

600 Master’s Project Report

Diabetic detection by ML Model building utilizing the health indicator diabetes dataset

by

Saket Reddy Chintala

02046463

Advisor

Hua Fang

Date

28/11/2023

Department of Computer and Information Sciences

University of Massachusetts Dartmouth

Diabetic detection by ML Model building utilizing the health indicator diabetes dataset

Saket Reddy Chintala

Department of Computer and Information Science

University of Massachusetts Dartmouth

Advisor

Dr. Hua Fang

Department of Computer and Information Science

University of Massachusetts Dartmouth

*Abstract* — In this project, we have built ML model on Health Indicator Diabetes Dataset allows us to illustrate which indicator is more relevant to cause diabetes in the common people with different economic, education, and other kinds of day-to-day habit which may cause or impact the body to trigger to our immune system to stop release of the insulin which cause us to become Diabetes. We have built the model with balanced dataset with 70,692 patients. We have evaluated the multiple ML model performance by measuring accuracy and time taken for training the model. We have found that Logistic Regression based model is most accurate with accuracy 81.42% and time taken in training is 0.27ms. Other two model Random Forest and Xgboost model was also quite close in accuracy. These models with selected indicators provide the outcomes for a person to become a diabetic if there is no change in our day-to-day lifestyle.

Keywords — Numpy, Feature Selection, Outlier detection, Scikit-learn, Xgboost.

# Introduction

Diabetes is a chronic disease that affects how the body processes glucose, which is the primary source of energy for the cells. Diabetes has become increasingly common in recent years, and it is estimated that millions of people worldwide are affected by this disease. In the United States, according to the Centers for Disease Control and Prevention (CDC), as of 2018, 34.2 million Americans have diabetes, and 88 million have prediabetes. However, the CDC estimates that many people with diabetes or prediabetes are unaware of their risk. In this project, we have selected the datasets generated by B such as 1 in 5 people with diabetes and roughly 8 in 10 people with prediabetes are unaware of their condition.

There are several reasons why people may be unaware of their risk for diabetes or prediabetes. One reason is that the early symptoms of diabetes can be mild or even absent, which can make it difficult to detect the disease in its early stages.

In this project, we have performed studies of 3 samples from health indicator diabetes datasets, and selected the one sample which is balanced, has only two categories i.e., diabetes, non-diabetes. We have cleaned data, studies each indicator to understand the correlation of each feature with diabetes and as well as each other too. After understand each indicator, we have performed the feature selection, and generated five machine learning model, to train and test to get the accuracy of the model. We have presented results in experiment, results and conclusion sections.

# Related Work

Machine learning techniques have been used for diabetes classification for decades. In the paper [3] by Juyoung Shin focus on to develop valid and applicable DM prediction models using an electronic medical record database. The data were extracted from electronic medical records of the Health Promotion Center of Seoul St. Mary’s Hospital between 2009 and 2018. Four prediction model develop in this study to ensure accuracy. Model-1 and Model-2 were predicting the Diabetes development after 2 and 1 years, respectively (normal or prediabetic). Model-3 was a model for predicting the development of DM after 1 year in prediabetic subjects. Model-4 was a model for predicting the development of DM after 1 year in prediabetic subjects after who knows the difference between one and 2 years before diabetes diagnosis. Gradient boosting algorithms were used in Model-1, Model-2, and Model-3, and random forest algorithms were used in Model-4. The non-diabetic and diabetic subjects adjusted with same number in each model. Two set of variables used to develop prediction model. One set comprises 62 easily access variables from hospital and other model developed by 27 variables out of 62 from national health checkup records. The variables were age, sex, medication use, underlying diseases, family history, physical examinations etc. The information collected from medical records and found 3952 individuals were diagnose with diabetes. The test included self-reported information, physical examinations, and laboratory tests, pulmonary functional test, fasting glucose, HbA1c levels. These parameters make model more effective with AUC value approximately 0.9. This model help to health providers, clinicians to encourage people to follow healthy lifestyle.

