Diabetes Health Indicators

Icon

Description automatically generated

Mohamad Nayal

IST 707 Project Report

December 21, 2021

# Introduction

According to the CDC 10.5% of the population of the population of the United States has diabetes and of those people 7.3 million people have diabetes and are undiagnosed ("National Diabetes Statistics Report, 2020 | CDC", 2021). Diabetes is a condition in which either a patient does not produce enough insulin or they produce insulin but their bodies are resistant to insulin. Both type of diabetes result in the patient having an abnormal high level of blood sugar. If untreated diabetes can be fatal and patients. An acute state of either hypoglycemia or hyperglycemia (either extremely high blood sugar or extremely low blood sugar) can directly lead to a coma or even death.

The nature of diabetes is that it alters a patient’s blood chemistry. This results in an impact on every major organ system throughout the body. Medline points out there are some well-known complications of chronically untreated or mismanaged diabetes and some less well known. It is common knowledge that having diabetes increases a patient’s risk for heart disease and stroke, but diabetes can cause nerve damage to the point, it may cause skin conditions, kidney disease as a result of high blood pressure, and may cause eye disease (Topics, 2021). The debilitating extreme scenarios in these situations include the need for an amputation, skin infections, loss of kidney function requiring dialysis or a transplant, and even blindness.

Diabetes is difficult disease that requires a lot of effort to manage. Patients with diabetes need to stay in contact with their physicians. Physicians will likely prescribe some combination of slow acting insulin, fast acting insulin intended for mealtime, and non-insulin medications. According to the American Diabetes Association, Physician switching of medications is common especially for fast acting insulin ("What to Do When Insurance Switches Your Medications | ADA", 2021). The American Diabetes Association indicates that at times the cause of switching medications may be insurance related to save costs but is also often caused by the physician working to find a more effective treatment or the efficacy of the treatment has changed over time.

# Motivation

Aside from the serious chronic health effects of diabetes the monetary costs of diabetes treatments have a major impact on individuals, health insurance companies, and government health care spending. The American Diabetes Association estimates that in 2017 the cost of diabetes treatment un the US was $237 billion dollars ("The Cost of Diabetes | ADA", 2021). The ADA lists that 30% of these costs are associated with hospital inpatient care and another 30% is associated with the prescription medications specifically to treat complications of diabetes ("The Cost of Diabetes | ADA", 2021). The remaining costs are associated with physician office visits and anti-diabetic agents and supplies. This means that in addition to greatly improving the quality of life of these patients properly managing their diabetes early on could save $142 billion dollars in health care costs in the United States.

The current gold standard for testing of diabetes is an HB1ac test which measures the amount of glycosylated hemoglobin in a patient’s blood. This means that to identify if a person is at risk for diabetes they need to search for a doctor, schedule an appointment, take time away from work, transport themself to the doctor’s office, get a referral for a blood test, schedule a blood test, take time off work, get transportation to the blood draw lab, and then have their blood drawn. Any one of these steps could be enough of hurdle that a patient may decide their risk levels are too low to bother. If a tool were available for people to enter basic health questionnaire data and tell them whether they are at elevated risk for diabetes this may increase the likelihood that those at risk would see a doctor. In addition, health insurance companies would benefit from having an algorithm to incentivize patients to be screened for diabetes based on health data the insurance companies already have. Keeping individuals healthy is much cheaper for these companies than having to pay for expensive medical care.

# About the Data

The data being used for the purpose of building this diabetes early screening tool is a the Diabetes Health Indicators Dataset available on Kaggle < <https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset> >. This dataset is a cleaned and filtered version of the Behavioral Risk Factor Surveillance System Dataset on Kaggle < <https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system> >. This is a collection of data that the Centers for Disease Control collects through phone surveys about the health data of individuals residing within the United States. The Diabetes Health Indicators Dataset used this data and selected all potentially relevant attributes related to diabetes and removed created a subset of individuals that have complete survey information. The first step in working with this dataset was to determine the dimensions of the dataset. The number of rows and columns are printed below.

dim(diabetes\_complete)

[1] 253680 22

## Number of rows in dataset: 253680

The Behavioral Risk Factor Surveillance System Dataset originally contained 441,455 entries. This version of the data has clearly been cleaned. To check that that no entries with missing attributes are still present all instances with one or more missing attributes were dropped and the number of remaining rows was printed.

