Assessing the Influence of Social, Demographic, and Health Factors on Diabetes Risk

Kate O'Rourke (Team Lead)*
Stephanie Garofalo (Recorder)*
13 January 2024

* Each author contributed equally to the design, coding & development, analysis, and writing of this project.

Synopsis

This project explores how social factors, demographic characteristics, and health conditions influence the likelihood of developing Type 2 diabetes, prediabetes, or borderline diabetes. Using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) survey, machine learning models were applied to uncover the relationships between key variables such as BMI, age, income, and hypertension. The results revealed that variables representing social factors, demographic characteristics, and health conditions all contribute to diabetes risk, emphasizing the need for public health strategies that address not only individual behaviors but also broader socio-economic disparities. This study offers clear insights and reliable models to help create targeted interventions for reducing and treating type 2 diabetes. Further studies are recommended to include clinical indicators of type 2 diabetes to fully characterize type 2 diabetes as a variable allowing models to predict the severity of the disease in an individual.

Table of Contents:

Background & Question

Background:

Type 2 diabetes remains one of the most pressing public health issues globally. The World Health Organization (2023) reports that the number of individuals with type 2 diabetes increased from “108 million in 1980 to 422 million in 2014”. “In 2019, [type 2] diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to [type 2] diabetes occurred before the age of 70 years” (World Health Organization, 2023). Rajendra et. al. is a study in 2021 that looked at the incidence of type 2 diabetes in women over 21 years of age. This study, among many others, focused on general variables that could be indicators of type 2 diabetes and trained the model. While significant research, such as the study by Rajendra et. al. (2021), has explored the clinical and biological risk factors for diabetes, there remains a gap in understanding how social determinants, demographic factors, and health conditions interact to influence diabetes prevalence. This project aims to specifically compare social factors, demographic characteristics, and health conditions. Addressing this gap is crucial for public health agencies to implement effective prevention and intervention strategies.

The 2023 Behavioral Risk Factor Surveillance System (BRFSS) dataset, sourced from the CDC, offers a unique opportunity to explore these connections. This survey collects data on health behaviors, demographics, and medical conditions from individuals across the United States, making it one of the most comprehensive tools available for studying population-level health trends. By examining the combined

influence of social factors (e.g., income, education), demographic characteristics (e.g., age, sex), and health conditions (e.g., hypertension, BMI), this project aims to provide insights that can guide targeted diabetes prevention initiatives. Any mention of diabetes in the remainder of this paper should be taken as indicating type 2 diabetes, prediabetes, or borderline diabetes as will be explained in the dataset section, unless otherwise indicated.

Research Question:

The central research question guiding this analysis is: *How do social factors, demographic characteristics, and health conditions influence the likelihood of developing Type 2 diabetes, prediabetes, or borderline diabetes?*

This question is motivated by the growing need for a holistic understanding of diabetes risk factors that extend beyond clinical settings. Public health initiatives often focus narrowly on biological markers and individual behaviors, overlooking the broader socio-environmental contributors. By integrating social determinants into diabetes research, this project aims to uncover nuanced patterns that can inform more equitable and effective interventions.

Hypothesis:

It is hypothesized that social factors (e.g., income, education), demographic characteristics (e.g., age, gender), and health conditions (e.g., high blood pressure) together influence the likelihood of developing type 2 diabetes, prediabetes, and borderline diabetes.

Prediction:

The data will show that certain demographic groups (e.g. older individuals or those from lower-income backgrounds) who also have specific health conditions (e.g. hypertension) are more likely to develop type 2 diabetes. Additionally, social factors, such as limited access to healthcare and education, will likely increase the risk of developing diabetes. This research aims to not only validate these predictions but also uncover additional patterns that could inform future public health policies and interventions.

Data

Data Acquisition:

The data for this project come from the 2023 **Behavioral Risk Factor Surveillance System (BRFSS)**, a yearly survey run by the Centers for Disease Control and Prevention (CDC). This survey collects information about the health, lifestyle, and demographics of adults across the United States. It is one of the most comprehensive sources for understanding health patterns, making it a great choice for answering the question: *"How do social factors, demographic characteristics, and health conditions affect the chances of developing diabetes?"*

The dataset includes the following predictor variables representing the three categories in the research question:

- Social factors:
    - PHYSHLTH: Indicates how many days the person experienced poor physical health in the last 30 days.
    - MENTHLTH: Indicates how many days the person experienced poor mental health in the last 30 days.
    - PERSDOC3: Indicates if the individual has at least 1 personal doctor.
    - MEDCOST1: Indicates if the individual has been unable to afford a doctor when needed in the last 12 months.
    - SMOKE100: Indicates if the individual has smoked at least 100 cigarettes in their lives.
    - DRNK3GE5: Indicates how many times during the past 30 days the individual had 5 or more drinks for men or 4 or more drinks for women on an occasion.
    - EXERANY2: Indicates if the person has exercised, aside from any exercise from a job, in the last month.
- Demographic characteristics:
    - EDUCA: Individuals indicate the highest education they have received.
    - EMPLOY1: Individuals indicate their current employment status.
    - INCOME3: Individuals indicate their annual income from provided ranges.
    - WEIGHT2: Individuals indicate their weight in pounds or in kilograms.
    - HEIGHT3: Individuals indicate their height in feet and inches or meters.

    - SEXVAR: Individuals indicate their sex.
    - _AGE80: Individuals give their age which is then collapsed over 80 years of age.
- Health conditions:
    - BPHIGH6: Individuals indicate if they have ever been told by a doctor, nurse or other health professional that they have high blood pressure.
    - TOLDHI3: Individuals indicate if they have ever been told by a doctor, nurse or other health professional that their cholesterol is high.
    - CVDINFR4: Individuals indicate if they have been told they had a heart attack, also called a myocardial infarction.
    - CVDCRHD4: Individuals indicate if they have been told they had angina or coronary heart disease.
    - CVDSTRK3: Individuals indicate if they have been told they had a stroke.
    - CHCKDNY2: Individuals indicate if they have been told they had kidney disease, not including kidney stones, bladder infection or incontinence.

These variables, shown in the codebook in **appendix 1**, help us explore how different factors combine to influence diabetes risk. The key outcome in this study is whether someone has been diagnosed with Type 2 diabetes, prediabetes, or borderline diabetes. The outcome variable is defined using information from the DIABETE4 and DIABTYPE columns, grouped into two categories: 1 (has diabetes, prediabetes, or borderline diabetes) and 0 (does not). Individuals with type 1 diabetes or who had diabetes only during pregnancy were completely excluded from the dataset.

The dataset originally contained 433,323 responses, but after removing incomplete or irrelevant data in accordance with the cleaning steps outlined in the following section and **appendix 1**, 128,356 rows remained. While the dataset is detailed, there are some caveats. For example, since the survey relies on self-reported answers, there could be inaccuracies due to memory errors or personal bias. These issues were considered when analyzing and forming conclusions about the data. All variable details, including how they were handled, are listed in the appendix for reference.

Cleaning:

Cleaning the data was a critical step to prepare the dataset for analysis. First, any rows where answers were missing, unclear, or marked as "Refused" or "Don't Know" were removed. EMPLOY1, PERSDOC3, and BPHIGH6 were simplified or recoded as shown in the final column of the table in **appendix 1.** PHYSHLTH, MENTHLTH, and DRNK3GE5 converted 88 to 0 as 88 refers to none in the survey. The rows where age is 80 years old have been removed as the CDC condensed all later ages over 80 into the value of 80. Employ1 was collapsed to employed, out of the workforce, and unemployed to allow for an ordinal variable to be applied. These definitions are used by the U.S. Bureau of Labor Statistics to group individuals. BMI was calculated from height and weight values using the equation from **appendix 1**. All data cleaning steps are outlined in detail for each variable in the last column of the table in **appendix 1**. The cleaned data provides a solid foundation for exploring and predicting diabetes risk.

Exploratory Data Analysis

Methods:

A variety of visualizations and summary tables were used to examine the relationships between diabetes prevalence and variables. The graphs were created using ggplot2 and dplyr and the tables were created using calculated summary statistics and gtsummary in R. ggcorrplot in R was used to examine potential patterns and relationships among variables in tables in **appendix 4**. Together, the visualizations and tables offer accessible insights into lifestyle, demographic, and health conditions associated with diabetes. **Tables 2, 3, and 4** can be compared to the table in **appendix 3** that contains descriptions of the dataset prior to cleaning.

