STA 325 Final Project Report

Emily Mittleman & Julia Rosner

2022-12-18

Introduction

Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood. There are different types of diabetes, but Type II diabetes mellitus is the most common. If left untreated, Type II diabetes can cause major health complications, including heart attack, kidney failure, stroke, and eye damage. In fact, Type II diabetes was the 7th leading cause of death in 2019; and unfortunately, its prevalence is rapidly increasing worldwide (CDC). According to the CDC, more than 37 million people in the United States have diabetes, and 1 in 5 people are unaware that they have it. Furthermore, approximately 96 million Americans (1 in 3) have prediabetes, and a shocking 80% of those Americans are unaware of their risk (CDC). Type II diabetes and prediabetes often begin as silent conditions, and so they often go undiagnosed for years with no clear symptoms, until serious health complications develop.

Although diabetes is an irreversible disease, it is largely preventable. The risk of developing diabetes can be reduced significantly through early detection of prediabetes and lifestyle interventions. While type 2 diabetes and prediabetes can be easily diagnosed through glucose blood testing, many people fail to test regularly. Access to healthcare and health insurance plays a large role in testing, diagnosis, and risk factors. Without it, Type II diabetes is difficult to detect early on. As a consequence, research shows that diabetes affects racial and ethnic minority and low-income adult populations in the U.S. disproportionately (Briggs). Therefore, evaluating diabetes risk through metrics other than glucose levels can prove to be extraordinarily valuable.

The prevalence of type II diabetes varies by age, education, income, other social determinants of health, risk behaviors, and chronic health conditions. For our project, we use these indicators to build a predictive model that aims to (1) identify individuals with diabetes, who could otherwise go undiagnosed, and (2) indicate individuals who are at high risk for diabetes. Our model is meant to be implemented by any medical professional, and used in all healthcare settings, from clinics to private practices. A major differentiator for our model is that it is accessible to all individuals – including those without a regular physician and who have limited health records. Clinicians can then implement our model as a part of any and all healthcare visits. If a clinician sees a result that indicates diabetes risk, then they can proceed with a glucose level test to determine whether or not there is a diagnosis. If there is a diagnosis they can proceed with the medical protocols/advice established. However, what separates our model is even if there is no diagnosis, then the model still indicates that the patient was at risk, and so the patient can then be proactive in lowering their risk for Type II diabetes, and implement preventative measures.

Data

Our data was obtained from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey collected annually by the CDC and gathers responses from over 400,000 Americans on health-related risk behaviors and chronic health conditions. For this project, a CSV of the dataset available on Kaggle was used. This original dataset contains responses from 441,455 individuals and has 330 features.

This dataset is sufficient in meeting our project goals since it has a significantly large number of observations, and a large number of predictors that can all easily be measured noninvasively in clinical settings.

The response variable is an indicator of whether someone does not have diabetes (0) or has diabetes (1).

Data Cleaning

The dataset originally had 330 features, but based on diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions, only 21 select features are included in this analysis. We then dropped any observations containing null values and duplicate observations from the dataset. After removing null and duplicate values, we were left with ~230,000 observations.

The original diabetes response variable has 3 levels that indicates an individual's diabetes diagnosis. Level 0 indicates that the individual does not have diabetes, level 1 indicates that the individual has prediabetes, and level 2 indicates that the individual has diabetes. We found that the data distribution across these three levels was very unbalanced. The no diabetes class (level 0) made up a very large majority of the observations. Meanwhile, there were hardly any observations for the prediabetes class (level 1). Since there were so few observations for prediabetes and prediabetes is very similar biologically to diabetes, we decided to merge the prediabetes and diabetes classes into one.

There are 21 predictors: most are binary indicators, and some are discrete data such as age, BMI, health over the past month, etc. All of these predictors are noninvasive measurements commonly taken in medical settings, and can easily be collected by doctors at physical checkups to be able to run our predictive algorithm in order to determine diabetes risk.

