# Inferential and Predictive Models for Diabetes

## Abstract:

According to the Global Disease Burden report [3], diabetes was responsible for 1.5 million deaths. As high as 48% of all deaths due to this disease occured before the age of 70. About 8.5% of adults (18+) reported to have diabetes. Therefore it becomes a question of vital importance to investigate what explains diabetes in a human-being. Further, a model of predicting diabetes may be meaningful to allocate health resources to deal with this challenge. We, in this project, use regression techniques to explain the factors responsible for diabetes and then provide a framework for prediction. We use the Diabetes Health Indicators Dataset available on Kaggle [1]. This dataset is originally based on Behavioral Risk Factor Surveillance System survey. Our tentative project outline is divided into four major tasks. In this first task we motivate the research question using standard existing references e.g. research articles from google scholar and WHO reports etc. EDA includes data visualization analysis– bar charts, scatter plots between two variables, and summary tables. Our second task is dedicated to inferential modeling i.e. to understand what set of factors explains the probability of getting diabetes the most. We will employ the techniques learnt in the class such as multivariate linear regressions and logistic regressions. We will fit regression models, interpret the coefficients and discuss the limitations. The third task is to do predictive modeling. In this section we use regression trees, we fit highly non-linear trees to predict diabetes. Our goal in this task is to get maximum predictive capabilities on test data. The final task is to aggregate all the results, summarize the outcomes, and tell a story based on the evidence found in the earlier tasks.

## 1. Introduction

Diabetes has become one of the most challenging health problems in the United States due to its prevalence and high treatment cost. According to The World Health Organization (WHO), “Today, more than 420 million people are living with diabetes worldwide. This number is estimated to rise to 570 million by 2030 and to 700 million by 2045” [6]. Millions of Americans are impacted by this chronic disease yearly, causing a higher risk for other complications and financial burdens while leaving no space for care for those in the low-middle social classes. Not only are millions of people being diagnosed with diabetes each year but also the mortality rates from people who died because of a consequence of diabetes are increasing in alarmingly high numbers. WHO states that during the year 2019 more than 1.5 million deaths were caused by diabetes, almost half of them coming from people who were younger than 70 years old [5]. This has been a concern for the whole country and organizations have been trying to take matters into their own hands by starting programs that create awareness so that people can start noticing the problem that the United States is facing.

A lot of people in the United States of America are not even aware of what the symptoms are or how they feel when they have developed a type of diabetes. Symptoms and signs of diabetes may vary but some of the most common ones include frequent urination, excessive thirst, blurry vision, low healing of cuts, and weight loss [4]. Even though these are only some symptoms or signs, there are many more symptoms that might not seem important but that should be checked because of the high risk of diabetes that the population is facing right now. This is just one more reason why learning about the importance of diabetes is important, thus, why organizations like WHO have started to create awareness about this disease that is alarmingly increasing each year.

The factors that increase the chances of developing diabetes vary from person to person but a lot of them are commonly found in lifestyles that the population has adopted over time. People have developed bad habits like smoking and consuming fried food and ones with high levels of sugar, as well as not exercising and developing obesity because of those choices. Obesity is one of the number one leading factors of diabetes alongside the consumption of certain types of foods. Therefore, it is important to create awareness about diabetes, the major factors that contribute to it and what lifestyle choices are bad, and what they can do to start developing a lifestyle that will be healthier for them. This will help these new generations adopt a life that will focus on their overall health, but most importantly, will help them decrease the chance of developing any type of diabetes at a very young age.

Everyday we can learn a new thing and diabetes keeps giving us more reason to continue with research about it. Till this day, there is still ongoing research on diabetes, why people are more prone to it and especially on what determinants lets us see who is at higher risk of diabetes. Hill-briggs et al state, "Decades of research have demonstrated that diabetes affects racial and ethnic minority and low-income adult populations in the U.S. disproportionately, with relatively intractable patterns seen in these populations' higher risk of diabetes and rates of diabetes complications and mortality" [2]. Genetics is not the only cause of diabetes; Social Determinants of Health (SDOH) also play a big role in determining the status of a patient's health and in this case, if people have a higher risk of developing diabetes. Some of these SDOH are the conditions in which these patients were born into, have grown, live, go to school and even work at.

The aim of this research is to create awareness and show why it is important to learn about diabetes and the most important risk factors of diabetes as well as the ones that contribute to other chronic diseases caused by diabetes like heart failure. The data set that is going to be used is the Diabetes Health Indicators Dataset that is available on Kaggle [1]. This dataset was based on the Behavioral Risk Factor Surveillance System survey that was done by the Center of Disease Control (CDC) in the United States of America. The survey was done by telephone and asked people questions about health risk behaviors, chronic illnesses, and preventive measures. The dataset shows a total of 22 variables along with more than 250,000 observations that are going to be analyzed. With the help of data analysis tools and techniques, like regression models, bar charts, scatter plots, and predictive modeling we will analyze the information that was gathered from the Center of Disease Control and plan to determine the maximum predictive capabilities on test data, aggregate all the results and outcomes, and answer our questions based on the findings from all of our analyses.

## 2. Exploratory data analysis

The dataset provided has 22 variables and 253650 observations. The variables of education, Genhealth, mental health, and physical health are continuous data; the remaining ones are categorical data that are measured on a nominal scale. Exploratory data analysis is divided into two parts– descriptive data analysis and inferential analysis. Descriptive analysis entails data visualization and analysis both graphically and numerically. The graphical analysis includes scatter plots and box plots while inferential entails correlation and skewness.