Results:

Figure 1. Prevalence of Diabetes by Exercise and BMI Category

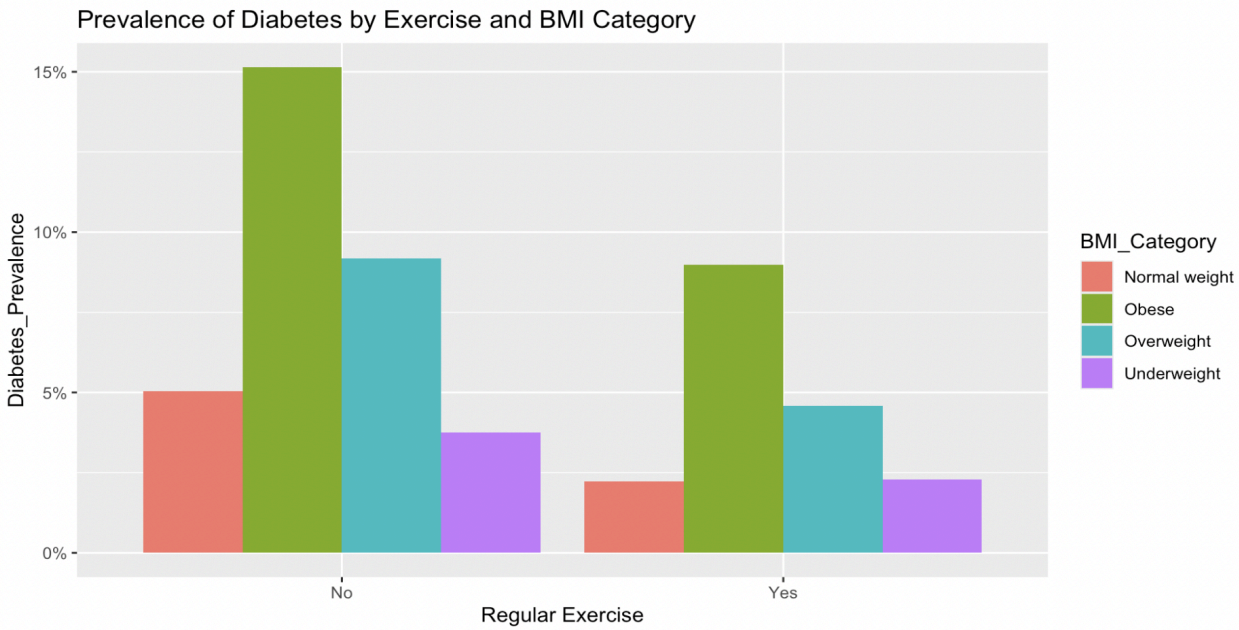


Figure 1 highlights how regular exercise and body weight relate to diabetes risk, serving as a clear message for public health agencies. It shows that regular physical activity reduces diabetes prevalence across all BMI categories, yet obesity remains a significant risk factor. The key takeaway is that, while maintaining a healthy weight is important, regular exercise can benefit everyone in managing diabetes risk regardless of BMI category. This visualization highlights the power of lifestyle choices, motivating us to prioritize physical activity as a preventive measure against diabetes.

Figure 2. Relationship Between Education, Employment Status, and Type 2 Diabetes

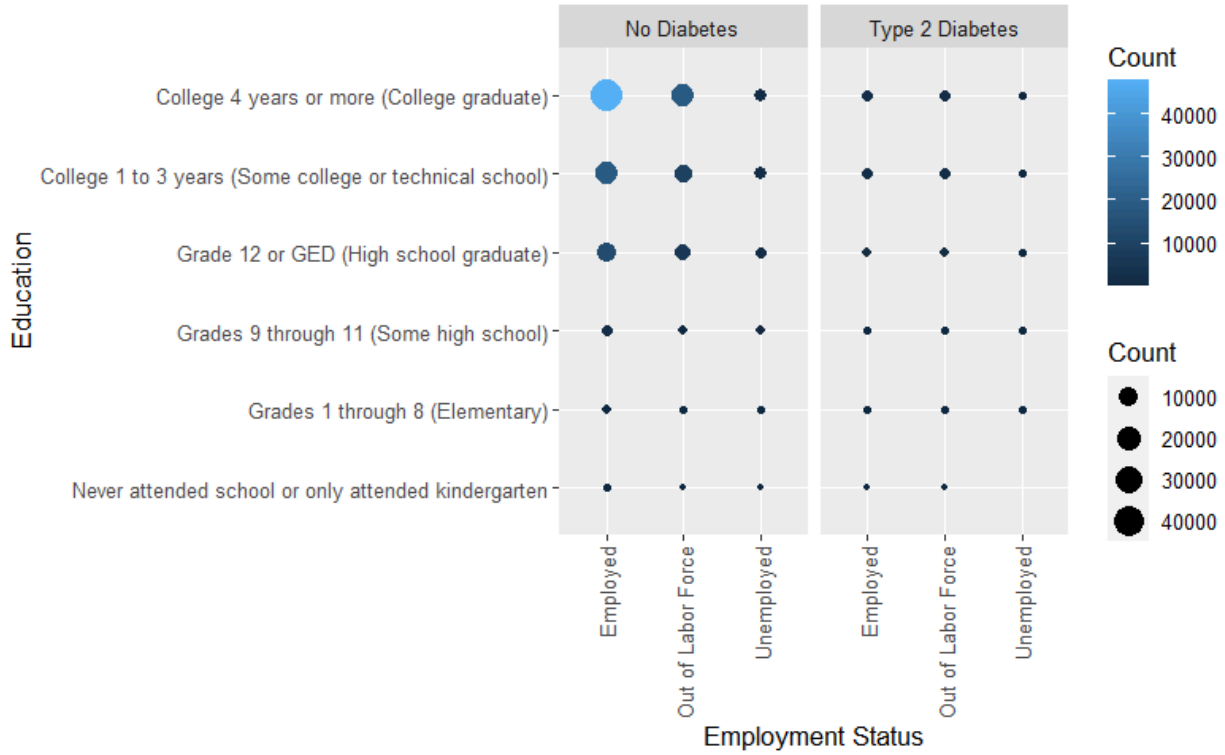


Figure 2 is important for public health agencies as it shows the relationship between education level, employment status, and type 2 diabetes. It is shown that higher education levels correspond to higher employment rates for individuals with and

without type 2 diabetes, just as lower education levels are associated with higher levels of unemployment.

Figure 3. Difference in BMI According to Type 2 Diabetes

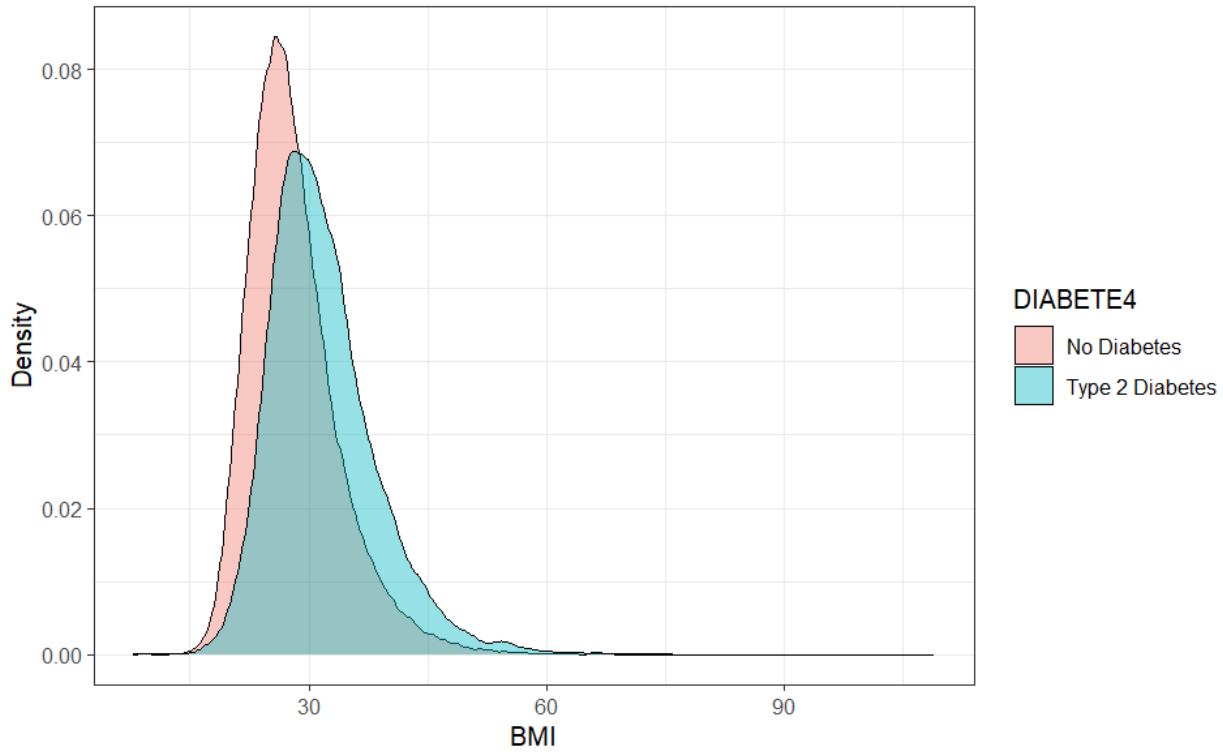


Figure 3 illustrates the differences in distribution between individuals with and without type 2 diabetes. The peak of the density curve occurs at a higher BMI for those with type 2 diabetes than for those without, indicating that individuals with type 2 diabetes are likely to have a higher BMI than those without.