## Number of rows after removing NAs: 253680

Next the attributes were examined by printing the first few rows of the dataset.

head(diabetes\_complete)

## Diabetes\_012 HighBP HighChol CholCheck BMI Smoker Stroke HeartDiseaseorAttack  
## 1 0 1 1 1 40 1 0 0  
## 2 0 0 0 0 25 1 0 0  
## 3 0 1 1 1 28 0 0 0  
## 4 0 1 0 1 27 0 0 0  
## 5 0 1 1 1 24 0 0 0  
## 6 0 1 1 1 25 1 0 0  
## PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost  
## 1 0 0 1 0 1 0  
## 2 1 0 0 0 0 1  
## 3 0 1 0 0 1 1  
## 4 1 1 1 0 1 0  
## 5 1 1 1 0 1 0  
## 6 1 1 1 0 1 0  
## GenHlth MentHlth PhysHlth DiffWalk Sex Age Education Income  
## 1 5 18 15 1 0 9 4 3  
## 2 3 0 0 0 0 7 6 1  
## 3 5 30 30 1 0 9 4 8  
## 4 2 0 0 0 0 11 3 6  
## 5 2 3 0 0 0 11 5 4  
## 6 2 0 2 0 1 10 6 8

From the above table the attributes available are Diabetes\_012, which is the label for each patient as 0 meaning no-diabetes, 1= prediabetes, 0r 2 diabetes. The attributes that follow are whether the person has high blood pressure, whether a person has high cholesterol, whether the person has had their cholesterol checked within the last 5 years, their body mass index, whether a person is a smoker, whether a person has ever had a stroke, whether the person has heart disease, whether a person has had physical activity in the last 30 days, whether they consume fruits on a daily basis, whether they consume vegetables on a daily basis, whether they consume more than 14 drinks per week for men or 7 drinks per week for men, whether they have health care coverage, whether they have avoided seeing a doctor in the past year due to costs, their self-assessed general health on a scale from 1 to 5, the number of days in the last month that they had poor mental health, the number of days when physical health was poor in the last month, whether they have difficulty walking, their sex, their age in one of 13 categories, their level of education on a scale from 1 to 6, and finally their income on a scale of 1 to 8 with 1 being under 10,000 dollars and 8 being more than 75,000 dollars.

The label in column 1 is of the highest importance because all of the other attributes will be used to predict the label. The number of individuals of each of the three categories is printed below.

## Number of healthy individuats: 213703

## Number of Pre-Diabetic individuats: 4631

## Number of Diabetic individuats: 35346

The counts of individuals above show that there are more than 40 times the number of individuals in the category no-diabetes as the pre-diabetic category. To ensure equitable representation across the categories when training all future models, a balanced dataset was created by splitting the dataframe into 3 one for each label category and then randomly sampling 4630 individuals from each subset and then combining them all into one dataset.

#create combined balanced dataset  
#get all healthy individuals  
healthy <- diabetes\_complete[diabetes\_complete$Diabetes\_012==0,]  
#sample 4630 healthy individuals by getting range from 1 to length of healthy, sampling and then using them as #index  
healthy <- healthy[sample(1:nrow(healthy), 4630, replace = FALSE),]  
  
#repeat for prediab  
preDiab <- diabetes\_complete[diabetes\_complete$Diabetes\_012==1,]  
preDiab <- preDiab[sample(1:nrow(preDiab), 4630, replace = FALSE),]  
  
#repeat for Diabetics  
diabetics<- diabetes\_complete[diabetes\_complete$Diabetes\_012==2,]  
diabetics <- diabetics[sample(1:nrow(diabetics), 4630, replace = FALSE),]  
  
#combine to form balanced dataframe  
balanced <- rbind(healthy, preDiab)  
balanced <- rbind(balanced, diabetics)  
  
head(balanced)

## Diabetes\_012 HighBP HighChol CholCheck BMI Smoker Stroke  
## 18017 0 1 0 1 30 0 0  
## 75760 0 0 0 1 28 0 0  
## 171401 0 0 1 1 27 0 0  
## 186819 0 0 0 1 31 1 1  
## 86443 0 0 1 1 24 0 0  
## 13123 0 0 1 1 28 1 0  
## HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump  
## 18017 0 1 0 1 0  
## 75760 0 1 1 1 0  
## 171401 0 1 1 1 0  
## 186819 0 0 1 1 0  
## 86443 0 1 1 1 0  
## 13123 1 1 0 0 0  
## AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age  
## 18017 1 0 3 20 5 1 0 11  
## 75760 0 0 2 0 0 0 0 12  
## 171401 1 0 3 0 30 1 0 4  
## 186819 1 0 2 0 0 0 1 9  
## 86443 1 0 1 0 0 0 1 10  
## 13123 1 0 3 4 5 0 1 12  
## Education Income  
## 18017 6 8  
## 75760 5 4  
## 171401 6 7  
## 186819 4 1  
## 86443 6 8  
## 13123 3 5

The counts of the individuals in each category are printed below.