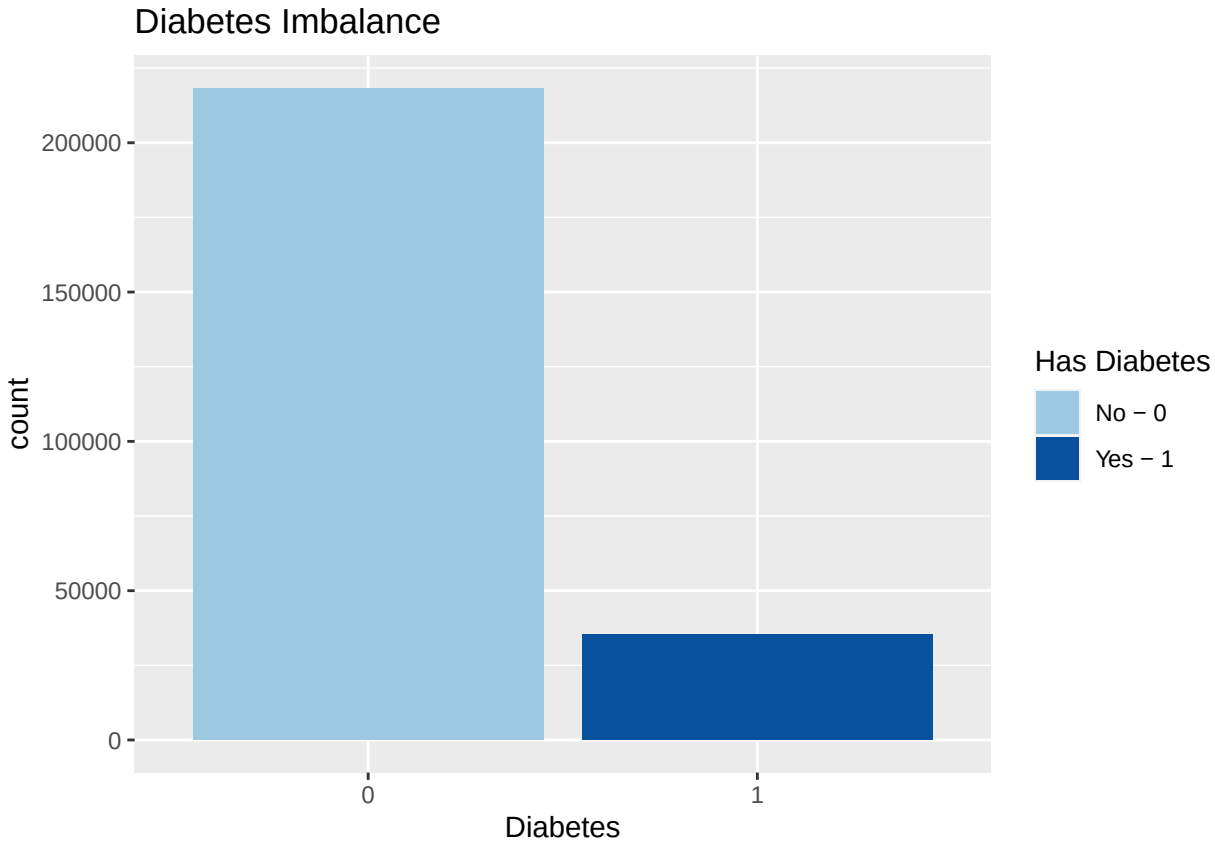
Table 1: Predictor Descriptions

Predictor	**Description**	**Data Type**
High BP	Has high blood pressure	Binary
High cholesterol	Ever had high cholesterol	Binary
Cholesterol check	Cholesterol check within past five years	Binary
BMI	Body Mass Index	Discrete (1-98)
Smoker	Smoked at least 100 cigarettes in entire life	Binary
Stroke	Ever had a stroke	Binary
Heart disease	Ever had coronary heart disease	Binary
Physical activity	Exercised within the past 30 days	Binary
Fruits	Consume fruit 1 or more times per day	Binary
Vegetables	Consume vegetables 1 or more times per day	Binary
Heavy alcohol use	Men: >14 drinks weekly, Women: >7 drinks weekly	Binary
Any healthcare	Has any kind of health care coverage	Binary
No doctor (cost)	Needed doctor in past year but couldn't go due to cost	Binary
General health	Scale of 1-5	Discrete (1-5)
Mental health	Days of poor mental health in past 30 days	Discrete (1-30)
Physical health	Physical illness or injury in past 30 days	Discrete (1-30)
Difficulty walking	Difficulty walking or climbing stairs	Binary
Sex	Male or Female	Binary
Age	Which age group (18-24, 24-30,...)	Discrete (1-13)
Education	Highest level of education (None, elementary,...)	Discrete (1-6)
Income	Annual income bracket (<$10k, $10k-$15k,...)	Discrete (1-8)

EDA

To help guide our model selection, we investigated our data further in our exploratory data analysis.

We looked at variable distributions and correlations to understand our dataset better. First, we explored the distribution of the response variable diabetes. Even after combining the multiclass variable into a binary response of 0 = no diabetes and 1 = diabetes or prediabetes, there was a very uneven split between the positive and negative classes.

