STA 325 Final Project Report

Emily Mittleman & Julia Rosner

2022-12-05

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. 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. After removing observations with missing values, we were left with 253,680 observations in our dataset.

The response variable is a binary indicator of whether someone does not have diabetes (0), or they do have diabetes or prediabetes (1).

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

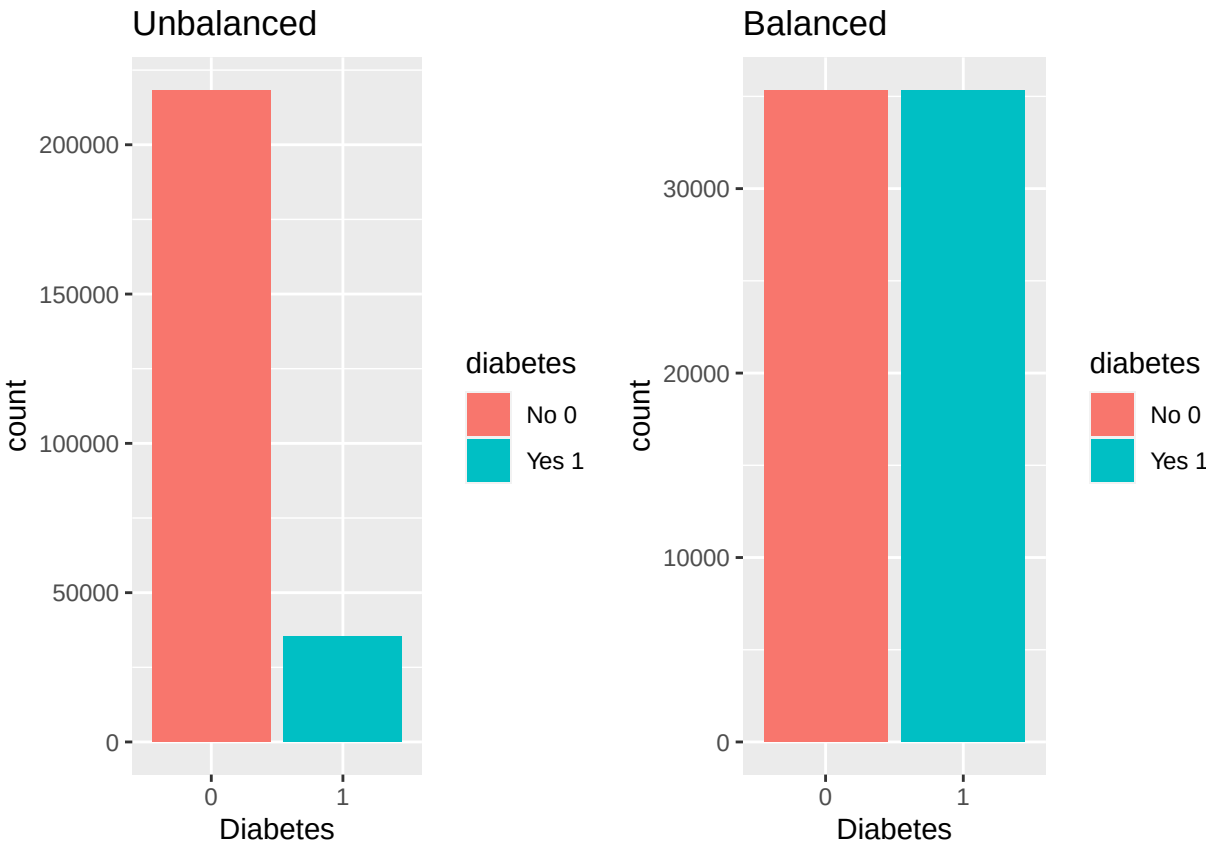
Predictor	**Description**	**Data Type**
High BP	Has high blood pressure	Binary
High cholesterol	Ever had high cholesterol	Binary
Cholesterol check	Cholesterol check within past five years	Binary
BMI	Body Mass Index	Discrete (1-98)
Smoker	Smoked at least 100 cigarettes in entire life	Binary
Stroke	Ever 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)

This dataset is sufficient in meeting our project goals since it has a significantly large number of observations (253,680), and a large number of predictors that can all easily be measured noninvasively in clinical settings.