Figure 4. Education Level Distribution and Proportion of Type 2 Diabetes

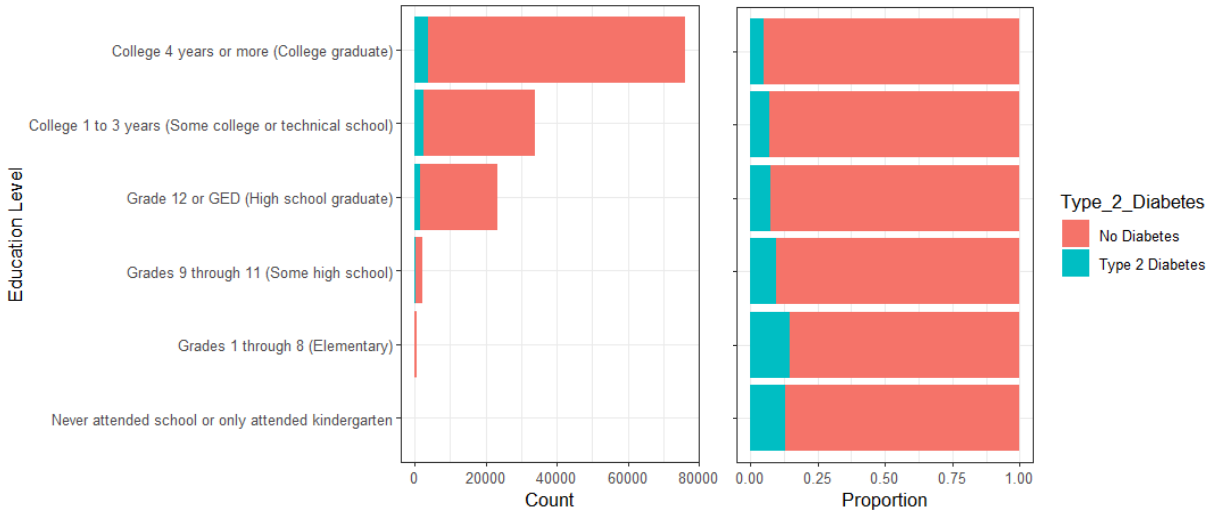


Figure 4 explores the distribution of education levels and their connection to type 2 diabetes. The left panel shows the count of individuals by education level, with most people being high school or college graduates. The right panel shows the proportion of individuals with and without type 2 diabetes within each education group. Across all levels, the proportion of those without type 2 diabetes is higher, but we can see a slightly higher proportion of type 2 diabetes among those with lower education levels. This suggests a possible link between lower education and higher diabetes risk.

Figure 5. Distribution of Age Based on Type 2 Diabetes

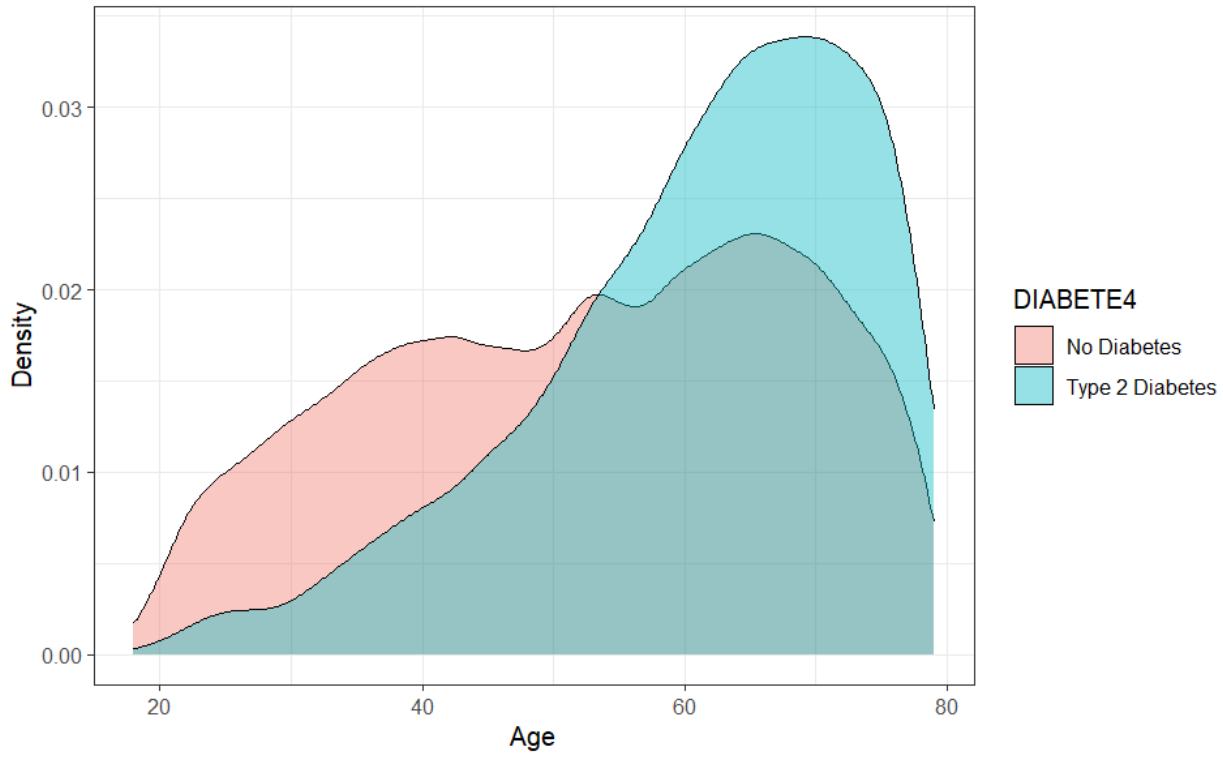


Figure 5 shows the difference in age distribution between individuals with and without type 2 diabetes. This graph shows a link between type 2 diabetes and older individuals. This likely indicates that individuals with type 2 diabetes are more likely to develop this disease at an older age though this graph cannot be used solely to conclude this statement.

Table 1. Summary Statistics for BMI and Age

Measure	BMI	Age (column name: _AGE80)
Minimum	7.915	18
Q1	24.208	40
Mean	28.202	52.92
Median	27.289	55
Q3	31.133	66
Maximum	111.381	79
Kurtosis	1.355	2.000
Skewness	7.731	-0.268

Table 1 provides statistics for BMI and age. BMI shows an average of 28.2, which is slightly above the healthy range, and a median of 27.3. With a skewness of 7.7, most individuals have lower BMIs, but a small amount of higher values influence the overall average. The BMI kurtosis of 1.355 means the BMI values are close to a normal distribution, with fewer extreme highs and lows. For age, the skewness shows that the age distribution is approximately symmetrical. The value of 2.000 for kurtosis indicates a platykurtic distribution.

Table 2. Summary Statistics for Social Factors Separate by Presence of Type 2 Diabetes

Characteristic	N	No Diabetes N = 120,949[1]	Type 2 Diabetes N = 7,407[1]
Physical_Health	128,356	0 (0, 2)	0 (0, 5)
Mental_Health	128,356	0 (0, 5)	0 (0, 5)
Personal_Doctor	128,356		
More than one		69,483 (57%)	4,414 (60%)
No		51,466 (43%)	2,993 (40%)
Not_Afford_Doctor	128,356		
No		112,845 (93%)	6,831 (92%)
Yes		8,104 (6.7%)	576 (7.8%)
Smoked_100_Pkgs	128,356		
No		74,781 (62%)	3,890 (53%)
Yes		46,168 (38%)	3,517 (47%)
Binge_Drinking	128,356	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)
Exercised	128,356		
No		16,668 (14%)	1,978 (27%)
Yes		104,281 (86%)	5,429 (73%)

[1] Median (Q1, Q3); n (%)

Table 2 highlights differences in social and lifestyle factors between individuals with and without Type 2 Diabetes. Those with Type 2 Diabetes report slightly poorer physical health (Q3), are more likely to have a history of smoking, and are less likely to engage in regular exercise compared to those without diabetes. Healthcare access and binge drinking show less variation, though slightly more individuals with diabetes report affordability issues and less binge drinking. This summary underscores potential behavioral and social influences on diabetes risk, suggesting areas for targeted health interventions.

Table 3. Summary Statistics for Demographics Separate by Presence of Type 2 Diabetes

Characteristic	N	No Diabetes N = 120,949[1]	Type 2 Diabetes N = 7,407[1]
Education	128,356		
College 4 years or more (College graduate)		68,762 (57%)	3,300 (45%)
College 1 to 3 years (Some college or technical school		29,673 (25%)	2,258 (30%)
Grade 12 or GED (High school graduate		20,177 (17%)	1,571 (21%)
Grades 9 through 11 (Some high school)		1,809 (1.5%)	187 (2.5%)
Grades 1 through 8 (Elementary)		495 (0.4%)	87 (1.2%)
Never attended school or only attended kindergarten		33 (<0.1%)	4 (<0.1%)
Employment_Status	128,356		
Employed		79,902 (66%)	3,533 (48%)
Out of Labor Force		34,753 (29%)	3,201 (43%)
Unemployed		6,294 (5.2%)	673 (9.1%)
Income	128,356	8.00 (7.00, 9.00)	7.00 (6.00, 9.00)
Sex	128,356		
Female		58,807 (49%)	3,127 (42%)
Male		62,142 (51%)	4,280 (58%)
BMI	128,356	27.3 (24.2, 31.0)	30.8 (27.2, 35.3)
Age	128,356	54 (40, 66)	63 (54, 71)

[1] n (%); Median (Q1, Q3)

Table 3 shows demographic differences between people with and without Type 2 Diabetes. Those with diabetes tend to have slightly lower education levels, higher unemployment or are more often out of the labor force, and show a higher median BMI and age. Men make up a larger portion of the diabetes group, and income is generally lower among individuals with diabetes. These patterns highlight socioeconomic factors linked to diabetes risk.

