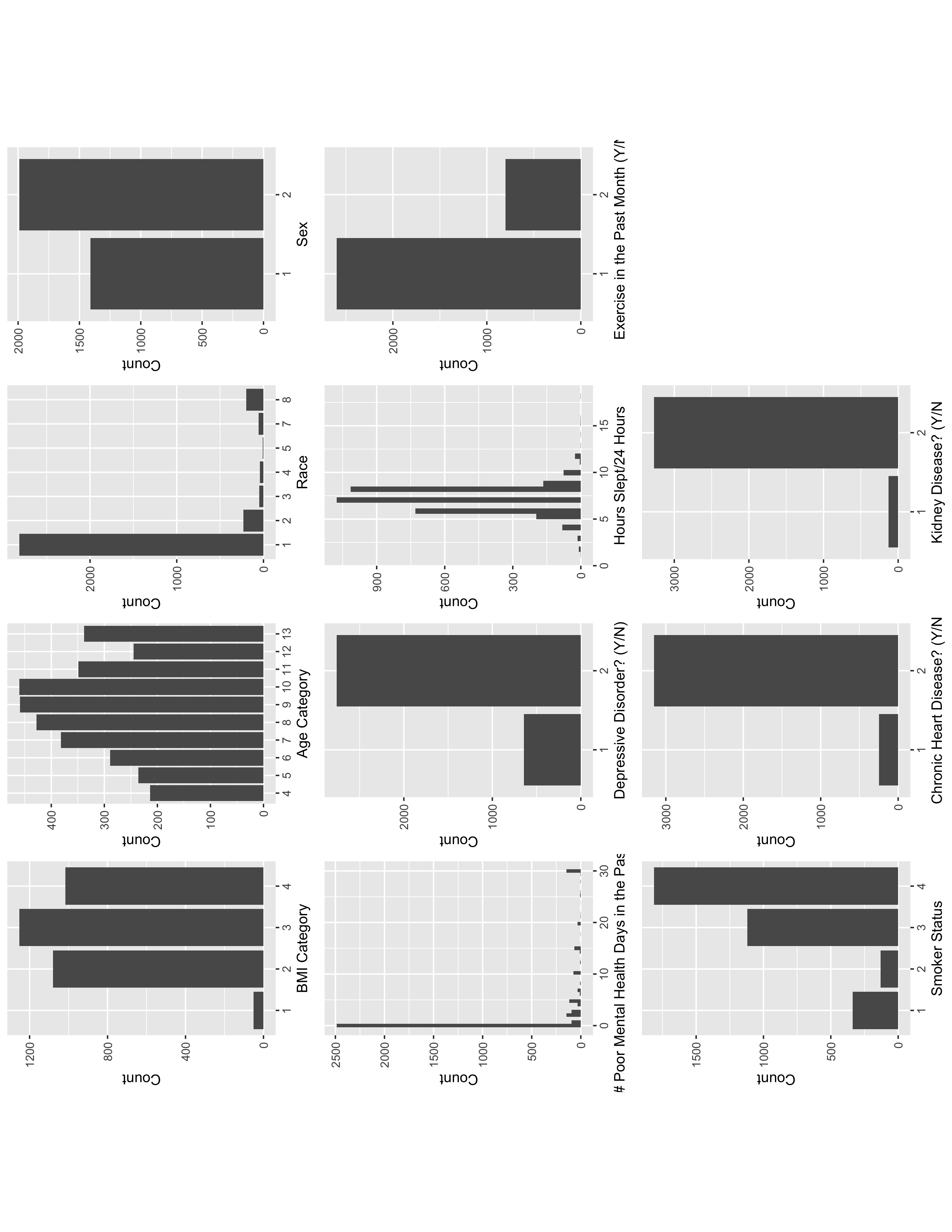
**Introduction:**

Diabetes is a chronic condition that is associated with increased risk of [stroke, nerve damage, and renal disease](https://www.ijncd.org/article.asp?issn=2468-8827;year=2016;volume=1;issue=1;spage=3;epage=8;aulast=Roglic). This report develops three predictive models for diabetes using data from the Behavioral Risk Factor Surveillance System survey and uses the most accurate one to propose diabetes risk factors. We start by outlining 1) the data, 2) the process of choosing, cleaning, and rescaling variables and 3) three different models we use and how we evaluated them. We then choose the most accurate model and outline key takeaways. The findings provide useful knowledge for the initial efficient screening for diabetes, facilitate early intervention, and potentially reduce prevalence while mitigating complications associated with diabetes. Throughout the study, we assume screening for Type 2 diabetes because 90-95% of diabetes diagnoses are