## Number of healthy individuats: 4630

## Number of Pre-Diabetic individuats: 4630

## Number of Diabetic individuats: 4630

With most of the attributes being choice of two options either a 0 or 1 as in the example attribute Smoker 0= non-smoker, Smoker 1= smoker all of these attributes had to be converted into factor data types. An example of the code to do this for the Smoker attribute is below.

balanced$Smoker <- as.factor(balanced$Smoker)

The success of making all the attributes mentioned factors was checked and the structure of the remaining attributes was checked. The output is below.

str(balanced)

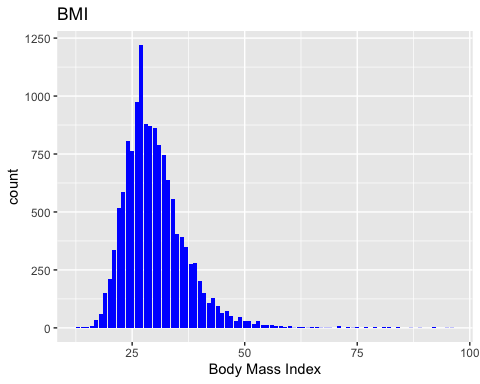
## 'data.frame': 13890 obs. of 22 variables:  
## $ Diabetes\_012 : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ HighBP : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 2 2 ...  
## $ HighChol : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 1 2 2 1 ...  
## $ CholCheck : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ BMI : num 30 28 27 31 24 28 22 31 27 26 ...  
## $ Smoker : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 2 1 ...  
## $ Stroke : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...  
## $ HeartDiseaseorAttack: Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...  
## $ PhysActivity : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 1 2 2 ...  
## $ Fruits : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 1 1 2 ...  
## $ Veggies : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 1 2 2 2 ...  
## $ HvyAlcoholConsump : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AnyHealthcare : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...  
## $ NoDocbcCost : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GenHlth : Factor w/ 5 levels "1","2","3","4",..: 3 2 3 2 1 3 4 1 2 3 ...  
## $ MentHlth : Factor w/ 30 levels "0","1","2","3",..: 20 1 1 1 1 5 1 1 1 1 ...  
## $ PhysHlth : Factor w/ 31 levels "0","1","2","3",..: 6 1 31 1 1 6 1 1 4 1 ...  
## $ DiffWalk : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 2 1 ...  
## $ Sex : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 1 1 1 2 ...  
## $ Age : Factor w/ 13 levels "1","2","3","4",..: 11 12 4 9 10 12 10 4 11 11 ...  
## $ Education : Factor w/ 6 levels "1","2","3","4",..: 6 5 6 4 6 3 4 5 4 6 ...  
## $ Income : Factor w/ 8 levels "1","2","3","4",..: 8 4 7 1 8 5 3 8 3 6 ...

GenHlth stands out from the rest of the attributes because it is one of the few attributes with more than 2 factors. In order to get a better idea of what this attribute looks like the bar plot below was created.



As this is an attribute based on a survey of what the individual believes is the quality of their general health it is unsurprising that the shape of the bar plot is a normal distribution or a bell curve, especially since there are tens of thousands of individuals in the balanced data set.

The next attribute that stands out from the list is BMI or body mass index. This is the only numerical attribute at this stage of working with the data. A histogram was created of all of all the possible BMI’s within this dataset. The resulting histogram is below.



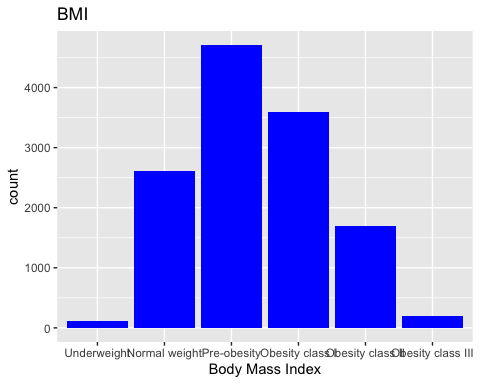
The histogram is skewed left with some outliers that are much larger than the median of the dataset. The mode BMI of individuals within this group is just below 30 and the median appears to slightly above 30. Some outliers are in the high 90s. In order to avoid the the outliers from having too large an effect on classifying individuals the BMI attribute was grouped in discrete bins. The bins were made according to the CDC’s published categories. According to the a BMI below 18.5 is underweight, 18.5 to 24.9 is healthy, 25 to 29.9 is overweight, 30.9 to 34.9 is obesity class I, 34.9 to 39.9 is obesity class II, and above 39.9 is obesity class III ("All About Adult BMI", 2021). The code below was used to create these discrete categories. The BMI attribute was the plotted once again in the bar plot below.

