Caleb Carr

Professor Eleish

ITWS 4600

10 December 2024

NHANES Analysis for Predicting Diabetes Diagnosis

During winter break of my first year of college, I was diagnosed with type I Diabetes, and not knowing the signs and symptoms to look for, I was quite shocked by it. When taking COGS 4210: Cognitive Modeling a little over a year after my diagnosis, for a final project a classmate used neural networks to analyze data from the National Health and Nutrition Examination Survey (NHANES) to do diagnosis prediction for a few different conditions. It certainly stimulated my interest in the subject so given the new techniques that I have learned in the last few months, I felt that a deep-dive into diabetes diagnosis would be insightful. NHANES data spans a broad range of information, ranging from demographics to questionnaire data about physical activity to bloodwork results from laboratory analysis. Given all of these factors, I figured that there must be some way to determine whether a person had a diabetes diagnosis, other than their A1C, which is the leading factor to most diagnosis[1]. My predictions were that, of course, A1C values would have the highest impact on this diagnosis but that there would be other factors, mainly those in the categories of demographics, physical activity, nutrition, and income, that would lead to an accurate binary classifier of the data.

NHANES was a clear choice for a dataset because of the sheer amount of data that they collect, having about 100 sub-datasets in 6 categories; I would be able to aggregate hundreds of attributes and assign them to a single person identified by a numerical sequence. The dataset that

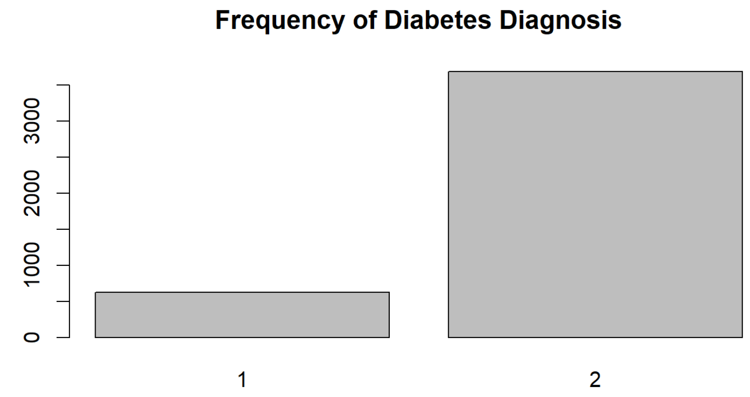
[1] https://diabetes.org/about-diabetes/a1c

I focused on was the 2017-Pre Pandemic collection[2], combining information from several sub-datasets, including Demographics[3], Diet[4], Blood Pressure and Cholesterol[5], Body Measures[6], Diet Behaviour[7], Income[8], Physical Activity[9], Weight History[10], Glycohemoglobin[11] (A1C), and of course, Diabetes[12]. This resulted in an original count of a total of 358 variables, though after removing all variables such that more than 33% or more of the values were null, and only including rows such that the subject's diabetes diagnosis was known, the variable count was reduced to 90 numeric variables and 78 categorical variables. Many of the selected sub-datasets were from questionnaire data, and participants were able to leave questions blank, sometimes because they were not given every question, and other times because they did not know the answer or refused to respond. Many of these null values stemmed from questions on a questionnaire that led to other questions depending on the answer selected, for example, the question "Were you ever breastfed" leads to "At what age did you stop breastfeeding" if yes but "How many days old were you when you were first fed formula" if no, so the variable DBD030 - Age stopped breastfeeding(days) had thousands of null values.

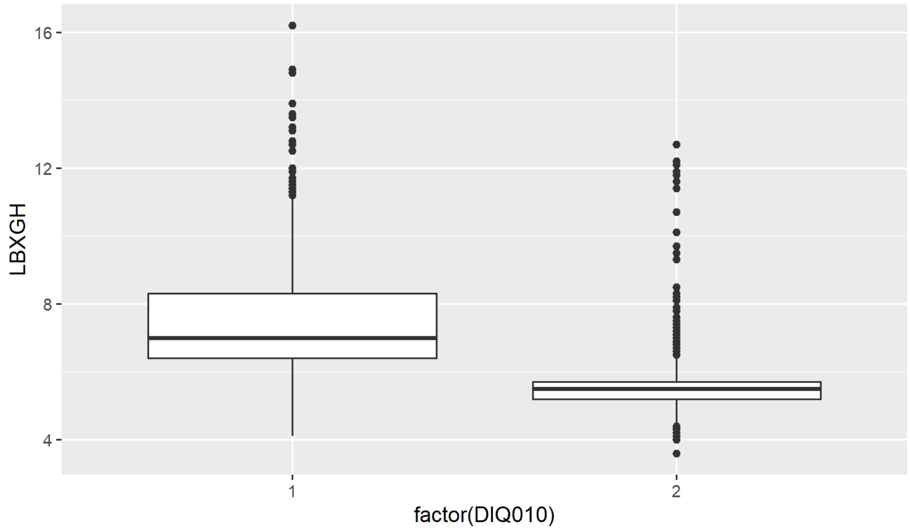
After this some preliminary analysis was done to see how skewed the data was. A small part of the population has diabetes, so it would be expected that national survey data would reflect this, and that the data would be skewed. With a positive diagnosis being the 1 case, and a