**Section 1:** **Data**

I used the Behavioral Risk Factor Surveillance System (BRFSS) from 2014. BRFSS is an observational study conducted by the Center for Disease Control (CDC) via phone. The sample contains 5000 unique individuals and 33 other variables, 1 of which is an indicator for whether an individual has diabetes. Four of these are numeric count variables (number of adults and children in a household, visits to the doctor in the past year, and weight), twenty-one are categorical (such as race, marital status, etc), and nine are ordinal. After cleaning the variables and adjusting values (I explain this process more in the next section), I visualize the variables and explore their basic characteristics.

Overall, there are more men than women, more individuals who attended and/or completed college or technical school than those who didn’t, and more current non-smokers than smokers. When comparing against other data sources like the [US Census](https://www.statista.com/statistics/241495/us-population-by-sex/), we find that this data is biased towards women, so we should be wary of differential accuracy of our models for the under and overrepresented groups. There are also a few rarely occurring indicator variables like history of kidney disease, chronic heart conditions, and history of depression.

Of the numeric variables visualized, respondents report an average of 3 poor mental health days in the past 30 days and the data is highly positively skewed (as is evident in the image). With respect to hours slept in 24 hours, respondents slept an average of 7 hours with a standard deviation of 1.35 hours. This variable is symmetric and is less spread out than the former.

Before cleaning and dropping observations, 41% of the sample is men. From a racial perspective 77% of the sample is white, 7.2% is black, 1.7% are Native American and Asian, 7.2% are Hispanic, 2% are multiracial and less than one percent are Native Hawaiian and other races. Additionally, 85% of the sample is not diagnosed with diabetes before cleaning the data, 12.5% have a diabetes diagnosis.

**Section 2: Choosing Variables**

While we know that biological data (blood pressure, sugar levels, etc) are the best predictor of diabetes, but we do not have access to that data in BRFSS. Instead, I conducted a review of empirical literature to explore sociodemographic factors for diabetes to proxy for some biological data. Demographically, I have data on [BMI](https://diabetesjournals.org/care/article/30/6/1562/30745/Effect-of-BMI-on-Lifetime-Risk-for-Diabetes-in-the), [age](https://www.researchgate.net/profile/Jothydev-Kesavadev/publication/8450026_Diabetes_in_Old_Age_An_emerging_epidemic/links/0912f50c8d89aeeb02000000/Diabetes-in-Old-Age-An-emerging-epidemic.pdf), [sex](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890267/), and [race](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3830901/). Lifestyle refers to [how stressed a respondent is, if they are depressed, if they smoke, how much sleep they get, and how much they exercise](https://www.researchgate.net/profile/Analava-Mitra/publication/228757143_Diabetes_and_Stress_A_Review/links/02e7e527cdc1e737d5000000/Diabetes-and-Stress-A-Review.pdf). Medical history data is relatively limited and the two most relevant variables are [heart or kidney disease diagnoses](https://www.cdc.gov/kidneydisease/publications-resources/link-between-ckd-diabetes-heart-disease.html). I therefore choose to include all the above variables based on the pre-existing research on this topic.

In order to see which sociodemographic variables could be most useful in predicting diabetes, I use the Boruta algorithm to select features that are predictors of diabetes as opposed to predictors that may be random but consistent with diabetes occurrence. The [Boruta algorithm](https://www.kaggle.com/code/residentmario/automated-feature-selection-with-boruta/notebook) adds copies of all variables then shuffles the copies, which are called shadow attributes. Then, the algorithm runs a random forest classifier on the extended data. After computing the Z scores, the algorithm finds the maximum Z score among shadow attributes, ranks them, then uses a two-sided test of equality with the maximum Z score to decide which variables are unimportant, important, and tentatively important. Finally, the algorithm removes the shadow attributes and repeats this until each variable has a label.