In the paper [6], Josiane Schneiders focus on the study, that was to evaluate care quality indicator in diabetes patients at primary and tertiary health care level in Brazilian health care system. The analysis was evaluated based on the comparison of these healthcare system. The study was based on the assessment of quality care for the diabetes in these healthcare centers. According to ADA, the diabetic patient with control glycemic value should have at least two HbA1c tests and the patient with no well control glycemic should have at least 4 HbA1c tests. Only 50% of the people tested in primary health care and one fifth patient in tertiary health care. The diabetic neuropathy feet assessment test rate was very low especially in primary care. The study also found that lower percentage of patients had undergone retinopathy test. The reason might be lack of facilities in health care center to examine the problem and due to presence of lack of professionals. The reason is expected because these centers refer to the patients to the ophthalmologist. Nephropathy is the most common quality indicator and easily assessed by creatinine test for glomerular filtration test included in the study. The quality care indicator assessment was at lower rate in health care centre according to this study. If the findings of the study considered, it could increase the prevalence of quality care indicator for the patient’s best possible care.

# Data Set and Features

For understanding the diabetes, I have browsed through many data maintaining website i.e. data-world, Kaggle, few hospital website as well which keeps the information of diabetes patients who are facing diabetes as challenges in their life.  We have selected the datasets, “Diabetes Health Indicators dataset”, was downloaded from Kaggle. This dataset was generated by health-survey named “The Behavioral Risk Factor Surveillance System” BRFSS that is collected in 2015 by the “Central of Disease Control” (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. 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. This dataset contains 3 files:

1.diabetes\_012\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC’s BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 feature variables and is imbalanced.

2.diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 70,692 survey responses to the CDC’s BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced. (selected for model building)

3.diabetes\_binary\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC’s BRFSS2015. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

For the selected dataset, we have the written the following important facts:

* It has many independent variables (20 feature) and 1 dependent variable (diabetes binary)
* It has many numbers of samples in each kind of dataset files 70692 (selected sample File). It is good for predictive model to have sufficient number of samples for training and testing.
* It has the features i.e. Diabetes\_binary, HighBP, HighChol, CholCheck, BMI, Smokes, Stroke, HeartDiseaseorAttack, PhysActivity,Fruits,Veggies,HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.
* We have added the demographics information about the dataset in the Appendix.

# Methods

For processing the datasets, we have considered to following steps to explore the dataset understanding:

### Cleaning the datasets for NA/NULL Values

Many machines learning classifier doesn’t work on the dataset which has null values in their samples as it produces error or generate inaccurate prediction, So It’s better to clean the dataset before performing any step towards model building. We have applied NumPy library to check and remove the sample consist the null values.

### Statistical analysis on each health indicator

We have utilized the Chi-Square test to perform the statistical analysis of each marker vs Diabetes to prove Null Hypothesis False. Null Hypothesis states that any relevance feature vs Diabetes is just luck by chance, there is no direct relevance between under the significance value.

### Outliers detection and removal for sample

We have utilized the box-plot to visualize the outlier of non-binary columns i.e., BMI, Education, MenHlths, Income. Removal of these outlier was performed by utilizing the inter-quartile-range (IQR) to remove the outliers’ samples from the datasets. This may change our dataset percentage of Diabetes Vs Non-Diabetes little off balance, but It was the necessary to get the highest accuracy.

### Feature Selection by Feature Importance

In our dataset, we have 21 health markers is not directly connect with diabetes, but still they have relation which can be seen in heatmap plot of the datasets. For reducing the dimension of the dataset for better prediction accuracy. We have tried many methods i.e. PCA/ Random Forest Feature Elimination / Extra Tree Classifier Feature Elimination / Recursive Feature Elimination. In most of the methods, we need to specify the dimension which we want to reduce to find the appropriate feature for building the model which can improve the classifier performance.

As the features selection is applied to select the prominent features to build ML model. We have selected five Supervised ML model for the experiment.

## Logisitic Regression

Logistic Regression is a statistical Machine Learning algorithm that is used for classification problems. It is based on the concept of probability. It is used when the dependent variable (target) is categorical. It is widely used when the classification problem at hand is binary; true or false, yes or no, etc. Logistics regression uses the sigmoid function to return the probability of a label.