Table 4. Summary Statistics for Diseases Separate by Presence of Type 2 Diabetes

Characteristic	N	No Diabetes N = 120,949[1]	Type 2 Diabetes N = 7,407[1]
HighBP	128,356		
No		78,803 (65%)	2,402 (32%)
Yes		42,146 (35%)	5,005 (68%)
High_Cholesterol	128,356		
No		77,171 (64%)	2,709 (37%)
Yes		43,778 (36%)	4,698 (63%)
Myocardial_Infarction	128,356		
No		117,585 (97%)	6,733 (91%)
Yes		3,364 (2.8%)	674 (9.1%)
Heart_Disease	128,356		
No		117,302 (97%)	6,699 (90%)
Yes		3,647 (3.0%)	708 (9.6%)

Stroke	128,356		
No		118,351 (98%)	7,012 (95%)
Yes		2,598 (2.1%)	395 (5.3%)
Kidney_Disease	128,356		
No		118,356 (98%)	6,875 (93%)
Yes		2,593 (2.1%)	532 (7.2%)

[1] n (%)

Table 4 summarizes the presence of other diseases among people with and without Type 2 Diabetes. Those with diabetes have a significantly higher prevalence of related conditions, including high blood pressure (68% vs. 36%), high cholesterol (64% vs. 37%), and kidney disease (7.9% vs. 2.5%). Heart conditions, such as myocardial infarction and heart disease, as well as stroke, are also more common in the diabetes group. This highlights a strong association between Type 2 Diabetes and an increased likelihood of health conditions.

Models

Preprocessing and Dimensionality Reduction / Feature Engineering:

The CDC_2023_ceaned.csv data file contains unused columns that were removed including an index column, WEIGHT2, HEIGHT3, DIABTYPE, and height_inches. All columns apart from EDUCA, EMPLOY1, INCOME3, PHYSHLTH, MENTHLTH, DRNK3GE5, _AGE80, and BMI were converted to factors. All numeric columns (the columns noted in the previous sentence) were normalized by centering around the mean and scaling by the standard deviation. The test and training split was

calculated by a calcSplitRatio.r file created by Katherine S. Geist, PhD from Merrimack College. The training:test ideal ratio was found to be 0.77:0.23. In order to ensure that the proportion of type 2 diabetics is preserved between the test and training dataset, the createDataPartition function from the caret library was used to split the dataset.

Dimensionality reduction was explored using Lasso regression, a technique that automatically identifies the most important variables by shrinking or removing the coefficients of less relevant ones. This method was chosen over other methods, such as ridge regression, because it works well with datasets that have many predictors and ensures that only the most useful features are kept. Cross-validation was used to determine the best penalty value (lambda) for Lasso regression, ensuring the final set of features was both effective and interpretable. The lasso regression did not eliminate any variables from the model as shown by the table of the lasso regression coefficients in **appendix 2**.

Algorithm Selection:

Two supervised machine learning algorithms were chosen: Random Forest and XGBoost. These models were selected because they handle classification tasks effectively when working with imbalanced datasets like this one, where only 6% of cases were diabetic, by adding case weights to control for the imbalance.

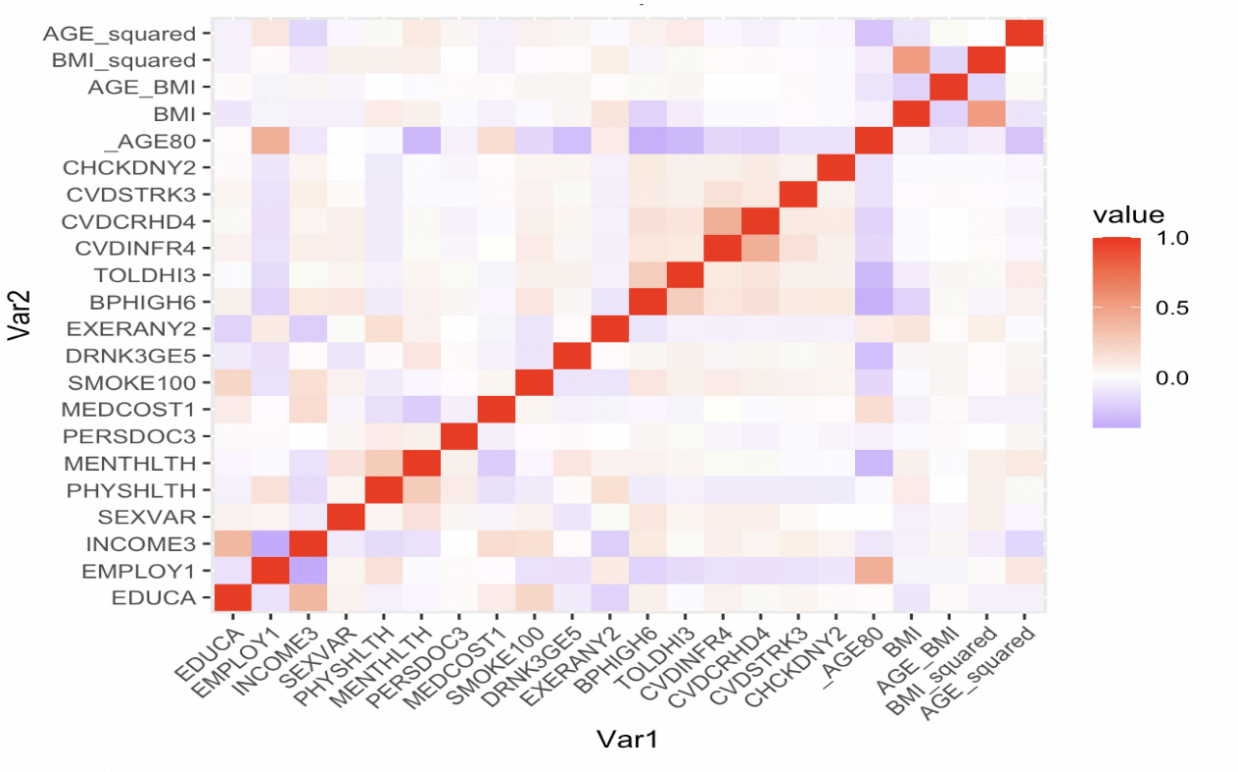
1. Random Forest: This algorithm creates multiple decision trees and combines their predictions, reducing the risk of overfitting and capturing complex relationships in the data. For Random Forest, the parameters of number of features to consider at each split (mtry) and the minimum number of samples

required at a node (min_n) were tuned by trial and error to ensure the model had high sensitivity and the calibration curve was sufficient (i.e. did not have any major drop offs indicating more significant underfitting). To address class imbalance, case weights were applied at 18:1 to account for the ratio of class imbalance. randomForest and ranger libraries were used in R to create a random forest model using the tidymodels framework. The code used to run this model is contained in the Random_Forest.rmd file in the GitHub repository.

2. XGBoost: This gradient boosting algorithm builds trees sequentially, improving predictions with each step. It is known for its efficiency and strong performance with large datasets. For XGBoost, we optimized tree depth, learning rate, and the number of trees parameters using 5-fold, stratified cross-validation. Case_weights were adjusted to 18:1 to account for the ratio of class imbalance. This model was run using the xgboost library in R. The code used to run this model is contained in the xgBoost_All_Predictors.rmd file in the GitHub repository.

Both algorithms assume that the data is collected independently and that there is no multicollinearity between predictors. We tested these assumptions by reviewing the CDC's data collection methods and repeating the correlation heatmap from the exploratory data analysis section with all predictor variables in **figure 6** to check for multicollinearity. **Figure 6** did not show any significant correlation that would impact these models.

Figure 6. Correlation Matrix Heatmap



The data from the BRFSS Survey, according to the CDC, was collected randomly using random digit dialing. It would be hard to determine if there is any bias related to whether or not an individual decides to answer the survey. The methodology of the BRFSS survey indicates that the data is collected randomly and independently and test:training splits are done randomly while controlling to ensure equal proportions of people with diabetes are contained in each set.

Overfitting was not an issue in this analysis, as the models are underfit (as seen in the following section and in the calibration curve in **appendix 5**). While the calibration curves suggested underfitting, this was likely due to the complexity of type 2 diabetes as a variable and the variations between how individuals are able to control their diabetes rather than an issue with the model directly.

Final Model

Random forest and xgBoost models were run. The final results showed that both algorithms performed well in predicting diabetes risk. The Random Forest model achieved an accuracy of 76.77% and an AUC of 0.7893, while the XGBoost model had an accuracy of 66.80% and an AUC of 0.7808. While both models showed acceptable predictive performance, Random Forest slightly outperformed XGBoost in this context as shown in **table 5**. While XGBoost had a slightly higher sensitivity (0.7600 vs. 0.6518), indicating better identification of positive diabetes cases, it fell short in terms of accuracy (66.80% vs. 76.77%) and AUC value (0.7808 vs. 0.7893). The Random Forest model demonstrated a more balanced performance across all metrics, making it the better choice for addressing the research question and predicting diabetes risk.