[2] https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020
[3] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DEMO.htm
[4] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DR1TOT.htm
[5] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_BPQ.htm
[6] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_BMX.htm
[7] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DBQ.htm
[8] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_INQ.htm
[9] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_PAQ.htm
[10] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_WHQ.htm
[11] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_GHB.htm
[12] https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_DIQ.htm

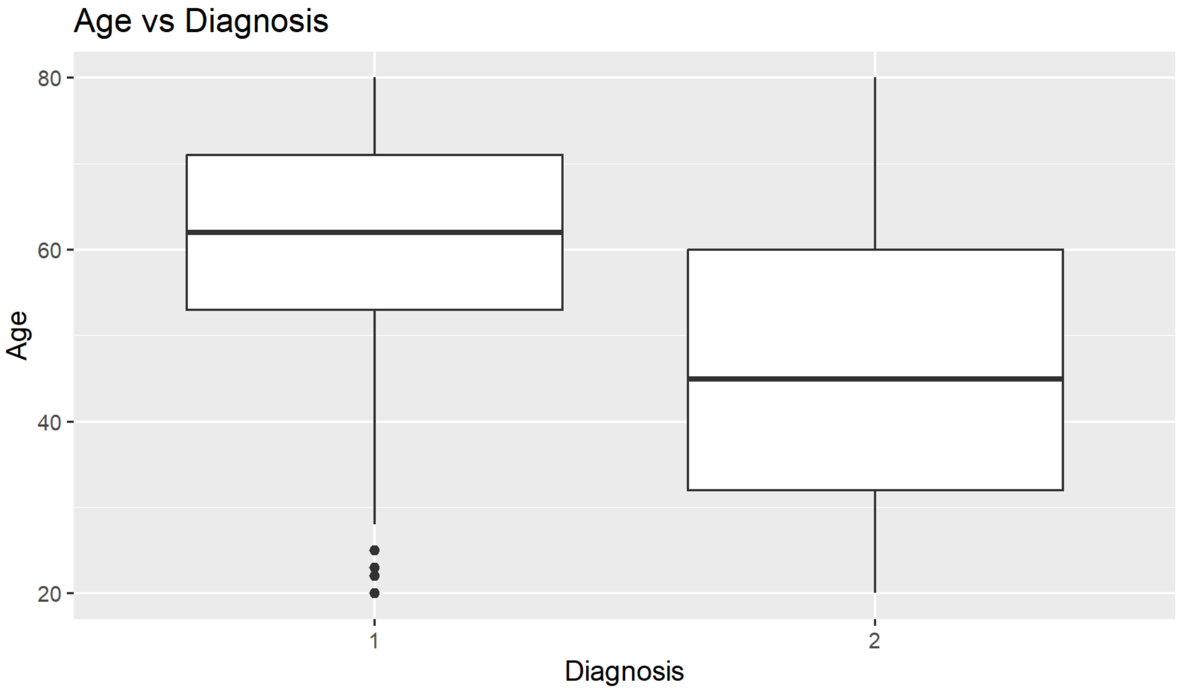
non-diagnosis being the 2 case, it can be clearly seen that the data has a large skew towards non-diabetics, and it will likely need to be mitigated during future analysis.



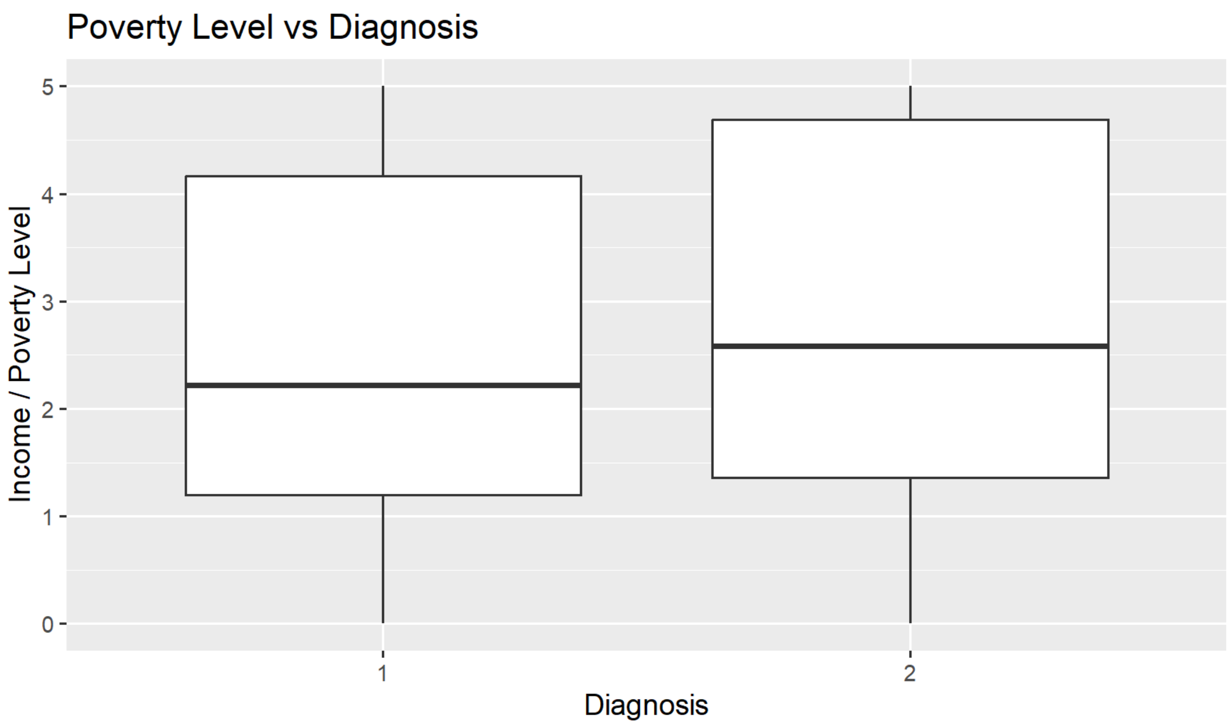
Also, given that it is known that A1C is used in most diagnoses, it is worth looking to see what the distributions are for the diabetic and non-diabetic cases. After performing this analysis, it can be seen that there is a clear difference, so much so that any point in the IQR for diabetic would be considered an outlier for the non-diabetic case.



