DSC 530 – Term Project Write-Up

Hanh Tran

I chose to explore for myself, a now famous dataset in terms of EDA, Diabetes: Pima Indian. Based off of the dataset, I wanted to see which variables has the highest influence or correlation to being able to diagnose a diabetic and how accurate would that prediction be? I wanted to see whether the effects we see in this sample are likely to appear in the larger population. An EDA showed that a couple descriptive variables have a significant correlation to the outcome variable of being diabetic, specifically Glucose level and BMI, with Glucose as the most influential factor in a diabetes diagnosis. The distribution for these variables where only slightly skewed, almost approximately normal so the analytic model was able to capture, for the most part, the relevant aspects of the diabetes data. By quantifying the size of the variable’s effect on outcome through a test statistic, testing a correlation. The outcome of the EDA showed that the Glucose variable is statistically significant at p-value < 0.05 in predicting the outcome of a diabetic. Some of the issues in the dataset is limited as the dataset contained variables that were from only women and are 21 and older. I would have like to take a look at preventative variables that could have helped the diagnosis of diabetes pertaining to nutrition or vitamin deficiency levels as poor nutrition is often the case with a nation with an alarmingly increasing number of diabetics. Income per household might be an interesting variable as it typically indicates access to better resources of food and nutrition. However, it looks like this dataset is from specifically Pima Indian women so it may be from a native American reservation where the community as a whole may be responsible for a self-sustaining food source. One of the assumptions I made that I felt were incorrect pertains to insulin as a factor influencing diabetes. When I ran the initial correlation matrix on the original dataset without cleaning out the zeros or NULL values on insulin, the correlation coefficient showed a lower value than other variables. However, I think this variable insulin could have potentially a higher correlation coefficient after cleanup because it contained the most NULL zero values. Additionally, analyzing insulin may have helped with differentiating between Type 1 and Type 2 diabetes since that information did not seem to be explicitly stated. Some of the challenges I faced were recognizing which packages to use that could have made the calculations easier. Other challenges include determining how to handle outliers as it requires some domain knowledge in diabetes and human health in general such as 17 pregnancies.