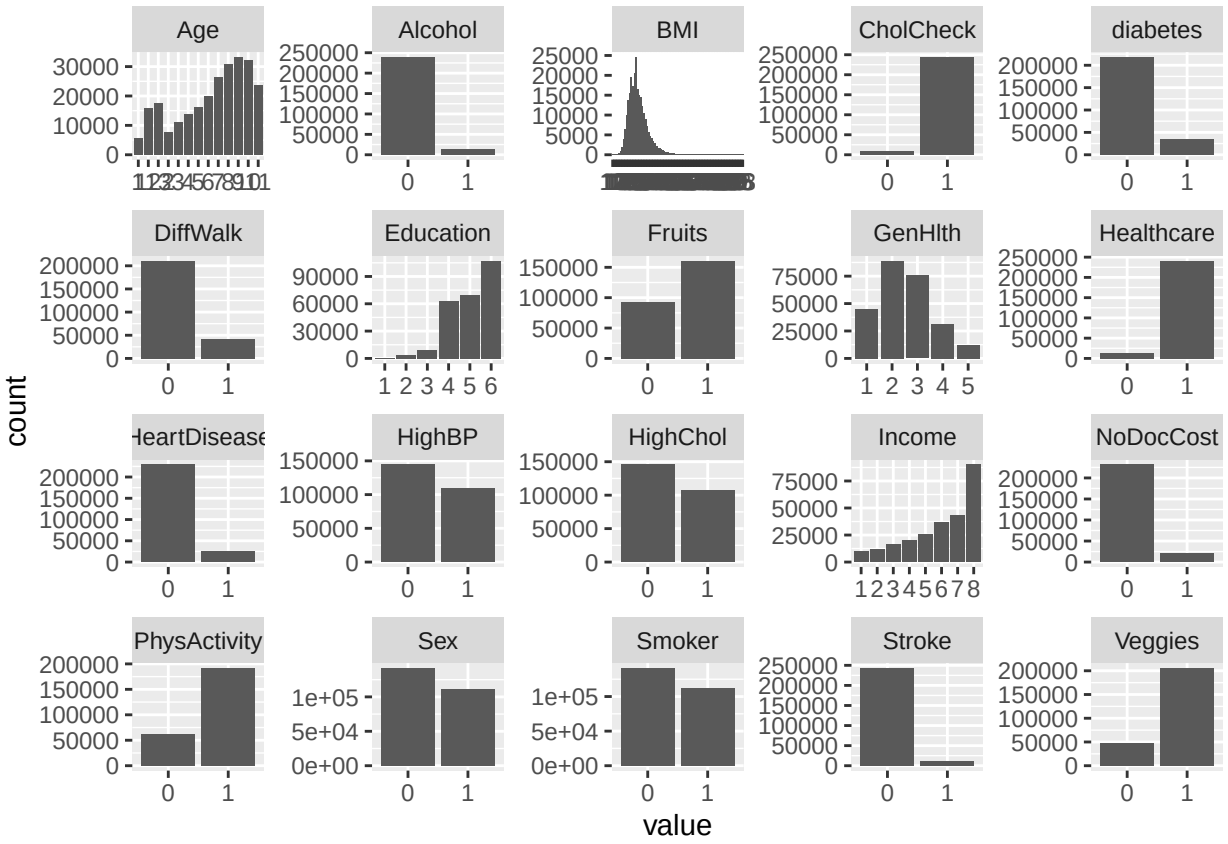


We were concerned that this uneven split would hurt our model's prediction accuracy, which is the core objective of our model. Therefore, we attempted two methods to handle the imbalanced data: the first method was to undersample the negative class, and the other was to oversample the positive class.

1) In undersampling the data, we removed observations that were classified as not having diabetes in order to make the number of observations classified as having diabetes equal to the number not having diabetes. This resulted in an equal split of the dataset consisting of 35,346 observations classified as diabetes = YES and 35,346 observations classified as diabetes = NO, for a total of 70,692 observations in the dataset. The downside to undersampling is we remove a lot of useful data from the negative class in order to balance out the two classes, resulting in a smaller dataset and less observations to train models on.

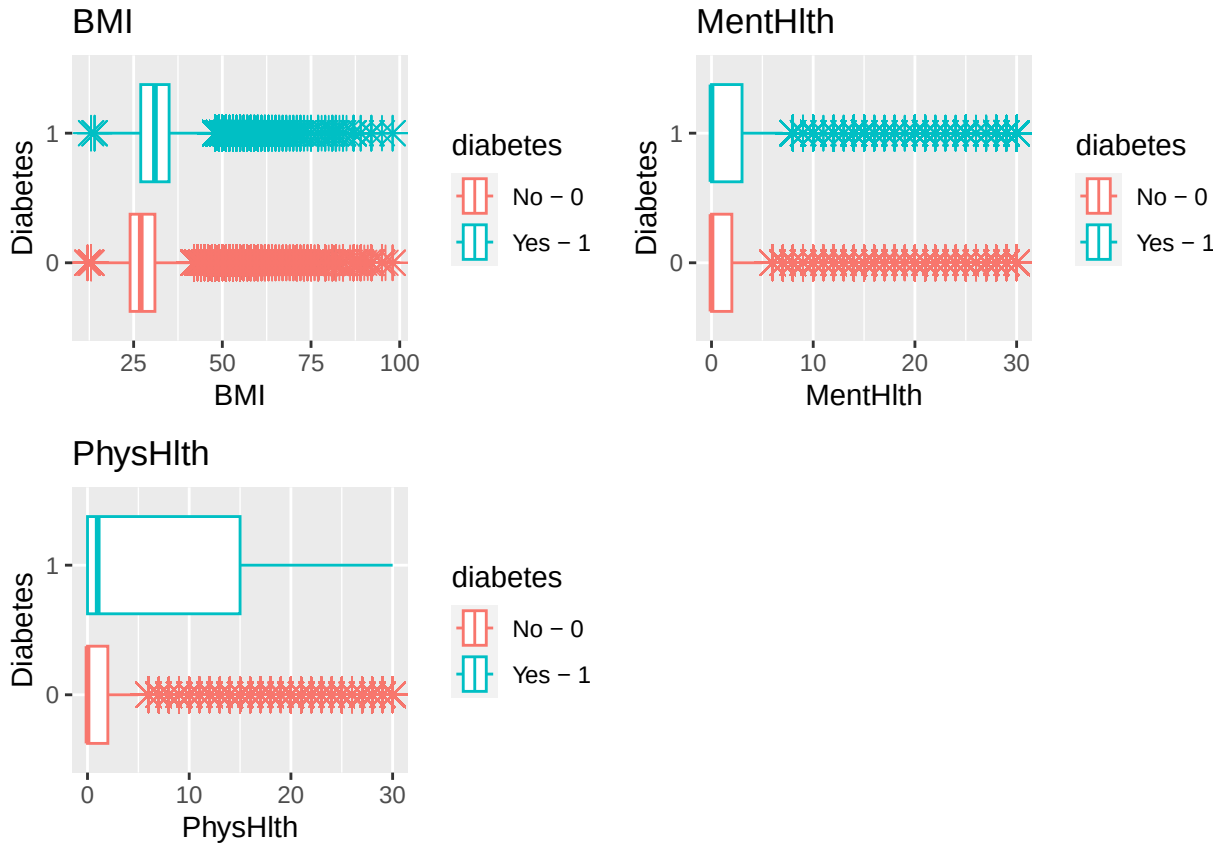
2) For the second method, we tried to oversample the data. We upsampled observations that were classified as diabetes = YES in order to make the number of observations classified as having diabetes equal to the number not having diabetes. This resulted in an equal split of the dataset consisting of 194,377 observations classified as diabetes = YES (which now contains duplicates) and 194,377 observations classified as diabetes = NO, for a total of 388,754 observations in the dataset. The benefit of oversampling is we get to utilize all negative class observations so we have a larger dataset, but the downside is we have duplicate values for the positive class observations. To combat any issues with duplicate values that can arise when training and validating models, we first split up the original dataset into training and test sets, and then only oversample the training set. This is very important because if there are tons of duplicate values between the training and test sets, then we won't be able to accurately access the model on the holdout test set because the model would have been trained on many of those observations.

