Modeling Diabetes Risk using the 2015 BRFSS Survey

Alex Teboul
Depaul University, IL 60604 USA

This paper was written for DSC 510 – Health Data Science at DePaul University with professor Stephanie Besser.
Python Code Link:
https://colab.research.google.com/drive/1HUYgcxhmgzv5zELcsnM0da_gwo1Zuiuo?usp=sharing

Abstract

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. This paper explores the use of machine learning methods towards diabetes classification using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. Specifically, Random Forests, Gradient Boosting, AdaBoost, and Neural Networks are applied to the binary classification task of predicting a type II diabetes diagnosis based on survey responses to 21 of the questions asked in the BRFSS. Our study supports previous research that has explored the use of BRFSS datasets prior to 2015 for type II diabetes prediction. Additionally, we present an open source Google Colaboratory notebook for cleaning BRFSS datasets and running machine learning models on them. The Neural Network with logistic activation and adam solver performed the best on the balanced dataset, and had 5 variables selected out of the 21 assessed. Using variables for High Blood Pressure, High Cholesterol, BMI, Age, and self-reported General Health, the Neural Network achieved accuracy of 74.0%, sensitivity of 79.6%, specificity of 69.3%, positive predictive value (PPV) of 72.2%, and negative predictive value (NPV) of 77.2%. These results represent an improvement upon previous survey-based diabetes risk prediction models in the literature and offer the possibility of a low-cost, diabetes risk screening tool.

1. Introduction

Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy.[3] After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed. Complications like heart disease, vision loss, lower-limb amputation, and

kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes.[4] While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients.[4] Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes.[1] Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk.[1] While there are different types of diabetes, type II diabetes is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health.[3] Much of the burden of the disease falls on those of lower socioeconomic status as well. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approxing $400 billion dollars annually.[3]

In this paper, we explore the efficacy of different machine learning techniques for type II diabetes prediction in the general population. To this end, cross-sectional survey data collected in 2015 through the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System is used. The aim of this project was to build a concise model using the most predictive survey questions that would also have relatively high Accuracy, AUC, Precision, and Recall. This predictive model could then serve as an awareness-tool for those at high risk of diabetes, especially considering that the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk.[1]

Initial hypotheses were high BMI and low physical activity would be strongly predictive of diabetes risk, as the Centers for Disease Control and Prevention indicates that 89% of diabetics are overweight and 38% were physically inactive.[1] Self-evaluation survey questions regarding general health and wellbeing were also anticipated to be important to the models as these are often the first signs for many people that something is awry before seeking medical attention and receiving a diagnosis. Age is another variable that was expected to factor into the model, as type II diabetes, like other chronic diseases mostly presents in adulthood. Given the large dataset, diverse feature set, and a review of prior literature on diabetes model building using BRFSS datasets, the models were expected to yield relatively high accuracies. The class imbalance problem in our dataset was expected to lead to some challenges with obtaining high recall (sensitivity), as about 15% of the individuals in the dataset had diabetes versus 85% that did not. Highly accurate models could still have low recall values, which would be a red flag for the use of a subset of BRFSS survey questions to serve as an awareness tool. By undersampling the non-diabetic group and including the few number of survey participants labelled as prediabetic in with the general population, this problem was solved.

2. Literature Review

Predictive and statistical models for diabetes risk have been explored by multiple groups in recent years.[2,5-8] Kavakiotis et al. conducted a systematic review of machine learning techniques being used in diabetes research in 2017 and found applications in diagnosis, genetic background, health care management, and laboratory settings.[7] They also found that most research used clinical datasets and algorithms like naive bayes, logistic regression, support vector machines, random forests, and neural networks. Hippisley-Cox & Coupland describe a different set of approaches to large scale diabetes risk modeling in England in 2017. Specifically, in England they identify risk factors and model using a cox proportional hazards model.[7] As diabetes is a global problem and all demographics are susceptible to diabetes disease, methods applied and risk factors identified in England can reasonably be assumed to be valid in the U.S. context as well.

