DIABETES PREDICTION

Yash Yaragarla

December 3, 2023

ABSTRACT

A dataset with information about people with diabetes and people without diabetes was used to generate a model using rpart to predict diabetes in patients. Analysis was done to see if there were any clear differences in the two population, finding that the population with diabetes had higher frequencies of behaviors tied to diabetes as reported by the WHO and CDC. The rpart model focused on self reported general health, age, BMI, and blood pressure as key predictors of diabetes. It converged upon the same ideas as the CDC and WHO and was able to predict diabetes with around 70% accuracy.

INTRODUCTION

Since the 1980's, the number of people with diabetes has nearly tripled. Diabetes can lead to heart attacks, strokes, blindness, and loss of limbs. Early detection can allow medical professionals to better treat individuals with diabetes. Prevention and the identification of strongly correlated variables would be useful. The WHO tied sedentary lifestyle, poor diet, smoking, and BMI to diabetes. An analysis on a dataset to see if these can be used as predictors could help increase early detection of diabetes.

MATERIALS AND METHODS

I used data from a dataset called "Diabetes, Hypertension, and Stroke Prediction." I used a subset of the diabetes section, which provides around 70,000 observations that document 18 variables.

I want to apply the rpart and Naive Bayes methods of classification and determine which would be a better fit, as well as determine what variables are most useful for predicting diabetes in a patient.

According to the WHO, avoiding smoking, staying active, and eating healthy are good ways to prevent Type II Diabetes.

VARIABLES

Age = Age of patient, separated into blocks of 5 years

Sex = Sex of patient

HighChol = High Cholesterol

CholCheck = Recent Cholesterol Check

BMI = BMI

Smoker = Does the patient smoke often?

HeartDiseaseorAttack = Has the patient had heart trouble?

Stroke = Has the patient had a stroke recently?

PhysActivity = Is the patient active?

Fruits = Does the patient eat fruits often?

Veggies = Does the patient eat veggies often?

HvyAlcoholConsump = Does the patient drink often?

GenHealth = Self Evaluation of health from 1-5, with 5 being the worst

MenHealth = Days of poor mental health in the last month

PhysHealth = Days of poor physical health in the last month

DiffWalk = Difficulty walking up stairs

PREPARATION OF DATA

There were no missing observations in the data, so there was no need to omit any values. I converted many of the variables into factors since they were left in a numerical form.

I found that my data was made up of half patients with diabetes and half without, so this was not a sample of the general population. I could not make any statements about the distribution of qualities over the whole population, but I could do that for those two groups.

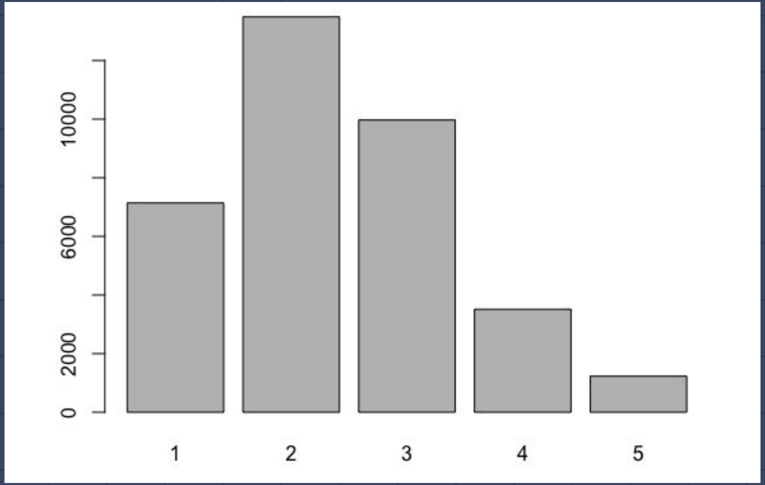
Both data sets had about the same distribution for age. Low physical activity, high BMI, high cholesterol, not eating fruits and veggies, smoking, drinking, history of poor health, and difficulty walking were found in higher frequencies in the subset with diabetes than the subset without.