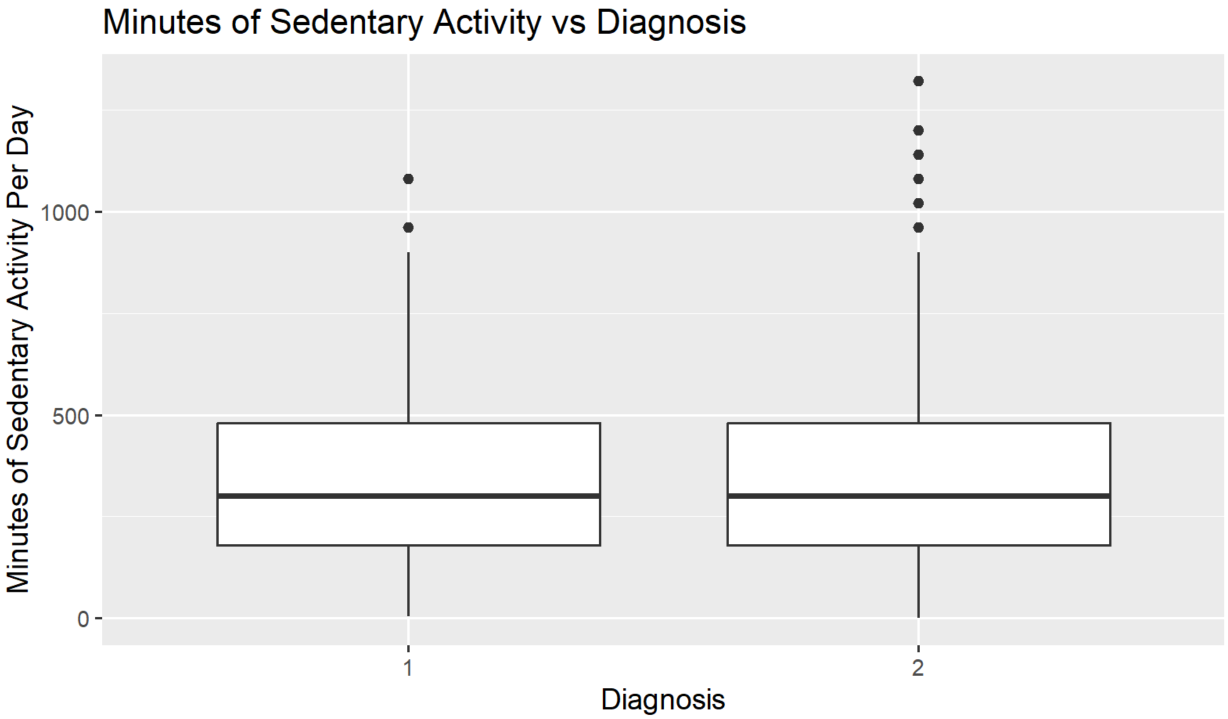
While this is useful information, it is worth looking at the data to see if there are any other factors that play a role, so that one does not need to know their A1C to determine if they are at a higher risk factor. Due to the goal of the analysis being to create a binary classifier, a linear regression on different variables doesn’t quite work, so boxplots make the most sense to compare different distributions. To test my initial predictions about variables that are led from or lead to a diabetes diagnosis, box plots were made for several numerical variables from the areas of physical activity, age, weight, and nutrition (Note that 1 is Diabetes = True and 2 is Diabetes = False).