Table 5. Comparison of Random Forest Models and xgBoost Models

Metric	Final Random Forest	xgBoost
AUC Value	0.7893	0.7808
Accuracy	76.77%	66.80%
Sensitivity	0.6518	0.7600
Specificity	0.7748	0.6623

Three key figures and one table provided the characterization of the random forest model:

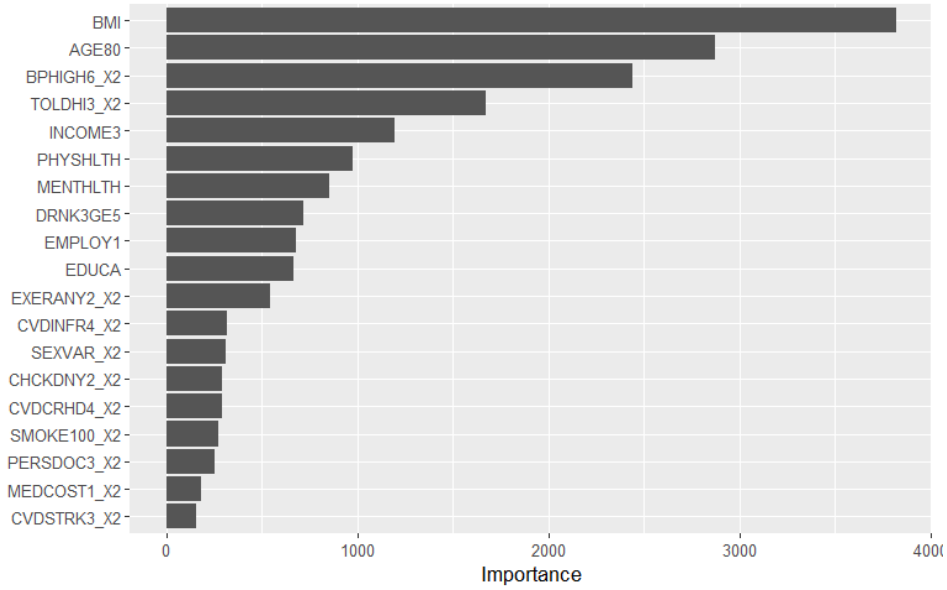
1. **Confusion Matrix:** The confusion matrix shows that the model is predicting individuals with and without diabetes in accordance with the accuracy, sensitivity, and specificity mentioned in **table 5**.

Table 6. Confusion Matrix for Random Forest Model

		Predicted	
		Type 2 Diabetes	No Type 2 Diabetes
Actual	Type 2 Diabetes	1,110 (True Positive)	593 (False Negative)
	No Type 2 Diabetes	6,265 (False Positive)	21,553 (True Negative)

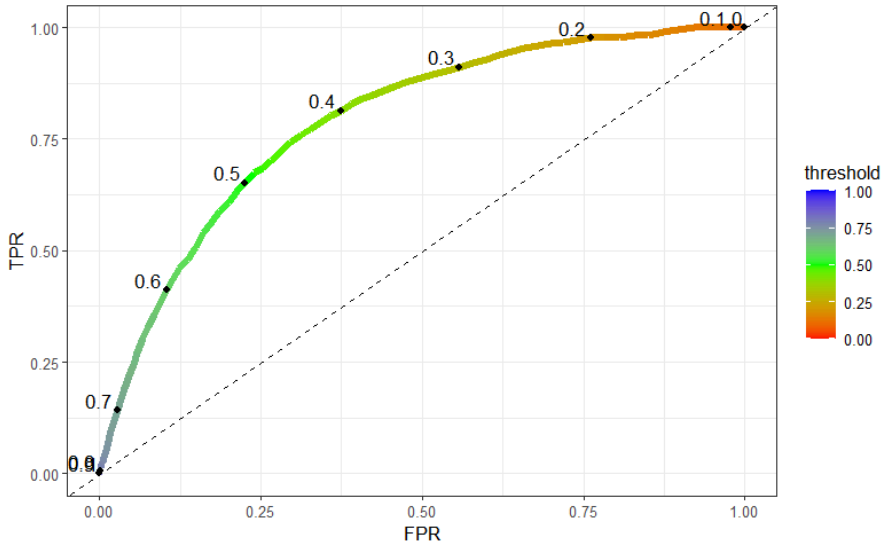
2. **Feature Importance Plot**: The feature importance plot highlighted the most critical predictors for each model, with BMI, age, high blood pressure, high cholesterol, and income emerging as key factors. Though these were the five most important predictors in this model, predictors from all three categories of social factors, demographics, and diseases were indicated as important in this model.

Figure 8. Feature Importance for Random Forest



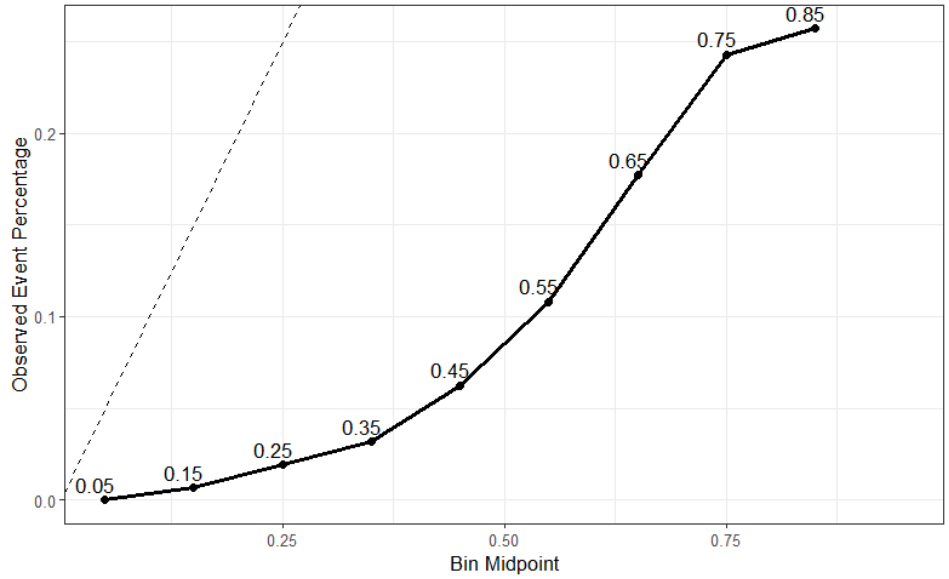
3. **ROC Curves**: The ROC curve illustrates the trade-off between sensitivity and specificity over different threshold values shown in the graphical legend, showing that this model performs significantly better than random guessing (illustrated by the dotted line).

Figure 9. ROC Curve for Random Forest



4. **Calibration Curve**: This curve shows that the model is underfitting the data throughout. This is likely due to the complexity of type 2 diabetes as an outcome variable.

Figure 10. Calibration Curve for Random Forest



The model provides a clear and actionable understanding of the factors influencing diabetes risk and creates a strong foundation for future refinement and application in public health strategies.

Conclusions

This project sought to answer the question: *How do social factors, demographic characteristics, and health conditions influence the likelihood of developing Type 2 diabetes, prediabetes, or borderline diabetes?* Using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), machine learning techniques were applied (Random Forest and XGBoost) to uncover the relationships between diabetes risk and the feature variables previously described. The methods included thorough data cleaning, feature engineering using Lasso regression, and careful handling of class imbalance to ensure robust and accurate predictions.

The results confirmed the hypothesis: individuals with higher BMI, older age, or specific health conditions like hypertension and high cholesterol were more likely to have diabetes. Social factors, such as lower income and limited education, also contributed significantly to risk. BMI, age, high blood pressure, high cholesterol, and income were indicated as the type five most important predictor variables in our random forest and xgBoost models. The models provided valuable predictive insights, with Random Forest slightly outperforming XGBoost in accuracy and overall fit.

These findings not only validate the importance of integrating social and demographic factors into public health interventions but also emphasize the need for tailored prevention strategies. For example, promoting regular physical activity and improving access to healthcare in lower-income populations could play a part in reducing diabetes risk. The results reinforce the broader understanding in the field that addressing diabetes requires a holistic approach, combining clinical care with efforts to reduce socio-economic disparities. This study serves as a foundation for further research into multi-factorial models and targeted public health strategies. Research studies such as the study performed by Mansoori et. al. (2023) showed high levels of accuracy, AUC values, sensitivity, and specificity when using clinical data such as hematocrit, platelet count, *r*ed cell distribution width, platelet distribution width, *r*ed blood cell, mean corpuscular hemoglobin concentration, *w*hite blood cell, and hemoglobin in conjunction with the variables in this study could help to more fully illustrate risk factors for type 2 diabetes.