EDA

To help guide our model selection, we investigated our data further in our exploratory data analysis. We began our EDA by searching for any null and duplicate values. We then dropped all of the null and duplicate values from our dataset.

Next, we looked at variable distributions and correlations to understand our dataset better. First, we explored the distribution of the response variable diabetes. The original diabetes 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, we decided to merge the prediabetes and diabetes classes into one. Even still, there was a very uneven split between the diabetes and no diabetes classes.

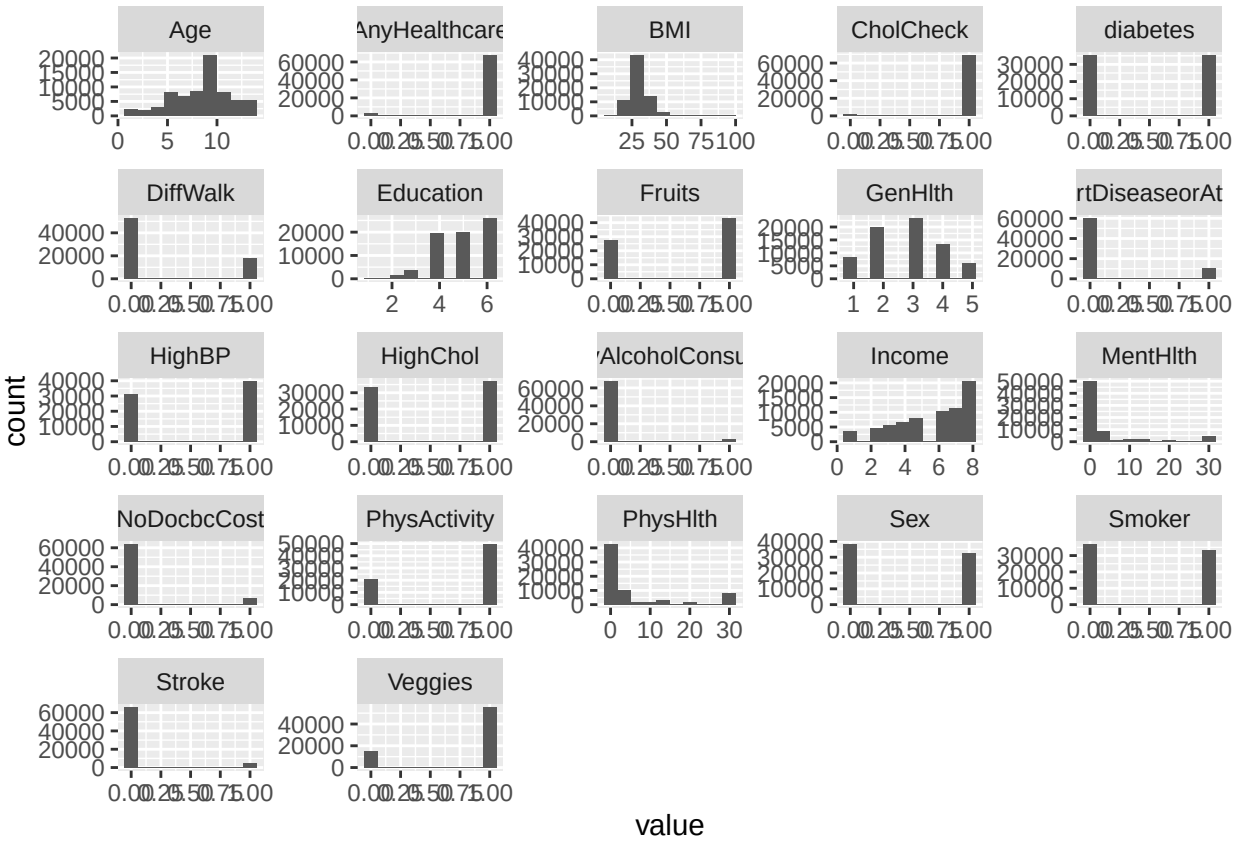


We were concerned that this uneven split would hurt our model's prediction accuracy, which is the core objective of our model. Therefore, we attempted two methods to handle the imbalanced data: the first method was to undersample the negative class, and the other was to oversample the positive class.

