Decision Tree Analysis on the Pima Indian Diabetes Dataset

Theodore Fitch
Department of Data Analytics, University of Maryland Global Campus
DATA 630: Machine Learning
Dr. Ami Gates
July 6th, Summer 2021

Introduction:

The goal of this analysis is to use a decision tree model to predict whether a person will receive a diabetes mellitus diagnosis within 5 years. Obesity and diabetes have become an epidemic in the U.S. Obesity is currently at 42% in adults where the rate was only 30% in 2000 (C.D.C., 2021). With the estimated medical costs of $147 billion due to obesity related illnesses, there is an excellent reason to pay close attention to preventing it (C.D.C., 2021). Additionally, the current diabetes rate is 10% with 33% of the population having pre-diabetes (C.D.C., 2020). While 5-10% of cases are related to type 1 diabetes and are unpreventable, the other 90-95% related to type 2 diabetes are completely preventable (C.D.C., 2020). The major risk factors include being over 45 years of age, being physically inactive and sedentary, having a family history of diabetes, and being obese (C.D.C., 2020). Consequently, the total cost of diabetes within the U.S. is $327 billion (between healthcare costs and lost wages for those affected)(C.D.C., 2020). Type 2 diabetes has been well-characterized as mostly preventable by: living an active lifestyle, maintaining a diet high in fruits and vegetables and low in sugary/processed foods, and monitoring one's metrics. It is critical to note that both obesity and type 2 diabetes have genetic components; however, both of them are also largely preventable with the right lifestyle choices (Baier & Hanson, 2004).

Type 2 diabetes fundamentally is the inability to correctly regulate sugar levels in one's bloodstream. Glucose levels too low will leave one lethargic and unable to function correctly while too high will trigger mechanisms to collect those blood sugars and store them into fat cells. When high glucose is detected in the bloodstream, beta cells in the pancreas will produce insulin which will cause cells to convert the sugars into fat to store for later (Mayo Clinic Staff, 2021). Those who have type 2 diabetes cannot make enough insulin to keep up with their blood sugar

levels leading to a host of pathologies: increased thirst, frequent urination, increased hunger, unintended weight loss, fatigue, blurred vision, slow-healing sores, frequent infections, numbness and/or tingling in the hands and feet, and areas of darkened skin (Mayo Clinic Staff, 2021). While the risk factors are well known and characterized, the exact mechanism of cause in developing type 2 diabetes is unknown (Mayo Clinic Staff, 2021).

The Pima Indians of Arizona have been well-studied since 1965 when it comes to diabetes because they have the highest incidence in the whole world (with exclusivity towards type 2)(Baier & Hanson, 2004). Studying this population has been instrumental in understanding the heritability of diabetes since there has been minimal variance in their diet (little to no introduction of highly sugary or processed foods) and minimal addition to their gene pool over the past decades. Thus, the incidence of diabetes in this population is primarily attributable to genetics. This has helped scientists confirm that at least one major gene is attributable to the onset of diabetes. This is important for medical researchers to understand because preventing diabetes may also include gene therapy for those who have strong genetic influence by this "major gene".

In light of all of this, this analysis is attempting to predict the diagnosis of diabetes using known risk factors via a decision tree model. The known risk factors determined by the literature will be compared against the predicted risk factors. This along with the statistics provided by the model will be used to assess how accurate the model is.

Analysis and Model Demonstration:

Data Information:

The dataset used in this analysis was taken from the Kaggle website (UCI Machine Learning Lab, 2016). All individuals in the dataset were females of Pima Indian descent that

were at least 21 years old. It was a .CSV file generated from real-world data. The original study that created this dataset pull the data from a much larger dataset and only kept variables which were known risk factors for diabetes. From now on, the dataset shall be referred to as "PimaD" for brevity.

Exploratory Data Analysis:

```
> setwd("C:/Users/soari/Documents/Assignments/Data Analytics/UMGC/Summer 2021 Data 630/Assignment 3")
> # Turn file into an object
> PD <- read.csv("pima_diabetes.csv", head =TRUE, sep=",", as.is=FALSE)
> str(PD)
'data.frame':   768 obs. of  9 variables:
 $ preg : int  6 1 8 1 0 5 3 10 2 8 ...
 $ plas : int  148 85 183 89 137 116 78 115 197 125 ...
 $ pres : int  72 66 64 66 40 74 50 0 70 96 ...
 $ skin : int  35 29 0 23 35 0 32 0 45 0 ...
 $ insu : int  0 0 0 94 168 0 88 0 543 0 ...
 $ mass : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ pedi : num  0.627 0.351 0.672 0.167 2.288 ...
 $ age  : int  50 31 32 21 33 30 26 29 53 54 ...
 $ class: Factor w/ 2 levels "tested_negative",..: 2 1 2 1 2 1 2 1 2 2 ...
```

Figure 1. Using the command Structure on PimaD showed there are 9 variables and 768 rows of data.

PimaD contained 9 variables with 768 observations. The variable of interest, "class", represents whether the individual would test positive or negative within 5 years for Diabetes. "Preg" represents the number of pregnancies the individual has had. "Plas" is short for plasma glucose concentration at the 2-hour timepoint during an oral glucose tolerance test. "Pres" represents the blood pressure of the individual in mmHg. "Skin" represents skin thickness (mm) at the Triceps skin fold. "Insu" indicates the 2-hour serum insulin level (mu U/mL). "Mass" is the body mass index (BMI) of the person ((mass in kg / height in m)^2). "Pedi" indicates the diabetes pedigree function which is a calculated risk factor based on family member diagnoses. This includes the influence of those who have been diagnosed with diabetes and those who have not. The equation can be found in the original study that developed this dataset from a larger dataset (Smith et al., 1988). Lastly, "age" simply represents age in years. It is critical to note that

all variables are integer types with the exception of class (which is a factor with 2 levels) and mass & pedi (which are numeric).