### 2.1 The Data Preparation

2.1.1 Survey: who, how the survey was conducted, and what is the response rate?

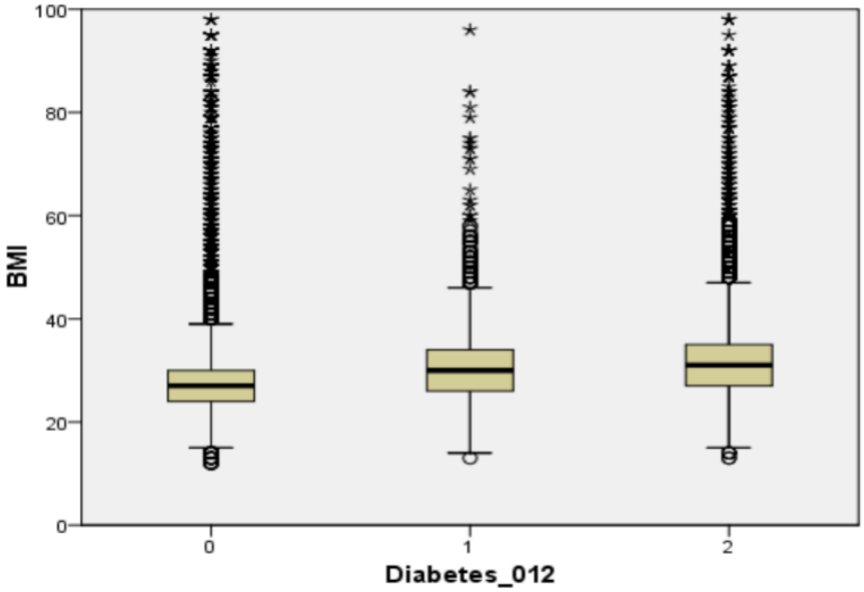
The data was collected through phone interviews by the researchers. The respondents were male and female, which was a factor that was considered to ensure the validity and reliability of the results: that is, the result was not biased on one gender. The survey focused on adults, both male and female, over the age of 18 years. The research sample size is large (n = 253, 680), proving that the response rate was high/ positive, and collecting large data would reduce the error margin, increasing the results' reliability. The researcher collected data through a set of questions which were closed questions: this is important as it enables the researcher to code the data and analyze it quantitatively.

2.1.2 Write more about the data: like dimensions and description of the variable, whether ordinal or nominal.

The primary variable of the research is diabetes and is reported as a categorical variable under the measured nominal. According to the coded format of the data, there are three categories of people on which the researcher aims to focus. Having collected the data and analyzing the different variables regarding the respondents' health enabled the researcher to categorize variable diabetes into three major categories according to the coded data in the dataset. The primary difference observed in the categorical data enables the researcher to focus on a particular aspect of the research: determining the diabetes level of the respondents.

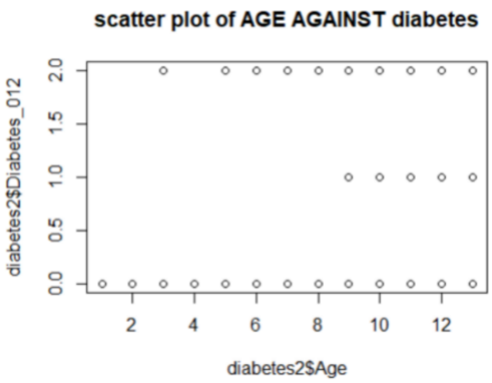
## 2.2 Main Variable

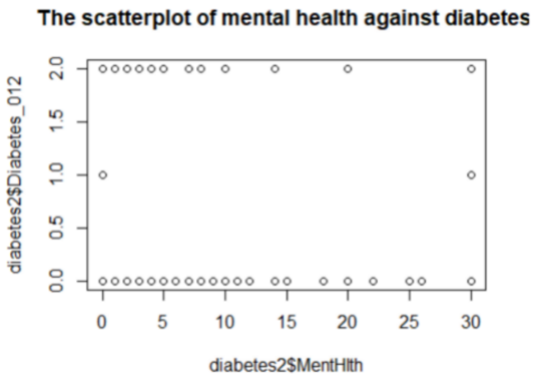
- **Box plot of diabetes_012 against BMI**

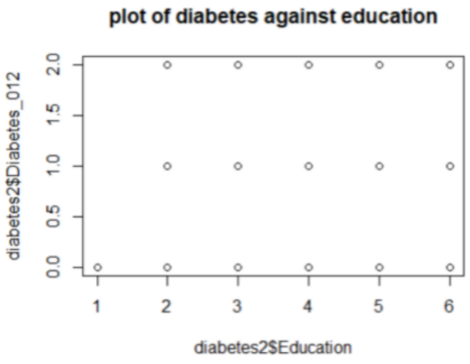


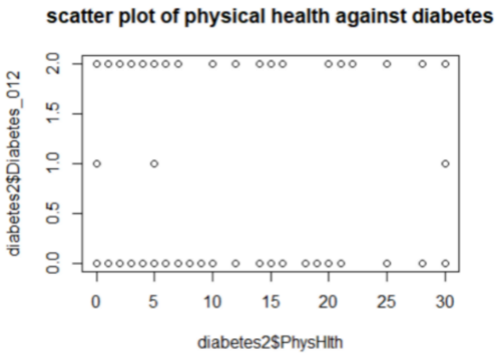
According to the boxplot, it has three major divisions where each respondent was categorized under the diabetes variable. The box plot shows the upper whisker, upper quartile, median, lower quartile, and lower whisker of every category. The first category of diabetes respondents is lower than the other two categories which are evidence that there is a high likelihood of BMI difference among the respondents. Additionally, the three box plots are comparatively short; this means that there is a probability of diabetes in all three categories.