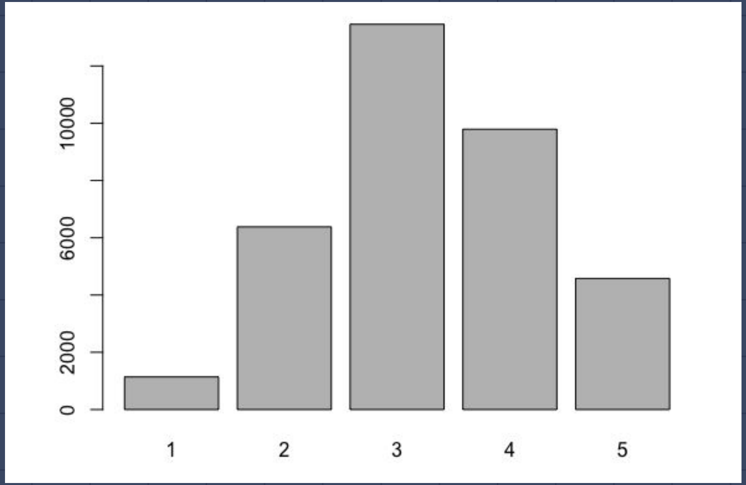
```
> summary(subset(diabetes_data, diabetes_data$Diabetes == "Diabetes"))
      Age           Sex                  HighChol               CholCheck           BMI              Smoker     
 65-69  :6558   Female:18411   Low or Normal Chol:11660   No Recent Check:  241   Min.   :13.00   Nonsmoker:17029  
 60-64  :5733   Male  :16935   High Chol         :23686   Recent Check   :35105   1st Qu.:27.00   Smoker   :18317  
 70-74  :5141                                                                     Median :31.00                    
 55-59  :4263                                                                     Mean   :31.94                    
 75-79  :3403                                                                     3rd Qu.:35.00                    
 80+    :3209                                                                     Max.   :98.00                    
 (Other):7039                                                                                                      
       HeartDiseaseorAttack      PhysActivity                  Fruits                      Veggies     
 No Heart Trouble:27468     Not Active:13059   Does not eat fruits:14653   Does not eat veggies: 8610  
 Heart Trouble   : 7878     Active    :22287   Eats Fruits        :20693   Eats veggies        :26736  




            HvyAlcoholConsump     GenHlth         MentHlth         PhysHlth                        DiffWalk    
 Does not drink heavily:34514   Min.   :1.000   Min.   : 0.000   Min.   : 0.000   No Difficulty Walking:22225  
 Drinks heavily        :  832   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 0.000   Difficulty Walking   :13121  
                                Median :3.000   Median : 0.000   Median : 1.000                                
                                Mean   :3.291   Mean   : 4.462   Mean   : 7.954                                
                                3rd Qu.:4.000   3rd Qu.: 3.000   3rd Qu.:15.000                                
                                Max.   :5.000   Max.   :30.000   Max.   :30.000                                

       Stroke          HighBP             Diabetes    
 No Stroke:32078   Min.   :0.0000   No Diabetes:    0  
 Stroke   : 3268   1st Qu.:1.0000   Diabetes   :35346  
                   Median :1.0000                      
                   Mean   :0.7527                      
                   3rd Qu.:1.0000                      
                   Max.   :1.0000                      
```

```
> summary(subset(diabetes_data, diabetes_data$Diabetes != "Diabetes"))
      Age            Sex                   HighChol                CholCheck            BMI              Smoker     
 60-64  : 4379   Female:19975   Low or Normal Chol:21869   No Recent Check: 1508   Min.   :12.00   Nonsmoker:20065  
 55-59  : 4340   Male  :15371   High Chol         :13477   Recent Check   :33838   1st Qu.:24.00   Smoker   :15281  
 65-69  : 4298                                                                     Median :27.00                    
 50-54  : 3784                                                                     Mean   :27.77                    
 45-49  : 2906                                                                     3rd Qu.:31.00                    
 70-74  : 2903                                                                     Max.   :98.00                    
 (Other):12736                                                                                                      
       HeartDiseaseorAttack      PhysActivity                  Fruits                      Veggies     
 No Heart Trouble:32775     Not Active: 7934   Does not eat fruits:12790   Does not eat veggies: 6322  
 Heart Trouble   : 2571     Active    :27412   Eats Fruits        :22556   Eats veggies        :29024  




            HvyAlcoholConsump     GenHlth         MentHlth         PhysHlth                        DiffWalk    
 Does not drink heavily:33158   Min.   :1.000   Min.   : 0.000   Min.   : 0.000   No Difficulty Walking:30601  
 Drinks heavily        : 2188   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.: 0.000   Difficulty Walking   : 4745  
                                Median :2.000   Median : 0.000   Median : 0.000                                
                                Mean   :2.383   Mean   : 3.042   Mean   : 3.666                                
                                3rd Qu.:3.000   3rd Qu.: 2.000   3rd Qu.: 2.000                                
                                Max.   :5.000   Max.   :30.000   Max.   :30.000                                

       Stroke          HighBP             Diabetes    
 No Stroke:34219   Min.   :0.0000   No Diabetes:35346  
 Stroke   : 1127   1st Qu.:0.0000   Diabetes   :    0  
                   Median :0.0000                      
                   Mean   :0.3742                      
                   3rd Qu.:1.0000                      
                   Max.   :1.0000                      
```