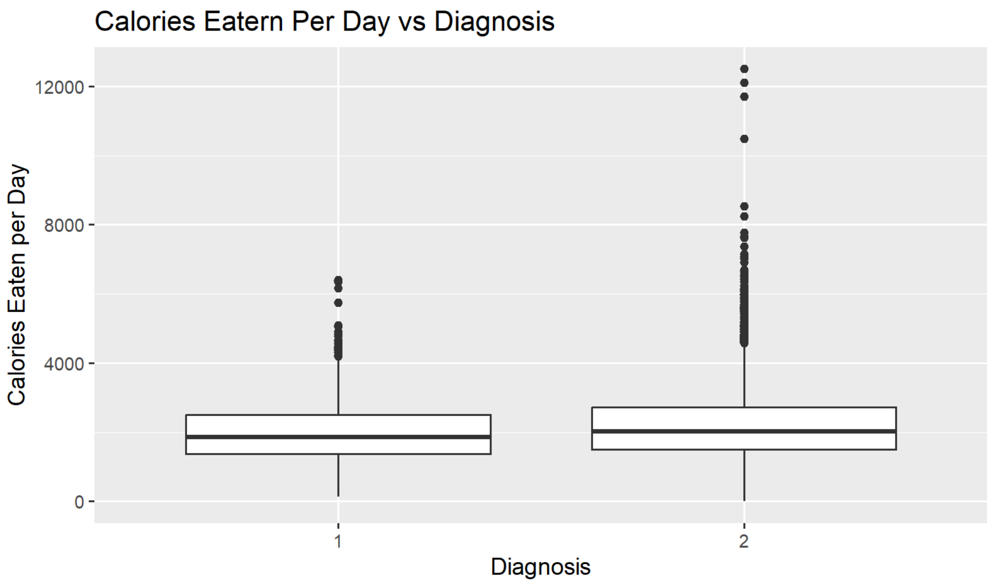
With a p-value < 2.2e-16, it can be confirmed that these do not come from different distributions, and that it can be reasonably assumed that age is correlated with diagnosis. This, at the most basic makes sense because as you age, you have more time to get diagnosed.

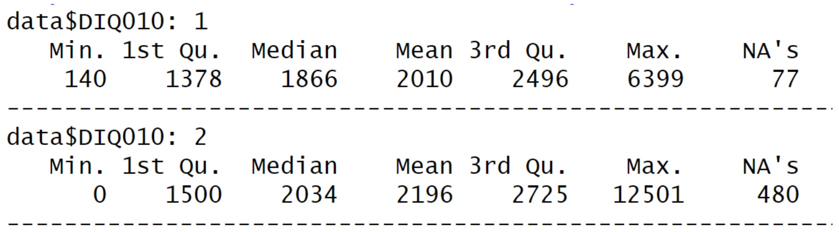


In this case, while not as drastic as the age factor, a small difference can be observed between the distributions with people with a positive Diabetes diagnosis being more likely to fall closer to the poverty line than those without a diabetes diagnosis. One may theorize that this difference is due to people with lower income being forced to eat a less nutritious diet than their richer counterparts. This difference in distribution is supported by a p-value of 0.004563.

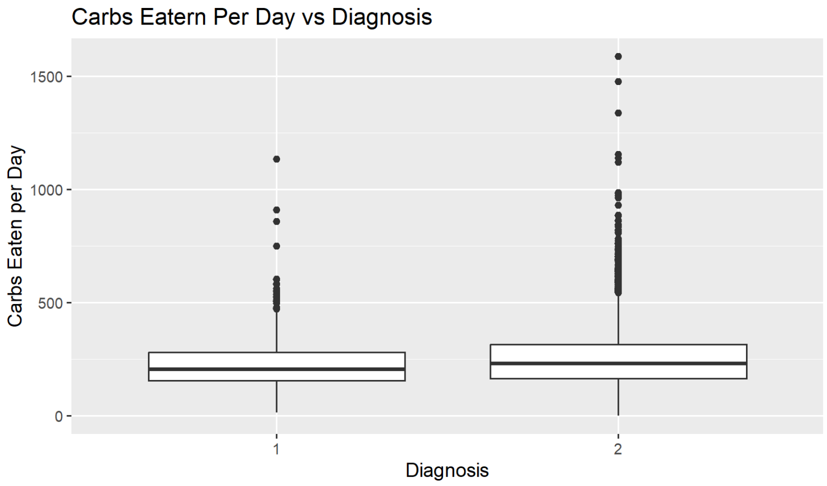


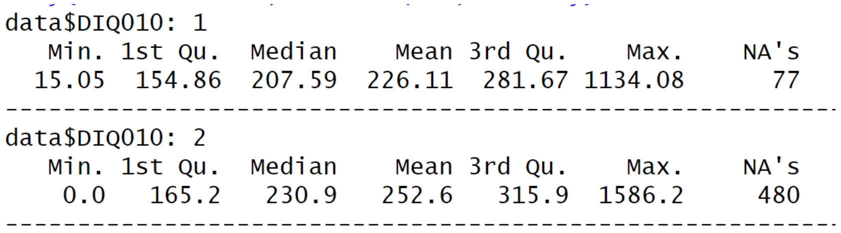
While these boxplots look very similar (they have the same 1st and 3rd quartiles, the distributions are slightly different, with people with Diabetes=True having about 23 more minutes a day of sedentary minutes per day: 361.8 vs 338.2 on average. While according to the wilcox.test, this is statistically significant with a p-value of 0.003496, I find this surprising.