Discussion & Next Steps

Key Takeaways

Using data from the CDC's BRFSS survey, the analysis highlighted the critical role of both clinical and non-clinical factors in predicting diabetes risk. The best-performing model, Random Forest, demonstrated an accuracy of 76.77% and an AUC of 0.7893, effectively capturing the relationships between key predictors and diabetes prevalence. Variables like BMI, age, high blood pressure, and income emerged as the most significant contributors with variables from social factors, demographics, and health conditions contributing to the predictions confirming the hypotheses set forth at the start of the project. The model's ability to answer the research question is strong, as it integrates diverse predictors to provide a nuanced understanding of diabetes risk. The ROC curve (**Figure 2**) confirmed the model's robustness in distinguishing between diabetic and non-diabetic individuals, making it a valuable tool for targeting high-risk populations.

Recommendations and Future Directions

Based on the findings from this model, several recommendations can be made for management and policy. Public health organizations should prioritize interventions that address both health conditions (e.g., hypertension, high BMI) and social determinants (e.g., income and education disparities). Strategies could include community health programs focused on promoting exercise, as well as policy initiatives aimed at increasing healthcare access in underserved areas.

To further investigate these findings and to account for the underfitting of the model likely occurring from the complexity of type 2 diabetes as an outcome variable, more studies should be completed. For public health organizations, the recommendation is to validate these models using data from previous years of BRFSS suveys to confirm their accuracy and continued relevance. If the models perform well, future studies should focus on the most influential variables identified by the feature importance plots—such as high blood pressure, BMI, age, high cholesterol, and income—while incorporating additional clinical predictors like A1c levels and family history. This approach could enhance the model's ability to identify individuals at risk and guide targeted public health interventions.

Caveats and Concerns

While the models performed well, several limitations should be acknowledged. The BRFSS dataset relies on self-reported data, which introduces potential biases such as inaccurate recall or underreporting of sensitive information. Moreover, the imbalanced nature of the dataset, with only 6% of cases being diabetic, required adjustments that may impact generalizability to populations with different distributions. Finally, while the models highlight correlations, they do not establish causation, and further research is needed to confirm the causal pathways suggested by these findings.

In conclusion, this study provides a strong foundation for understanding diabetes risk through a combination of social, demographic, and health factors. With targeted improvements and additional data, the models and insights developed here can significantly contribute to public health efforts to reduce the burden of diabetes.

Code Availability

https://github.com/KatePOr/Diabetes_Capstone/tree/main

References:

Ahlqvist, E., Prasad, R. B., & Groop, L. (2020). Subtypes of type 2 diabetes determined from clinical parameters. *Diabetes*, *69*(10), 2086-2093.

CDC. (2023) Behavioral Risk Factor Surveillance System (BRFSS) 2023 Data [Dataset]. CDC. https://www.cdc.gov/brfss/annual_data/annual_2023.html

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), 24.

Rajendra, P., & Latifi, S. (2021). *Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1*, 100032.

U.S. Department of Health and Human Services. (2024). *Drinking levels and patterns defined*. National Institute on Alcohol Abuse and Alcoholism. https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking#:~:text=The%20Substance%20Abuse%20and%20Mental,or%20within%20a%20couple%20of

Zhu, T. (2020, August). Analysis on the applicability of the random forest. In *Journal of Physics: Conference Series* (Vol. 1607, No. 1, p. 012123). IOP Publishing.

Appendix 1. Codebook

Column Name	Question	Values	Data Cleaning/Processing
EDUCA (Education)	What is the highest grade or year of school you completed?	1: Never attended school or only kindergarten 2: Grades 1 through 8 (Elementary) 3: Grades 9 through 11 (Some high school) 4: Grade 12 or GED (High school graduate) 5: College 1 year to 3 years (Some college or technical school) 6: College 4 years or more (College Graduate) 9: Refused	People who refused or did not answer this question will be removed from the dataset.
EMPLOY1 (Employment_status)	Are you currently…?	1: Employed for wages 2: Self-employed 3: Out of work for 1 year or more 4: Out of work for less than 1 year 5: A homemaker 6: A student 7: Retired 8: Unable to work 9: Refused	The following categories will be created: 1 (employed): 1, 2 2 (out of labor force): 5, 6, and 7 3 (unemployed): 3, 4, 8 People who refused or did not answer this question will be removed from the dataset.
INCOME3 (Income)	Is your annual household income from all sources: (If respondent refuses at any income level, code ´Refused.´)	1: Less than $10,000 2: $10,000 to <$15,000 3: $15,000 to <$20,000 4: $20,000 to <$25,000 5: $25,000 to <$35,000 6: $35,000 to <$50,000 7: $50,000 to <$75,000 8: $75,000 to < $100,000 9: $100,000 to <$150,000 10: $150,000 to <$200,000 11: $200,000 or more 99: Refused	People who refused or did not answer this question will be removed from the dataset.
WEIGHT2 (combined to create BMI)	About how much do you weigh without shoes? (If respondent answers in metrics, put a 9 in the first column)[Round fractions up.]	50 - 0776: Weight in pounds (0 _ _ _ = weight in pounds) 7777: Don't know/Not Sure 9023 - 9352: Weight in kilograms (The initial ´9´ indicates this was a metric value.) 9999: Refused	BMI will be calculated using the following equations after editing the 9 out of the metric measurements and calculating height in inches: ● For imperial: Weight * 703 / (height in inches)^2 ● For metric: (Weight in kg) / (Height in m)^2 Remove 7777 and 9999 values from the dataset. These two columns were removed once the BMI column was created.
HEIGHT3 (combined to create BMI)	About how tall are you without shoes? (If respondent answers in metrics, put a 9 in the first column)[Round fractions down.]	200-711: Height (0 _ / _ _ = feet / inches) 7777: Don't know/Not sure 9061-9998: Height (meters/centimeters) 9999: Refused	
SEXVAR (Sex)	Sex of Respondent	1: Male (48.5%) 2: Female (51.5%)	

_AGE80 (Age)	Age value collapsed after 80 years old.	18 - 80 years old.	80: value will be removed as 80 contains all individuals aged 80 or over.
PHYSHLTH (Physical_Health)	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?	1 - 30: Number of days 88: None 77: Don't know/Not sure 99: Refused	People who refused or did not answer this question will be removed from the dataset. The following categories will be created: 0: Zero days when physical health not good (88) 1-30: Number of days when physical health is not good 9: Don't know/Refused/Missing (99)
MENTHLTH (Mental_Health)	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?	1 - 30: Number of days 88: None 77: Don't know/Not sure 99: Refused	People who refused or did not answer this question will be removed from the dataset. The following categories will be created: 0: Zero days when mental health not good (88) 1-30: Number of days when mental health is not good 9: Don't know/Refused/Missing (99)
PERSDOC3 (Personal_Doctor)	Do you have one person (or a group of doctors) that you think of as your personal health care provider?	1: Yes, only one 2: More than one 3: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset. 1 and 2 will be combined to indicate the person has at least one personal doctor (1) and 3 will be changed to 2 indicating that the person does not have a personal doctor.
MEDCOST1 (Not_Afford_Doctor)	Was there a time in the past 12 months when you needed to see a doctor but could not because you could not afford it?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
SMOKE100 (Smoked_100_Pkgs)	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	1: Yes 2: No 7: Don't know/Unsure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
DRNK3GE5 (Binge_Drinking)	Considering all types of alcoholic beverages, how many times during the past 30 days did you have 5 or more drinks for men or 4 or more	1 - 76: Number of times 88: None 77: Don't know/Not Sure 99: Refused	People who refused or did not answer this question will be removed from the dataset. The following categories will be applied to this column: ● 1-76: Number of times the individual binge drank in past 30 days

	drinks for women on an occasion?		• 0: 88 (None)
EXERANY2 (Exercised)	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
BPHIGH6 (HighBP)	Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure? (If ´Yes´ and respondent is female, ask ´Was this only when you were pregnant?´.)	1: Yes 2: Yes, but female told only during pregnancy 3: No 4: Told borderline high or pre-hypertensive or elevated blood pressure 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset. 1: Yes (1, 4) 2: No (2, 3)
TOLDHI3 (High_Cholesterol)	Have you ever been told by a doctor, nurse or other health professional that your cholesterol is high?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
CVDINFR4 (Myocardial_Infarction)	(Ever told) you had a heart attack, also called a myocardial infarction?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
CVDCRHD4 (Heart_Disease)	(Ever told) (you had) angina or coronary heart disease?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
CVDSTRK3 (Stroke)	(Ever told) (you had) a stroke.	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.
CHCKDNY2 (Kidney_Disease)	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?	1: Yes 2: No 7: Don't know/Not sure 9: Refused	People who refused or did not answer this question will be removed from the dataset.

DIABETE4 (Type_2_Diabetes)	(Ever told) (you had) diabetes? (If ´Yes´ and respondent is female, ask ´Was this only when you were pregnant?´. If Respondent says pre-diabetes or borderline diabetes, use response code 4.)	1: Yes 2: Yes, but female told only during pregnancy 3: No 4: No pre-diabetes or borderline diabetes 7: Don't know/Not sure 9: Refused	People answered 2, 7 or 9 will be removed from the dataset. People who answered 1 or 4 will be analyzed in DIABTYPE. The column will be converted into two groups: 0: does not have diabetes (3) 1: has type 2 diabetes or pre-diabetes or borderline diabetes (1 and 2 for DIABTYPE or 4)
DIABTYPE	According to your doctor or other health professional, what type of diabetes do you have?	1: Type 1 2: Type 2 7: Don't know/Not sure 9: Refused	If a person answered yes to DIABETE4 and 2 to DIABTYPE, they will be kept as a positive to having type 2 diabetes. If a person answered yes to DIABETE4 and 1, 7, or 9 to DIABTYPE, they will be excluded from the dataset as we are predicting the likelihood of being diagnosed with type 2 diabetes. This column was removed after DIABETE4 was transformed.