Then, we looked at the distributions of our predictor variables.

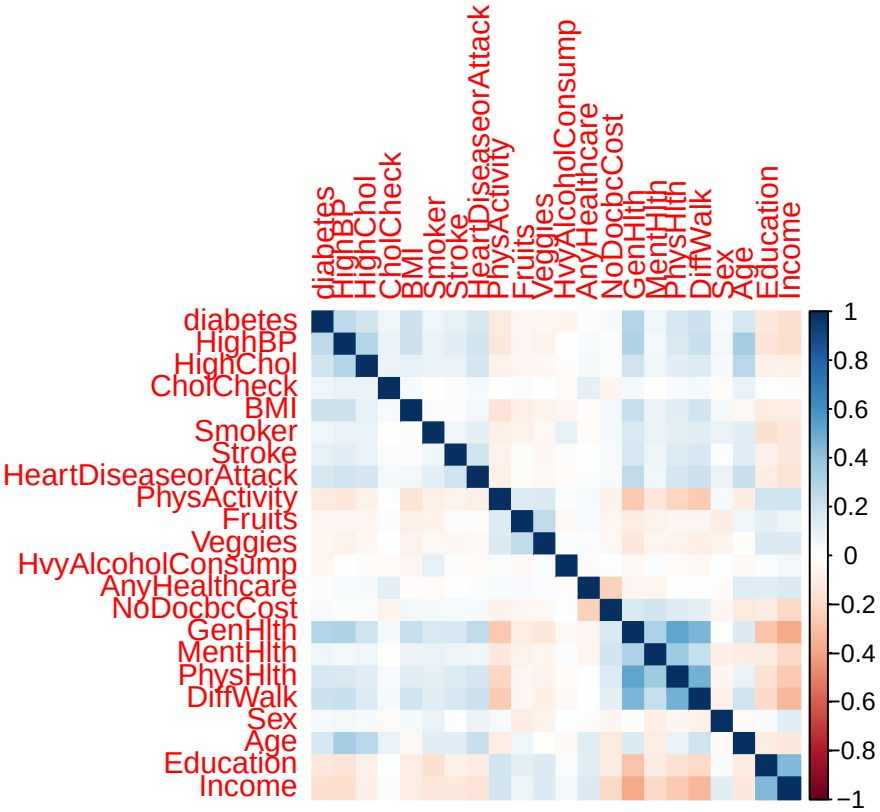


We can see that all of our variables are binary or discrete. Out of the discrete variables, only the Age, BMI, and GenHlth variables follow a roughly normal distribution (Education and Income do not). Also, many of the binary variables have very non-uniform distributions, including Alcohol, Cholesterol check, Difficulty walking, Stroke, etc. These non-normal and non-uniform variable distributions could pose a problem in our modeling if they are highly correlated with diabetes.