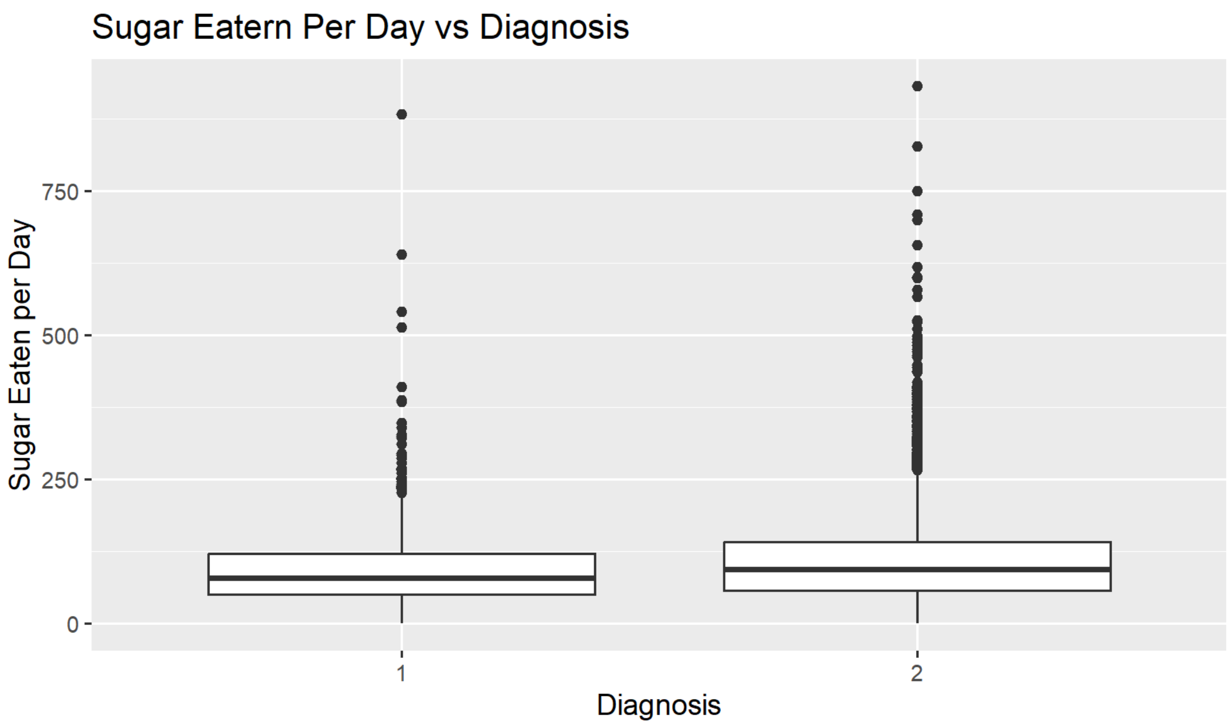


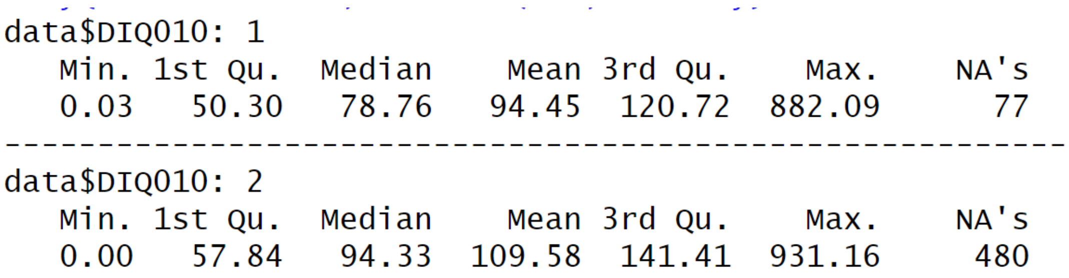
```
data$DIQ010: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    140    1378    1866    2010    2496    6399      77 
---------------------------------------------------------
data$DIQ010: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
      0    1500    2034    2196    2725   12501     480 
---------------------------------------------------------
```





```
data$DIQ010: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  15.05  154.86  207.59  226.11  281.67 1134.08      77 
---------------------------------------------------------
data$DIQ010: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    0.0   165.2   230.9   252.6   315.9  1586.2     480 
---------------------------------------------------------
```

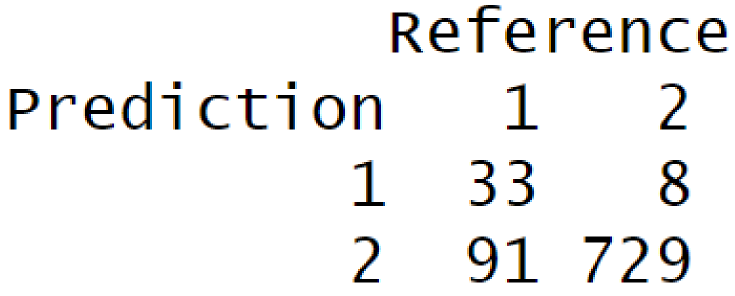




```
data$DIQ010: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.03   50.30   78.76   94.45  120.72  882.09      77 
---------------------------------------------------------
data$DIQ010: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   57.84   94.33  109.58  141.41  931.16     480 
```