When considering the above variables, it is known that women who have 4 or more pregnancies are at higher risk for developing diabetes (Lv et al., 2019). Normal glucose tolerance test results are 140 mg/dL, prediabetes levels are 141-199 mg/dL, and diabetic levels are greater than 200 mg/dL (Mayo Clinic Staff, 2020). Blood pressure is indicative of diabetes but not necessarily a causal factor. Age becomes a risk factor for those over 45 years of age (C.D.C., 2020). Each variable was explored in order to understand the distributions of their values (Figure 2). Insulin levels was the only variable which showed a mean significantly higher than its median (indicating it was right skewed)(Figure 3). There were a few outliers detected. Skin thickness had 1 row with a value of 99 with no values coming close to this (Figure 5). In addition, BMI showed 11 rows that had values of 0 (which is impossible by the BMI scale)(Figure 6). At least 9 of these values also had zeroes listed for blood pressure, skin thickness, and insulin levels (Figure 7). These values were left in in order to make a model with outliers and one without (to show the improvement). There were nearly double the number of negative diabetic diagnoses as there were positive. Based on the previously mentioned metric, all values from the 3rd quartile of "plas" to the maximum value fall into the pre-diabetes category.

```
> summary(PD)
      preg              plas             pres             skin             insu             mass      
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10  
      pedi             age                class      
 Min.   :0.0780   Min.   :21.00   tested_negative:500  
 1st Qu.:0.2437   1st Qu.:24.00   tested_positive:268  
 Median :0.3725   Median :29.00                        
 Mean   :0.4719   Mean   :33.24                        
 3rd Qu.:0.6262   3rd Qu.:41.00                        
 Max.   :2.4200   Max.   :81.00                        
```

Figure 2. The summary command shows that insu is the only variable which has a mean significantly higher than its median (indicating it is right-skewed).

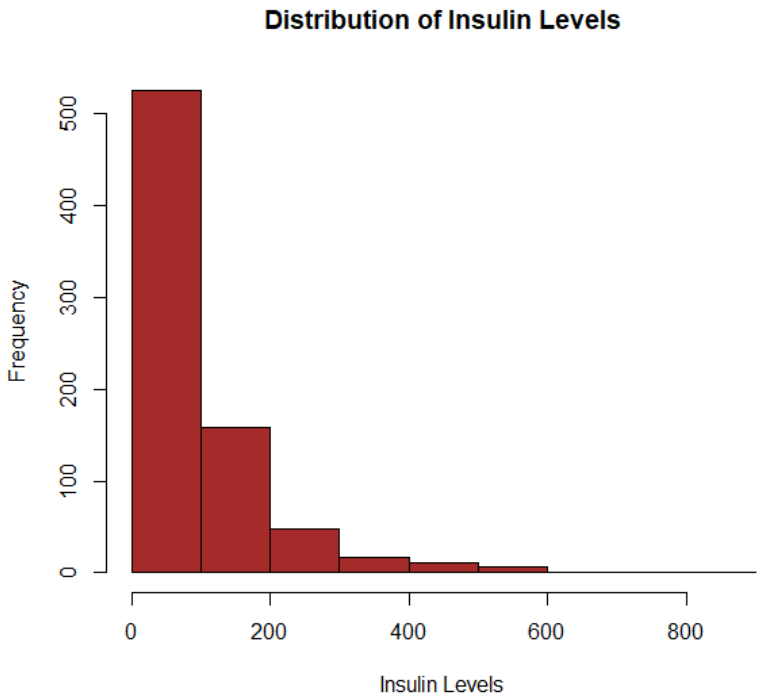

Figure 3. Histogram of Insulin levels shows they are right skewed with the majority of values falling below 200.

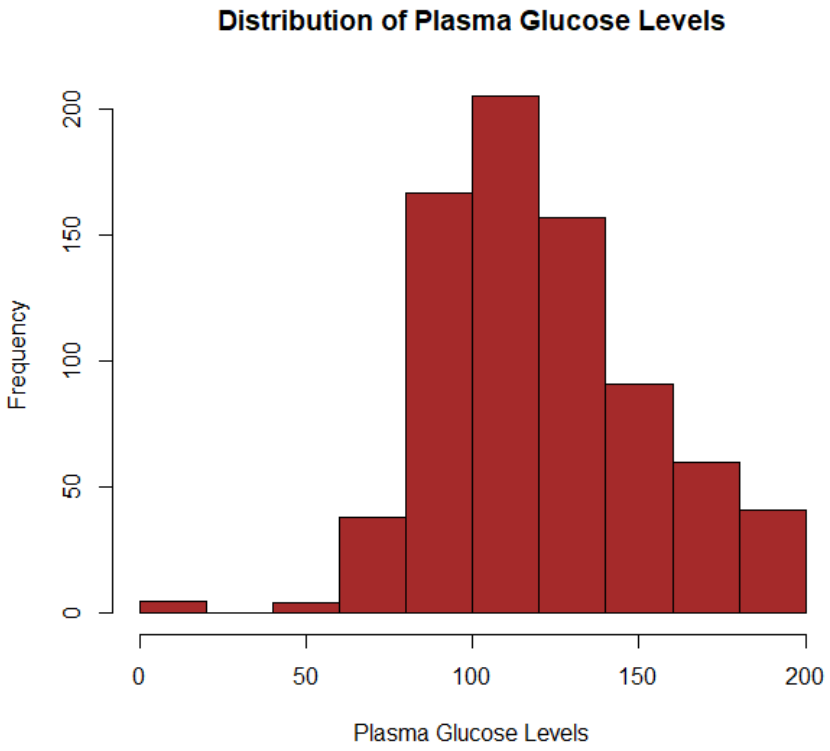

Figure 4. Distribution of plasma glucose levels shows a semi-normal distribution; (somewhat right-skewed).