Researchers have demonstrated that models of high accuracy can be developed to predict diabetes mellitus. For example, Zou et al. in 2018 were able to use random forests to achieve roughly 80% accuracy in diabetes prediction on a dataset collected in Luzhou, China.[2] Though they did not produce any other necessary performance metrics. A more robust set of predictive models was created by Yuvaraj & SriPreethaa in the Journal of Cluster Computing in 2019 using a dataset from the National Institute of Diabetes of 75,664 patients.[8] Importantly, their dataset had detailed clinical data on the patients like their 2-hour serum insulin, which is not available in the BRFSS dataset used in this project. In the paper, precision, recall, F-measure, and accuracy were reported for naive bayes, random forest, decision trees and other algorithms using Hadoop. Performance metrics all fell within the 77-94 range indicating relatively strong performance in predicting diabetes risk.[8] That we are able to achieve similar performance metrics on the survey style BRFSS dataset compared to a clinical/lab value dataset is promising for its use as an awareness tool.

Finally, as we are using the 2015 BRFSS survey data in this project it is useful to consider it in the context of a recent publication by CDC researchers Xie et al. in 2019 which explored the 2014 BRFSS for type II diabetes prediction.[2] They found that for a cross-sectional group of 138,146 survey participants from 2014, neural networks could predict Type II Diabetes incidence with 82.4% accuracy, 90.2% specificity, and 0.795 AUC.[2] Unfortunately, not all of the variables they used in their final model were available in the 2015 BRFSS dataset. In particular, sleep was missing from the new dataset, which they had identified as an important risk factor. The performance metrics in the literature served as important benchmarks as we tuned our own models, opting to accept some accuracy loss in exchange for greater recall.

3. Methods

This project draws upon existing research into the risk factors that impact diabetes risk to build binary classifiers using Random Forests, Gradient Boosting, AdaBoost, and Neural Networks.

Our primary research questions:

1. **Predicting Diabetes:** To what extent can a subset of survey questions from the BRFSS be used to effectively predict type II diabetes risk?
2. **Research Tool:** Could this serve as a screening tool and can we produce an open source Google Colab notebook to allow researchers or students to clean BRFSS datasets and run machine learning models on them?

The methods used to accomplish this are as follows:

1. Dataset and Feature Selection
2. Data Preprocessing
3. Undersampling to Create a 50-50 balanced Dataset
4. Model Building

3.1 Dataset and Feature Selection

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

From this original dataset, features were selected that matched the risk factors identified in existing diabetes research and presented in the literature review. To help understand what the variables meant, the response options, and original questions, the BRFSS 2015 Codebook was consulted. This Codebook is available in the Appendix along with the BRFSS 2015 Kaggle dataset link. To aid in this process, we referenced the variables discussed in the paper by Xie et al. for *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques* using the 2014 BRFSS.[2] We aimed to validate the results of this existing research by achieving similar performance metrics on the more recent dataset using similar techniques.

Selected variables can be found below in *Table 1: Model Features*. Our dependent variable was Diabetes. From the BRFSS survey we selected 21 variables, which were originally questions posed to participants. These variables were selected from the original 330 available variables in the dataset, and cover High Blood Pressure, High Cholesterol, BMI, Smoking History, related Chronic Health Conditions for Heart Disease and Stroke, Physical Activity, Diet, Alcohol Consumption, Healthcare Access, General Health and Wellbeing, and Demographics. Most of the variables are binary, others are categorical but ordinal, and only BMI is metric. More

information on what these variables mean and their original names in the BRFSS dataset can be found in the *Variable Specifics* Table in the Appendix.

