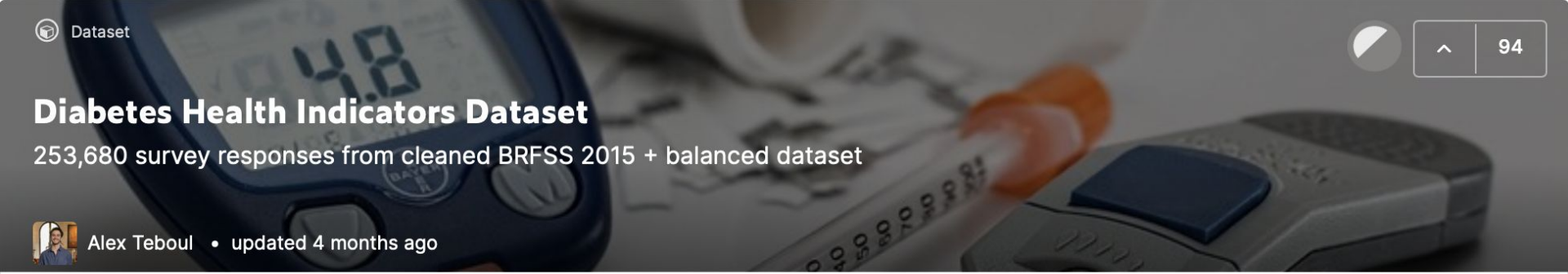
Predicting Diabetes Risk using the 2015 BRFSS Survey

DSC 510 - Health Data Science

Alex Teboul



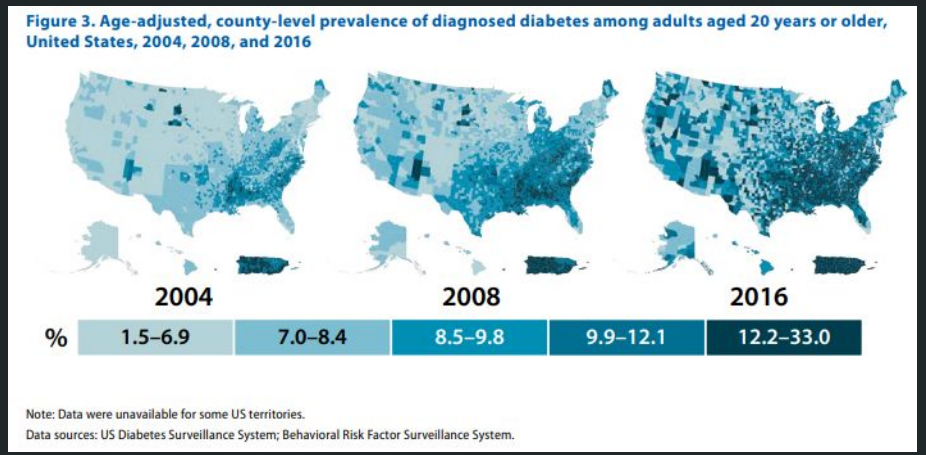
Links

1. Diabetes Health Indicators Dataset - Kaggle Open Source
2. Data Cleaning Notebook - Kaggle Open Source
3. Full Code Notebook with Model Building
4. Paper

Introduction - Diabetes

- **34.2 million** Americans have diabetes[2]
- **88 million** American adults have prediabetes[2]
- **7th leading cause of death** in the United States[2]
- **$400+ Billion** annually[5]

- **1 in 5 are unaware** they are diabetic[2]



(Centers for Disease Control and Prevention, 2020)

Literature Review - Machine Learning in Diabetes

Diabetes Machine Learning Systematic Review (Kavakiotis et al., 2017)

- Applications in diagnosis
- Clinical datasets
- Naive bayes, logistic regression, support vector machines, random forests, and neural networks.

Diabetes Healthcare Systems Hadoop ML (Yuvaraj & SriPreethaa, 2019)

- Robust Clinical Dataset
- 94% Accuracy (RF)
- 94% PPV (RF)
- 88% Sensitivity (RF)
- 91% F-measure (RF)
- Not applicable as screening tool.