Comparing the subsets' perception of their own health, people with diabetes ranked it higher by 1 rank.

```
> tapply(diabetes_data$GenHlth, diabetes_data$Diabetes, mean)
No Diabetes    Diabetes
   2.383183    3.290981
> table(diabetes_data$GenHlth, diabetes_data$Diabetes)

    No Diabetes Diabetes
  1        7142     1140
  2       13491     6381
  3        9970    13457
  4        3513     9790
  5        1230     4578
>
```

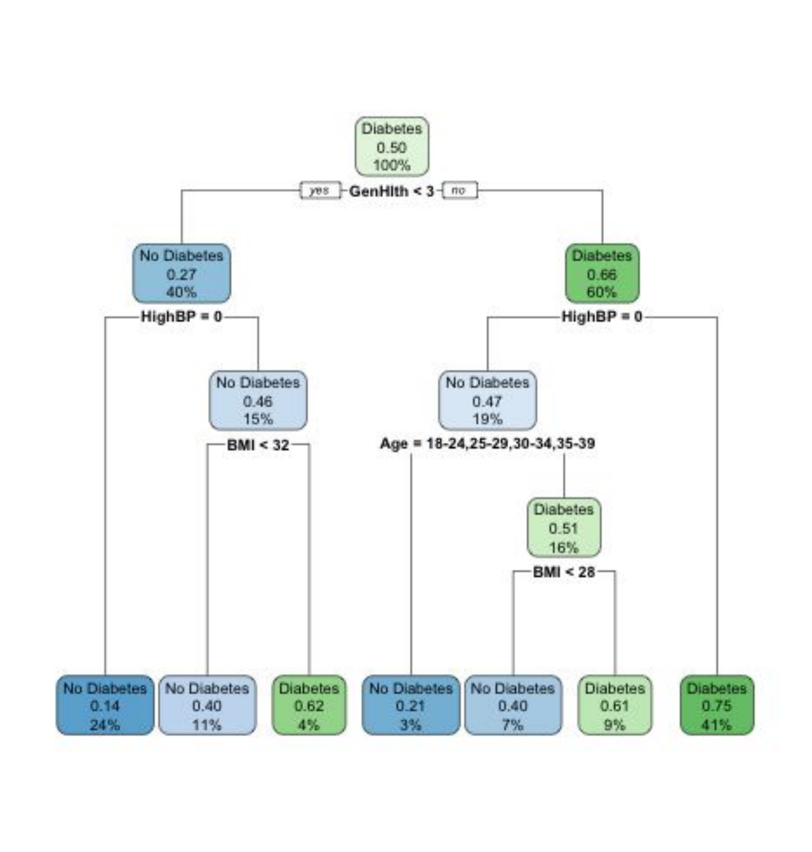




- I used the create_train_test function discussed in class to split my data into an 80/20 train-test split.
- I then used rpart and Naive Bayes functions to create models. I allowed all variables to be used as predictors in order to see what the models would pick up on.

RESULTS

Based on the rpart plot, the model picked up on a correlation between high blood pressure, self-reported health, BMI, and age as predictors of diabetes.



ACCURACY TESTING

Upon creating a Naive Bayes model and testing both models for accuracy, I found that the rpart accuracy was 0.7239, and the Naive Bayes model accuracy was 0.7208. These are not highly accurate, but the models seem to be comparable.

The Confusion Matrices are very similar in their distribution.

Rpart Confusion Matrix

pred_rpart	No Diabetes	Diabetes
No Diabetes	4885	1620
Diabetes	2284	5350

Naive Bayes Confusion Matrix

	No Diabetes	Diabetes
No Diabetes	5469	2248
Diabetes	1700	4722

DISCUSSION

A CDC study also found a similar connection between blood pressure, age, and diabetes. Their cutoff for age was around 45, and ours was 40.

The importance of a patients self-perception of their general health being a major predictor was unsurprising given the dataset, but may be largely useless. A person feeling unhealthy is more likely to have a disease than someone who feels healthy. When removing that column from the dataset, the tree created by rpart loses most of its branches.

Neither model was highly accurate, so a larger set of variables could allow a clearer pattern to be identified.

Literature Cited

Chuks, Prosper. “Diabetes, Hypertension and Stroke Prediction.” *Kaggle*, 19 Dec. 2022, www.kaggle.com/datasets/prosperchuks/health-dataset.

“Diabetes.” *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 3 Dec. 2023.

“Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 9 Mar. 2017, www.cdc.gov/pcd/issues/2017/16_0244.htm#.