#Making BMI Discrete USing CDC BMI categories  
#https://www.cdc.gov/healthyweight/assessing/bmi/adult\_bmi/index.html  
balanced$BMI <- cut(balanced$BMI, breaks = c(10,18.5,24.9,29.9,34.9,39.9,40),   
 labels=c("Underweight", "Normal weight", "Pre-obesity", "Obesity class I", "Obesity class II","Obesity class III"))  
  
balanced <- na.omit(balanced)

balanced <- na.omit(balanced)

summary(balanced$BMI)

## Normal weight Pre-obesity Obesity class I Obesity class II   
## 2604 4708 3589 1701



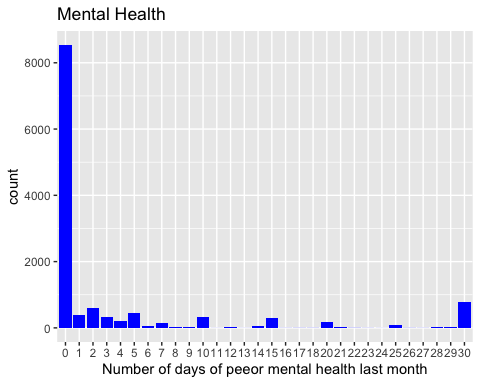
The barplot above shows that the distribution is no longer clearly skewed. This should decrease the chance of having a large outlier effect in the following models.

The next attribute to address is the MentHlth. This attribute asked individuals how many days they felt that the had poor mental health in the last 30 days. They mentioned some possible examples as having depression, anxiety, or other poor mental health. A plot of the original MentHlth attribute is below.

MenHlthPlot <- ggplot(balanced, aes(x=MentHlth)) +   
 geom\_bar(fill = "blue") +  
 ggtitle("Mental Health")+  
 xlab("Number of days of peeor mental health last month")  
 ylab("Number of People")+  
 theme(plot.title = element\_text(hjust = 0.5))

## NULL

MenHlthPlot

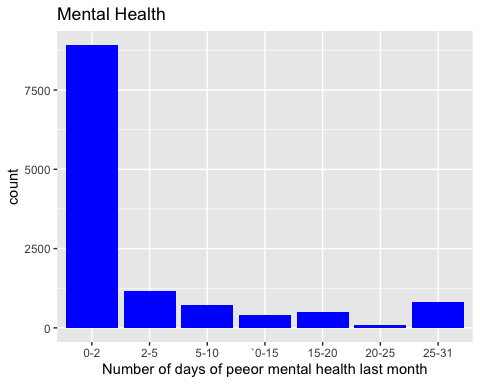


The plot shows that the majority of individuals reported they had poor mental health 0 days. The distribution is skewed very far right with a small number of individuals reporting up to 30 days of poor mental health. This attribute was binned into discrete groupings using the code below and then it was re-plotted.

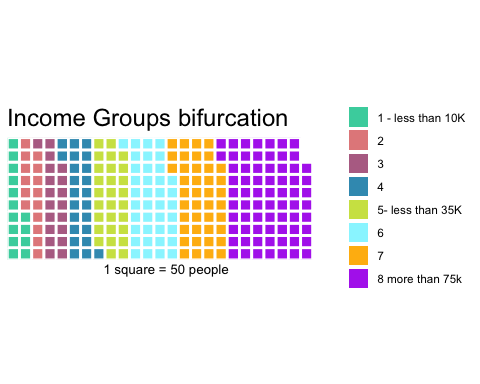
balanced$MentHlth <- as.numeric(balanced$MentHlth)  
balanced$MentHlth <- cut(balanced$MentHlth, breaks = c(0,2,5,10,15,20,25, 31),   
 labels=c("0-2", "2-5", "5-10", "`0-15", "15-20","20-25", "25-31"))  
  
MenHlthPlot2 <- ggplot(balanced, aes(x=MentHlth)) +   
 geom\_bar(fill = "blue") +  
 ggtitle("Mental Health")+  
 xlab("Number of days of peeor mental health last month")  
 ylab("Number of People")+  
 theme(plot.title = element\_text(hjust = 0.5))

## NULL

MenHlthPlot2



The next attribute to investigate is income, as it has it factor levels. One limitation of this dataset is not having a clear idea of the income groupings used. The source data mentions the income levels as group 1 having less than $10,000 a year, group, the fifth group has individuals that earn less than $35,000 a year, and the eighth group has individuals earning more than $75,000 a year. To visualize the distribution of income across this dataset the waffle plot below was created. Interestingly the largest category of respondents belonged to the largest income earners group.



To determine if there was an approximately even split between males and females the pie chart below was created. Interestingly there were more males than females. If sex turns out to be a major factor in deciding diabetes state a new subset of data that has a balanced number of males of females will be taken. Chart, pie chart

Description automatically generated

The two attributes relating to diet, whether an individual regularly consumes fruit and whether they regularly consume vegetables stood out as behavioral questions that might be directly linked to BMI. If that is the case the fruit and vegetable attributes may need to be removed as they may be inflating the effect of BMI on classifying an individual with diabetes. To test this a t test was run to compare the BMI of individuals who eat vegetables compared to the BMI of individuals that do not regularly eat vegetables. Not this t test was conducted on the dataset diabetes\_complete dataset before BMI was changed from a numeric variable into a discretized set of groups.