The algorithm concluded that the following variables are confirmed as predictive to diabetes: number of children in a household, general health, age, BMI, education, nightly sleep, mental health status, race, employment status, chronic heart disease, kidney disease, equipment usage for medical reasons, depression, physical activity in the last 30 days, difficulty with memory, concentration, or decision-making, marital status, and sex. Frequency of checkups, smoking status, and exercise in the last 30 days are listed as tentative predictors. This means that their predictive power is a gray area and it is up to the analyst to include them. I choose to exclude reporting physical activity in the last 30 days (totinda) because no physical exercise (totinda=2) is fully correlated with responses from another question about participating in running, calisthenics, golf, gardening, or walking as exercise, which was a confirmed predictor. It would be redundant for the model to include both, so I excluded the one that was listed as tentative.

**Section 3:** **Data Cleaning**

To clean the data, I delete duplicate columns and convert for missing or 0 values as provided by the data documentation. For example, if the number of children in a household is 99, the data are missing so we replace the value with NA. There are also observations where the weight is in kilograms rather than pounds like the rest of the data (this is marked in the documentation), so we convert those kg values to lbs.

After this, we see that for number of adults in a household, health coverage status, and metropolitan code, about 35% of the data are missing. Additionally doctor visit counts and income indicators have 30% and 15% missingess respectively. I omit these variables before selecting features because the missingness would make us throw out a large chunk of our observations when modeling and there are no straightforward ways to impute those missing values.

I filter the data to include respondents who are 35 or older because the [US Preventative Service Task Force recommends that screenings for diabetes should start at age 35](https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/screening-for-prediabetes-and-type-2-diabetes#:~:text=Recommendation%20Summary&text=The%20USPSTF%20recommends%20screening%20for,prediabetes%20to%20effective%20preventive%20interventions.). Therefore, it is unlikely that anyone under 35 is screened for Type 2 diabetes. Finally, I omit all other incomplete observations. Because the missing values do not seem to follow a pattern, and each variable included only has ~5% of observations missing, I feel comfortable deleting those observations without drastically changing the result. Our final sample has 3417 observations.

**Section 4:** **Rescaling the Data**

We first split the data into training and test dataset based on a 75%-25% split. It is worth noting that in the cleaned data we use, 86% of the sample is not diagnosed with diabetes. I was initially worried that the rarity of our predicted event would make it hard for a classifier model to accurately pick up the signal in our data and be able to accurately predict diabetes occurrence. However, I decided against any statistical models that would under or over sample the data to correct for this. [However, data imbalance is typically only a problem if the model is poorly specified or if we are interested in good performance or the model itself (not both)](https://www.sciencedirect.com/science/article/abs/pii/S0378375800001154). Further, with oversampling, we risk overfitting because the algorithm makes exact copies of minority classes. This implies that the model will construct rules that seem accurate but rely on replicated/memorized data. With undersampling, we delete observations from the majority class, which, resulting in losing potentially valuable information. Additionally, with our size sample, we have more than 600 observations in our minority class. Given that we can typically fit about 1 parameter per 15 members of the minority class without overfitting, we have enough flexibility in the sample. Lastly, according to the [National Diabetes Statistics Report (2014)](https://www.silverbook.org/reference/national-diabetes-statistics-report-2014-estimates-of-diabetes-and-its-burden-in-the-united-states/), 9.3% of the U.S. population has diabetes, meaning that we are not necessarily aiming for a 50/50 split and our sample is relatively representative of the US. Because we assume a well-specified model and the data is representative, we do not have to correct.

**Section 5: Building and Assessing the Model**

I then built three models and chose the one with the most predictive power: a logistic regression model, a decision tree, and a random forest. Logistic regression is essentially a simple linear regression but with corrections to account for a binary predicted variable. Decision trees are principally flow charts that are formed by dividing the data into smaller and smaller chunks and yield a prediction. Random forests randomly choose some of the specified features, uses those to build multiple decision trees, and then repeats the process and takes averages. We opt to use these in addition to decision trees, as they are less prone to overfitting.