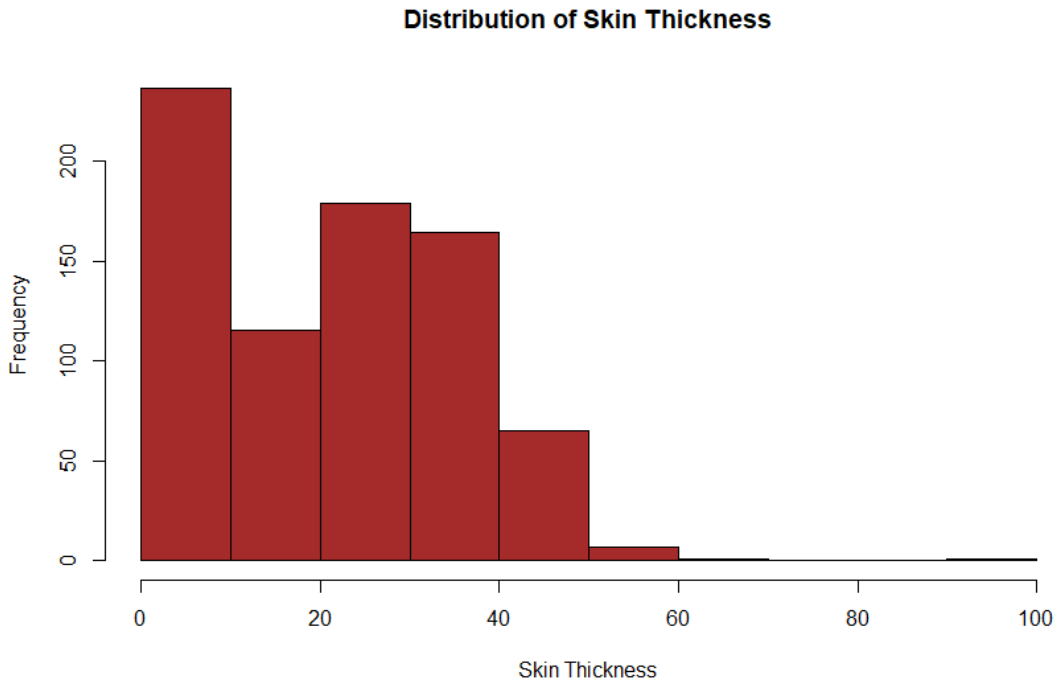

Figure 5. Distribution of skin thickness shows 1 outlier with a value of 99.

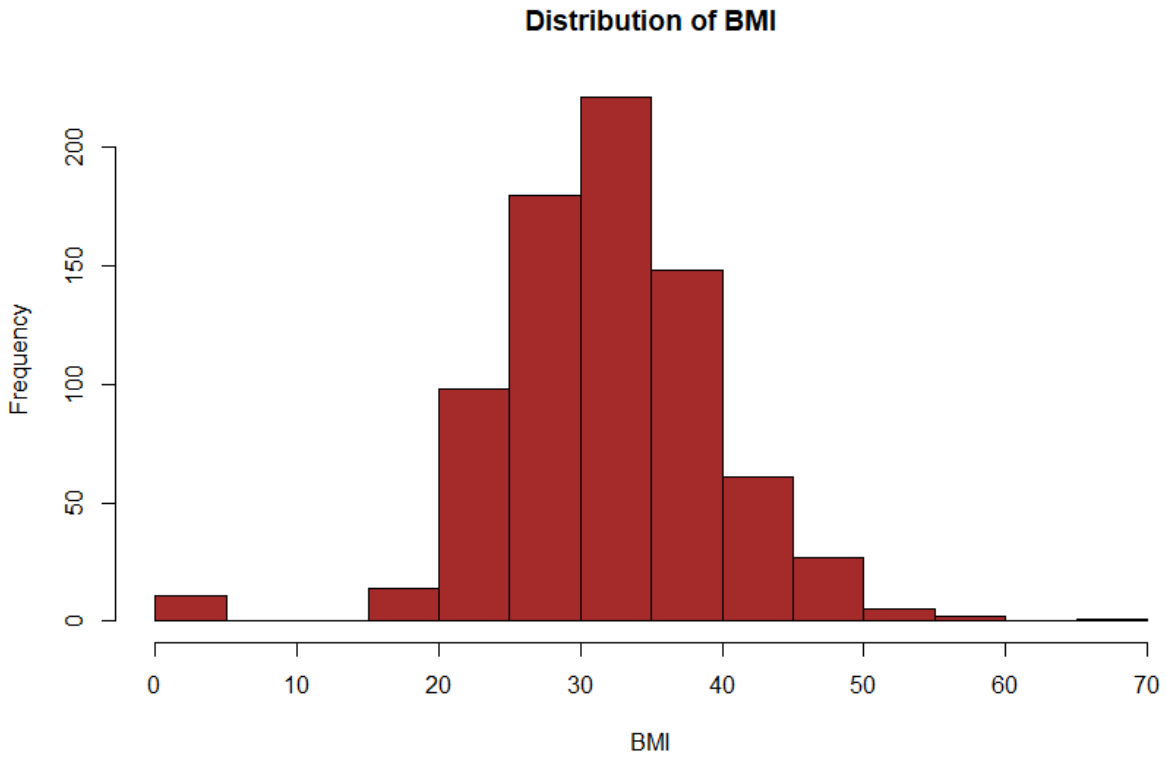

Figure 6. Distribution of BMI shows 11 outliers with values of 0 (which is impossible to have).