#running a t test to see if difference in BMI based off those who eat veggies  
  
t.test(diabetes\_complete$BMI[diabetes\_complete$Veggies==0],diabetes\_complete$BMI[diabetes\_complete$Veggies==1])

##   
## Welch Two Sample t-test  
##   
## data: diabetes\_complete$BMI[diabetes\_complete$Veggies == 0] and diabetes\_complete$BMI[diabetes\_complete$Veggies == 1]  
## t = 30.181, df = 68691, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.9837803 1.1204299  
## sample estimates:  
## mean of x mean of y   
## 29.23606 28.18396

The null hypothesis of this test would be that there is no significant difference between the BMI of an individual that regularly eats vegetables and one that does not. The p value is 2.2e-16 which is significantly less than the most commonly used alpha values of 0.1 or 0.05. This would indicate that we should reject the null hypothesis and there is a statistically significant difference in the BMI of veggie eaters and non-veggie eaters. However, below it shows the mean of x 29.2 and the mean of y is 28.1 which are not far apart. To investigate further a side by side box plot was used to look at the distributions as well as the bar plot below.

noveggielist <- diabetes\_complete$BMI[diabetes\_complete$Veggies==0]

veggielist <- diabetes\_complete$BMI[diabetes\_complete$Veggies==1]

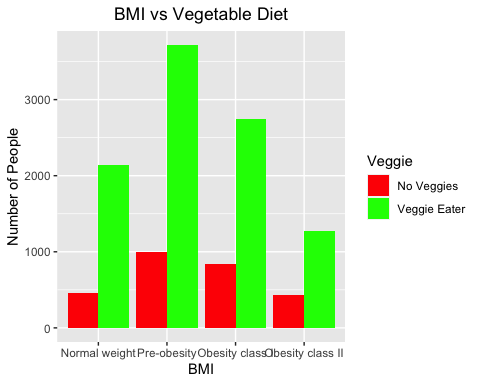
par(mfrow=c(1,1)

boxplot(noveggielist,veggielist, main = "BMI of veggie eater vs non veggie eater", ylab = "BMI",names = c("Non-Veggie eaters", "Veggie eaters"))

bmiVeg <- ggplot(balanced,aes(x=BMI, fill=Veggies))+  
 geom\_bar(position="dodge")+  
 ggtitle("BMI vs Vegetable Diet")+  
 ylab("Number of People")+  
 theme(plot.title = element\_text(hjust = 0.5))+  
 scale\_fill\_manual(name="Veggie",values=c("red", "green"),labels=c("No Veggies", "Veggie Eater"))  
bmiVeg

Chart, box and whisker chart

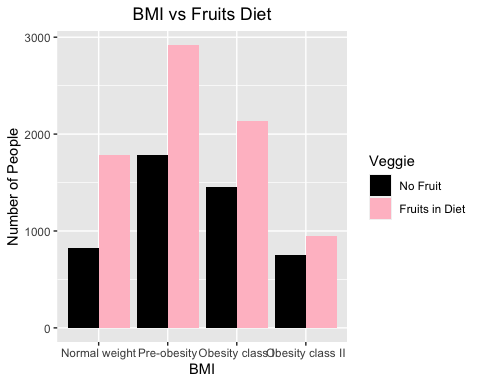
Description automatically generated



The box plots show that non-veggie eaters have a slightly higher median BMI but the two are extremely close. The bar-plots show that there are similar distributions among the groups but that there are far more veggie eaters than non-veggie eaters which might explain the surprising t-test results. All of these results together give strong evidence that the veggie eater attribute is different from the BMI attribute and would not artificially inflate the use of BMI as an indicator of diabetes prediction.

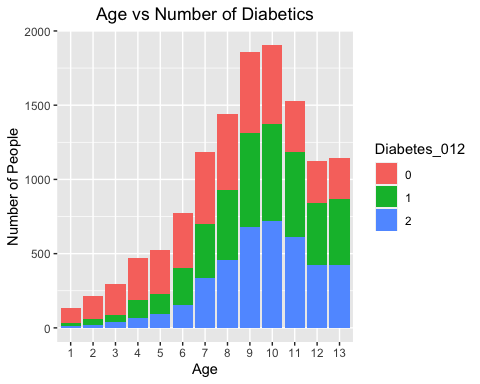
A similar investigation the BMI individuals who regular consume fruit was conducted. The t-test, boxplots, and bar plots are below.

## Welch Two Sample t-test  
##   
## data: diabetes\_complete$BMI[diabetes\_complete$Fruits == 0] and diabetes\_complete$BMI[diabetes\_complete$Fruits == 1]  
## t = 43.288, df = 180846, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.146486 1.255230  
## sample estimates:  
## mean of x mean of y   
## 29.14402 27.94316Chart, box and whisker chart

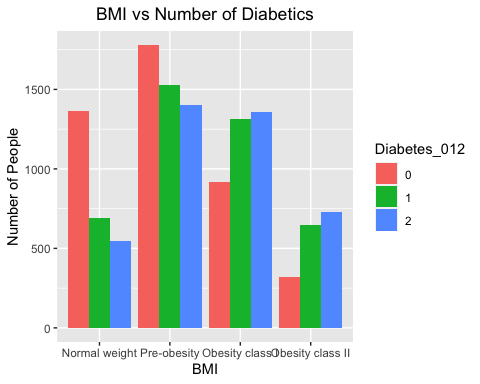
Description automatically generated 

The results are similar with all three methods. The t test shows there is a statistically significant difference but the difference is very slight. Interestingly the difference is a BMI difference of 2 rather than 1 based on fruits. The bar plot shows similar distributions but that there are more individuals consume fruit than individuals that do not. Again all of this evidence together suggests that the fruit attribute can remain in further analysis.