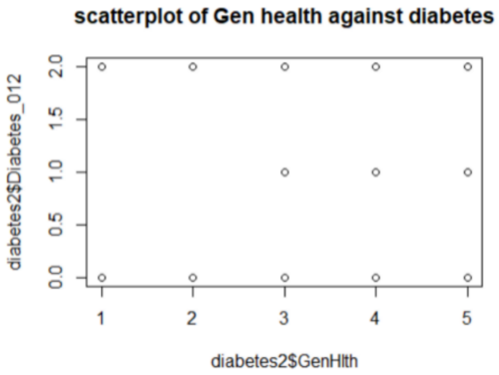
- **Scatter plot of diabetes with other variables**









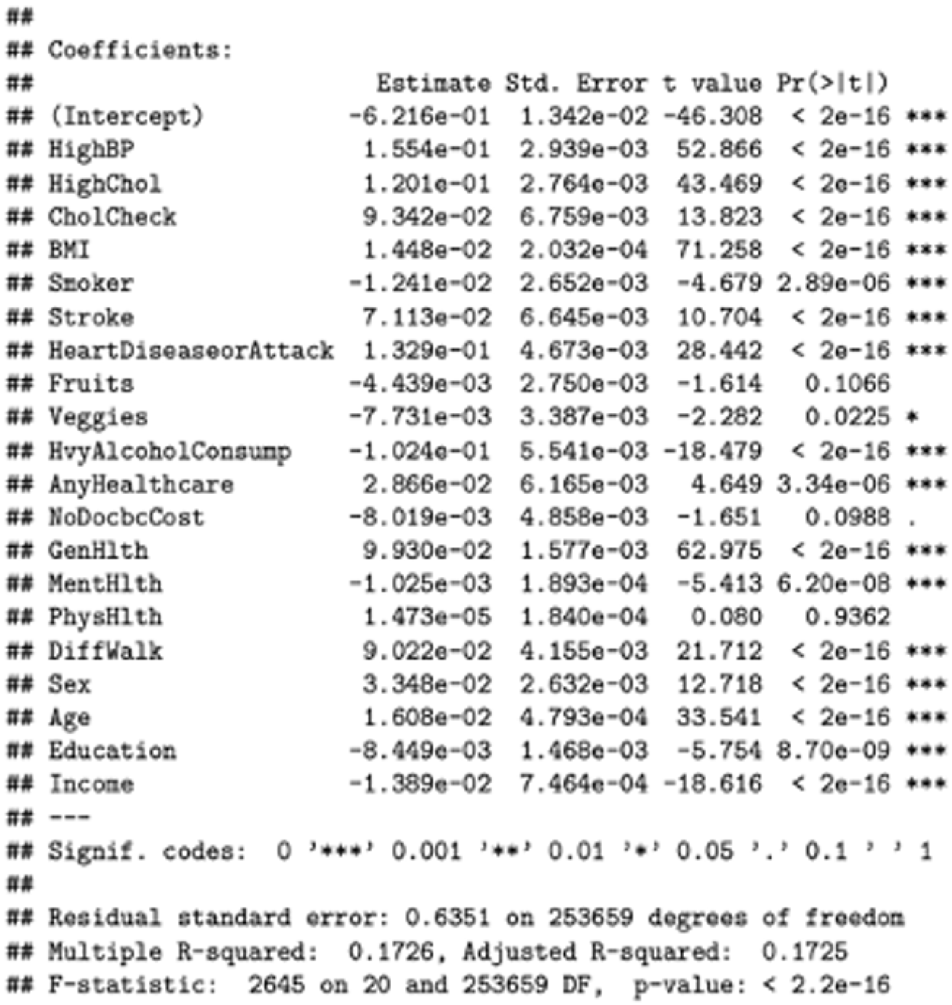


According to the scatter plot of diabetes and various variables, there is no linear or nonlinear correlation between diabetes and age, physical and mental health and education level. Therefore, we can draw the conclusion that the increase of one variable will not affect the other variables.

## 3. Inferential Modeling

### 3.1 Multivariate Linear Regressions

We estimate Model 1 as (the output is attached as following):



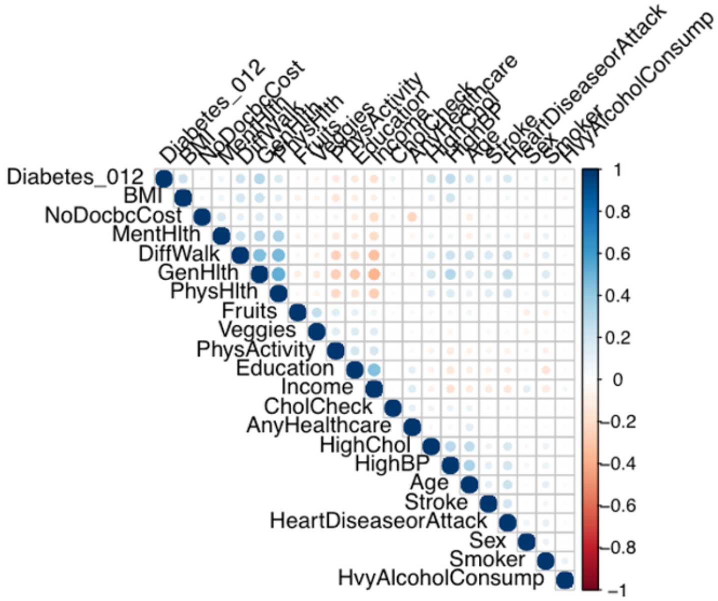
```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6.216e-01  1.342e-02 -46.308  < 2e-16 ***
## HighBP                 1.554e-01  2.939e-03  52.866  < 2e-16 ***
## HighChol               1.201e-01  2.764e-03  43.469  < 2e-16 ***
## CholCheck              9.342e-02  6.759e-03  13.823  < 2e-16 ***
## BMI                    1.448e-02  2.032e-04  71.258  < 2e-16 ***
## Smoker                -1.241e-02  2.652e-03  -4.679 2.89e-06 ***
## Stroke                 7.113e-02  6.645e-03  10.704  < 2e-16 ***
## HeartDiseaseorAttack  1.329e-01  4.673e-03  28.442  < 2e-16 ***
## Fruits                -4.439e-03  2.750e-03  -1.614   0.1066
## Veggies               -7.731e-03  3.387e-03  -2.282   0.0225 *
## HvyAlcoholConsump     -1.024e-01  5.541e-03 -18.479  < 2e-16 ***
## AnyHealthcare          2.866e-02  6.165e-03   4.649 3.34e-06 ***
## NoDocbcCost           -8.019e-03  4.858e-03  -1.651   0.0988 .
## GenHlth                9.930e-02  1.577e-03  62.975  < 2e-16 ***
## MentHlth              -1.025e-03  1.893e-04  -5.413 6.20e-08 ***
## PhysHlth               1.473e-05  1.840e-04   0.080   0.9362
## DiffWalk               9.022e-02  4.155e-03  21.712  < 2e-16 ***
## Sex                    3.348e-02  2.632e-03  12.718  < 2e-16 ***
## Age                    1.608e-02  4.793e-04  33.541  < 2e-16 ***
## Education             -8.449e-03  1.468e-03  -5.754 8.70e-09 ***
## Income                -1.389e-02  7.464e-04 -18.616  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6351 on 253659 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:  0.1725
## F-statistic:  2645 on 20 and 253659 DF,  p-value: < 2.2e-16
```