	preg	plas	pres	skin	insu	mass	pedi	age	class
10	8	125	96	0	0	0.0	0.232	54	1
50	7	105	0	0	0	0.0	0.305	24	0
61	2	84	0	0	0	0.0	0.304	21	0
82	2	74	0	0	0	0.0	0.102	22	0
146	0	102	75	23	0	0.0	0.572	21	0
372	0	118	64	23	89	0.0	1.731	21	0
427	0	94	0	0	0	0.0	0.256	25	0
495	3	80	0	0	0	0.0	0.174	22	0
523	6	114	0	0	0	0.0	0.189	26	0
685	5	136	82	0	0	0.0	0.640	69	0
707	10	115	0	0	0	0.0	0.261	30	1
419	1	83	68	0	0	18.2	0.624	27	0
439	1	97	70	15	0	18.2	0.147	21	0
527	1	97	64	19	82	18.2	0.299	21	0
240	0	104	76	0	0	18.4	0.582	27	0

Figure 7. Rows with 0 in the BMI column also have zeroes in several other columns likely indicating there are some missing values.

Preprocessing:

PimaD required minimal data preprocessing. Outliers were left in the dataset to first assess the model created with the outliers included. Class values were transformed from "tested_negative" and "tested_positive" to "0" and "1" respectively for the sake of brevity.

Decision Tree Method:

Decision trees are an analytical classification method designed to use values to predict which class a value will belong to (Han et al., 2011). Whereas regression analyses are a numeric prediction method (i.e., "a value of A in X category will likely have a value of B in Y category"), decision trees are a classification prediction method such that the input and outputs must be classes. It deals with questions like: "Is this loan likely safe or risky?" or "Should this patient receive treatment A or treatment B?". In this case, this method was used in attempting to predict which individuals would receive a positive diagnosis for type 2 diabetes within 5 years based on their current metrics. The decision tree method can also use numeric variables as inputs or outputs so long as they have been converted to "classes" like a range of values (for example, "age of > 45 years or < 45 years"; or "ages 30-45, 45-60, 60-75 as classes"). Like several other prediction methods, decision tree methods require that a dataset be split into two parts: training data (70%) and test data (30%). The model is formed based on the training data and then tested for accuracy against the test data. Essentially, the program will attempt to split the data based on the variables and will choose the variables with the starkest splits. So, for example, the method would examine PimaD and may determine that BMI has the starkest split where those > 35 BMI are likely to test positive and those < 35 will test negative. Then, the program will attempt to

split those nodes even further until the split has the best contrast possible between the classes the model is attempting to predict.

Model 1:

The results are made reproducible by setting a seed value at 1,234; this causes the split between the training and test datasets to occur in the same way every time (otherwise different rows would end up in the training dataset and test dataset every time this program was run). Once the split was performed, the structure command was run on both datasets to ensure it ran correctly (Figure 8). Approximately 69.5% of the rows were committed to the training dataset and 30.5% to the test dataset.

```
> str(train.PD)
'data.frame':   534 obs. of  9 variables:
 $ preg : int  6 1 8 1 5 3 10 2 8 4 ...
 $ plas : int  148 85 183 89 116 78 115 197 125 110 ...
 $ pres : int  72 66 64 66 74 50 0 70 96 92 ...
 $ skin : int  35 29 0 23 0 32 0 45 0 0 ...
 $ insu : int  0 0 0 94 0 88 0 543 0 0 ...
 $ mass : num  33.6 26.6 23.3 28.1 25.6 31 35.3 30.5 0 37.6 ...
 $ pedi : num  0.627 0.351 0.672 0.167 0.201 0.248 0.134 0.158 0.232 0.191 ...
 $ age  : int  50 31 32 21 30 26 29 53 54 30 ...
 $ class: Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 2 1 ...
> str(test.PD)
'data.frame':   234 obs. of  9 variables:
 $ preg : int  0 1 7 10 1 13 4 2 4 7 ...
 $ plas : int  137 189 100 125 97 145 103 90 111 105 ...
 $ pres : int  40 60 0 70 66 82 60 68 72 0 ...
 $ skin : int  35 23 0 26 15 19 33 42 47 0 ...
 $ insu : int  168 846 0 115 140 110 192 0 207 0 ...
 $ mass : num  43.1 30.1 30 31.1 23.2 22.2 24 38.2 37.1 0 ...
 $ pedi : num  2.288 0.398 0.484 0.205 0.487 ...
 $ age  : int  33 59 32 41 22 57 33 27 56 24 ...
 $ class: Factor w/ 2 levels "0","1": 2 2 2 2 1 1 1 2 2 1 ...
```

Figure 8. Structure command of the training and test dataset shows that they respectively have 534 rows and 234 rows (out of 768 total).