The next visualizations used was to look at how many diabetics there are different in each age groups. The X axis of the age vs Diabetes plot is not ages but age groups. Similar to the limitation with income groupings in this dataset it is not clear what exact ranges below to each of the 13 age groups below. The diabetes vs age plot show that there is an increased risk of becoming pre-diabetic or diabetic as a person gets older.



The final plot looked at how many diabetics there in each BMI group. This plot shows that group 0 in red were there no diabetics was skewed right with most individuals in the normal weight and pre-obesity not having diabetes and a small number of obese individuals without diabetes. The other two diabetes categories had a more bell-shaped curve across BMI levels. This indicates that having a BMI lower than needed to be part of the category obese decreases an individual’s risk of being pre-diabetic or diabetic.



# Analysis

The most useful model for from a patient outreach or public health perspective is one that a doctor might be able to explain to a patient and not necessarily one that has the highest accuracy. The ability to create a visual guide to patients on what are the most important risk factors for diabetes was used. The complexity of how the model was created is not of major concern because patients are not as likely to ask about the foundational research behind medical practices as much as how to use the medical guidelines. An example of this is the long list of directions and contradictions found on prescription medications and the absence of medical journal articles on how the medication came to be designed on prescription bottles. After creating this type of model further models meant to for the use organizations such as health insurance providers or the government in the case of Medicare coverage analysis, which prioritize accuracy of readability will be created.

The first step to create a decision tree model is to split the balanced dataset into a training data set and a testing dataset. Two thirds of the dataset was used for training and one third was reserved for testing. The code below was used to make the first decision tree model and fancyplot below.

Graphical user interface, text, application

Description automatically generated

train\_tree <- rpart(Diabetes\_012 ~., data = train, method="class", control=rpart.control(cp=0, minsplit = 50, maxdepth = 5))   
  
fancyRpartPlot(train\_tree)



An immediately apparent issue with this model is the top deciding factor for classifying diabetes state is a subjective question asked to the individual. The general health attribute is the top attribute for splitting this tree is whether an individual said their health was excellent or not. Having such a subjective question at the top of this tree would does not fulfill the requirement of having a tool to show patient’s their risk factors and what potential behavioral changes are recommended. In order to account for this, a new decision tree was created that excludes the attribute GenHlth. In the new fancyplot below the top attribute to split is high blood pressure.

#repeating a decision tree to exclude subjectivity from patient question on their general health  
# Decision Tree  
train2 <- subset(train,select=-c(GenHlth))  
  
  
train\_tree2 <- rpart(Diabetes\_012 ~., data = train2, method="class", control=rpart.control(cp=0, minsplit = 50, maxdepth = 5))   
  
fancyRpartPlot(train\_tree2)



The training accuracy of both versions of the decision tree were investigated to see if removing the GenHlth caused a reduction in accuracy of the model. The accuracy of the two decision trees are very similar to one another, neither of the two decision trees had a high level of training accuracy. Because both of these had a training accuracy around 50% there was no need to measure the testing accuracy.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1 2  
## 0 1845 507 563  
## 1 712 852 1152  
## 2 438 621 1627  
##   
## Overall Statistics  
##   
## Accuracy : 0.5199   
## 95% CI : (0.5091, 0.5307)  
## No Information Rate : 0.4018   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.2794   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Statistics by Class:  
##   
## Class: 0 Class: 1 Class: 2  
## Sensitivity 0.6160 0.4303 0.4868  
## Specificity 0.7989 0.7059 0.7871  
## Pos Pred Value 0.6329 0.3137 0.6057  
## Neg Pred Value 0.7871 0.7986 0.6954  
## Prevalence 0.3601 0.2381 0.4018  
## Detection Rate 0.2218 0.1024 0.1956  
## Detection Prevalence 0.3505 0.3266 0.3230  
## Balanced Accuracy 0.7075 0.5681 0.6370