**The limitations**

**The adjusted R-squared is 0.1725.** Low adjusted r-squared suggests that model 1 is not accounting for much variance in the outcome. If we add more and more useless variables to a model, adjusted r-squared will decrease. If we add more useful variables, adjusted r-squared will increase. Let's check the correlation between the parameters and target (Diabetes_012).. The output is shown as Plot 1 The correlation between the parameters.

**Observations from correlation:**

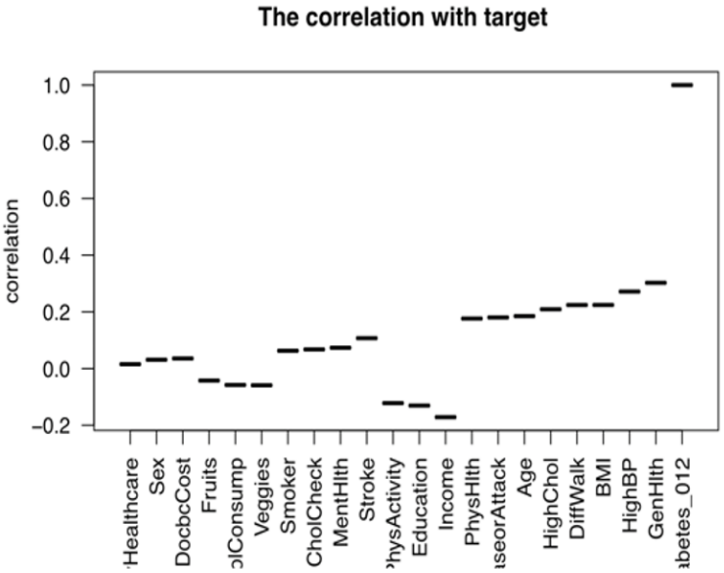
> *"Education" and "Income" have a high correlation coefficients of 0.45.*
> *"GenHlth" and "PhysHlth" have high correlation coefficients of 0.52.*
> *"GenHlth" and "DiffWalk" have high correlation coefficients of 0.46.*
> *" PhysHlth" and "DiffWalk" have high correlation coefficients of 0.48.*



Plot 1 The correlation between the parameters

**Observations from correlation:**

- Healthcare, Sex and DocboCost are the least correlated with the target variable.
- All other variables have a significant correlation with the target variable.



Plot 2 The correlation with target

### 3.2 Why can’t we use linear regression

“Diabetes” is treated as ordinal, which is classified and ranked according to certain features or characteristics. For example, consider diabetes as no diabetes (0), pre-diabetes (1), diabetes (2). The ranking of the levels does not necessarily mean the intervals between them are equal. We summarize the categories and their respective ratings in Table 1 Diabetes Ratings.

| **Category** | **Rating** |
|---|---|
| **No diabetes** | 0 |
| **Pre-diabetes (higher than normal blood sugar level)** | 1 |
| **Respondent has diabetes.** | 2 |

Table 1 Diabetes Ratings

### 3.3 Ordinal Logistic Regressions-

### Dataset Preparation

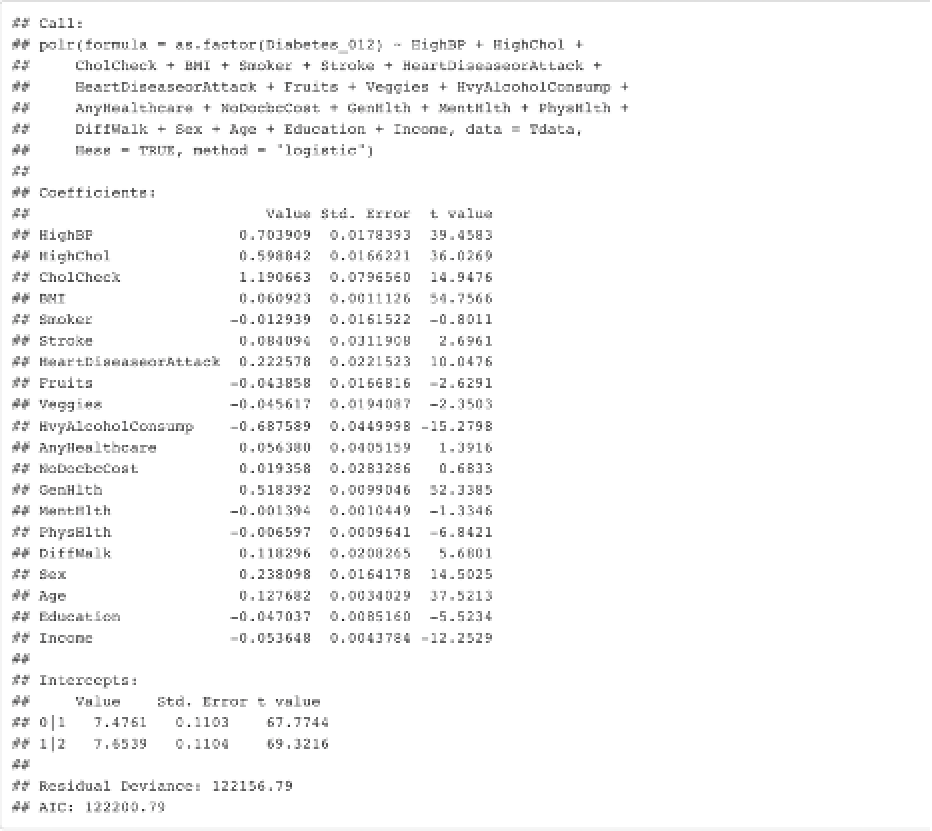
For this data set, we’re predicting an ordinal outcome (diabetes diagnosis). So, we are using ordinal logistic regression rather than linear regression (to predict a continuous variable). The value of diabetes is ranked as 0, 1 and 2. We would like a nonlinear specification that constrains the predicted probability between 0 and 1.