The first model was created using the first two lines seen in Figure 9. This created a variable which was attempting to predict “class” using all other variables (as denoted by the . in line 1). Then, the decision tree was made based on the training data (“train.PD”). The tree had 6 terminal nodes (in which the values of the predicted variable lie). While the tree had 5 splits, only 4 variables were used with plas used twice (once splitting >144 and $\leq$144, and once splitting >114

and ≤114)(Figure 10, 11). It is critical to highlight the p-values (Figure 10-11) and the criterion (Figure 9). The criterion is simply 1 minus the p-value and any p-value less than 0.05 is considered significant (the lower the number, the more significant it is). The smallest bucket is node 7 with only 7 values while the largest bucket was node 10 with 167 values. These trees are interpreted sequentially. Node 11 has a 71.3% percent chance of testing positive (1) for individuals with a plas value of > 144. However, node 10 shows a 37.1% chance for testing positive for individuals with a plas value of > 114 and ≤144. Thus, each subsequent node down the tree requires having multiple conditions.

```
> myFormula<-class~.
> pima_ctree <- ctree(myFormula, data = train.PD)
> print(pima_ctree)

         Conditional inference tree with 6 terminal nodes

Response:  class
Inputs:  preg, plas, pres, skin, insu, mass, pedi, age
Number of observations:  534

1) plas <= 144; criterion = 1, statistic = 109.381
  2) plas <= 114; criterion = 1, statistic = 25.366
    3) preg <= 6; criterion = 1, statistic = 16.694
      4) mass <= 30.9; criterion = 0.995, statistic = 11.788
        5) age <= 39; criterion = 0.999, statistic = 15.932
          6)*  weights = 117
        5) age > 39
          7)*  weights = 7
      4) mass > 30.9
        8)*  weights = 85
    3) preg > 6
      9)*  weights = 43
  2) plas > 114
    10)*  weights = 167
1) plas > 144
  11)*  weights = 115
```

Figure 9. Printing the decision tree in word form shows the model has 6 terminal nodes with 5 internal nodes.

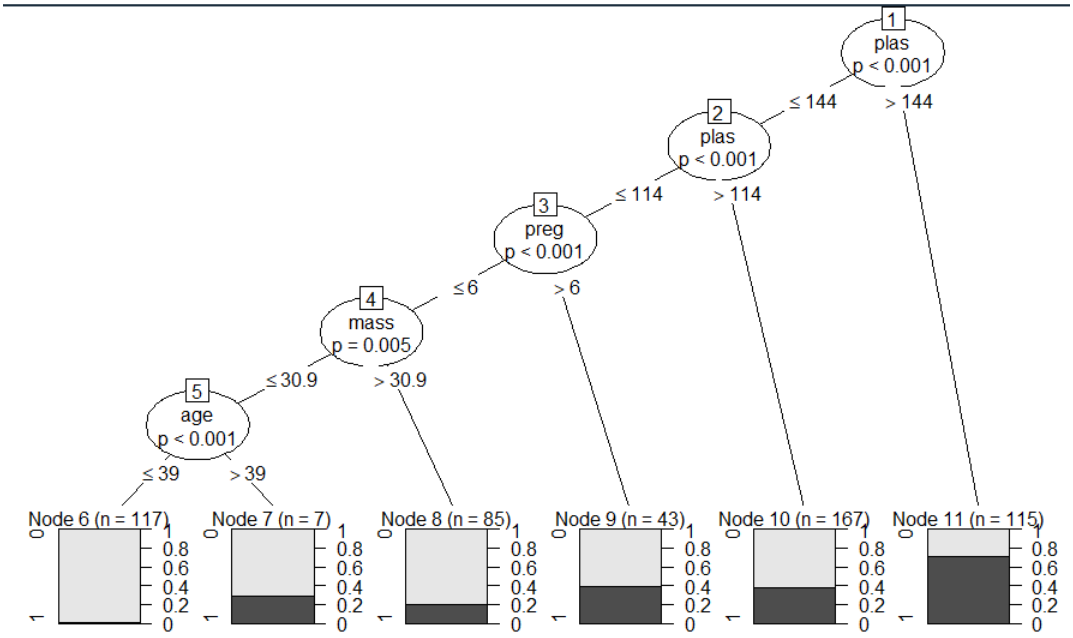

Figure 10. Decision tree of model 1 shows 6 terminal nodes using plas, preg, mass, and age to form the tree.

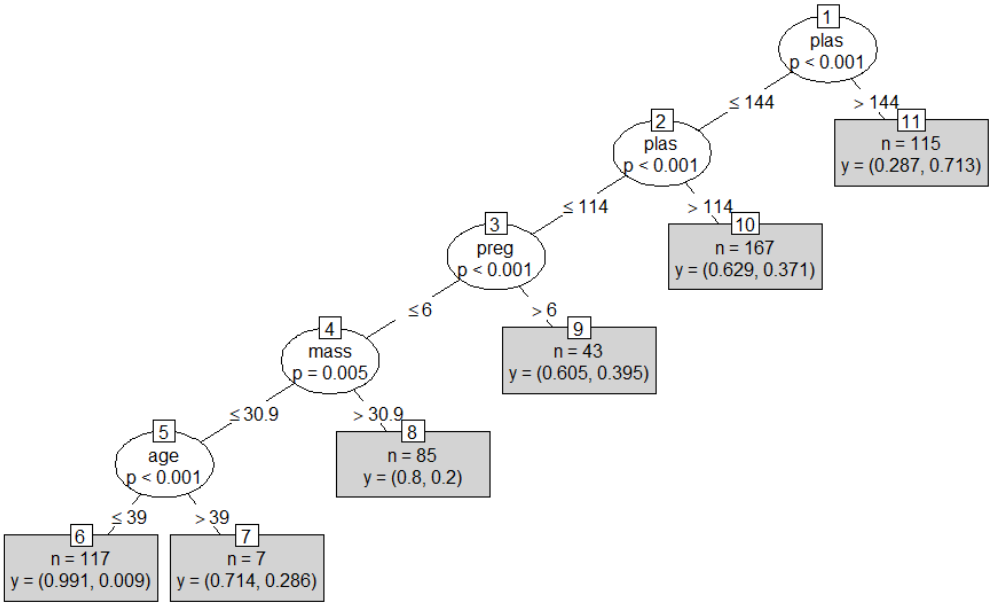

Figure 11. Simplified graph of decision tree of model 1 shows 6 terminal nodes using plas, preg, mass, and age to form the tree.