## Gaussian Naïve Bayes

Naïve Bayes classifier used the Bayes Theorem for prediction of samples. The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

Picture 5

where P(c|x) is the Posterior Probability of the response (target) variable given the training data inputs, P(c) is Prior probability of the class (target), P(x|c) is Probability of the predictor (x) given the class/target (c), and P(x) is Prior probability of the predictor (x).

## Decision Tree Classifier

## Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Decision trees can perform both classification and regression tasks. Decision is kind of an umbrella term, the intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the dataset until you isolate all data points belonging to each class. With this process you’re organizing the data in a tree structure. Every time you ask a question, you’re adding a node to the tree. And the first node is called the root node. The result of asking a question splits the dataset based on the value of a feature, and creates new nodes. If you decide to stop the process after a split, the last nodes created are called leaf nodes.

## Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.

## XGBoost Classifier

XGBoost is an ensemble learning algorithm meaning that it combines the results of many models, called base learners to make a prediction. Just like in Random Forests, XGBoost uses Decision Trees as base learners. XGBoost Library is parallelizable which means the core algorithm can run on clusters of GPUs or even across a network of computers. This makes it feasible to solve ML tasks by training on hundreds of millions of training examples with high performance. Due to Its speed and performance are unparalleled and it consistently outperforms any other algorithms aimed at supervised learning tasks.

# Application

## Libraries:

## **Numpy:**

Numpy is a library for python programming language. Numpy is developed by Travis Oliphant in 2006[10]. It is a fundamental package for computing, and it contains among other things

* a powerful N-dimensional array object
* sophisticated (broadcasting) functions
* tools for integrating C/C++ and Fortran code
* useful linear algebra, Fourier transform, and

random number capabilities

**Matplotlib:**

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MATAB.

**Scikit-learn:**

Scikit-Learn is a Python module for machine learning built on top of SciPy. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

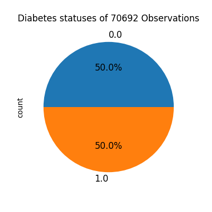
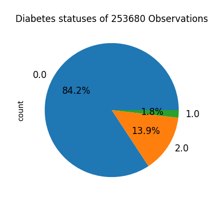
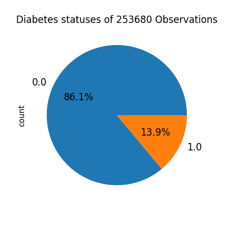
## Development Enviornment

## **Jupyter Notebook**

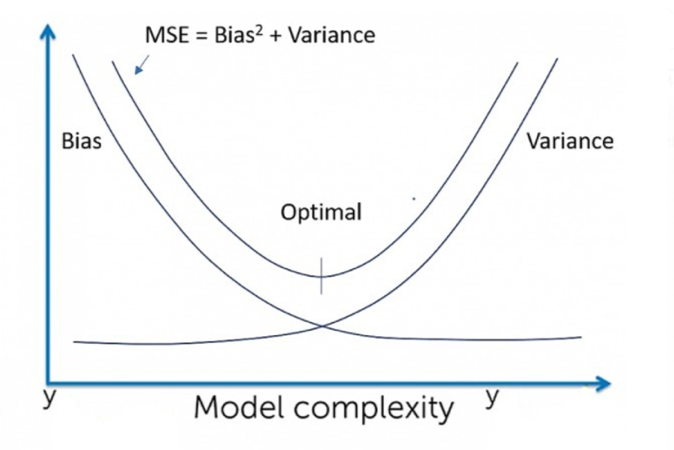
Project Jupyter is a project to develop open-source software, open standards, and services for interactive computing across multiple programming languages. It was spun off from IPython in 2014 by Fernando Pérez and Brian Granger. Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis.

