some basic exploratory analysis, including histograms of all columns:

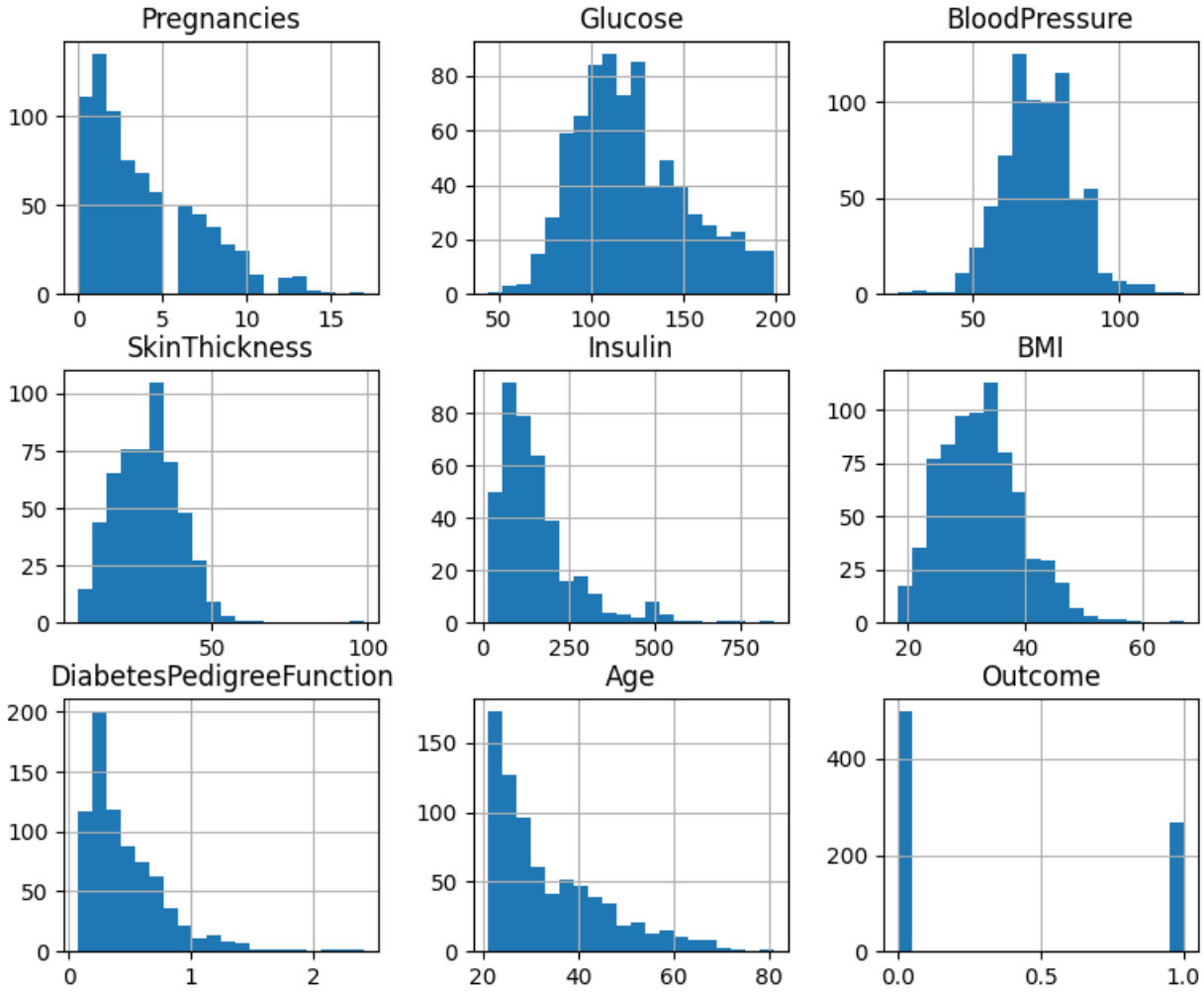

Figure 1 – Histograms of dataset

I created a correlation heatmap that shows that blood glucose has the highest correlation to outcome:

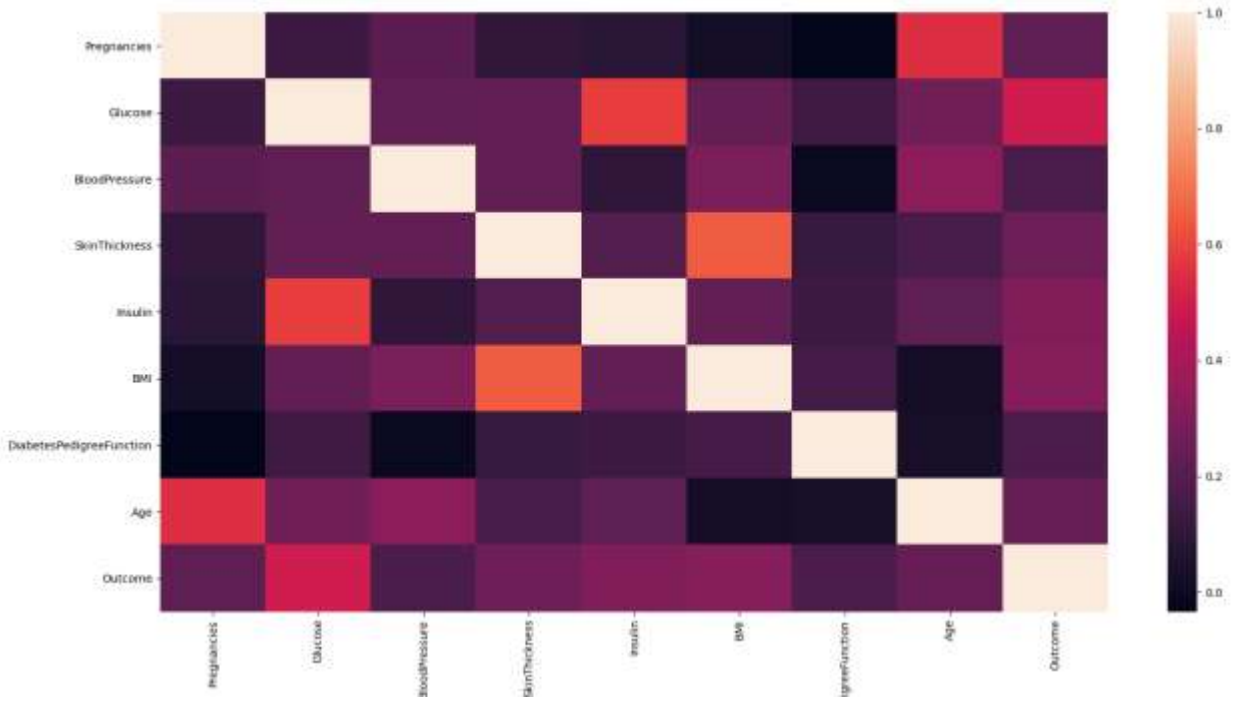

Figure 2 – Correlation Heatmap

I then created new features to aid in analysis. BMI_CAT describes the BMI in categories (underweight, healthy, overweight, and obese). INSULIN_CAT describes if the blood insulin level is normal or abnormal. I then created a new heatmap with the new categories:

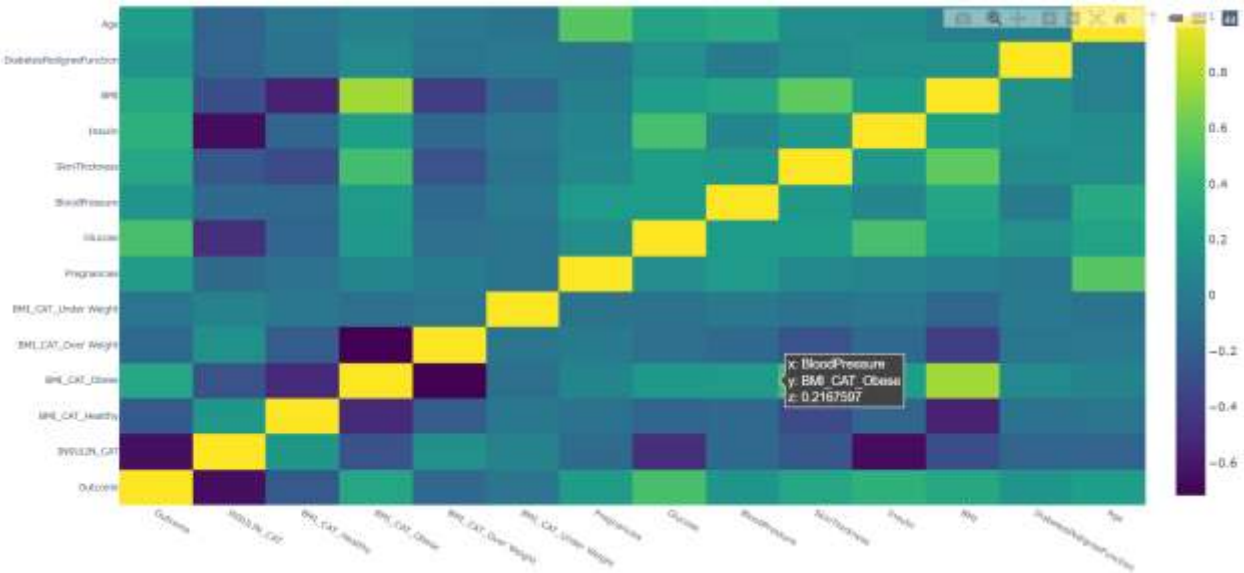

Figure 3 – Correlation Heatmap with Features

This new heatmap shows outcome has a low correlation with INSULIN_CAT and a high correlation with BMI_CAT Obese, glucose, insulin, and BMI.

Using code from Avinash, I tested 10 different models with the data, including linear regression, KNN (K-nearest neighbor), random forest, and GBM (Gradient Boosting Machine) (2020). The GBM model had the highest accuracy at 89%:

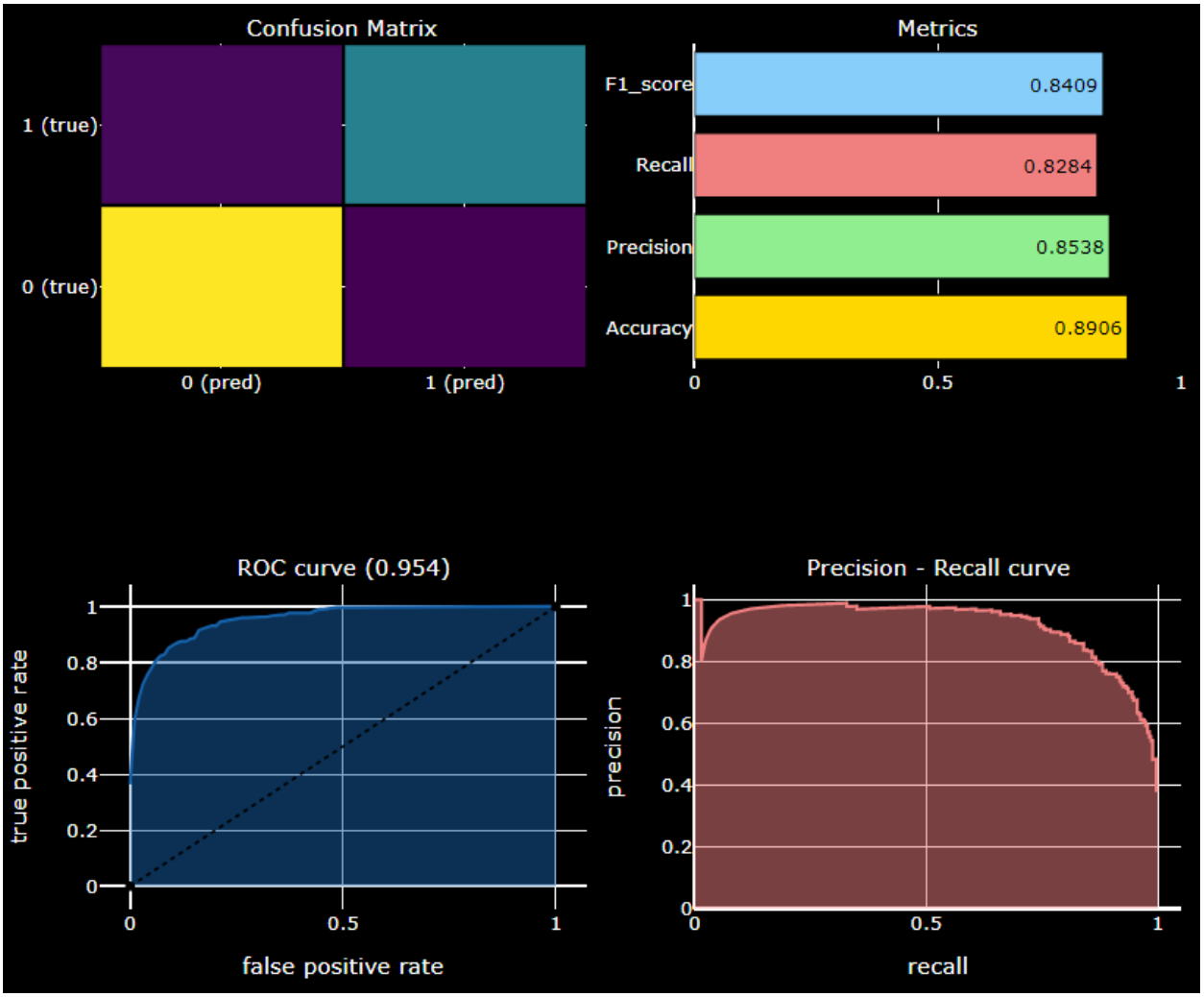

Figure 4 – GBM model performance (5 folds)

Conclusion

The GBM model returns 89% accuracy. As the data only includes females over 21, this model can only be used for that demographic. Further research needs to be done as to whether this model will work for the general population of America, or just for the Native American population, who has a higher rate of diabetes. If this model holds true to the general public, it can be used to identify diabetic and pre-diabetic patients earlier for intervention.

References

American Diabetes Association. (n.d.). Statistics About Diabetes. Retrieved October 31, 2020, from https://www.diabetes.org/resources/statistics/statistics-about-diabetes#:~:text=Overall%20numbers%201%20Prevalence%3A%20In%202018%2C%2034.2%20million,million%20seniors%20%28diagnosed%20and%20undiagnosed%29.%20More%20items...%20

Avinash, L. (2020). End to End ML Project. Retrieved October 31, 2020, from https://www.kaggle.com/avinashlalith/end-to-end-ml-project

Baier, L.J. & Hanson, R.L. (2004). Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians. Retrieved October 31, 2020, from https://diabetes.diabetesjournals.org/content/53/5/1181

Britannica. (n.d.). Pima. Retrieved October 31, 2020, from https://www.britannica.com/topic/Pima-people

Kaggle. (2016). Pima Indians Diabetes Database. Retrieved October 31, 2020, from https://www.kaggle.com/uciml/pima-indians-diabetes-database

NIH. (n.d.). Diabetes. Retrieved October 31, 2020, from https://www.niddk.nih.gov/health-information/diabetes

NIH. (n.d.). Health Statistics. Retrieved October 31, 2020, from https://www.niddk.nih.gov/health-information/health-statistics#diabetes

Radcliffe, S. (2018). Diabetes Care Could Top $336 Billion by 2034. Retrieved October 31, 2020, from https://www.healthline.com/health-news/diabetes-could-top-336-billion-by-2034

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

WebMD. (n.d.). Type 2 Diabetes Causes and Risk Factors. Retrieved October 31, 2020, from https://www.webmd.com/diabetes/diabetes-causes