After fitting the models on both the training and testing data, I assessed each based on accuracy, specificity, sensitivity, positive prediction value, negative prediction value, and usefulness. I also considered using accuracy as a metric, but decided against it because it can be misleading in cases of imbalanced data like we have here. Each metric used answers the following question:

* Positive Prediction Value: What is the probability that someone predicted to have diabetes actually does?
* Negative Prediction Value: What is the probability that someone who is predicted to not have diabetes actually does not have diabetes?
* Area Under Curve: On a scale from zero to 100%, how well is the model able to distinguish between if someone has diabetes? 0.5 suggests no ability to diagnose based on the model. 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding

|  |  |  |  |
| --- | --- | --- | --- |
| Model Type | Logistic Regression | Decision Trees | Random Forests |
| Positive Prediction Value | 0.8642 | 0.8573 | 0.8630 |
| Negative Prediction Value | 0.4571 | 0.4000 | 0.4828 |
| Area Under Curve | 0.5499 | 0.5232 | 0.5448 |

Based on these factors, there is an 86.4% probability that someone who designated by the model as having diabetes actually does. However, the probability of correctly that predicting someone does not have diabetes is much lower at 45.7%. Area Under the Curve (AUC) is highest for logistic regression relative to decision trees and random forests (though not by much). The AUC results indicate that none of the models are able to diagnose/distinguish between if someone does or does not have diabetes. However, it is encouraging that the positive prediction values are high for all of the models. Given the results in the table above, I choose the logistic regression model because it has the highest positive prediction value and area under the curve. Logistic regressions are also more explainable to the general public, which we value when discussing results with non-technical stakeholders.

**Section 6: Results from the most successful model**

After selecting the logistic regression model, I retrained this model on the full dataset. I can use the same breakdown of demographics, lifestyle, and medical history to explain the results from the logistic regression. Demographically, women are 0.321 less likely to be diagnosed with diabetes relative to men. With respect to age, people in age groups 50-54 and 55-59 are nearly twice as likely to receive a diabetes diagnosis relative to adults between 35-40. All ages between 60-64, 64-70, 70-74, 75-79, and 80 or older are nearly twice as likely to receive a diabetes diagnosis relative to adults between 35-40. Obese individuals are 1.2 times as likely to develop diabetes relative to underweight people. Race is also a significant risk factor; Black, Hispanic, and Native Americans are nearly 0.75 times more likely to receive a diabetes diagnosis relative to their white peers. Asians are over 2.5 times more likely.

With respect to lifestyle, marriage and exercise are not significant risk factors. However, individuals who report “very good”, “good”, “fair”, or “poor” health are 0.938, 1.669, 2.078, and 2.381 times more likely to receive a diabetes diagnosis relative to individuals who report “excellent” health respectively. Relatedly, individuals who report going to a doctor for a routine checkup are 0.479 times less likely to be diagnosed relative to those who have seen a doctor within 2 or 5 years and more than 5 years.

While we do not have comprehensive medical history information, we can conclude that individuals who are not diagnosed with chronic health conditions are 0.616 times less likely to receive a diabetes diagnosis relative to people with chronic health conditions. Similarly, individuals without chronic kidney conditions are 0.566 times less likely to be diagnosed with diabetes relative to those with chronic kidney conditions. The key takeways from this model are:

1) Black, Native, and Asian Americans are at highest risk of receiving a diabetes diagnosis, as well as individuals who are obese and over the age of 50.

2) Marriage and exercise are not significant in diagnosing diabetes

3) Chronic heart and kidney disease are significant risk factors in receiving a diabetes diagnosis

**Conclusion:**

Overall, logistic regression is the most effective model to predict diabetes occurrence, with sex, race, obesity, routine checkups, chronic heart condition diagnosis, chronic kidney condition diagnoses, and health status as identifiable risk factors for diabetes. Luckily, these factors are also easily identifiable by a basic screening by medical professionals or county health departments who are looking to reduce diabetes prevalence.

This report was limited in the data used. I only use point-in-time data from 2014, and relationships between populations and risk factors may change over time. Methodically, I do not explore causality in this report, meaning the risk factors outlined are merely correlative to diabetes. Further work should also explore how diagnoses level change for each race across different education, employment, and reported health and/or weight levels. With respect to the methods, future work should explore different metrics to assess the models because given the current metrics and what we know about random forests being able to generally perform better than decision trees the results are strange.