3.2 Data Preprocessing

We took the original 441,455 individuals in the dataset and removed missing values as well as responses that were either "*don't know"* or "*refused to answer*". This brought us to 253,680 survey responses collected from the 2015 BRFSS survey for use in this study. Of these, 35,346 self-reported having been diagnosed with diabetes. Prediabetes participants were added to the participants who had reported never having been diagnosed with diabetes or prediabetes. We did this because there were only a couple thousand prediabetic individuals in the dataset. This was significantly lower than would be expected given the CDC's given prevalence of prediabetes in the U.S. population.[1] We also wanted our model to be trained on data that would best approximate the general public in the United States. Additional preprocessing steps can be found in the open source Google Colab notebook found in the Appendix, as each of the 22 variables in the dataset were adjusted and tested.

3.3 Undersampling to Create a 50-50 Balanced Dataset

Machine learning models were initially run on the clean dataset, but we found that models achieved high accuracy and AUC but low precision and recall. We realized the need for either undersampling or oversampling due to the large class imbalance that existed between diabetics and nondiabetics in the dataset. We opted for randomly undersampling the nondiabetic group and forming a 50-50 balanced dataset, though SMOTE was used in the literature to oversample the diabetic group. We found that both methods produced similar performance metrics, but the smaller, balanced dataset trained faster. This dataset consisted of all 35,346 diabetics and 35,346 randomly selected nondiabetics. See *Table 1: Cleaned Datasets* below for summary.

Table 1: Cleaned Datasets

Dataset	Participants with Diabetes	Participants without Diabetes	Total
Binary Unbalanced	35,346	218,334	253,680
Binary Balanced	35,346	35,346	70,692

3.4 Model Building

Once the datasets were cleaned, the 4 models were tested on each dataset with different parameter settings. All models were trained on 70% of the data, with 30% reserved for testing.

Cross validation was also applied with 5-fold cross validation as some form of cross validation was common throughout the research papers identified in the literature review. Performance was measured in terms of Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). The random forests, gradient boosting, AdaBoost, and neural network models were tested both with and without feature selection. It was determined that feature selection did not negatively impact performance metrics, so we report only feature selection models here. For the random forests, we used feature selection with entropy, yielding 8 selected features. For the other models we used a wrapper selection method and gradient boosting classifier set to deviance loss for feature selection, yielding 5 selected features. *Table 4: Final Model Parameters* in the Appendix lists the specifications for each model displayed in the Results section.