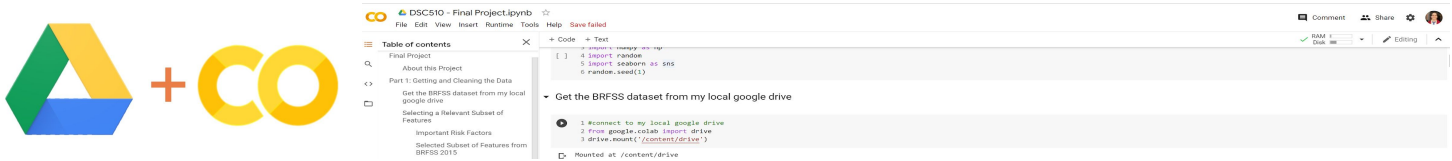
BRFSS 2014 Type II Diabetes Prediction, CDC (Xie et al., 2019)

- 82.4% Accuracy (NN)
- 90.2% Specificity (NN)
- 37.8% Sensitivity (NN)
- 51.6% Sensitivity (DT)
- Concluded decision tree model could provide initial population screening.

Methods - 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?



Methods - Dataset

BRFSS™
Behavioral Risk Factor Surveillance System

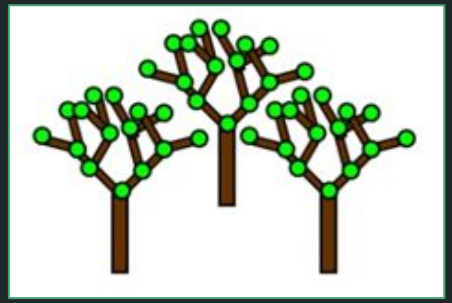
A health-related telephone survey that is collected annually by the CDC.

- **Diabetes** as Dependent Variable
- **21 Other Variables** selected from survey
- **Data Cleaning** in Colab Notebook

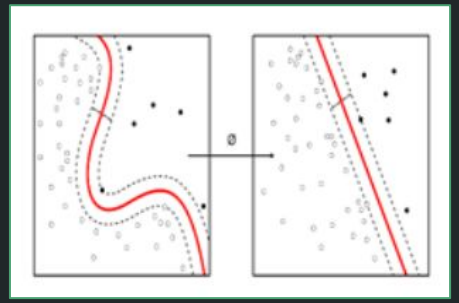
Table 2: Cleaned Datasets			
Dataset	**Participants with Diabetes**	**Participants without Diabetes**	**Total**
Binary Unbalanced	35,346	218,334	253,680
Binary Balanced	35,346	35,346	70,692

Models Tested

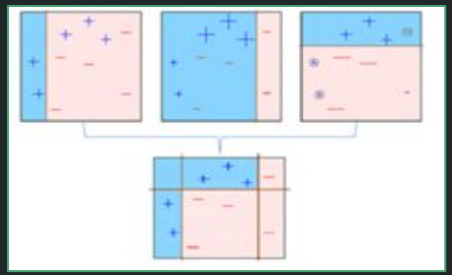
Random Forest



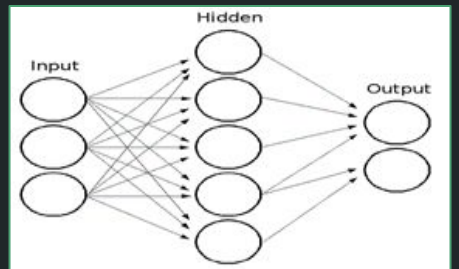
Gradient Boosting



AdaBoost



Neural Network



Model Testing

- 5-fold cross validation
- Feature selection tested with Random Forest Entropy and Gradient Boosting
- Unbalance & Balanced Dataset
- Performance Metrics
  - Accuracy
  - Sensitivity
  - Specificity
  - Positive Predictive Value (PPV)
  - Negative Predictive Value (NPV)

Table 3

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

Results - Predicting Diabetes

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%

- High accuracy and specificity but low sensitivity ➔ Bad
- Need to optimize Sensitivity and PPV if applicable as screening tool.

Results - Predicting Diabetes

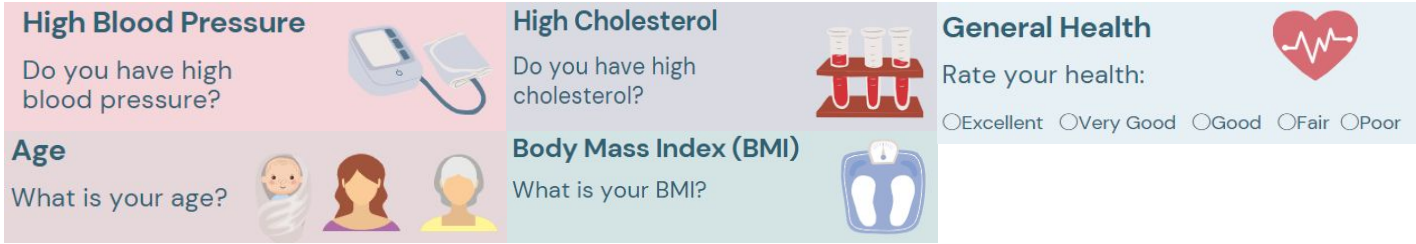
Table 3: Binary Balanced Dataset - Model Results					
Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Random Forest	71.0%	74.5%	68.3%	70.2%	72.8%
Gradient Boosting	74.0%	78.9%	70.0%	72.4%	76.9%
AdaBoost	74.0%	76.7%	71.9%	73.2%	75.5%
Neural Network	74.0%	79.6%	69.3%	72.2%	77.2%