Once the model was formed, it was tested on the training data and test data for accuracy for which 4 tables/matrices were produced. A confusion matrix shows how many of each value

were correctly and incorrectly identified. The first showed 320 values of 0 were accurately identified as 0 and 82 values of 1 were accurately identified as 1 (Figure 12). In the second matrix, these values were then displayed as percentages by dividing them by the total number of values in the training dataset (534). Thus, the total accuracy of the model on the training data was 75.3%. The model was then applied to the test dataset which yielded an accuracy result of 73.9% (comparable to the training dataset)(Figure 12).

```
> table(predict(pima_ctree), train.PD$class)

     0   1
  0 320  99
  1  33  82
> prop.table(table(predict(pima_ctree), train.PD$class))

            0          1
  0 0.59925094 0.18539326
  1 0.06179775 0.15355805
> testPred <- predict(pima_ctree, newdata = test.PD)
> table (testPred, test.PD$class)

testPred   0   1
       0 133  47
       1  14  40
> prop.table(table (testPred, test.PD$class))

testPred          0          1
       0 0.56837607 0.20085470
       1 0.05982906 0.17094017
```

Figure 12. Decision tree results show model has a 75.3% accuracy on the training data and a 73.9% accuracy on the test data.

Model 2:

In order to make the model more accurate, the previously mentioned outliers were removed from the dataset. 12 values were removed leaving behind 756 total values for model 2. The same methods were applied for model 2 as were applied for model 1: the dataset was split into the training and test datasets (524 values and 232 values); the model was created using the

training dataset; and then the model accuracy was assessed using confusion matrices on the training and test datasets (Figure 13-16).

```
> myFormula<-class~.
> pima_ctree <- ctree(myFormula, data = train.PD)
> print(pima_ctree)

         Conditional inference tree with 4 terminal nodes

Response:  class
Inputs:  preg, plas, pres, skin, insu, mass, pedi, age
Number of observations:  524

1) plas <= 127; criterion = 1, statistic = 109.928
  2) age <= 30; criterion = 1, statistic = 21.568
    3)*  weights = 205
  2) age > 30
    4)*  weights = 122
1) plas > 127
  5) mass <= 29.9; criterion = 1, statistic = 16.211
    6)*  weights = 50
  5) mass > 29.9
    7)*  weights = 147
```

Figure 13. Printing the decision tree in word form shows model 2 has 4 terminal nodes with 3 internal nodes.

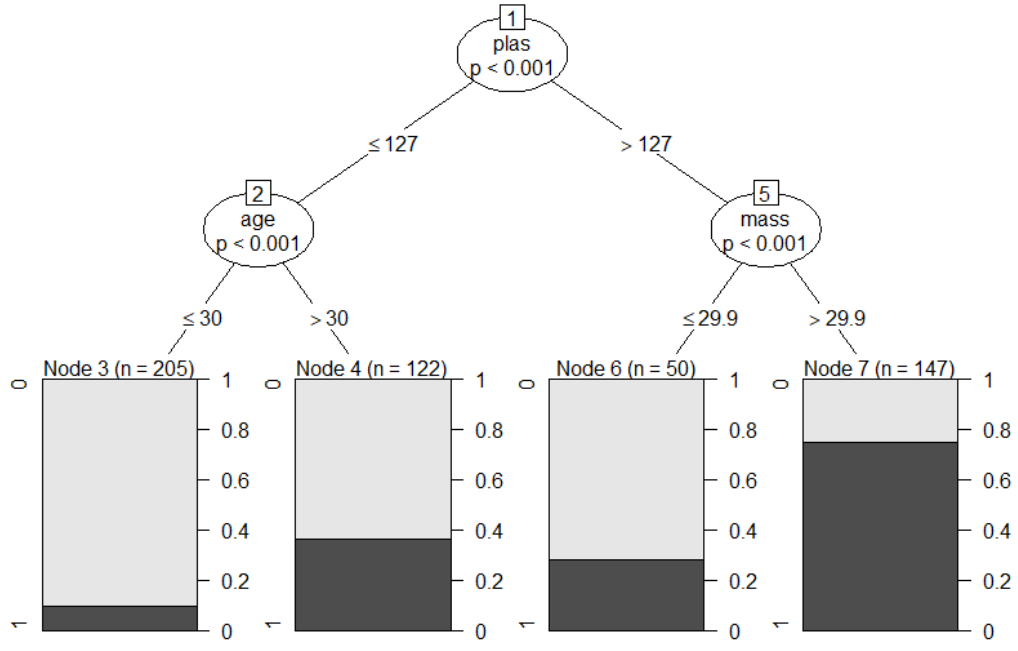

Figure 14. Decision tree of model 2 shows 4 terminal nodes using plas, age, and mass to form the tree.