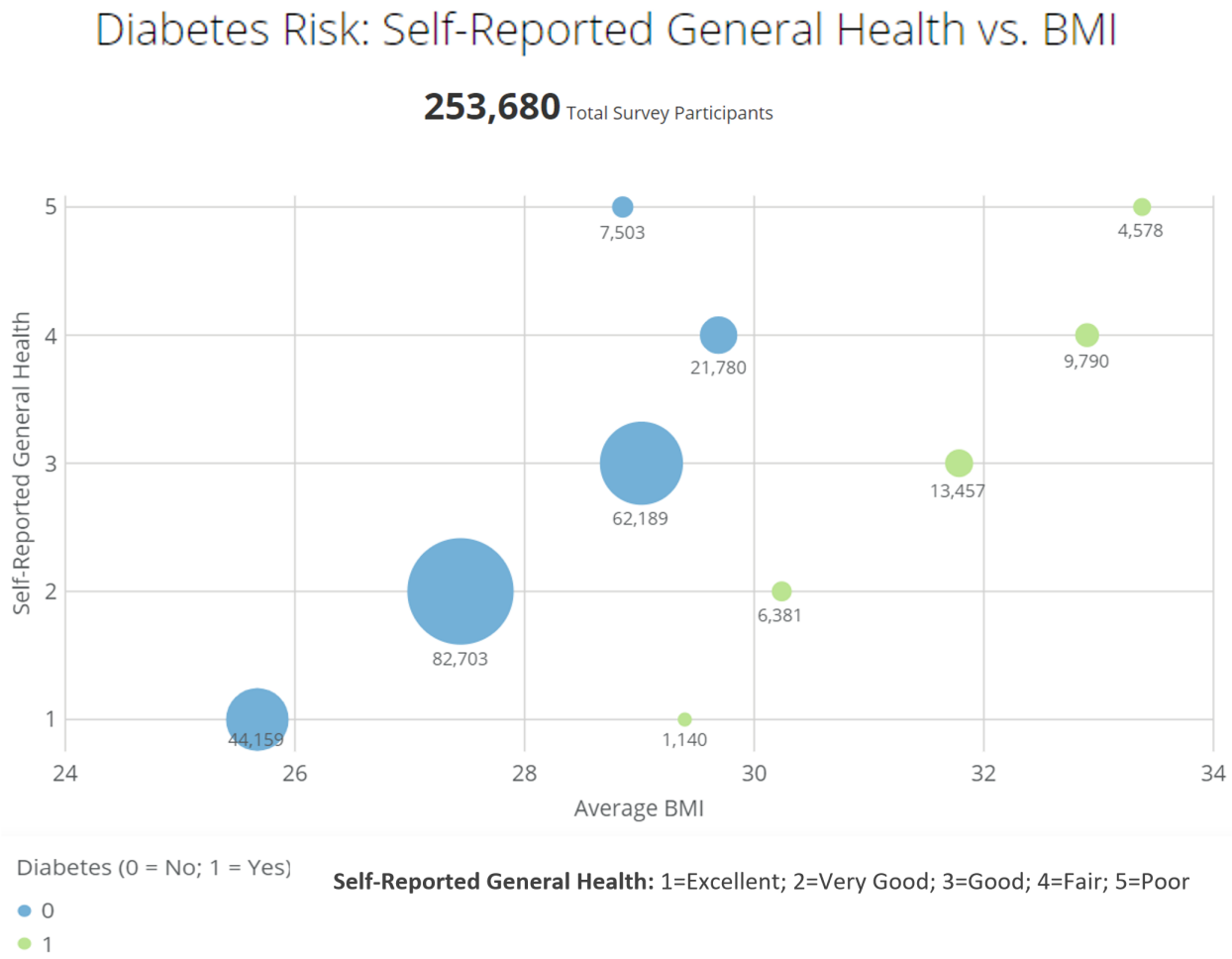
4. Results

While all 4 models originally had high accuracy (85.0% - 87.0%), high specificity (95.8% - 98.5%), and high NPV (87.4% - 87.9%), they suffered from low sensitivity (12.3% - 18.5%) and low PPV (41.8% - 57.6%). The cause was identified as class imbalance. By randomly undersampling the non-diabetic group to create a balanced dataset of 35,346 diabetics to 35,346 non diabetics, we improved model performance. At the expense of accuracy (71.0% - 74.0%), specificity (68.3% - 71.9%), and NPV (72.8% - 77.2%), we were able to improve sensitivity (74.5% - 79.6%) and PPV (70.2 - 73.2%). We consider the drop in accuracy to be acceptable for the gain observed in sensitivity. As our application is to predict diabetes risk in the general population, higher sensitivity is critical.

From the feature selection with Random Forest entropy, models indicate that High Blood Pressure, BMI, General Health Rating, Mental Health Rating, Physical Health Rating, Age, Education, and Income are the 8 most predictive variables. Using the wrapper select method described in the methods section, the Gradient Boosting, AdaBoost, and Neural Networks used 5 important features - High Blood Pressure, High Cholesterol, BMI, Age, and self-reported General Health. High Blood Pressure and High Cholesterol are binary variables for individuals who have ever been told by a doctor that they meet the condition. BMI is a calculated variable based on survey participant self-reported weight and height. Age is a 14 level ordinal variable. Self Reported General Health is an ordinal variable on a scale from 1 as Excellent to 5 as Poor. Selected variables are in line with research into risk factors associated with diabetes risk.