- Improved sensitivity, but lower accuracy and specificity → Better
- CDC BRFSS Diabetes model 51.6% sensitivity & 82.4% accuracy (Xie et al., 2019)

Discussion - Predicting Diabetes

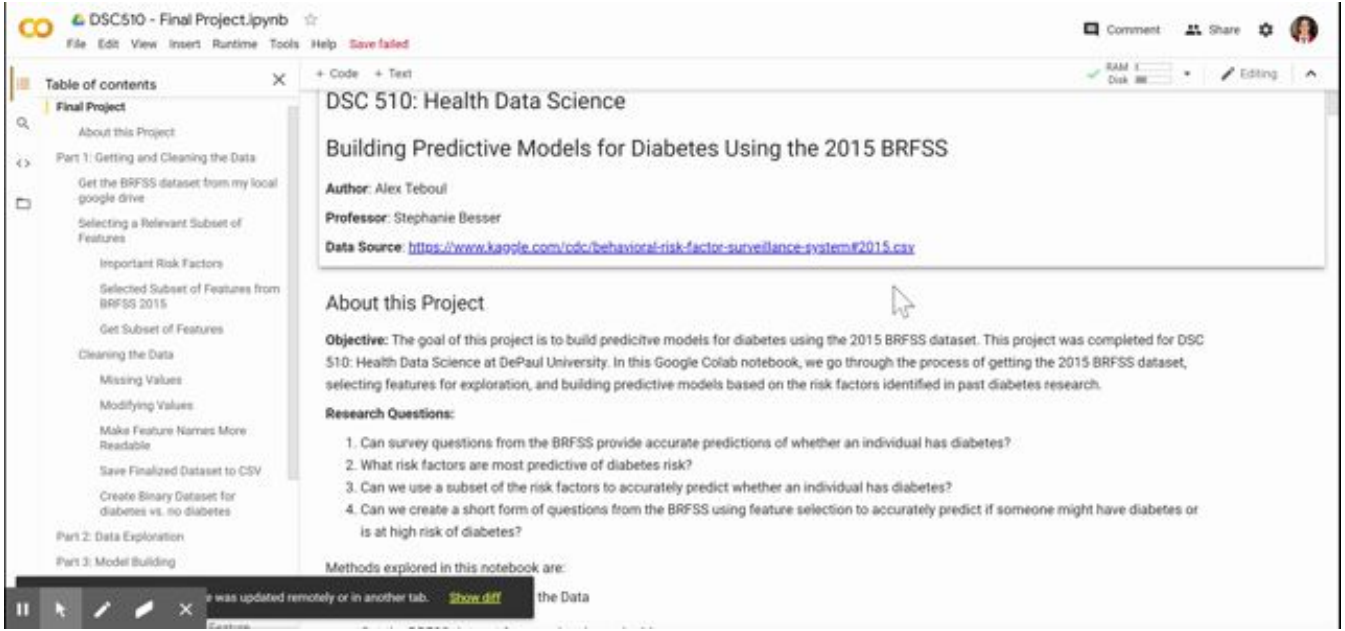
- To what extent can a subset of survey questions from the BRFSS be used to effectively predict type II diabetes risk?

- Performance not high enough to be used in place of medical diagnosis.
- Performance comparable to models with BRFSS survey data (Xie et al. 2019)
- High Blood Pressure, High Cholesterol, BMI, Age, and General Health.



Discussion - 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?



Limitations & Future Work

- Prediabetics included in non-diabetic group because small sample size and didn’t match CDC prevalence.
- Survey data is self-reported and subject to recall bias.
- Model is predicting if patient has been told by a doctor that they have diabetes - 1 in 5 are undiagnosed (Centers for Disease Control and Prevention, 2020).

- Future work could involve adding in race variables or improving modeling notebook.