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.

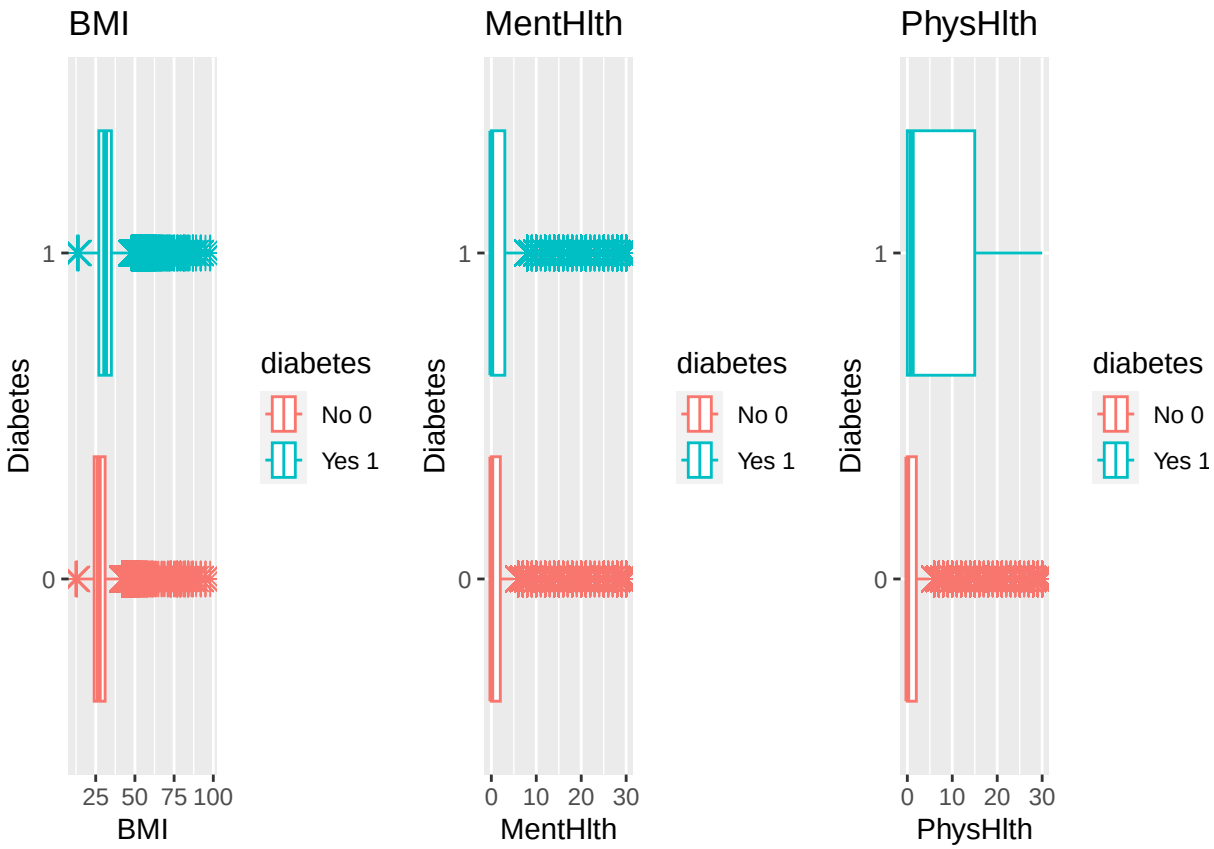
2) For the second method, we tried to oversample the data. We upsampled observations that were classified as diabetes = YES in order to make the number of observations classified as having diabetes equal to the number not having diabetes. This resulted in an equal split of the dataset consisting of 194,377 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.