#Accuracy without GenHlth  
trainPred = data.frame(predict(train\_tree2, train2))  
trainPred = as.data.frame(names(trainPred[apply(trainPred,1,which.max)]))  
colnames(trainPred) = 'prediction'  
trainPred$number = substr(trainPred$prediction, 2,2)  
trainPred = train2 %>% bind\_cols(trainPred) %>% select(Diabetes\_012, number) %>% mutate(label=as.factor(Diabetes\_012), number=as.factor(number))  
confusionMatrix(trainPred$Diabetes\_012, trainPred$number)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1 2  
## 0 1808 310 797  
## 1 693 638 1385  
## 2 460 422 1804  
##   
## Overall Statistics  
##   
## Accuracy : 0.511   
## 95% CI : (0.5002, 0.5218)  
## No Information Rate : 0.4793   
## P-Value [Acc > NIR] : 3.722e-09   
##   
## Kappa : 0.2665   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Statistics by Class:  
##   
## Class: 0 Class: 1 Class: 2  
## Sensitivity 0.6106 0.46569 0.4526  
## Specificity 0.7933 0.70088 0.7964  
## Pos Pred Value 0.6202 0.23490 0.6716  
## Neg Pred Value 0.7866 0.86931 0.6125  
## Prevalence 0.3560 0.16472 0.4793  
## Detection Rate 0.2174 0.07671 0.2169  
## Detection Prevalence 0.3505 0.32656 0.3230  
## Balanced Accuracy 0.7020 0.58329 0.6245

After creating a basic model with high readability the next goal is to create a model focused more on accuracy. The comparison between the two previous decision trees did not show a significant loss of accuracy from eliminating the GenHlth attribute so it will be excluded from all analysis going forward. After excluding the GenHlth attribute there are 20 attributes and one label. Support vector machines are known to have acceptable accuracy when working with large numbers of attributes and they are computationally as costly as running a k nearest neighbor when creating a model with this many attributes. The code to generate an svm model is below.

# Support Vector Machine (SVM)  
  
svm.model <- train(Diabetes\_012 ~ ., data = train2, method="svmRadial",metric=metric,trControl=control,  
 tuneLength = 5)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1 2  
## 0 2053 463 399  
## 1 579 1460 677  
## 2 431 538 1717  
##   
## Overall Statistics  
##   
## Accuracy : 0.6288   
## 95% CI : (0.6183, 0.6392)  
## No Information Rate : 0.3683   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4426   
##   
## Mcnemar's Test P-Value : 1.347e-06   
##   
## Statistics by Class:  
##   
## Class: 0 Class: 1 Class: 2  
## Sensitivity 0.6703 0.5933 0.6148  
## Specificity 0.8359 0.7855 0.8246  
## Pos Pred Value 0.7043 0.5376 0.6392  
## Neg Pred Value 0.8130 0.8213 0.8089  
## Prevalence 0.3683 0.2959 0.3358  
## Detection Rate 0.2468 0.1755 0.2064  
## Detection Prevalence 0.3505 0.3266 0.3230  
## Balanced Accuracy 0.7531 0.6894 0.7197

The svm model had a higher training accuracy than the previous model 62% is not sufficient for the purpose of this project. Looking back at the previous attributes, it is possible to filter out some extreme cases in the BMI attribute. The previous bar plot showed a small minority of individuals belong to the underweight group and the Obesity class III group. These groups were dropped from the dataset and BMI was re-plotted to see how this new distribution looks.

balanced <- subset(balanced, BMI != "Underweight")  
balanced <- subset(balanced, BMI != "Obesity class III")  
balanced$BMI <- droplevels(balanced$BMI)

## NULL

Chart, bar chart

Description automatically generated

To improve accuracy the training model was simplified further by grouping pre-diabetic individuals with diabetic individuals. The business case justification for this being that organizations are going to be using this tool to flag either health insurance subscribers or Medicare recipients to go see a doctor to further investigate their diabetes risk. The intended use case for this was not meant as a diagnostic aide for doctors to use in their practice. The code used group the pre-diabetics and diabetics together is below. The cross validation parameter remained set to 3 for all of these models.

#Trying to improve accuracy by making classification as either health or flag for diabetes

smp\_size <- floor(0.66 \* nrow(balanced))

balancedBinary <- balanced

balancedBinary$Diabetes\_012[balancedBinary$Diabetes\_012==2] <- 1

balancedBinary$Diabetes\_012 <- as.factor(as.numeric(balancedBinary$Diabetes\_012))

trainBinary\_ind <- sample(seq\_len(nrow(balanced)), size = smp\_size)

trainBinary <- balancedBinary[trainBinary\_ind, ]

testBinary <- balancedBinary[-trainBinary\_ind, ]

# Creating a control with cross validation of 3

control <- trainControl(method ='cv',number = 3)

# Metric for comparison will be accuracy for this project

metric <- "Accuracy"

This new training dataset with only two possible labels was used to create three new models a support vector machine which would be the most appropriate to use for the number attributes used, a random forest which is very effective at finding the most influential attribute among all of these attributes through the creation of many random decision trees, and a k nearest neighbors model. Of these the random forest ran the quickest because it was able to take advantage of parallel processing but the others were not. The knn model took the longest to run. The code for these models is below along with a dot plot showing the accuracy of each. One important note is that changing the label data type to numeric and back to factor causes the labels to change. From this step forward label 1 is no indication of diabetes and label 2 is either prediabetic or diabetic.