# Results and Discussion

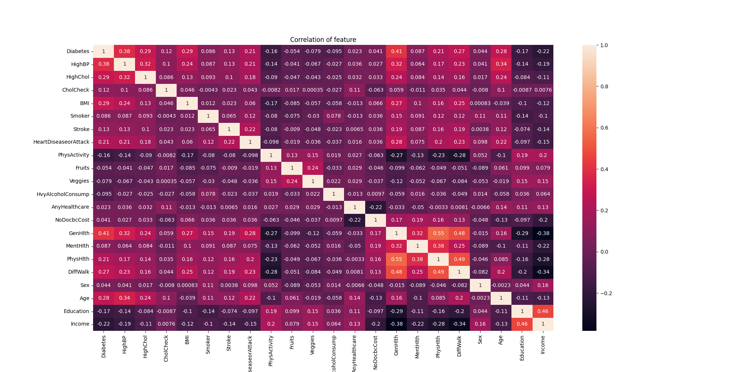
In this project, we have selected the sample by checking which has two classes and balanced dataset. We can see the figure 1 for each sample classes pie distribution. We have selected the third file, which has 50-50% sample for both the classes for diabetes/non-diabetes. By choosing this sample, we made the ML model which is not biased. We need to find balance via building the model between bias vs variance as model should find balance between optimal value of it.



*Figure 1: Three samples class distribution*

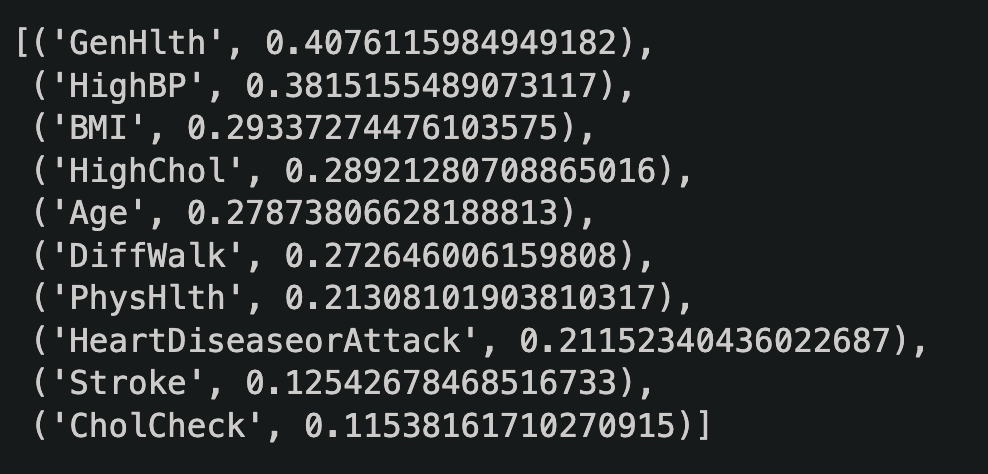


*Figure 2: Bias Vs Variance for building ML model*

**

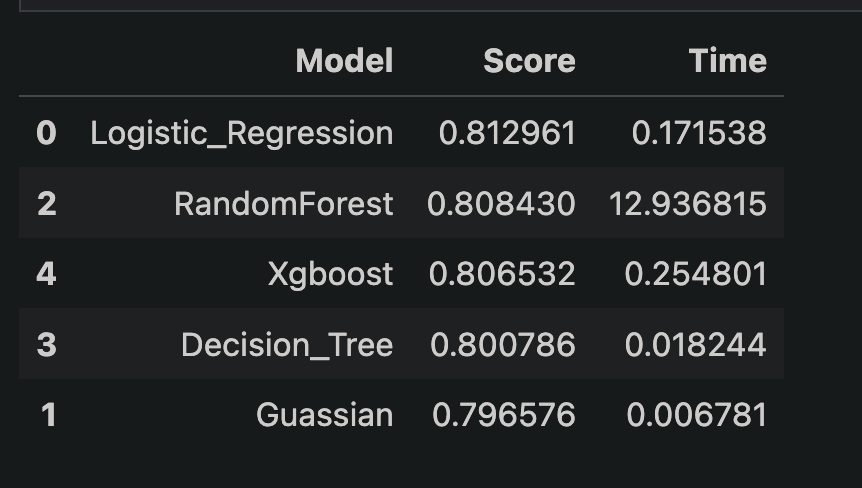
*Figure 3: Correlation matrix for each health marker*

We have explored each health indicator to understand how well, it is correlated to a patient become diabetic. we plotted the heatmap plot for correlation matrix in figure 3. From that plot. We have selected the prominent health indicator by which can build the ML model.