For this data set, we separate the data set into “training” and “test” sets. To train our model we will use 60% data. And we will use 40% data for testing.

## 3.4 Model

### Ordinal Logistic Regressions

First, we use the "polr" command to estimate an ordered logistic regression model. Then, we'll specify Hess=TRUE to let the model output show the observed information matrix from optimization which is used to get standard errors. The results are as shown as Plot 3 Ordinal Logistic Regression Model 1.



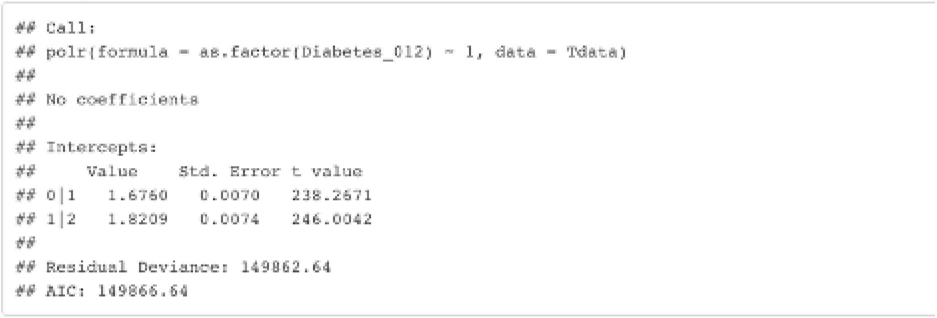
```
## Call:
## polr(formula = as.factor(Diabetes_012) ~ HighBP + HighChol +
##     CholCheck + BMI + Smoker + Stroke + HeartDiseaseorAttack +
##     HeartDiseaseorAttack + Fruits + Veggies + HvyAlcoholConsump +
##     AnyHealthcare + NoDocbcCost + GenHlth + MentHlth + PhysHlth +
##     DiffWalk + Sex + Age + Education + Income, data = Tdata,
##     Hess = TRUE, method = 'logistic')
##
## Coefficients:
##                          Value Std. Error  t value
## HighBP                0.703909  0.0178393  39.4583
## HighChol              0.598842  0.0166221  36.0269
## CholCheck             1.190663  0.0796560  14.9476
## BMI                   0.060923  0.0011126  54.7566
## Smoker               -0.012939  0.0161522  -0.8011
## Stroke                0.084094  0.0311908   2.6961
## HeartDiseaseorAttack  0.222578  0.0221523  10.0476
## Fruits               -0.043858  0.0166816  -2.6291
## Veggies              -0.045617  0.0194087  -2.3503
## HvyAlcoholConsump    -0.687589  0.0449998 -15.2798
## AnyHealthcare         0.056380  0.0405159   1.3916
## NoDocbcCost           0.019358  0.0283286   0.6833
## GenHlth               0.518392  0.0099046  52.3385
## MentHlth             -0.001394  0.0010449  -1.3346
## PhysHlth             -0.006597  0.0009641  -6.8421
## DiffWalk              0.118296  0.0208265   5.6801
## Sex                   0.238098  0.0164178  14.5025
## Age                   0.127682  0.0034029  37.5213
## Education            -0.047037  0.0085160  -5.5234
## Income               -0.053648  0.0043784 -12.2529
##
## Intercepts:
##     Value    Std. Error t value
## 0|1   7.4761   0.1103    67.7744
## 1|2   7.6539   0.1104    69.3216
##
## Residual Deviance: 122156.79
## AIC: 122200.79
```

Plot 3 Ordinal Logistic Regression Model 1

### Check the overall model fit

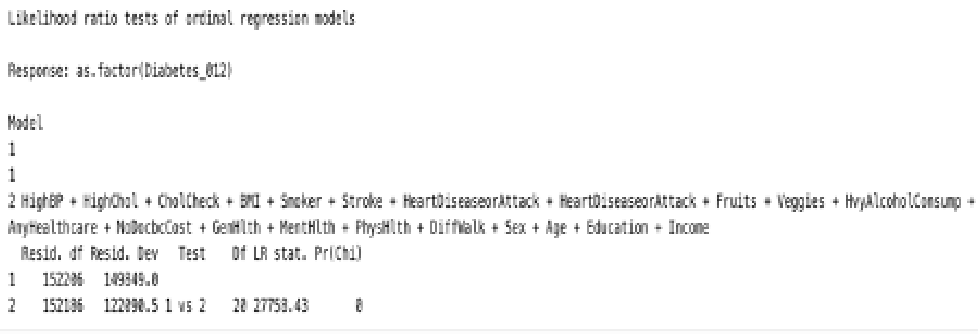
We use the “polr” command to run a test logistic regression model with “Only Intercept” shown as( Plot 4 Test Logistic Regression Model) and compare it to Logistic Regression Model 1(Plot 5 Likelihood Ratio Tests of Ordinal Regression Models).