This table shows the column names of all columns chosen from the CDC BRFSS dataset. The column names used during exploratory data analysis are contained in parentheses. The question column indicates the question asked during the BRFSS survey. The values column shows what the original values of the column indicate. Data cleaning/processing shows the steps that were taken to transform each respective column and what the new values mean (if new values were created). The colors in this table show columns representing social factors, demographic characteristics, health conditions, and outcome variable.

Appendix 2. Lasso Regression Output

Variable	Coefficient
Intercept	-1.0595
EDUCA	-0.0807
EMPLOY1	0.0030
INCOME3	-0.1465
SEXVAR2	-0.1402
PHYSHLTH	0.0665
MENTHLTH	0.0311
PERSDOC32	-0.0650
MEDCOST12	-0.2075
SMOKE100	-0.0056
DRNK3GE5	-0.1092
EXERANY2	0.2716
BPHIGH6	-0.6370
TOLDHI3	-0.6551
CVDINFR4	-0.3268
CVDCRHD4	-0.1747
CVDSTRK3	-0.1160
CHCKDNY2	-0.5672
_AGE80	0.4313
BMI	0.4415

Legend: This table indicates the lasso regression and the coefficients that the lasso regression returned for each column in this dataset.

Appendix 3. Table of Proportions of Variables Prior to Data Cleaning

Characteristic	N = 433,323[1]
EDUCATION LEVEL	
1	687 (0.2%)
2	8,324 (1.9%)
3	16,161 (3.7%)
4	106,613 (25%)
5	114,346 (26%)
6	184,867 (43%)
9	2,316 (0.5%)
Unknown	9
EMPLOYMENT STATUS	
1	177,871 (41%)
2	37,923 (8.8%)
3	7,668 (1.8%)
4	8,976 (2.1%)
5	17,521 (4.1%)
6	10,244 (2.4%)
7	139,949 (33%)
8	25,490 (5.9%)
9	4,713 (1.1%)

Unknown	2,968
INCOME LEVEL	8 (6, 10)
Unknown	8,075
REPORTED WEIGHT IN POUNDS	180 (150, 220)
Unknown	10,611
REPORTED HEIGHT IN FEET AND INCHES	507 (504, 511)
Unknown	11,649
SEX OF RESPONDENT	
1	203,782 (47%)
2	229,541 (53%)
NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD	88 (14, 88)
Unknown	3
NUMBER OF DAYS MENTAL HEALTH NOT GOOD	88 (10, 88)
Unknown	3
HAVE PERSONAL HEALTH CARE PROVIDER?	
1	234,388 (54%)
2	141,728 (33%)
3	52,973 (12%)
7	3,271 (0.8%)
9	960 (0.2%)
Unknown	3
COULD NOT AFFORD TO SEE DOCTOR	

1	37,198 (8.6%)
2	394,587 (91%)
7	1,174 (0.3%)
9	362 (<0.1%)
Unknown	2
SMOKED AT LEAST 100 CIGARETTES	
1	158,774 (38%)
2	251,981 (61%)
7	2,251 (0.5%)
9	643 (0.2%)
Unknown	19,674
BINGE DRINKING	88 (30, 88)
Unknown	221,634
EXERCISE IN PAST 30 DAYS	
1	325,227 (75%)
2	106,845 (25%)
7	927 (0.2%)
9	322 (<0.1%)
Unknown	2
EVER TOLD BLOOD PRESSURE HIGH	
1	176,222 (41%)
2	3,280 (0.8%)

3	247,855 (57%)
4	4,047 (0.9%)
7	1,312 (0.3%)
9	604 (0.1%)
Unknown	3
EVER TOLD CHOLESTEROL IS HIGH	
1	158,906 (42%)
2	219,333 (57%)
7	2,868 (0.8%)
9	404 (0.1%)
Unknown	51,812
EVER DIAGNOSED WITH HEART ATTACK	
1	23,451 (5.4%)
2	407,304 (94%)
7	2,314 (0.5%)
9	251 (<0.1%)
Unknown	3
EVER DIAGNOSED WITH ANGINA OR CORONARY HEART DISEASE	
1	23,454 (5.4%)
2	405,638 (94%)
7	3,936 (0.9%)

9	292 (<0.1%)
Unknown	3
EVER DIAGNOSED WITH A STROKE	
1	18,350 (4.2%)
2	413,499 (95%)
7	1,212 (0.3%)
9	258 (<0.1%)
Unknown	4
EVER TOLD YOU HAVE KIDNEY DISEASE?	
1	20,555 (4.7%)
2	410,876 (95%)
7	1,622 (0.4%)
9	267 (<0.1%)
Unknown	3
(EVER TOLD) YOU HAD DIABETES	
1	59,786 (14%)
2	3,253 (0.8%)
3	358,706 (83%)
4	10,594 (2.4%)
7	683 (0.2%)
9	296 (<0.1%)
Unknown	5

WHAT TYPE OF DIABETES DO YOU HAVE?	
1	1,958 (8.1%)
2	20,069 (83%)
7	2,199 (9.1%)
9	49 (0.2%)
Unknown	409,048
IMPUTED AGE VALUE COLLAPSED ABOVE 80	58 (41, 71)

[1] n (%); Median (Q1, Q3)

Legend: All numeric columns are described by a median and first and third quartile values in parentheses separated by a comma. All factor columns are represented by a percentage denoting the proportion. This table shows the data prior to data cleaning and processing.

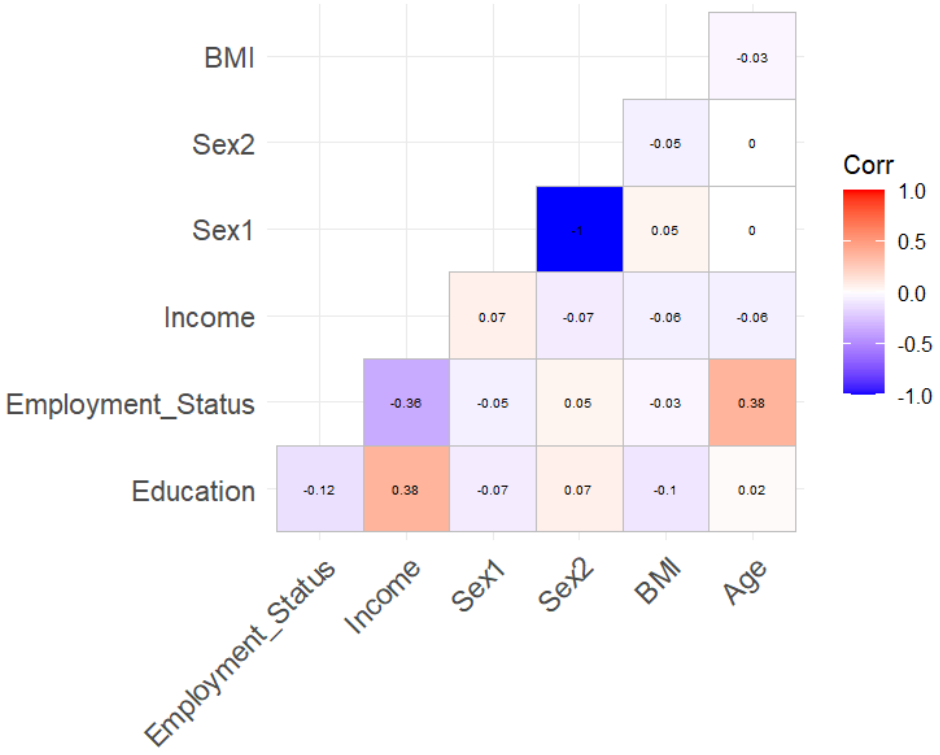
Appendix 4. Correlation Plots for Each Group of Predictor Variables

Legend for the following plots: The following plots show high correlation by a dark red for positive and dark blue for negative.