We further explored these variables by checking to see if they had any outliers. We found that only the predictors BMI, Mental health, and Physical health had a lot of outliers. However, in the end, we decided not to transform our variables for our final model because scaling is not necessary for random forests, and random forests are also not sensitive to outliers.



Next, we looked at the correlations between our predictor variables.



Several variables seem to be correlated to each other, however, random forests work by taking subsections of the predictors. Thus, our random forest model handles multicollinearity itself, and we do not need to take any action here.

Methodology

Our main goal of predicting whether an individual has diabetes or is at risk requires us to focus on making the most accurate predictive algorithm. Interpretability of the model is not important for our goals since doctors know what factors lead to higher risk of diabetes; we want a model that can take into account numerous medical, behavioral, and environmental factors of an individual and based on the combination of all of these things, flag to the doctor whether they are predicted to have a high risk of being a diabetic. We are not necessarily trying to have doctors learn from our model, but rather use our algorithm as a precaution in case doctors miss the signs of diabetes. Since we are not concerned with inference, we focused on making more complex models to achieve greater predictive power.

We want a highly accurate predictive model, but more specifically, we are looking for high accuracy in detecting when someone is at risk or has diabetes. Since the model will be used to flag cases of potential diabetics, we need to reduce the possibility of missing these cases as much as possible: it is better for the model to output more false positives than false negatives, because a false positive can be corrected by running blood tests on the individual to determine that they don't actually have diabetes, while a false negative means the doctor would miss this case and leave the individual undiagnosed which is extremely problematic and directly opposes the intention of this study. Since we are trying to optimize for correctly classifying all positive cases of diabetes, the most important metric we use to assess model fit is the sensitivity rate. In tandem, we also look at overall model accuracy, because a model that predicts a positive result for all inputs would have an extremely high sensitivity, but this means every individual would be predicted to have diabetes and lots of unnecessary testing which is bad. Overall, we want an accurate model that minimizes misclassifying positive cases as negative.

After exploring logistic regression, K-nearest neighbors, and decision tree classifiers, we were not satisfied with the accuracy of these models. Since the performance of these simpler models showed much room for improvement, we further explored Random Forest Classifiers. We expected random forests to perform better compared to decision trees because it uses multiple decision trees to get the optimal result by choosing the majority among them as the best value. Random forests have a smaller chance of overfitting since we are using multiple decision trees, and it has greater accuracy since it runs on a larger data set.

We chose Random Forest over using Support Vector Machine (SVM) because SVM is inefficient as the number of data points grows very large, and we have two data sets that we need to train every model on (constructed by under and oversampling the data due to class imbalances) which have 70,692 and 321,178 observations respectively. Since the number of observations is very large, we chose random forests over SVM.

The best performing random forest was trained using the undersampled dataset where we had removed negative class observations to be balanced with the positive class. After confirming that our model outperformed any other model we created thus far, as it had extremely high accuracy and sensitivity, we trained our final model. We did a 70/30 training-test data split in which we trained the model on 70% of our data, and then validated it on the remaining 30% of unseen data to evaluate its predictive performance. This gave us confidence in our model that it is able to predict new data points very well without being overfit, and this model missed the lowest number of positive cases giving us confidence that it won't misclassify diabetics as wrongly predicting they are not diabetic.

To be absolutely sure this model was the best random forest, we tried training on the full dataset without sampling, the undersampled data, and oversampled data. For each of these methods, we trained multiple models on different subsets of the predictors, but we found that including all predictors gave the most accurate predictions every time. After trying many iterations of different models, we were confident that we found the best performing model having the highest accuracy and sensitivity. We conclude our search for and training of our predictive model, and move on to discuss the results we found from our final predictive random forest model.

Table 2: Random Forest Accuracy Comparisons

Training Dataset	Accuracy	Sensitivity
Undersampled	90%	89.1%
Full	85.2%	54.8%
Oversampled	79.8%	38.1%

Results

We used our final predictive random forest model to make predictions for the hold-out testing dataset. To evaluate our model's predictive performance, we did a 70/30 training-test data split in which we trained the model on 70% of our data, and then validated it on the remaining 30% of unseen data to evaluate its predictive performance. We produced a confusion matrix to further assess the fit and accuracy of the model, which can be seen below.

(a) Confusion Matrix

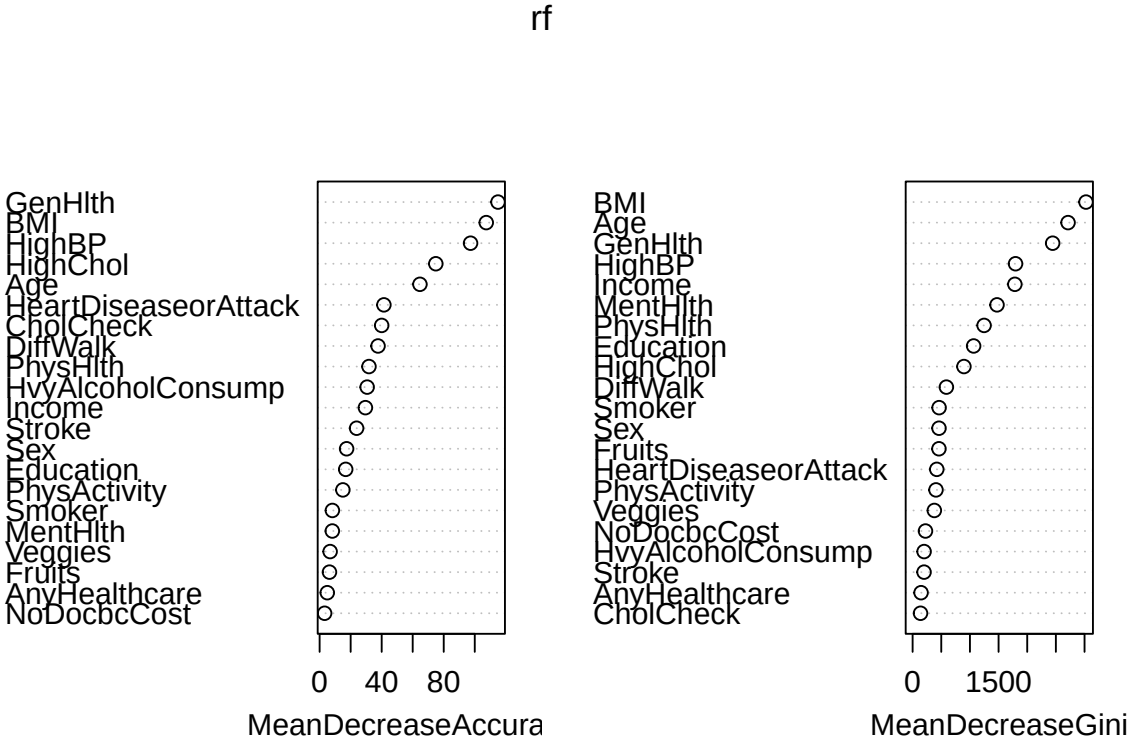
(b) Confusion Matrix Statistics

Figure 1: Model Analysis

The model was able to accurately classify 90% of the hold-out data set. On top of having a high accuracy, it has a sensitivity of 0.89, meaning that 89% of the diabetic cases were correctly predicted. One of our top goals is to have a very high proportion of diabetics correctly identified as being diabetic by the model, since we'd rather the model err on the side of false positives as opposed to false negatives, since it is significantly more detrimental for a patient to go undiagnosed as opposed to taking a blood test to find out it was a false positive.

The no-information rate is 0.5, which is the accuracy achievable by always predicting the majority class label. Since it is a balanced dataset, blindly choosing a class label gives us 50/50 odds of guessing correctly. The no-information rate is used as a baseline to know how useful our model is; since the odds are 50/50 by randomly guessing the label, then we know our model contains useful information since it achieves an accuracy of 0.9 which is significantly higher than the no-information rate.

The positive predictive value, otherwise referred to as precision, is 0.92, which is the number of correctly-identified diabetics divided by the total number of times the classifier predicted "diabetic", rightly or wrongly. 0.92 is a very high precision, meaning that the model was careful to avoid labeling non-diabetics as diabetics.



By looking at the graphs of the mean decrease in accuracy and mean decrease in Gini index (or node purity), we can deduce which predictors are the most and least important in predicting whether an individual has diabetes. The most important variables that have the strongest effect on the predictive power of the model are a patient's general health, BMI, age, and whether they have high blood pressure or high cholesterol. The lesser important variables that don't have as strong predictive power in the model are whether an individual eats any fruits or vegetables, needed to but didn't go to a doctor in the past year due to costs, has healthcare, or is a smoker.

The high accuracy and sensitivity of our model indicates that it has very good predictive power, and can be very effective for identifying individuals with diabetes and indicating whether an individual is at high risk for developing diabetes. These results support both of our initial goals of trying to make a model that will help flag undiagnosed or at-risk diabetic patients, for the doctors to then run further tests if the model predicts diabetes. If a clinician sees a result that indicates diabetes risk, then they can proceed with a glucose level test to determine whether or not there is a diagnosis. If there is a diagnosis they can proceed with the medical protocols/advice established. However, what separates our model is even if there is no diagnosis, then the model still indicates that the patient was at risk, and so the patient can then be proactive in lowering their risk for Type II diabetes, and implement preventative measures.

Conclusion

The high accuracy and high sensitivity of our model suggests that its predictive performance is very good. Therefore, our model can be very effective for identifying individuals with diabetes and indicating whether an individual is at high risk for developing diabetes. Since our data model relies on noninvasive health factors that can be gathered through a simple survey, our model can be applied in all healthcare settings, and is inclusive of all patients, regardless of their socioeconomic status. Since diabetes Type II disproportionately affects racial and ethnic minority and low-income adult populations in the U.S., our model can help people in these populations detect diabetes risk and take steps to prevent it or receive treatment sooner.

The implications of our model are significant, and therefore it is important that it be used properly. For example, it could be used for harm if we were to sell it to insurance companies, who could then significantly raise premiums for high risk insurance members. Such uses would increase the burden that diabetes causes for racial and ethnic minorities and low-income populations, and would further reduce equitable access to healthcare. Additionally, our model was designed to err on the side of caution by over-predicting cases of diabetes in order to avoid missing potential cases, so insurance companies would wrongly assume individuals are at risk for or have diabetes when they in fact do not. Therefore, we leave our model in the hands of medical professionals in healthcare settings only.

Furthermore, there are some limitations to our models capabilities. Firstly, since a random forest combines multiple decision trees, it is very difficult to interpret. Therefore, the model cannot provide a patient at high risk for diabetes with feedback on which health factors were the most influential in the patients classification. This limitation can easily be combated by healthcare professionals because the causes and indicators of type II diabetes have been extensively researched and studied, so any doctor who looks at the patient's health factors will be able to determine which ones are red flags that may have triggered a positive prediction. As a reminder, the intention of this algorithm is not for interpretability, but rather to flag potential individuals who are at-risk for or have diabetes for a doctor to then further look into. Secondly, while our model is very accurate in predicting diabetes, it is not a diagnosis and cannot be used as one. Diabetes is diagnosed by measuring glucose levels in the blood, and data on glucose levels are nowhere included in our model. Thirdly, since our model is a random forest composed of a large number of trees, the algorithm can be slow to run. This could be a drawback in clinical settings where many patients are being seen at once.