The chi-square is 27758.43and p = 0. Since the p value is less than 0.05, this means that we can reject the null hypothesis that the model without predictors is as good as the model with the predictors. In other words, there is a statistically significant relationship between diabetes and 20 variables.



```
## Call:
## polr(formula = as.factor(Diabetes_012) ~ 1, data = Tdata)
##
## No coefficients
##
## Intercepts:
##      Value    Std. Error t value
## 0|1   1.6760   0.0070   238.2671
## 1|2   1.8209   0.0074   246.0042
##
## Residual Deviance: 149862.64
## AIC: 149866.64
```

Plot 4 Test Logistic Regression Model



```
Likelihood ratio tests of ordinal regression models

Response: as.factor(Diabetes_012)

Model
1
1
2 HighBP + HighChol + CholCheck + BMI + Smoker + Stroke + HeartDiseaseorAttack + HeartDiseaseorAttack + Fruits + Veggies + HvyAlcoholConsump +
AnyHealthcare + NoDocbcCost + GenHlth + MentHlth + PhysHlth + DiffWalk + Sex + Age + Education + Income
  Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1    152206   149849.0
2    152186   122090.5 1 vs 2    20 27758.43      0
```

Plot 5  Likelihood Ratio Tests of Ordinal Regression Models

## 3.5 Results and Model Accuracy (Table 2 Confusion Table)

### Confusion matrix

**Accuracy**

The classification accuracy achieved was 84.73% using logistic regressions model 1. And the overall accuracy rate is computed along with a 95 percent confidence interval. It means 84.73% of the time the classifier is correct. Using the confusion matrix, we also find that the misclassification error for our model is 15.26%.

We don't achieve any predicted values of Pre-diabetes (higher than normal blood sugar level). It means it is very hard to use this model to predict the pre-diabetes.

Sensitivity:

This describes what proportion of patients with diabetes are correctly identified as having the correct level of diabetes. The sensitivity of No diabetes (0) is 0.98, which means 98% of the time it predicts true when it is true. It is high. So, we aren't missing many people with no diabetes. The sensitivity of Has diabetes (2) is 0.17. It is extremely low; we aren't missing many people with diabetes.

Specificity:

This describes the model's ability to predict true negatives of each available category. The specificity of No diabetes (0) is 0.16, which means 16% of the time it predicts false when it is actually false. It is extremely low. So, we have false positives and people will either incorrectly receive treatment. The specificity of Has diabetes (2) is 0.97, which means 97% of the time it predicts false when it is actually false.  So fewer cases of having diabetes are missed.

|  | Predicted Value |  |  |  |  |
|---|---|---|---|---|---|
| **True Value** | 0 | 1 | 2 | Specificity | Sensitivity |
| **0** | 83530 | 0 | 2032 | 0.16 | 0.98 |
| **1** | 1628 | 0 | 167 | 1.00 | 0.00 |
| **2** | 11665 | 0 | 2450 | 0.97 | 0.17 |

Table 2 Confusion Table

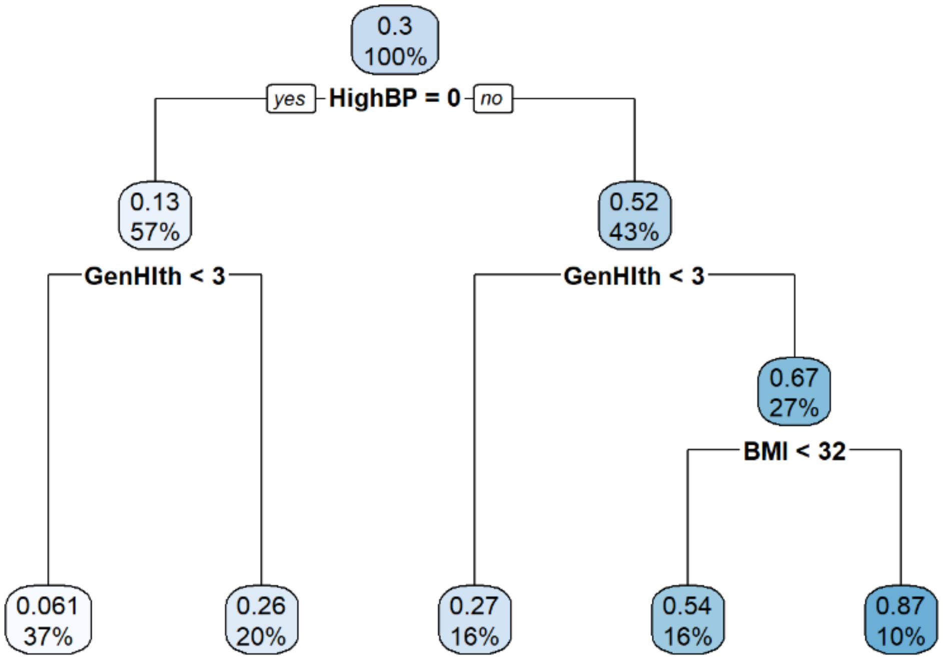
# 4. Predictive Modeling

### 4.1 Why do regression trees make sense?

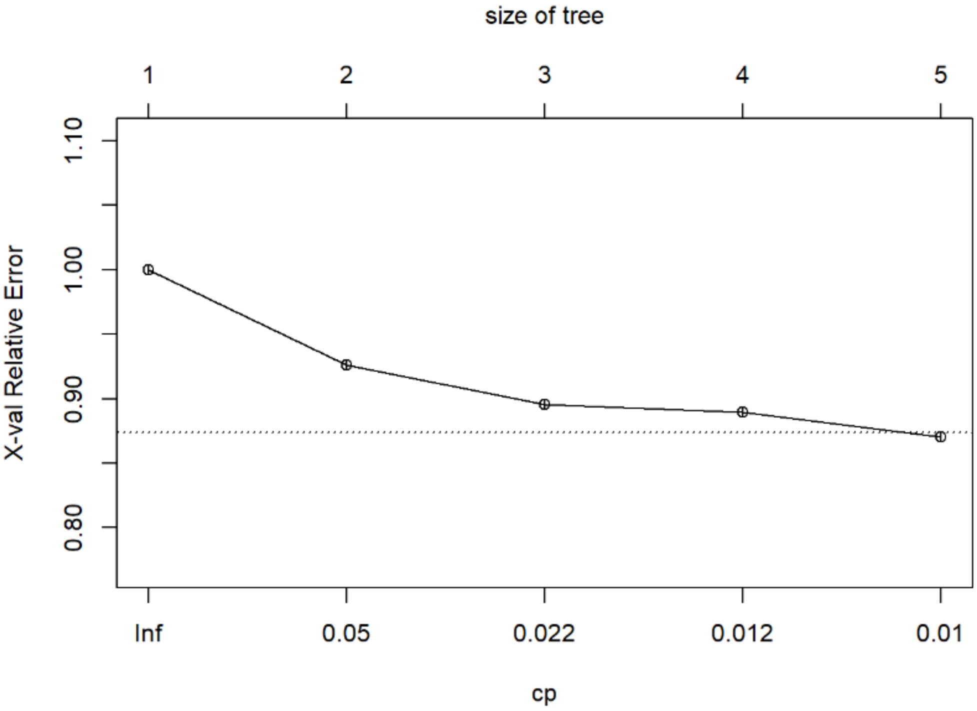
Regression trees make sense in finding out the health indicators for diabetes because it helps make the data more interpretable and allow us to make predictions faster. Additionally, the algorithms from the tree are fast and reliable. Since the diabetes data is a multilinear regression, the variance should be low, meaning the predictions are a lot more accurate. Our research is involved in the predictions rather than the causes, which makes using regression tree sensible to use.