Table 2: Binary Unbalanced Dataset - Model Results					
Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Random Forest	85.0%	18.5%	95.8%	41.8%	87.9%
Gradient Boosting	87.0%	14.0%	98.3%	57.2%	87.6%
AdaBoost	87.0%	15.0%	98.2%	56.1%	87.6%
Neural Network	87.0%	12.3%	98.5%	57.6%	87.4%

Table 3: Binary Balanced Dataset - Model Results					
Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Random Forest	71.0%	74.5%	68.3%	70.2%	72.8%
Gradient Boosting	74.0%	78.9%	70.0%	72.4%	76.9%
AdaBoost	74.0%	76.7%	71.9%	73.2%	75.5%
Neural Network	74.0%	79.6%	69.3%	72.2%	77.2%



An interesting outcome of this model building process was to observe that Self-Reported General Health and BMI were two of the most predictive variables for diabetes. When visualized together in the bubble chart above, we can observe an interesting association between quality of life and BMI, as well as their relationship with diabetes. Diabetics have higher BMIs than non-diabetics on average and they also tend to have lower quality of life as indicated by the self-reported general health scores.

5. Discussion

Our best model was a neural network that achieved accuracy of 74.0%, sensitivity of 79.6%, specificity of 69.3%, positive predictive value (PPV) of 72.2%, and negative predictive value (NPV) of 77.2%. For reference, the best models presented by Xie et al (2019) from the CDC also using a survey based BRFSS dataset had lower sensitivity. Their highest accuracy model was a neural network with 82.4% accuracy and 37.8% sensitivity.[2] A decision tree model of lower accuracy but 51.6% sensitivity was indicated as a possible screening tool for diabetes risk. We offer our model and parameter settings as an improvement on this related paper. A more sensitive predictive model is important in order to effectively designate individuals with diabetes as

positive. The more sensitive the test, the fewer false negatives results and therefore fewer diabetics would be missed by the model. Our dependent variable for type II diabetes risk is based on whether survey participants had ever been told by a doctor that they had diabetes. Given that the CDC estimates as many as 1 in 5 individuals that have diabetes are undiagnosed, we find our results to be within expectations for the general population.[1] In interpreting our results, we also find that the few number of variables selected in our best model is preferable to the large number of variables used in models built on clinical datasets or past survey-based datasets. Our model requires only 5 survey questions to be posed to achieve these metrics. These were HighBP, HighChol, BMI, Age, and General Health. The ability to ask few survey questions and gain relatively strong predictive power is beneficial for creating a low-cost, easily applicable screening tool for type II diabetes risk.

That said, while our best model did comparably to models built by the CDC team that studied diabetes with the BRFSS 2014, there remain a number of limitations to our study.[2] First, a potential limitation is that many of the diabetes specific questions in the BRFSS lacked sufficient responses to serve our goals of building a general purpose survey with predictive power for diabetes diagnoses. So questions about an individual's blood sugar, feet sores, times seen by doctor regarding diabetes, glycosylated hemoglobin, eye impairment, etc. that were specific to survey participants who have been diagnosed with diabetes, were left out. As a result, some of the more powerful risk factors or indicators of a patient suffering from diabetes were not included in our models. This had the unintended benefit of leading to a model built on features that are widely known and a model that does not require clinical data to make predictions. Another limitation of this research is that we were unable to use the same sleep variable that researchers Xie et al. (2019) demonstrated to be highly predictive of diabetes risk with the 2014 BRFSS dataset.[2] This sleep variable was unavailable in the 2015 dataset and other sleep variables that were included had a too low a response rate to be included in our final dataset. We also did not include race in our selected variables, though in any future work this would be worthwhile as it is a risk factor identified in the literature.

Additionally, a survey or subset of survey questions with strong predictive power like we have developed cannot be used in lieu of a traditional fasting blood sugar test or A1c blood tests for glycated hemoglobin.[4] A typical means of diagnosing diabetes is to have a patient fast for at least eight hours before measuring their blood sugar levels, and seeing how they compare to normal fasting blood sugar levels which are under 100 mg/dl. That said, we believe that a simple survey from a subset of the BRFSS questions could potentially prove useful as a component of an online awareness tool to help drive individuals towards getting tested and seeking medical diagnosis. Given that 1 in 3 individuals are prediabetic, and may become diabetic, it is important that individuals have access to medical care, nutritional support, and at least tests for diagnosis. In the Appendix, we have shown an infographic that could aid in an awareness campaign. It lists the 5 most significant risk factors identified in our models - high blood pressure, high cholesterol,

BMI, age, and general health rating. Future work could add in a variable for race and clean up the Google Colab Python code in order to make it more user friendly.