Figure 1. Correlation Plot for Social Factors Variables

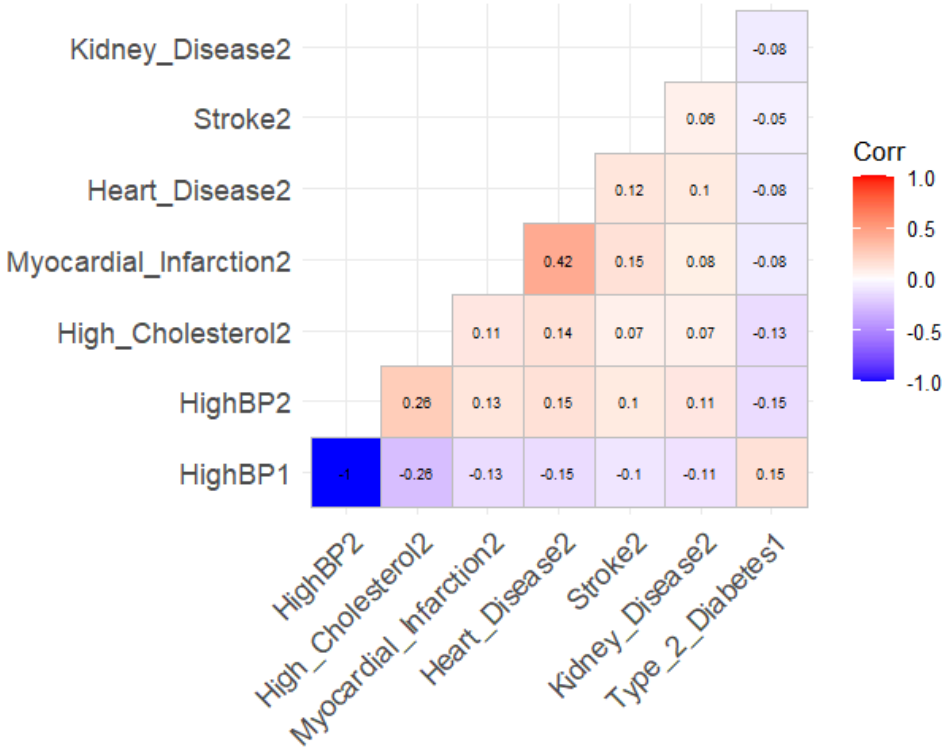
There do not appear to be any significant correlation coefficients between social factor variables. The only significant correlation coefficient occurs within the Physical_Health variable between Physical_Health0 and Physical_Health1. The correlation plot treats each factor within a variable as an individual variable.

Figure 2. Correlation Plot for Demographic Variables



There do not appear to be any significant correlation coefficients between demographic variables. The most significant correlation between Sex1 and Sex2 is due to the sex variable being treated as a factor resulting in female and male being treated as separate variables. The next highest correlation coefficients come from employment status, education, income, and age. The correlation coefficients are not significant enough to impact the chosen models (random forest and xgBoost).

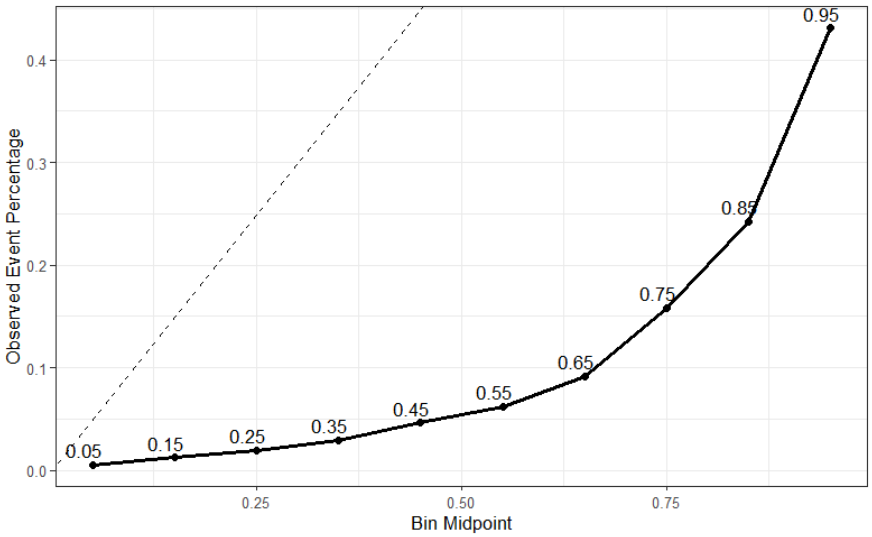
Figure 3. Correlation Plot for Disease Variables



There do not appear to be any significant correlation coefficients between demographic variables. The only significant correlation coefficient occurs within the HighBP variable between HighBP1 and HighBP2. The correlation plot treats each factor within a variable as an individual variable.

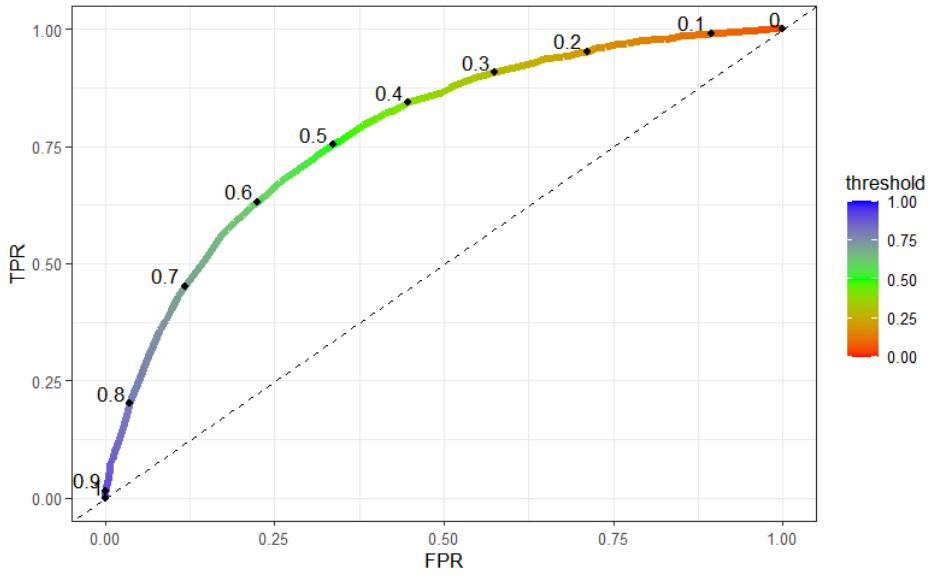
Appendix 5. Plots and Table Characterizing the xgBoost Model

Figure 1. Calibration Plot for xgBoost



The calibration plot shows that the model is underfitting as the curve is below the dashed line that represents a well fit model.

Figure 2. ROC Plot for xgBoost



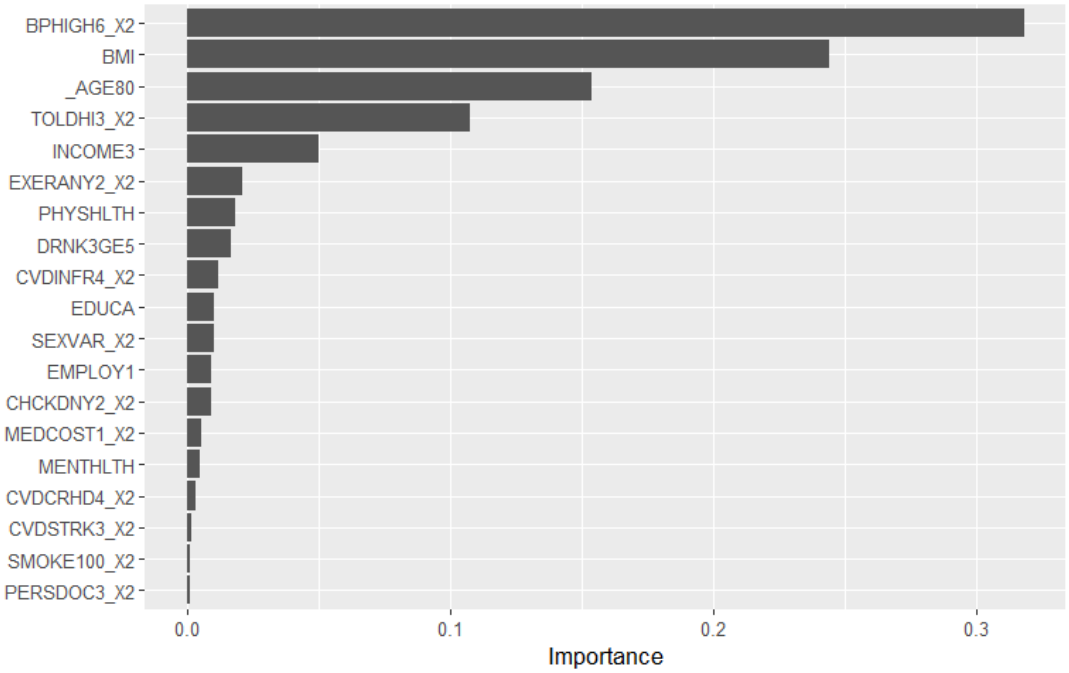
The ROC plot shows the false positive rate (FPR) and the true positive rate (TPR) for different potential threshold values. The threshold values are represented by the colors shown in the graphical legend. The dashed line represents a model that has the same predictive capabilities as random guessing.

Table 1. Confusion Matrix for xgBoost Model

		Predicted	
		Type 2 Diabetes	No Type 2 Diabetes
Actual	Type 2 Diabetes	1,290 (True Positive)	406 (False Negative)
	No Type 2 Diabetes	9,396 (False Positive)	18,430 (True Negative)

The confusion matrix shows the predictions of the model using test data under predicted and the actual determination of type 2 diabetes by the test dataset next to actual.

Figure 3. Feature Importance for xgBoost Model



The feature importance plot indicates the relative feature importance for the xgBoost model predictions.