Ryan H. Gonzalez
February 17, 2018
Dr. Christoph F. Eick

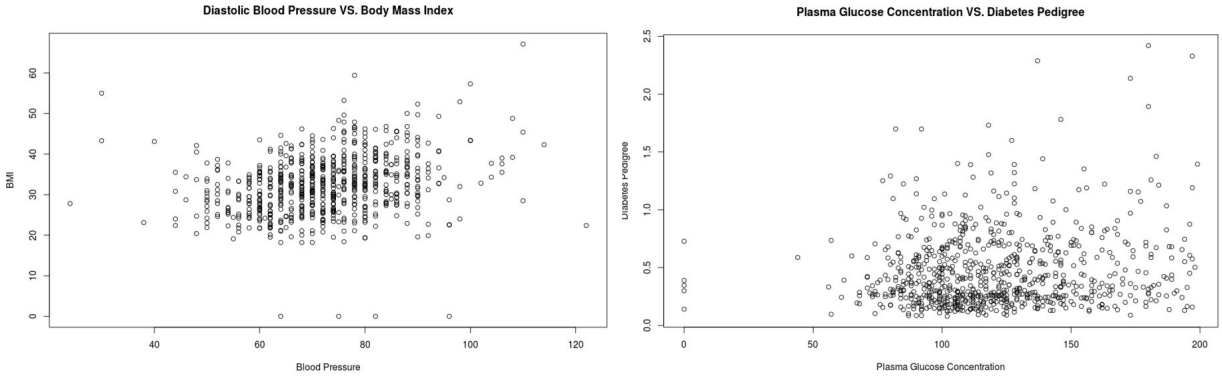
Data Mining - Report One

| | | | | | |
|---|---|---|---|---|---|
| corAtt2Att3 | numeric | 1 | 48 B | 0.221436208527777 |
| corAtt2Att4 | numeric | 1 | 48 B | 0.211353771621624 |
| corAtt2Att5 | numeric | 1 | 48 B | 0.580009931715283 |
| corAtt2Att6 | numeric | 1 | 48 B | 0.221071069458983 |
| corAtt3Att4 | numeric | 1 | 48 B | 0.226839067407822 |
| corAtt3Att5 | numeric | 1 | 48 B | 0.0982722994546555 |
| corAtt3Att6 | numeric | 1 | 48 B | 0.258735165879597 |
| corAtt4Att5 | numeric | 1 | 48 B | 0.184888420189759 |
| corAtt4Att6 | numeric | 1 | 48 B | 0.631958593879867 |
| corAtt5Att6 | numeric | 1 | 48 B | 0.22832812557456 |

2. *Legend at the bottom for explaining naming convention if needed.*

Based on the correlations calculated through R, the following have weak uphill (positive) linear growth: corAtt2Att3, corAtt2Att4, corAtt2Att6, corAtt3Att4, corAtt3Att6, and corAtt5Att6. The following have strong uphill (positive) linear growth: corAtt2Att5, and corAtt4Att6. When we look at corAtt3Att5 and corAtt4Att5 we notice that the correlation between the two attributes is pretty low, lower than +.20, with that being said, it can be determined that the correlation between those two attributes have little to no positive linear relationship.

3.




Looking at the two scatterplots, we notice a trend or pattern within the two. For diastolic blood pressure vs. body mass index we notice that the scatter plot averages out with people having a resting blood pressure between 50 - 90 with a body mass index of 25 - 45. There is a couple of outliers in the scatterplot but a majority of people fit in the average. With the plasma glucose concentration vs diabetes pedigree we see a huge concentration of people averaging their plasma glucose concentration between 80 - 150 and their diabetes pedigree between .1 - .7.

4. When viewing all of the histograms, it has become apparent that class 1 tends to have higher margin compared to class 0. However, class 0 has a bigger set of data, and its bell curve seems to be more normal, i.e less skewed left compared to class 1.

5. When I began to look at the box plots, I noticed that class 0 always had a box that was lower than that of class 1. Perhaps this is due to class 0 being a newer generation, (younger age). This however is my assumption, I also observed that class 0 has many more outliers than that of class 1.

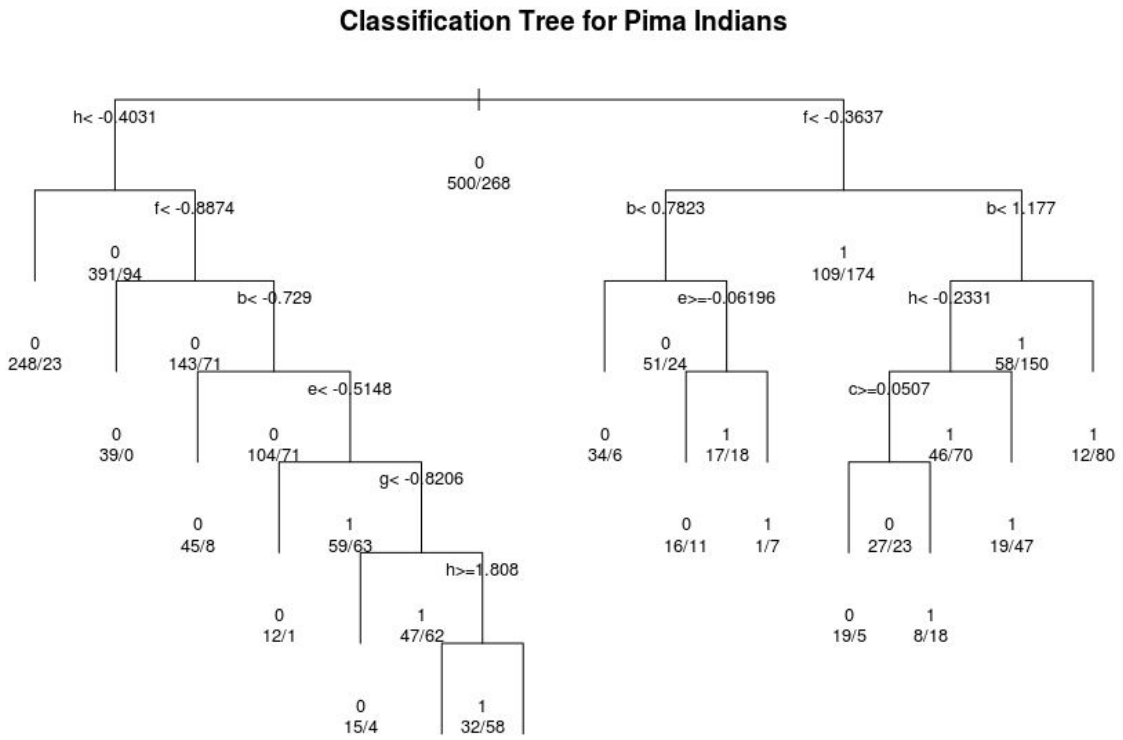
6. On the 3D scatterplots we notice that there is two constants, glucose and the body mass index. Based on the results that I found, it looks as though blood pressure and triceps skin fold thickness may have some similarity as the results without the outliers, are relatively in the middle of the plot. Based on this, BMI and blood pressure appears to be the most dense and consistent compared to the other attributes.

7. Evaluating the star plots, the plots for the most part skewed and weakly consistent. Based on these different sets, people of the same age have slightly similar results. Nothing else to determine from these plots.

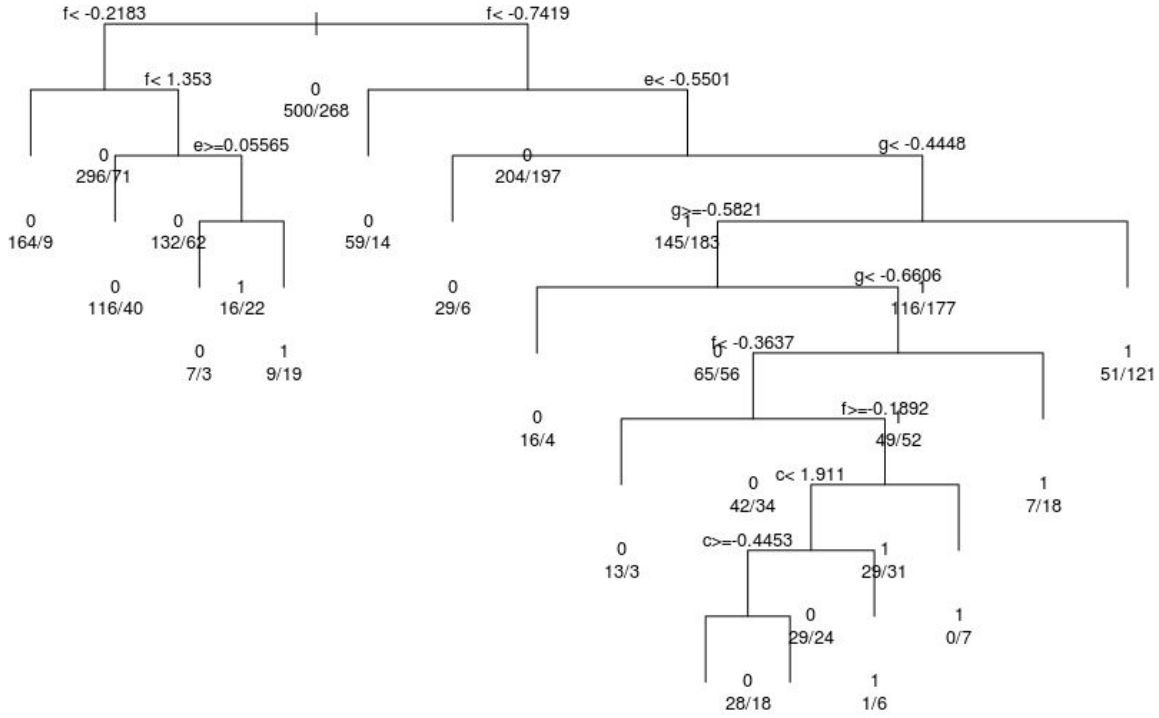
8. Insulin and blood pressure have a inverse relationship compared to the others which have a overall positive relationship trend. Because tricep skin fold thickness and insulin had the coefficients closest to 0 out of the other attributes, I decided to remove them in order to perform my second analysis. The results I produced was not the change I was looking for, infact the change was miniscule at best. All the other coefficients slightly became more positive from the previous test. At the end of my analysis, the two results proved that the people in class 0 have a healthier score compared to the people in class 1. I did notice that individuals in class 1 have slightly older aged subjects compared to class 0. Therefore, the conclusion i came up with earlier that class 0 is a newer generation incorrect. The biggest coefficient is glucose therefore the most significant difference would be the glucose levels.

9. The first decision tree that I generated showed that the attribute glucose had a significant impact as it is shown in four different instances and was the very first decision selected. With that being said, I decided to remove glucose and see what the outcome would be for the second tree. When I removed glucose from the formula, the first decision becomes the age coefficient attribute. The decision made the tree skew more to the right then from the previous tree. It's possible that the Indians have more old people but it looks as if the body mass index is a major deciding factor on deciding the splits as body mass index appears in 5 different locations of the tree. The last tree is decided using the formula from question eight where we had to remove two coefficients that are closest to 0, tricep skin fold thickness and insulin.