Conclusion

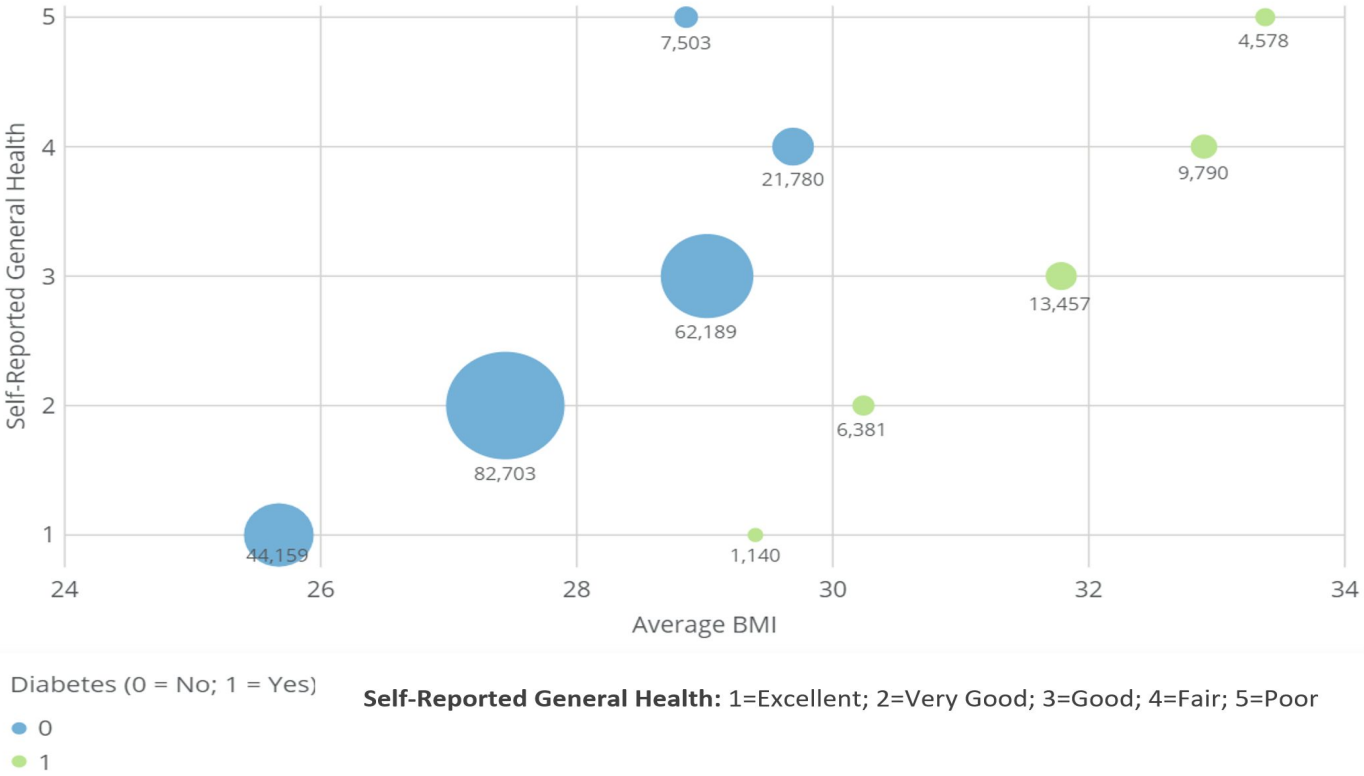
- Need for effective screening tools
- BRFSS survey predictive models comparable to clinical data models for Type II Diabetes risk.
- Kaggle & Open-source

References

1. American Diabetes Association. (2019). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. *Diabetes care*, *42*(Supplement 1), S13-S28.
2. 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
3. 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.
4. 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.
5. 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..
6. 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.
7. 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. 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.

Diabetes Risk: Self-Reported General Health vs. BMI

253,680 Total Survey Participants



Self-Reported General Health: 1=Excellent; 2=Very Good; 3=Good; 4=Fair; 5=Poor

Appendix

Table 3: 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

- **The original BRFSS 2015 .csv can be downloaded here:** https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system#2015.csv
- **The BRFSS 2015 Codebook is available here:** https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf
- **The open-source Google Colab notebook with Python code is available here:** https://colab.research.google.com/drive/1HUYgcxhmgzv5zELcsnM0da_gwo1Zuiuo?usp=sharing
- **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
- **Video Presentation Link**

Appendix

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** *BRFSS: _BMI5*	Body Mass Index (BMI)

= Dependent Variable

= Selected in best Model

Appendix

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**	Consume Fruit 1 or more times per day

Appendix

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?

Appendix

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.")