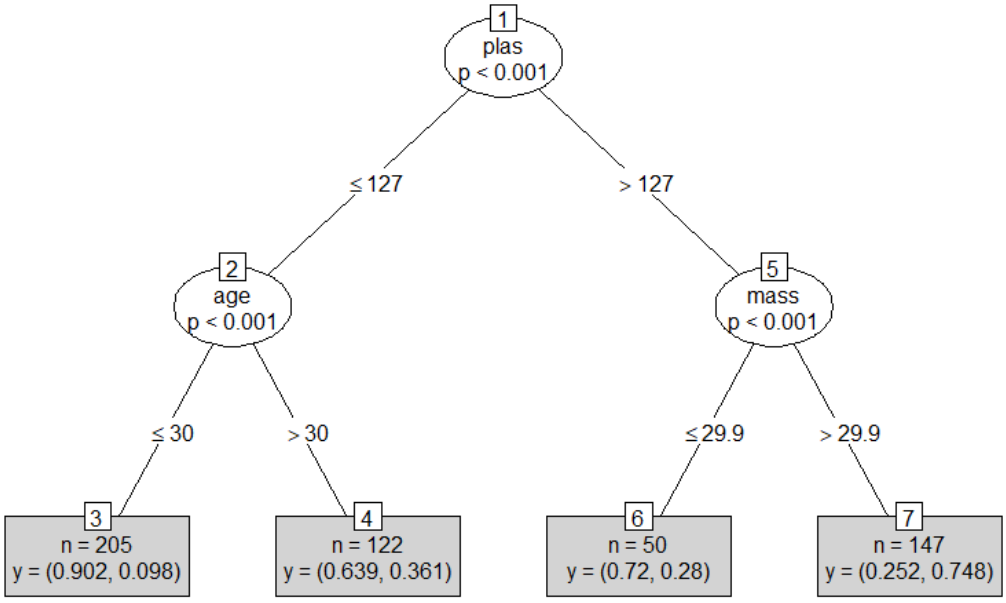

Figure 15. Simplified decision tree of model 2 shows 4 terminal nodes using plas, age, and mass to form the tree.

```
> table(predict(pima_ctree), train.PD$class)

     0   1
  0 299  78
  1  37 110
> prop.table(table(predict(pima_ctree), train.PD$class))

            0          1
  0 0.57061069 0.14885496
  1 0.07061069 0.20992366
> testPred <- predict(pima_ctree, newdata = test.PD)
> table (testPred, test.PD$class)

testPred   0   1
       0 135  38
       1  20  39
> prop.table(table (testPred, test.PD$class))

testPred         0         1
       0 0.5818966 0.1637931
       1 0.0862069 0.1681034
```

Figure 16. Decision tree results show model 2 has a 78.1% accuracy on the training data and a 75.0% accuracy on the test data.

Results and Model Evaluation:

In recap, two models were made. One with all datapoints included in the original PimaD dataset (model 1) and one where 12 outliers were removed (model 2). As previously mentioned, model 1 had a 75.3% accuracy on the training data and a 73.9% accuracy on the test data whereas model 2 had a 78.1% accuracy on the training data and a 75.0% accuracy on the test data. This was critical to highlight that just removing 12 values (which were likely incorrectly input into the dataset) improved the model by 2-3% accuracy. It also simplified the tree from 5 internal nodes and 6 terminal nodes to 3 internal nodes and 4 terminal nodes. A simplified model usually is less accurate but more robust (more applicable to more possibilities of possible real-world values). However, in this case, it was seen that the simplified model was more accurate which is the ideal situation. So long as data entries are full and complete (no missing entries), then model 2 will be more effective. Otherwise, model 1 may be more effective since it can use other variables to predict the diagnosis.

Another feature of note was that both models used the same variables except model 1 used preg and used plas to make two splits. All 4 variables are known risk factors in terms of type 2 diabetes. High plasma glucose levels at the 2-hour timepoint during an oral glucose tolerance test is indicative of type 2 diabetes. This is because the body is not increasing insulin levels in response to increased sugar to cause uptake into the cells (Mao Clinic Staff, 2020). While this can be used to diagnose diabetes, it can also be used to detect pre-diabetes (since insulin resistance is a gradual process). As previously discussed, age is a risk factor in particular when individuals are over years of age. BMI is interesting as a predictor since it likely speaks to the combined genetic and lifestyle risks in regard to obesity. BMI is a measure of obesity; how much fat has accumulated on one’s body. Individuals can have genetic predispositions towards

obesity based on how quickly they burn calories, how much ghrelin they produce, and how much baseline energy they have. On the other hand, obesity also occurs as the result of choice: how much sugar is ingested, how many processed foods are consumed, and how high the ratio is of calories consumed to calories burned through daily baseline expenditure and exercise (C.D.C., 2021). Both genetic predisposition and lifestyle choices play a part in obesity and thus BMI can act as a proxy for how a person has been affected by both of them.

Lastly, pregnancy can be a risk factor for type 2 diabetes. During pregnancy, gestational diabetes can occur which is a temporary form of diabetes. The body causes a mild form of insulin resistance in order to put on some normal extra weight so that the mother's body will have more resistance to sickness and can more easily sustain her child. However, the body may cause too much insulin resistance to occur, and her baseline levels of insulin may not return to normal. Thus, multiple pregnancies can cause continued increased insulin resistance and be a risk factor for permanent type 2 diabetes (Baptiste-Roberts et al., 2009). The program choosing preg, pres, age, and mass does not mean that the other variables (pres, skin, insu, and pedi) are not risk factors. As mentioned before, all the variables being used were known risk factors. However, it does mean that the training dataset showed the model that those 4 variables (preg, pres, age, and mass) were the most important in being able to classify the data in predicting whether an individual would be diagnosed with type 2 diabetes in 5 years. Perhaps different variables would be chosen if the timepoint was 1 year or 10 years. All p-values of the internal nodes were < 0.05 (the standard threshold of significance). This means that there was a >95% chance that the data would appear in the configuration it did (as a predictor of diagnosis) and not due to random chance. In fact, the highest p-value was 0.005 indicating there was a 99.5% chance that the data appeared as it did not due to random chance.

