Diabetes Mellitus Type II Predictions

Olivia Plamann

Millions of adults in the United States are at risk for Diabetes Mellitus Type II (Type II Diabetes), a chronic condition affecting how the body processes sugar. Because of the longevity of this condition, researchers spend time identifying factors correlated with Type II Diabetes. With my project, I hope to further identify these factors that may lead to Type II Diabetes so individuals could alter their habits in hopes of avoiding this chronic condition.

This data was created by Dr. John Schorling at the University of Virginia and web scraped from https://biostat.app.vumc.org/wiki/pub/Main/DataSets/diabetes.html. Dr. Schorling conducted a study interviewing 403 patients who were African American in central Virginia to better understand the prevalence of Type II Diabetes. The 403 participants were screened, and if their glycosylated hemoglobin was greater than 7.0, a positive diagnosis for Type II Diabetes was received.

I selected six features for my analysis on the patients' glycosylated hemoglobin levels: (1) age, (2) waist-to-hip ratio, (3) stabilized glucose level, (4) weight-to-height ratio, (5) the ratio of High-Density Lipoprotein (HDL) to overall Cholesterol, and (6) gender. There were only two gender options recorded in the study: male and female. I was able to compute the waist-to-hip and weight-to-height ratios from the original dimensions provided in the dataset.

I start exploring these features in **Figure 1**, where I show participants' stabilized glucose level versus HDL cholesterol ratio. The HDL to cholesterol ratio tends to be fairly consistent across all patients, varying from about 2.5 to 10 in males and 2.5 to 7.5 in females; however, there is one female outlier with a ratio of 19.3. This figure also shows that patients with lower stabilized glucose levels also tend to have glycosylated hemoglobin levels below the 7.0 threshold for Type II Diabetes, corresponding to the blue data points on the graph. Most of the individuals in this study with glycosylated hemoglobin levels below 7.0 have stabilized glucose levels of about 100; however, some males also have levels up to 200. Both the males and females with glycosylated hemoglobin levels above 7.0 have a wider range of stabilized glucose levels, from around 100 to upwards of 400 for males and 300 for females.

I continue my exploration with an analysis of how the number of clusters in the data impacts the inertia, the average squared distance from the data points to the nearest centroid. **Figure 2** shows as I increase the number of clusters from one to two, the inertia decreases by about 345,072. Again, as I increase the number of clusters to three, the inertia decreases by about 52,897. The inertia decreases substantially less as the number of clusters increases past three, leveling out at about 16,000 when there are twelve clusters, so three clusters will be a good assessment of this data. When thinking of the context of this data, one cluster could be individuals with diabetes and another could be those without, but what would be the third? Future analyses could aid in answering this question.

In order to use the six features to predict whether an individual has diabetes or not, I made a new column that obtained values of True if the individual's glycosylated hemoglobin level was greater than the 7.0 threshold and False otherwise. I then used a sklearn pipeline with a OneHotEncoder transformation on gender to predict with sklearn's LogsiticRegression whether a patient has diabetes or not. I used this model after splitting my dataset to be 75% training data and 25% testing data. This model has an accuracy of 87.9% with a false-negative rate of 14.3%. False-negative predictions from my model are arguably the most detrimental errors to make in the case of diabetes diagnosis, so although 14.3% is not terribly high, I would hope to decrease this further in future models.

The coefficient weights for each feature in my model are represented in **Figure 3**. The male gender has the smallest impact with a weight of -0.009. On the other hand, the waist to hip ratio and age features have the greatest impact in this model with weights of -0.42 and 0.42 respectively. Since the ratio weight is negative, a greater ratio of High-Density Lipoprotein Cholesterol is correlated with a lower glycosylated hemoglobin level and thus not having diabetes. Age has a positive coefficient, so a greater age value corresponds with a higher glycosylated hemoglobin level and thus having diabetes.

In all, I analyzed how six features impact Type II Diabetes in individuals. I found that higher stabilized glucose correlates with higher glycosylated hemoglobin levels. I also found that there were three main clusters in my data. Taking a further look into these clusters may provide significant insight in future analyses. I finished out my analysis by using the features to predict whether an individual has diabetes, which could be a useful tool in predicting whether an individual has Type II Diabetes.

Figure 1: HDL Cholesterol Ratio by Stabilized Glucose and Gender

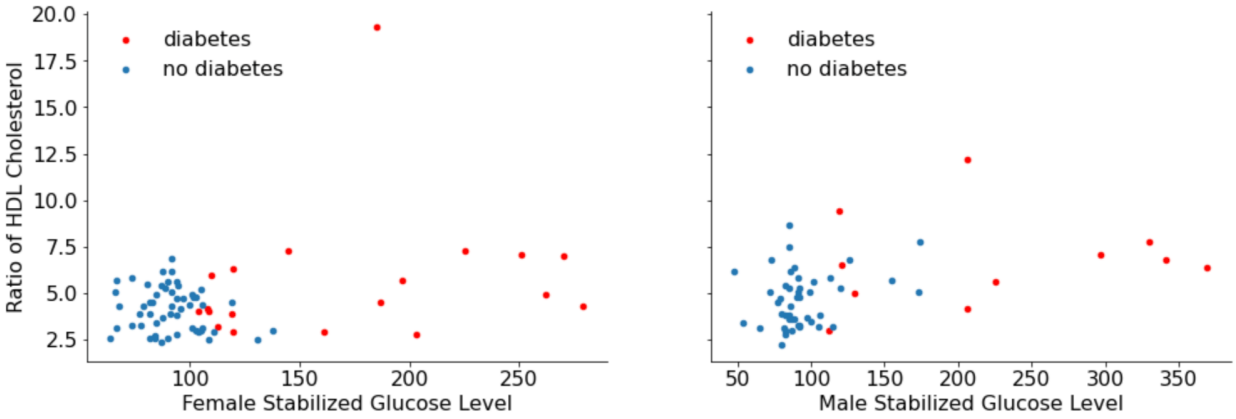

Figure 2: Inertia by Number of Clusters

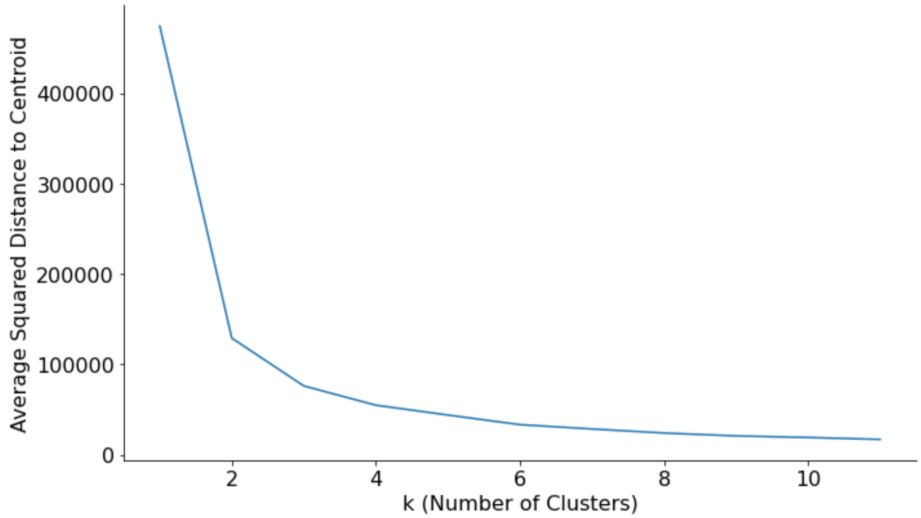

Figure 3: Logistic Regression Coefficients