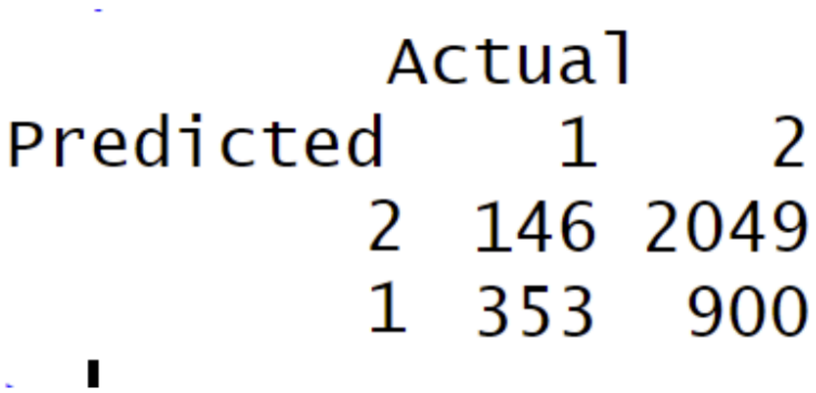
One may associate diabetes with people who eat too much or eat too unhealthy of a diet. However, the NHANES data would suggest otherwise. When looking at calories, carbohydrates, and sugars eaten per day, it can be seen that people with diabetes are actually on average having less of all three than their non-diabetic counterparts. The difference is clearly significant, with the p-value for each of the calories, carbs, and sugar being 3.866e-07, 2.825e-09, and 1.387e-10, respectively. This is one of the cases where I believe that the factors being different is not what caused the diagnosis but is in fact caused by the diagnosis. People with diabetes need to watch their sugar and their carbs, and often type 2 diabetics are able to reverse some of their symptoms through eating a healthier diet. Due to this, it makes sense that people who did have a positive diabetes diagnosis would be eating “better” than non-diabetics.

Once I had looked at some individual variables, all the numerical data in the dataset was run though models for binary classification and clustering. The first clustering method attempted was K-Nearest Neighbours. Skew was certainly an issue, with the model having a true negative rate of only 36%, however the true positive rate was 98.9%, with an overall accuracy of 88.5%, which is only slightly better than random chance, which would lead to an accuracy of 85.6%.



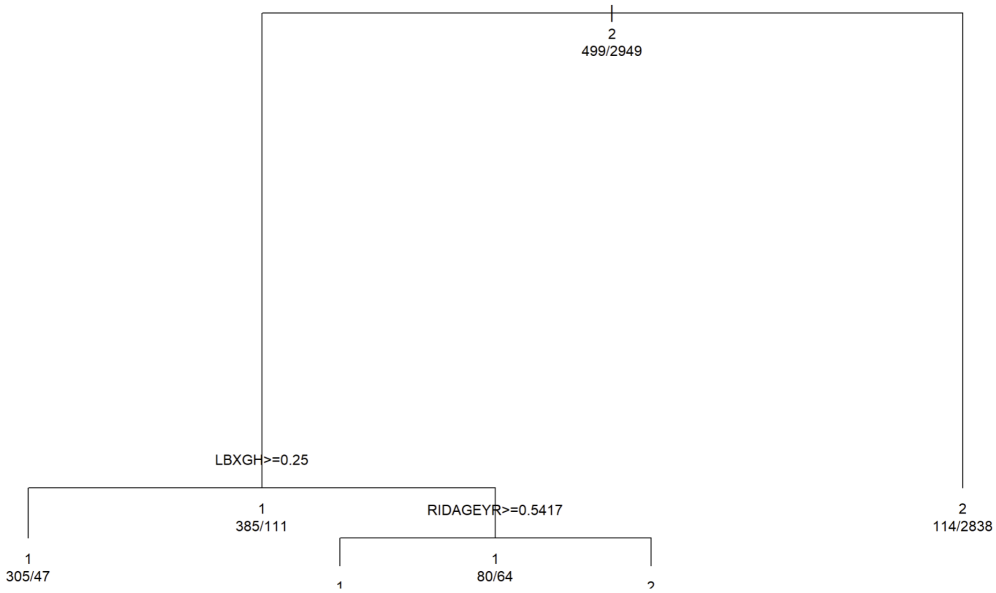
```
          Reference
Prediction   1   2
         1  33   8
         2  91 729
```

Following this, I ran a K-Means Clustering model on the numerical attributes and received worse results, with a true negative rate of 29.2%, true positive rate of 96.4%, and overall accuracy of 67.7%, worse than guessing. This suggests that clustering may not be the best method for classifying the data in this case. Given the number of attributes, A1C, Age, or other variables would not stick out as much.