6. Conclusion

With roughly 1 in 10 Americans diagnosed with diabetes and as many as 1 in 5 undiagnosed, there is a strong need for effective screening tools and educational outreach.[1] We believe that predictive models built on health-related survey data hold strong potential for use in these screening tools and could complement existing awareness efforts. Moreover, the Behavioral Risk Factor Surveillance System (BRFSS) survey has been demonstrated to be an effective dataset for use in predictive model building for chronic disease. While the majority of prior research in the space of type II diabetes modeling has focused on clinical data modeling, our survey-based models are comparable in predictive power and could serve as a complement to clinical data. Moving forward, we will post the python code on Kaggle to facilitate researchers and students interested in building predictive models for chronic diseases using BRFSS datasets. We hope that this study can serve as a springboard for more effective early screening for diabetes using predictive models and raise awareness of some of the key risk factors for the disease.

7. References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf
2. Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Peer Reviewed: Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Preventing chronic disease, 16.
3. O'Connell, J. M., & Manson, S. M. (2019). Understanding the economic costs of diabetes and prediabetes and what we may learn about reducing the health and economic burden of these conditions. *Diabetes care*, *42*(9), 1609-1611..
4. American Diabetes Association. (2019). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. *Diabetes care*, *42*(Supplement 1), S13-S28.
5. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515.
6. Hippisley-Cox, J., & Coupland, C. (2017). Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *bmj*, *359*, j5019.
7. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.
8. Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing*, *22*(1), 1-9.

8. Appendix

1. The original BRFSS 2015 .csv can be downloaded here: https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system#2015.csv
2. The BRFSS 2015 Codebook is available here: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf
3. The open source Google Colab notebook with Python code is available here.: https://colab.research.google.com/drive/1HUYgcxhmgzv5zELcsnM0da_gwo1Zuiuo?usp=sharing
4. The datasets used for model building, created in the Google Colab notebook are available in this Google Drive folder: https://drive.google.com/drive/folders/1yoEQqCn75TxKknWGkWQvDrVn2O_qYitI?usp=sharing

TABLE 1
MODEL FEATURES

#	Renamed Features	Categories
*1	Diabetes	Response Variable
---	---	---
2	HighBP	High Blood Pressure
3	HighChol	High Cholesterol
4	CholCheck	High Cholesterol
5	BMI	BMI
6	Smoker	Smoking History
7	Stroke	Chronic Health Conditions
8	HeartDiseaseorAttack	Chronic Health Conditions
9	PhysActivity	Physical Activity
10	Fruits	Diet
11	Veggies	Diet
12	HvyAlcoholConsump	Alcohol Consumption
13	AnyHealthcare	Health Care Access
14	NoDocbcCost	Health Care Access
15	GenHlth	General Health & Wellbeing
16	MentHlth	General Health & Wellbeing
17	PhysHlth	General Health & Wellbeing
18	DiffWalk	General Health & Wellbeing
19	Sex	Demographics
20	Age	Demographics
21	Education	Demographics
22	Income	Demographics

Table 4: Final Model Parameters	
Model	**Parameters**
Random Forest	**Parameters**: n_estimators = 200, max_depth = None, min_samples_split = 3, criterion = entropy, Train-Test Split: 70%-30%, cross validation = 5-fold **Feature Selection Parameters:** n_estimators = 200, max_depth = None, min_samples_split = 3, criterion= entropy **Selected Features:** HighBP, BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income
Gradient Boosting	**Parameters**: n_estimators = 200, loss = deviance, learning_rate = 0.1, max_depth = 3, min_samples_split = 3