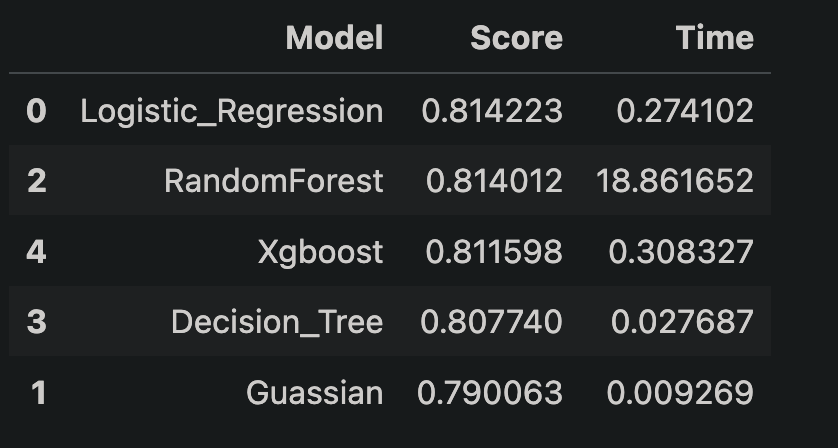
*Figure 4: Each marker correlated value w.r.to. Diabetes*

In this project, we have performed two experiments. In Experiment 1, I have removed the outliers from the data and build model on the remaining samples and compute the accuracy for the selected five classifiers i.e., Logistic Regression, Gaussian Naïve Bayes, Random Forest classifier, Decision Tree classifier, and XG Boost classifier. We were able to reach the performance of each model w.r.to time taken into training the model.



*Figure 5: Experiment 1 results*

In Experiment 2, I have didn’t remove the outlier as it may contain some information which may impact our accuracy. We built the model on the all samples and compute the accuracy for the selected five classifiers i.e., Logistic Regression, Gaussian Naïve Bayes, Random Forest classifier, Decision Tree classifier, and XG Boost classifier. We were able to reach the performance of each model w.r.to time taken into training the model.

**

*Figure 6: Experiment 2 results*

From these two experiment, we can observe that for experiment 2 has little higher accuracy w.r.to model accuracy (score) and training time is also least among the top three classifier. The models developed had good discriminatory power, accuracy ranging from 79% to 81%. These findings demonstrate the potential of machine learning algorithms in predicting the risk of diabetes and improving early detection and prevention strategies. As you observe, the Logistic Regression Classifier has similar accuracy as Random Forest and XGBoost classifier. But if you can observe, that it has the minimum converse time comparison to other two classifier. It is 10x time faster than XGBoost classifier and 39x faster than Random Forest Kernel where number of Decision tree taken in Random Forest was 100 trees.

# Conclusion and Future Work

In this project, we have explored the health marker which impacts a human life to trigger his system to become a Diabetes Person. We have developed diabetes prediction models using machine learning kernels and different dataset cases. The studies found that Age, BMI, family history of diabetes, physical activity, smoking status, waist circumference, systolic blood pressure, fasting plasma glucose, glycated hemoglobin (HbA1c), triglycerides, and total cholesterol were significant predictors of diabetes.

The very high correlated health marker towards a person to become Diabetes are CholCheck, Stroke, HeartDiseaseorAttack, PhyHlth, Diff-Walk, Age, HighCol,BMI, HighBP,and GenHlth. A person to keep maintaining health life balance and keep checking these parameter to stay non-diabetic.

##### References

[1] Zhang, Liying, et al. "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study." Scientific reports 10.1 (2020): 4406

[2] Laila, Umm E., et al. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." Sensors 22.14 (2022): 5247.

[3] Shin, Juyoung, et al. "Development of various diabetes prediction models using machine learning techniques." Diabetes & Metabolism Journal 46.4 (2022): 650-657.

[4] Rani, KM Jyoti. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305

[5] Schneiders, Josiane, et al. "Quality indicators in type 2 diabetes patient care: analysis per care-complexity level." Diabetology & metabolic syndrome 11 (2019): 1-9.