rf.modelbinary <- train(Diabetes\_012 ~ ., data = trainBinary, method="rf", metric=metric, trControl=control,

tuneLength = 10)

knn.modelbinary <- train(Diabetes\_012 ~ ., data = trainBinary, method="knn", metric=metric, trControl=control,

tuneLength = 5)

svm.modelBinary <- train(Diabetes\_012 ~ ., data = trainBinary, method="svmRadial",metric=metric,trControl=control,

tuneLength = 5)

results <- resamples(list(Random\_Forest=rf.modelbinary, knn=knn.modelbinary, SVM=svm.modelBinary))  
  
dotplot(results)

Chart

Description automatically generated with low confidence

The accuracy of these models is higher than that of any previous model. All three have similar accuracy with the SVM the highest. In this case the SVM had an accuracy of 75%. The testing accuracy of the newest SVM model was measured in the confusion matrix below.

svm.modelBinaryPRED = data.frame(predict(svm.modelBinary, testBinary))

colnames(svm.modelBinaryPRED) = 'prediction'

svm.modelBinaryPRED = testBinary %>% bind\_cols(svm.modelBinaryPRED) %>% select(Diabetes\_012, prediction) %>%

mutate(label=as.factor(Diabetes\_012), prediction)

confusionMatrix(svm.modelBinaryPRED$Diabetes\_012, svm.modelBinaryPRED$prediction)

Text

Description automatically generated with medium confidence

# The SVM shows 75% accuracy on the testing dataset. The confusion matrix shows that the model will flagged 30% of individuals that should not be and it will miss 23% individuals who should be flagged. Of these two types of errors missing individuals who should be checked by doctor is the problematic in this case. Further optimization of this model could be performed by re-balancing individuals how now belong to category one with the number of individuals number of individuals that belong to category two. Adjustment of the cross validation parameter could also yield a more accurate model in future aspects of this project.

Moving on two the decision tree model. The new dataset which has only two labels was used to create a new decision tree. Keeping in mind the goal of this model is to have high readability for ease of understanding by any individual looking to assess their personal level of risk for diabetes, the max depth of this tree was shortened to four.

train\_treeBinary <- rpart(Diabetes\_012 ~., data = trainBinary, method="class", control=rpart.control(cp=0, minsplit = 50, maxdepth = 4))

fancyRpartPlot(train\_treeBinary)

Graphical user interface, application

Description automatically generated

The training accuracy for this model is in the confusion matrix below.

Table

Description automatically generated

Finally, the test accuracy of the decision tree is below.

Table

Description automatically generated

The decision tree model showed a training accuracy of 71% and a testing accuracy of 72%. The accuracy of this model could be improved by altering the number of splits as well as the minimum number of individuals in each node.

# Conclusion

In this project the Diabetes Health Indicators Dataset available on Kaggle was explored. The feasibility of creating two separate models to predict risks of diabetes based on health indicators and behavioral habits. The first of these models was a decision tree intended to be a tool for educating patients about what aspects of their personal health maybe putting them at risk for developing diabetes. The testing accuracy of the final version of the decision tree had an accuracy of 72%.

The second model created was a support vector machine used to help institutions that pay for medical coverage to communicate with those who receive the care that they should see a doctor to be tested for diabetes. This model would use health records available to them and could send this information to the care receivers without necessarily explaining to them what it is about their health history caused them to be flagged diabetes test. The support vector machine had a 75% accuracy. In the case where the model was presented with an individual that should be flagged for a health screening it missed 23% of these individuals. This is problematic because the goal would be to have the highest number of individuals that need to be tested visit a doctor without wasting too many resources. The financial and quality of life cost of telling a healthy person to get their blood sugar tested is much lower than the cost of having a person who needs to seen by doctor missed. The current version of the model has some utility in that it can start to motivate some individuals that need to see a doctor to schedule a visit but there is room for improvement. Changing of how some variables are filtered, which variables should be excluded, and model parameters could improve the model accuracy.

# Citations

All About Adult BMI. (2021). Retrieved 18 December 2021, from https://www.cdc.gov/healthyweight/assessing/bmi/adult\_bmi/index.html

National Diabetes Statistics Report, 2020 | CDC. (2021). Retrieved 14 December 2021, from <https://www.cdc.gov/diabetes/data/statistics-report/index.html>

The Cost of Diabetes | ADA. (2021). Retrieved 14 December 2021, from https://www.diabetes.org/resources/statistics/cost-diabetes

Topics, H. (2021). Diabetes Complications | MedlinePlus. Retrieved 14 December 2021, from https://medlineplus.gov/diabetescomplications.html

What to Do When Insurance Switches Your Medications | ADA. (2021). Retrieved 14 December 2021, from https://www.diabetes.org/healthy-living/medication-treatments/insurance-switching-diabetes-medication