### 4.2 Modeling using Regression trees

For the regression tree, we had the data split between 80% for the training set and 20% for the testing set. As shown in this model, the regression tree reinforces the fact that High BP, general health, and BMI play the most important factors in determining the cause of diabetes. High BP is divided if whether it is high or low, which is divided towards general health of whether it is higher or lower than average, and then whether those in higher general health have BMI of higher or lesser than 32.



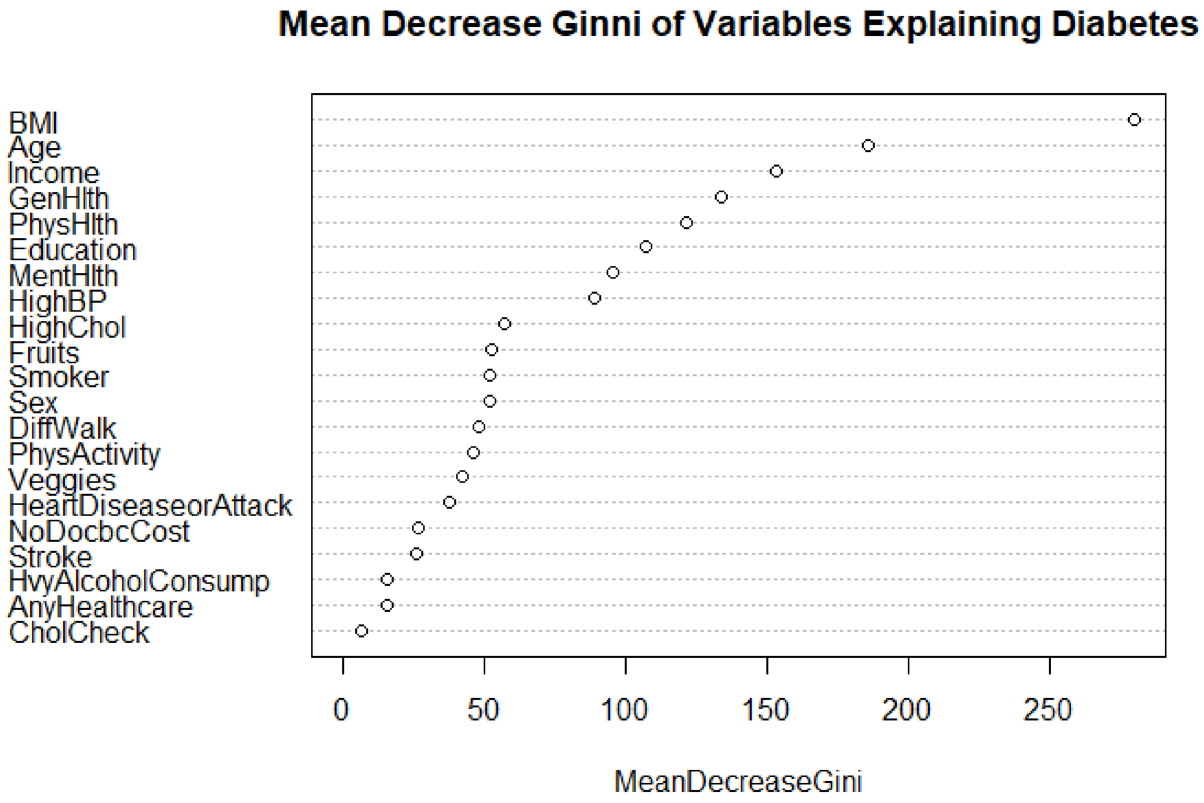
For the second model, the regression tree has been plotted out, with the dotted line to be the standard deviation. Diminishing returns are shown throughout the model, with the lower x-axis being the cost complexity value of the data, the upper axis data being the number of nodes left based on the regression tree, and the y-axis being the cross-validation error. The deeper the tree, the fewer diminishing returns of error reduction are left. The line shows that the point it crosses the dotted line is where it is to experience results within a small margin of error, meaning the 5th terminal node, is the minimal expected error. Therefore, the elbow rule is placed, in which there are no longer any substantial reduced error after the 5th terminal node.



### 4.3 Predictions using Random Forest

Regression trees are no doubt a good method to make predictions, but their prediction accuracy is limited. A great leap forward idea is to grow a large number of trees and take their average prediction. This idea gave impressive results. Since it is made up of a large number of trees, that is why it is called forest. The splitting at each node is done by randomly selected variables. That is why it has got its full name *random sample.*

We grow a random sample with 500 trees with the number of variables tried at each split equal to 4 in our problem. This resulted in 84.4% accuracy. Not only this, this method also provided some inferential insights. We got the relative ordering of importance of variables. In a figure below, we plot the explanatory variables based on their ranking in the importance of explaining Diabetes. Higher Mean Decrease Gini means a more important variable. We note that the inferential insights from this model is consistent with our logistic regression built in previous sections.