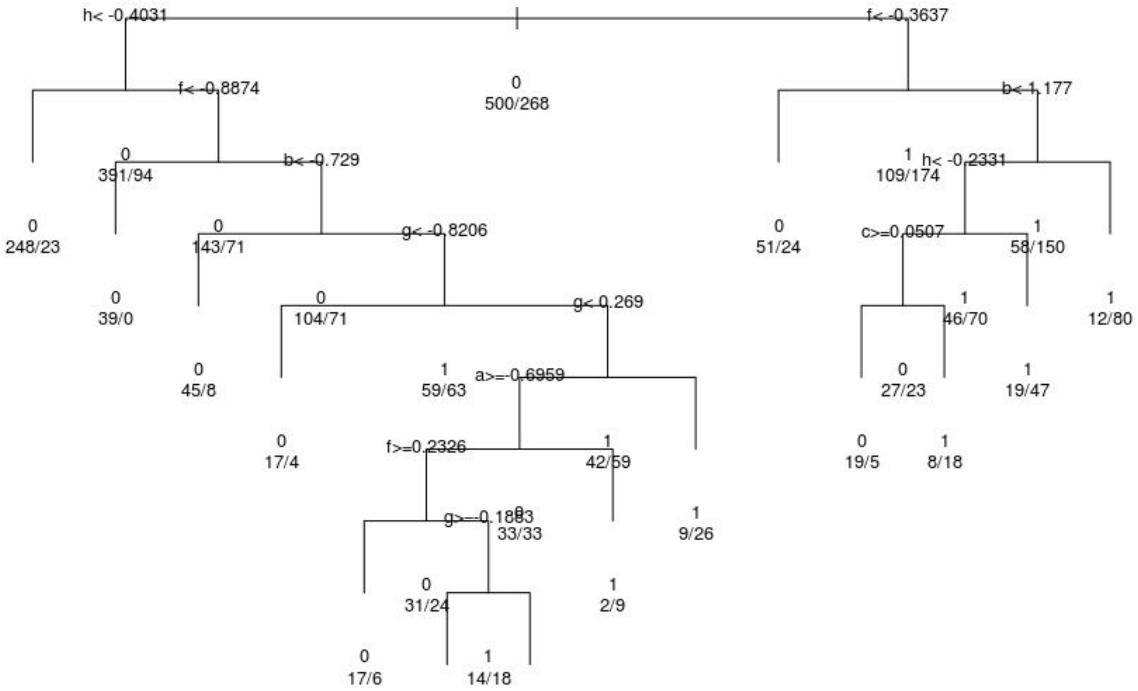
10. Finally, to conclude everything, it is obvious that the Indians have cases of diabetes and will still have some in the future if no action is taken to prevent it. Newer generations will still have diabetes because the younger generation of the Indians in the data set have diabetes or are already developing symptoms. Based on the data analysis, there is a positive relationship trend, which is considered bad in this case for the Indians to keep rising. The most troubling attribute would be their body mass index as their mean body mass index is 31.9. Body mass index is the number that determines if the individual is obese, normal or on the verge of becoming obese. The average states that they are considered obese, I say they're obese because they're considered obese by today's standards. By being obese, these indians are leading themselves to other health issues that are prominent as shown in this dataset. One very evident discover that supports that obesity leads to other health issues would be the mean blood pressure. The mean blood pressure of the Indians is a 72 and that is bad as it states the individual has low blood pressure. Having low blood pressure may lead to critical heart problems.



## Classification Tree for Pima Indians without Glucose




## Classification Tree without Triceps and Insulin



Legend:

| | |
|---|---|
| Att1 | Number of Times Pregnant |
| Att2 | Plasma Glucose Concentration |
| Att3 | Diastolic Blood Pressure |
| Att4 | Tricep Skinfold Thickness |
| Att5 | 2- Hour Serum Insulin |
| Att6 | Body Mass Index |
| Att7 | Diabetes Pedigree Function |
| Att8 | Age |
| Att9 | Class (0 and 1) |

Ex: corAtt2Att3 = correlation with plasma glucose concentration and diastolic blood pressure.