Lastly, it is important to highlight how to interpret the plots of the model (Figure 10-11, 14-15). The terminal nodes show graphs (Figure 10/14) or values (Figure 11/15) in order to show a probability of a diagnosis based on the conditions of the previous nodes. For example, node 11 of Figure 10-11 shows that there is a 71.3% chance an individual will be diagnosed with diabetes in 5 years if they have a plasma glucose level of > 144. Likewise, node 3 of Figure 14-15 shows that there is a 90.2% chance that an individual will not be diagnosed with diabetes in the next 5 years if they have both a plasma glucose level of $\leq$ 127 and $\leq$ 30 years of age. Thus, the terminal nodes do not show absolutes (it is not 100% certainty) however it does show the likelihood of diagnosis based on each precondition. It is also critical to highlight that the diabetes diagnosis is in 5 years. It is absolutely possible that patients would receive their metrics, understand their risk factors, and make quick lifestyle choices. That is one of the reasons why there is still a 25.2% chance that an individual in node 7 (Figure 14-15) will not receive a diagnosis despite having two risk factors (high BMI and high plasma glucose levels). Individuals may have some genetic resistance (where despite high metrics, they do not develop diabetes) or they may make critical healthier lifestyle choices which reverses the trajectory of their diagnosis.

Conclusion:

In summary, a dataset named PimaD was explored. It contained data regarding type 2 diabetes in a Pima Indian population for females 21 and older. The dataset was used to make a decision tree program which predicted whether a woman would likely be diagnosed with type 2 diabetes within 5 years based on the other diabetes-related metrics. The most important risk factors were determined to be plasma glucose levels (at the 2-hour timepoint during an oral glucose load test), age, number of pregnancies, and BMI. All 4 of these are also well-characterized by the literature as being risk factors. The other variables (pedigree (a measure of

how family members which have been diagnosed with diabetes may influence an individual), systolic blood pressure, insulin levels, and skin thickness (at the triceps fold in mm)) were all known risk-factors for type 2 diabetes, but they were not determined as important to predict which individuals would develop diabetes for this dataset. Thus, medical professionals attempting to provide care for pre-diabetic Pima Indian females should: warn them that their risk will increase with greater age and more pregnancies; use oral glucose load tests and monitor the numbers; and share risks involved with higher BMI numbers.

Limitations and Improvements:

There are a few areas which could still be explored. It would be interesting to prune this dataset and take away the top 4 variables used to make the models. It would be interesting to both see 1. the accuracy of the resulting model and 2. the variables the decision tree program would use to make the model. The resultant model should be less accurate. It would be interesting to explore the pedigree variable more because it was not selected as a variable in these models. In particular, the Pima Indians are said to be the population in which heritability is observed the most. So, it was surprising pedi did not appear in the models. Perhaps pedi is a strong causal factor on another factor which was used in the models (like plas or mass). In that case, that intermediate factor could be acting as a proxy for pedi. However, it is also possible that this dataset was too narrow to see the effects of pedigree on the model and with a large dataset, pedi would be one of the most influential factors. Lastly, a dataset with a negative control should be used in the future to ensure the model is working correctly. All 8 of the possible predictive variables were known risk-factors in developing type 2 diabetes. But at least one variable which does not relate to diabetes should be included in further analyses (like height which has no known risk-factor towards diabetes).

References:

Baptiste-Roberts, K., Barone, B. B., Gary, T. L., Golden, S. H., Wilson, L. M., Bass, E. B., & Nicholson, W. K. (2009). Risk factors for type 2 diabetes among women with gestational diabetes: a systematic review. *The American journal of medicine*, *122*(3), 207–214.e4. https://doi.org/10.1016/j.amjmed.2008.09.034

C.D.C. (2020). A Snapshot: Diabetes in the United States. Centers for Disease Control and Prevention. Retrieved from: https://www.cdc.gov/diabetes/library/socialmedia/infographics/diabetes.html

C.D.C. (2021). Adult Obesity Facts. Centers for Disease Control and Prevention. Retrieved from: https://www.cdc.gov/obesity/data/adult.html

Baier, L. J., & Hanson, R. L. (2004). Genetic studies of the etiology of type 2 diabetes in Pima Indians: hunting for pieces to a complicated puzzle. *Diabetes*, *53*(5), 1181–1186. https://doi.org/10.2337/diabetes.53.5.1181

Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. Elsevier. Retrieved June 6th, 2021 from:

http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Lv, C., Chen, C., Chen, Q., Zhai, H., Zhao, L., Guo, Y., & Wang, N. (2019). Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. *Menopause (New York, N.Y.)*, *26*(9), 1010–1015. https://doi.org/10.1097/GME.0000000000001349

Mayo Clinic Staff. (2020). Glucose Tolerance Test. Mayo Clinic. Retrieved from: https://www.mayoclinic.org/tests-procedures/glucose-tolerance-test/about/pac-20394296

Mayo Clinic Staff. (2021). Type 2 Diabetes. Mayo Clinic. Retrieved from: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Schulz, L. O., Bennett, P. H., Ravussin, E., Kidd, J. R., Kidd, K. K., Esparza, J., & Valencia, M. E. (2006). Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the U.S. *Diabetes care*, *29*(8), 1866–1871. https://doi.org/10.2337/dc06-0138

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.

UCI Machine Learning Lab (2016). Pima-indians-diabetes.csv. Kaggle. Retrieved from: https://www.kaggle.com/uciml/pima-indians-diabetes-database