[6] Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

[7] Website : [TowardDataScience](https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2) <https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2>

[8] Kaggle Dataset link : <https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2>

##### Appendix

**Demographics Information about Dataset**

|  |  |
| --- | --- |
| **Characteristics** | **Values** |
| **Age** Mean + Std (Range)  Divided into 13 age categories  1=18-24, 2=25-29, 3=30-34,4=35-39,5=40-44,6=45-49,7=50-54,8=55-59,9=60-64,10=65-69,11=70-74,12=75-79,13=80 or older | 8.58+2.85  (1-13) |
| **AnyHealthcare**Binary (1-0)  1 – Yes  0 No/ Don’t know/Not sure/Refused | 67508 (95.49%)  3184 (4.50%) |
| **BMI** Mean + Std  (Range) | 29.86 + 7.11  (12-98) |
| **CholCheck**  Binary (1-0)  1 – Yes, Cholesterol check in 5 years  0 – No Cholesterol check in 5 years | 68943 ( 97.52%)  1749 (2.47%) |
| **Diabetes**  Binary (1-0)  1 – Yes, Diabetes  0 – No, Non-Diabetes | 35346 (50%)  35346 (50%) |
| **DiffWalk**  Binary (1-0) :difficulty walking or climbing stairs ?  1 – Yes,  0 – No, | 35346 (74.72%)  35346 (25.27%) |
| **Education**  1 – Never attended School  2 – Grades 1 through 8 (Elementary)  3 – Grades 9 to 11 (some high school)  4 – Grades 12 or GED ( High School graduate)  5 – College 1 year to 3 years (some college )  6 – College 4 year or more (college graduate) | 75  (0.10%)  1647 ( 2.32%)  3447 (4.87%)  19473 ( 27.54%)  20030 ( 28.33%)  26020 ( 36.80%) |
| **Fruits**  Binary (1-0)  1 – Yes, Take Fruits regularly  0 – No, Didn’t eat fruits regularly | 27443 (38.82%)  43249 (61.17%) |
| **GenHealth**  1 – Excellent  2 – Very good  3 – Good  4 – Fair 12 or GED ( High School graduate)  5 – Poor / Not sure/ Refused | 8282 (11.71%)  19872 ( 28.11%)  23427 (33.13%)  13303 ( 18.81%)  5808 ( 8.21%) |
| **HeartDiseaseorAttack**  Binary (1-0)  1 – Yes,  0 – No, | 10449 (14.78%)  60243 (85.21%) |
| **HighBP**  Binary (1-0)  1 – Yes,  0 – No, | 39832 (56.34%)  30860 (43.65%) |
| **HighChol**  Binary (1-0)  1 – Yes,  0 – No, | 37163 (52.57%)  33529 (48.42%) |
| HvyAlcholConsump  Binary (1-0)  1 – Yes,  0 – No, | 3020 (4.27%)  67672 (95.72%) |
| **Income**  1 – Less than $10, 000  (0 to 10000)  2 – Less than $15,000   (10001 to 15000)  3 – Less than $20, 000  (15001 to 20000)  4 – Less than $25,000   (20001 to 25000)  5 – Less than $35, 000  (25001 to 35000)  6 – Less than $50,000   (35001 to 50000)  7 – Less than $75,000   (50000 to 75000)  8 – more than $75,000   (75000 to more) | 3611  (5.10%)  4498 ( 6.36%)  5557 (7.86%)  6658 (9.41%)  8010 ( 11.33%)  10287 ( 14.55%)  11425 (16.16%)  20646 (29.20%) |
| **MentHlth**  Mean + Std (Range)  Number of day of poor mental health scale | 3.75+8.16 (0-30) |
| **PhysActivity**  Binary (1-0)  1 – Yes,  0 – No, | 49699 (70.30%)  20993 (29.63%) |
| **NoDocbcCost**  Binary (1-0)  1 – Yes,  0 – No, | 6639 (9.39%)  64053 (90.60%) |
| **Sex**  Binary (1-0)  1 – Yes,  0 – No, | 32306 (45.69%)  38386 (54.30%) |
| **PhysHlth**  Mean + Std (Range)  Number of  day of poor mental health scale | 5.81 + 10.06  (0-30) |