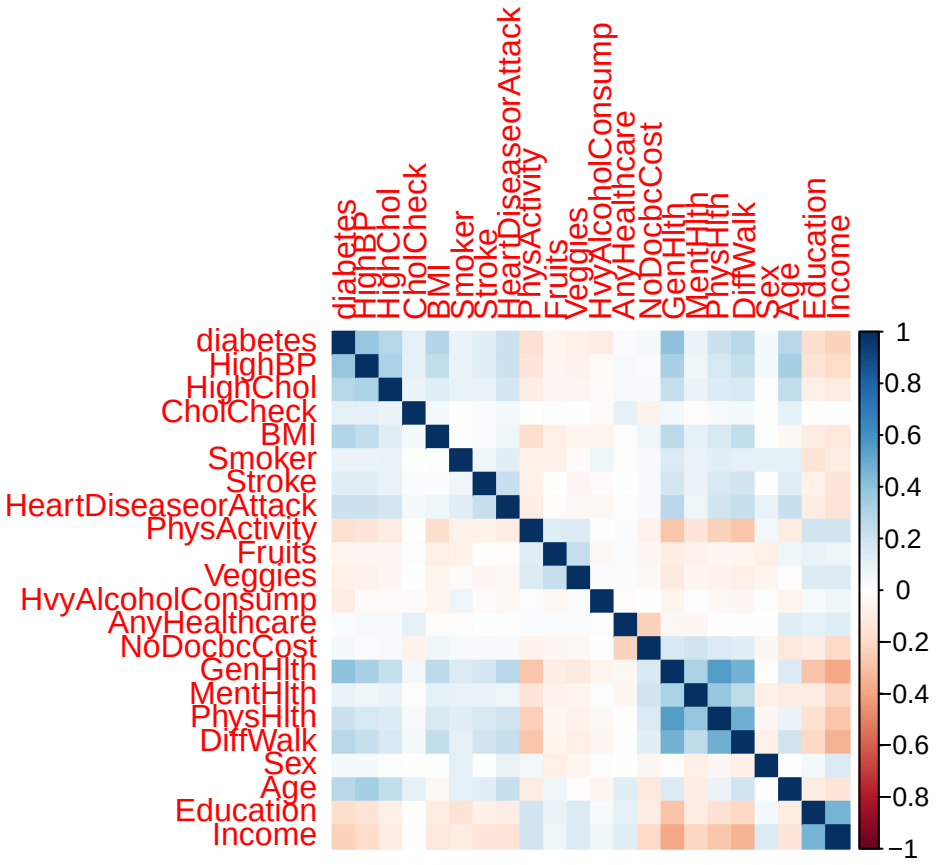


We can see that all of our variables are binary or discrete. Out of the discrete variables, only the Age, BMI, and GenHlth variables follow a roughly normal distribution (the Education, Income, PhysHlth, and MentHlth variables do not). Also, many of the binary variables have very nonuniform distributions, including AnyHealthcare, HeartDiseaseorAttack, Stroke, HvyAlcoholConsump, etc. These not normal and not 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 the only the predictors BMI, MntHlth, and PhysHlth had a lot of outliers, which are shown below. However, in the end, we decided not to transform our variables for our final model. This is 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 training set sizes of 48,221 and 272,030 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. 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. The model had 90% accuracy, and more importantly, a sensitivity of 89%. 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.

Results

Statistical analyses of the fitted model(s), and a translation of these findings into meaningful & understandable conclusions for the target audience (e.g., engineers, business managers, policy-makers, etc). See project rubric for details.

Conclusion

A summary of key findings and potential impacts of your project.