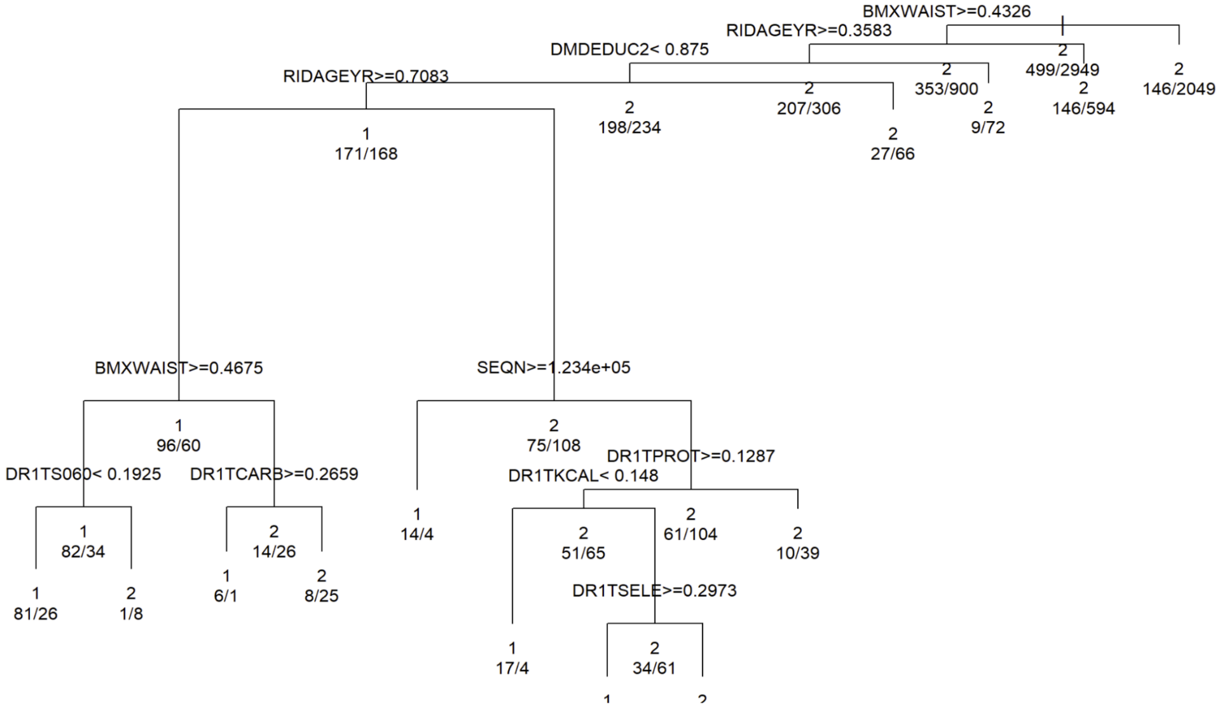


```
         Actual
Predicted    1    2
        2  146 2049
        1  353  900
```

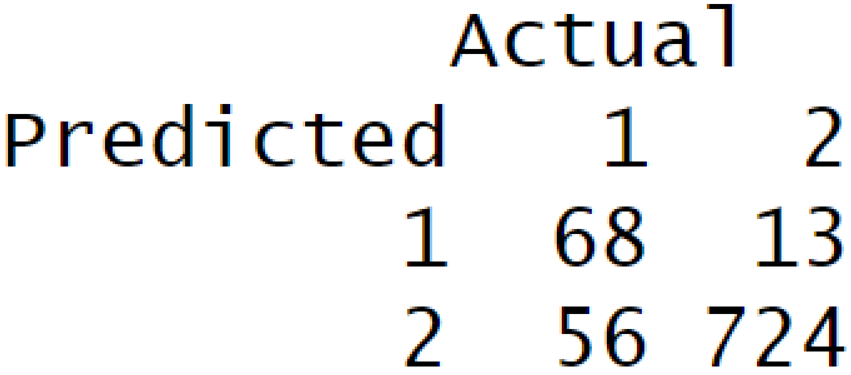
At first, because running random forests would take too long for the hundreds of variables, it was decided that creating classification trees would be the ideal way to select the most important variables in assessing the helpfulness of variables in classifying a patient as diabetic or not. The trees most often went based off of A1C and then based on age. This was still about as inaccurate as the K-Means model, but it was simple, so it did reveal that A1C was in fact an important factor in a diagnosis. Specifically, the A1C value of 6.15 was selected as being the cutoff, which is somewhat accurate to real life, with diabetics being said to have an A1C of over 6.5 with prediabetic being at or above 6.



Then, to find which of the attributes other than A1C had an impact on the classification, the A1C variable was temporarily removed from the dataset to make a few search trees that could not rely on it for its classification. This time, a much larger search tree was found and while the accuracy was 84.6%, it still had many misclassifications. It included variables like waist size, age, education level, protein levels, calories, carbs, and Selenium levels.

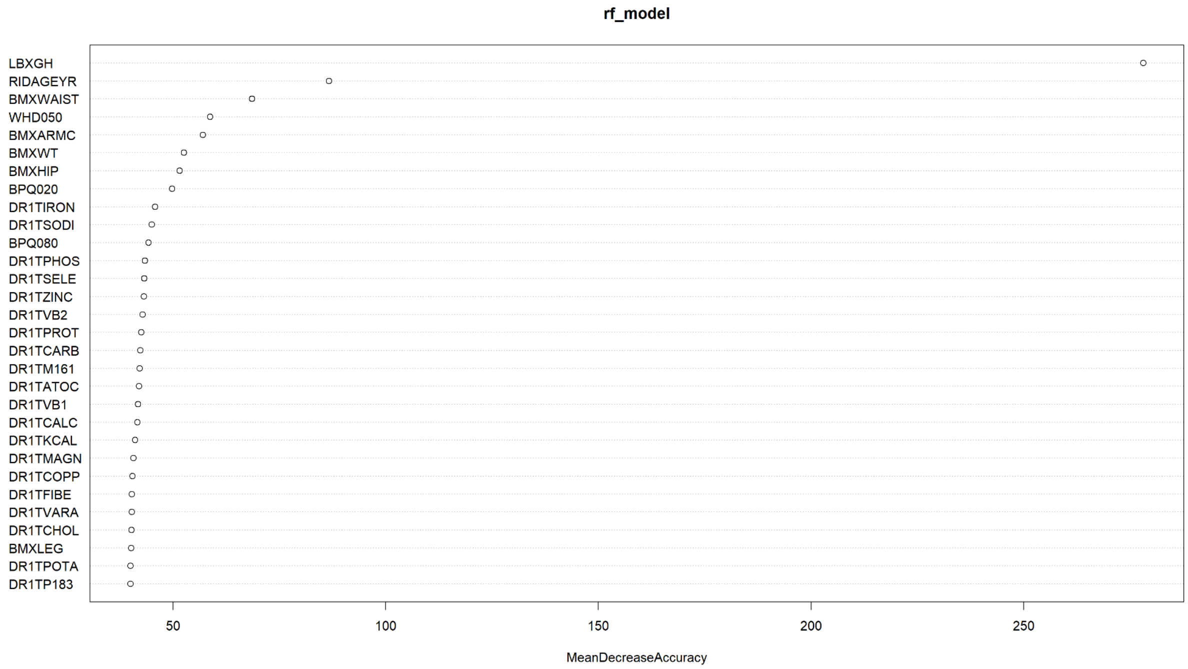


Following these trees it was decided that clustering was not the best method to separate the two datasets, so support vector machines may be able to separate them in a way that clustering could not. I did support vector machines for all of the linear, quadratic, radial, and sigmoid kernels, with linear performing the best of the four different kernels. For the first time in the analysis of the NHANES data, the accuracy of the model was over 90% with a total accuracy of 91.98%, true positive rate of 98.2% and a true negative rate of 54.8%, going above 50% for the first time. I assume that this ability to linearly separate the data has something to do with the fact that A1C was included in this model and therefore the support vector may have been perpendicular to the line separator of A1C=6.5



```
         Actual
Predicted   1   2
        1  68  13
        2  56 724
```

To determine whether or not this was the case, a random forest model was run on all of the numerical attributes to determine which of them had the highest importance (and which of them had the second most given it was almost certain that A1C would have the most.)



After running the random forests model, of course A1C was by far the most important factor. After this was age, waist circumference, weight one year ago, arm circumference, current weight, and then hip circumference, then blood pressure. These variables are also unsurprising because these are physical characteristics often associated with those with type II diabetes.

The NHANES data presented me with an obvious binary classification task with hundreds of variables, though creating an accurate model presented itself to be a more than difficult task. First, clustering seemed appropriate, though with the skew of the clusters, it was not conducive to accurate clustering. Realizing that the data may not be in clusters and instead may be linearly separable, SVM presented itself as a good choice of model, but was still not perfect. Given these models, it only goes to support the idea that A1C is the most important factor in determining a diabetes diagnosis. While other factors do have some impact, they are far less important than A1C is. Nevertheless, it is important to acknowledge the impact that these factors have when deciding whether or not a patient has the disease. Older age and lower income were associated with a higher risk of diabetes, potentially due to lifestyle factors, access to healthcare, and dietary habits. In addition, sedentary behavior showed a positive correlation with diabetes risk. It is also important to acknowledge that preconceived notions of what the disease means for people are not always true. For example, individuals with diabetes tended to have lower calorie, carbohydrate, and sugar intake, suggesting that dietary modifications may be a response to the condition rather than a causative factor. The merit of the classification is questionable, however, because one could always raise the question of if many of these variables are caused by the diagnosis, work towards causing the diagnosis, or are simply correlated. Given the ramifications of a false diagnosis, it is important that any model have an extremely high true negative and true positive rate, which none of the models throughout this analysis have. In the future, I think it would be worth it to work with some of the sub-datasets individually instead of as a whole to see which variables had the most impact. If I were to do a similar analysis again, I would be interested in researching the difference between these characteristics and type I vs type II diabetics, as I assume that their habits, pre-existing conditions, and bodily factors would differ greatly. I also think that it may be worth repeating some of these analyses with a balanced database to see if a model can be found that classifies an unbalanced dataset. Using real, skewed data was intentional because in real life the professionals who diagnose people with the disease have a skewed population to work with.