### 4.4 Prediction using K Nearest Neighbor (KNN)

KNN is a highly flexible model built upon the measures of similarities and differences between two data-points. Following is the algorithm adopted in this methodology:

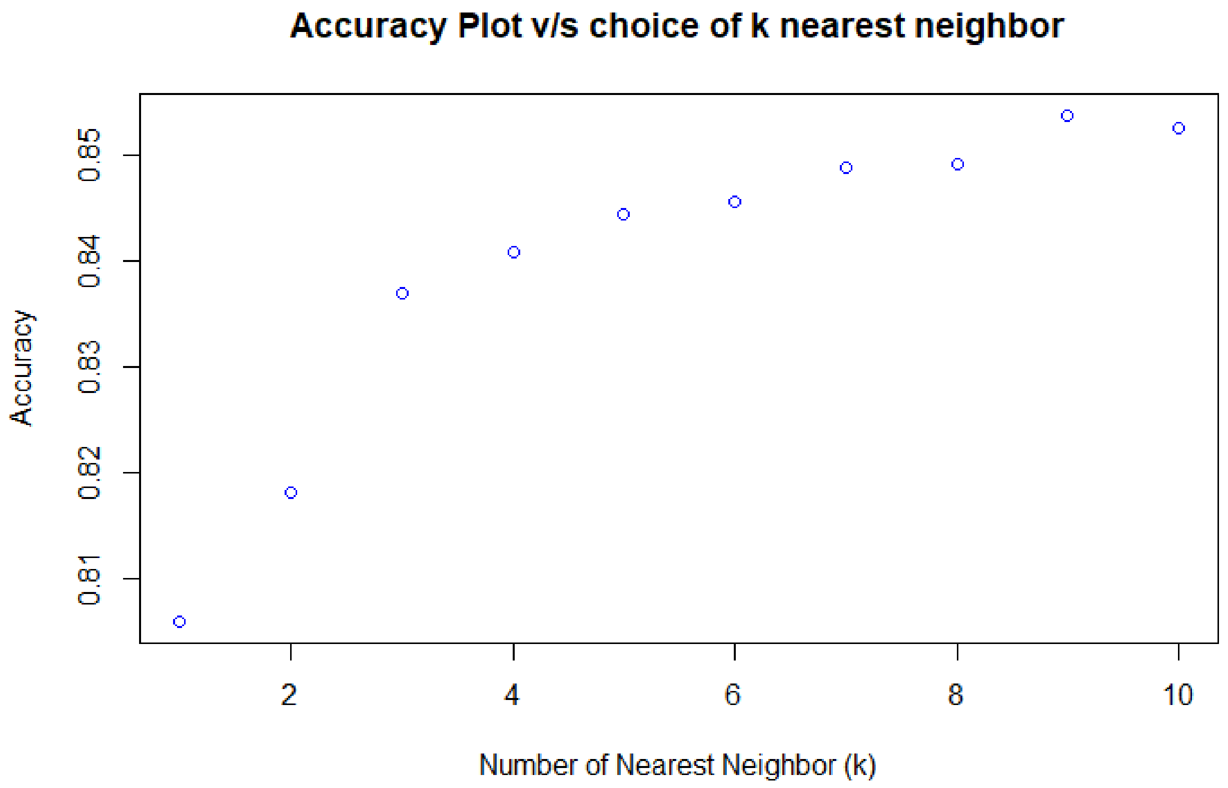
Step-1: Choose a positive integer K.

Step-2: Select any one random point. Calculate the root mean squared distance between this one random data point with all other remaining ones.

Step-3: Identify the K nearest neighbors in the sense that K is the number of points with minimum distance from this randomly selected initial point.

Step-4: Now take a majority vote of the class and assign that class to this randomly selected datapoint.

Choice of K is important in this problem. Below we illustrate a graph with choice of K and accuracy. We note that we get the highest accuracy for K=9. Therefore 9 nearest neighbors should be chosen in this methodology.



## 5. Conclusion

Even though diabetes is a high risk chronic illness, not a lot of attention has been put into creating awareness and showing people the risk factors and what diabetes really is. Millions die each year because of complications coming from diabetes, especially the elderly and how some people are not diagnosed because information about what diabetes consists of is not talked about. This research intended to create that awareness, as well as show with data analysis tools like confusion matrix and logistic regression models what factors are the ones that contribute more to a patient being at a higher risk of developing diabetes, what percentage of people with diabetes are correctly identified, what percentage of patients are missed, how regression trees help us find indicators for diabetes, and many more. The point here being is to show, based on the variables found in our data set, what factors help identify people with diabetes accurately, how some factors do not correlate with one another and that a change in one would not result in a change of another, and how some factors are greater indicators than others as to show that a patient is more likely to develop diabetes over time.

## 6. References

[1] Diabetes Health Indicators Dataset (2016). Behavioral Risk Factor Surveillance System. Retrieved October 28, 2022, from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook

[2] Felicia Hill-Briggs et al. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care* 1 January 2021. https://doi.org/10.2337/dci20-0053

[3] Global Burden of Disease Collaborative Network, Global Burden of Disease Study (2019). Institute for Health Metrics and Evaluation.

[4] The importance of understanding and preventing diabetes. *Genesis Medical Associates, Inc*. Retrieved October 28, 2022, from https://www.genesismedical.org/blog/the-importance-of-understanding-and-preventing-diabetes

[5] World Health Organization. *Diabetes*. World Health Organization. Retrieved October 26, 2022, from https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Another%20460%20000%20kidney%20disease,due%20to%20diabetes%20increased%2013%25

[6] World Health Organization. Draft Recommendations to Strengthen and Monitor Diabetes Responses Within National Noncommunicable Disease Programmes, Including Potential Targets. *WHO.* Retrieved 25 October 2021, from https://cdn.who.int/media/docs/default-source/searo/eb150---annex-2-(diabetes).pdf?sfvrsn=b01fa62_12&download=true