	Train-Test Split = 70%-30%, cross validation = 5-fold **Feature Selection Parameters:** Wrapper selection method using Gradient Boosting Classifier, n_estimators=200, loss=deviance, learning_rate=0.1, max_depth=3, min_samples_split=3 **Selected Features:** HighBP, HighChol, BMI, GenHlth, Age
AdaBoost	**Parameters**: n_estimators = 200, base_estimator = None, learning_rate = 0.1 Train-Test Split = 70%-30%, cross validation = 5-fold **Feature Selection Parameters:** Wrapper selection method using Gradient Boosting Classifier, n_estimators=200, loss=deviance, learning_rate=0.1, max_depth=3, min_samples_split=3 **Selected Features:** HighBP, HighChol, BMI, GenHlth, Age
Neural Network	**Parameters**: activation = logistic, solver = adam, alpha = 0.0001, max_iter = 1000, hidden_layer_sizes = (10,) Train-Test Split = 70%-30%, cross validation = 5-fold **Feature Selection Parameters:** Wrapper selection method using Gradient Boosting Classifier, n_estimators=200, loss=deviance, learning_rate=0.1, max_depth=3, min_samples_split=3 **Selected Features:** HighBP, HighChol, BMI, GenHlth, Age

Variable Specifics		
Index	**Variable**	**BRFSS Question**
0	**Diabetes_Binary** *BRFSS: DIABETE3*	(Ever told) you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?".
1	**HighBP** *BRFSS: _RFHYPE5*	Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional.
2	**HighChol** *BRFSS: TOLDHI2*	Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?
3	**CholCheck** *BRFSS: _CHOLCHK*	Cholesterol check within past five years
4	**BMI**	Body Mass Index (BMI)

	BRFSS: _BMI5	
5	**Smoker** *BRFSS: SMOKE100*	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
6	**Stroke** *BRFSS: CVDSTRK3*	(Ever told) you had a stroke.
7	**HeartDiseaseorAttack** *BRFSS: _MICHD*	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
8	**PhysActivity** *BRFSS: _TOTINDA*	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
9	**Fruits** *BRFSS: _FRTLT1*	Consume Fruit 1 or more times per day
10	**Veggies** *BRFSS: _VEGLT1*	Consume Vegetables 1 or more times per day
11	**HvyAlcoholConsump** *BRFSS: _RFDRHV5*	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
12	**AnyHealthcare** *BRFSS: HLTHPLN1*	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
13	**NoDocbcCost** *BRFSS: MEDCOST*	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
14	**GenHlth** *BRFSS: GENHLTH*	Would you say that in general your health is:
15	**MentHlth** *BRFSS: MENTHLTH*	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?

16	**PhysHlth** *BRFSS: PHYSHLTH*	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?
17	**DiffWalk** *BRFSS: DIFFWALK*	Do you have serious difficulty walking or climbing stairs?
18	**Sex** *BRFSS: SEX*	Indicate sex of respondent.
19	**Age** *BRFSS: _AGEG5YR*	Fourteen-level age category
20	**Education** *BRFSS: EDUCA*	What is the highest grade or year of school you completed?
21	**Income** *BRFSS: INCOME2*	Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.")

Modeling Diabetes Risk: BRFSS 2015

1 in 10 Americans have diabetes...

...or 34.2 million people [1]

High Blood Pressure

Do you have high blood pressure?

General Health

Rate your health:

◯Excellent ◯Very Good ◯Good ◯Fair ◯Poor

High Cholesterol

Do you have high cholesterol?

Body Mass Index (BMI)

What is your BMI?

Age

What is your age?

Using these 5 Questions....

- **Diabetes can be predicted with**
  - 74% (+/- 0.01) Accuracy
  - 0.82 (+/- 0.01) AUC
  - 0.78 (+/- 0.01) Recall
  - 0.71 (+/- 0.02) Precision
- **Models Tested**
  - Neural Networks
  - Random Forests
  - AdaBoost
  - Gradient Boosting
- **Model Specs**
  - 75,323 responses used in models
  - Undersampling used to create this balanced 50-50 dataset
  - 21 total variables assessed

1 Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf