Diabetes Prediction Using 2015 BRFSS Health Survey Data

# CS 559 Machine Learning Course Project

# Proposal

# Shreyoshi Ghosh

## Background

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. According to the CDC, the objective of the BRFSS is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases such as diabetes. The goal of this research project is to attempt to use data from the BRFSS survey to build a set of models to classify survey responders as diabetic or not.

## Goals

The 2015 BRFSS survey data contains enough information to build at least a binary classifier (with reasonable precision/recall), if not also a multiclass classifier. For binary classification, my goal is to build a classifier to distinguish diabetics from non-diabetics. For multiclass classification, my goal is to build a classifier to distinguish between diabetics, pre-diabetics, and non-diabetics.

## Data Source

The original BRFSS data can be found here, along with the data codebook: <https://www.cdc.gov/brfss/annual_data/annual_2015.html>

The data was then turned into a csv here: <https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system>.

I am taking advantage of two datasets in Kaggle that further narrow down the data to include only features that are relevant to diabetes detection. One of them is suitable for a binary classifier and can be found here: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv>

The other one can be used for multiclass classification and can be found here: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_012_health_indicators_BRFSS2015.csv>

## Dataset Details

Number of samples: 253680

Output Variable/Class : **Diabetes\_012 (0 = no diabetes 1 = prediabetes 2 = diabetes)**

Number of input variables: 22

|  |  |  |
| --- | --- | --- |
| Input Variable | Description | Type |
| HighBP | 0 = no high BP 1 = high BP |  |
| HighChol | 0 = no high cholesterol 1 = high cholesterol |  |
| CholCheck | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years |  |
| BMI | Body Mass Index |  |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes |  |
| Stroke | (Ever told) you had a stroke. 0 = no 1 = yes |  |
| HeartDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |  |
| PhysActivity | physical activity in past 30 days - not including job 0 = no 1 = yes |  |
| Fruits | Consume Fruit 1 or more times per day 0 = no 1 = yes |  |
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes |  |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes |  |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes |  |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes |  |
| GenHlth | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor |  |
| 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? scale 1-30 days |  |
| 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? scale 1-30 days |  |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes |  |
| Sex | 0 = female 1 = male |  |
| Age | 3-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older |  |
| Education | Education level (EDUCA see codebook) scale 1-6 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) |  |
| Income | Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more |  |

## Methodology

I am outlining the standard steps that I plan to take. The specific details and the outcomes of each step are to be determined once the experiment has been completed.

1. Exploratory data analysis
   1. Imputation
   2. Duplicate removal
   3. Correlation analysis
2. Feature selection based on statistical tests
3. Data preparation
   1. Scaling, if needed
   2. One hot encoding, if needed
   3. Splitting
4. Build and compare baseline models with hyperparameter tuning + cross validation
   1. random forest
   2. gradient boosted trees
   3. Neural Network
5. Model regularization as needed
6. Depending on model performance, either pick one model or create an ensemble, like a voting classifier
7. Repeat process for binary + multiclass classifiers

## Expected Result

1. Binary Classifier (diabetic vs. non-diabetic)

2. Multiclass Classifier (diabetic vs. pre-diabetic vs. non-diabetic)

3. Performance comparison of models in each of the above scenarios

## Evaluation

Since the dataset imbalance results may lead to misleading accuracy scores, my goal is to evaluate model performance using the F1 score metric (combining precision and recall) as well as the standalone precision and recall scores. I will also use the AUC for the binary classifiers as another metric to measure with.

## Potential Issues/Solutions

The dataset is highly imbalanced, with significantly higher numbers of negative cases (non-diabetic). This can lead to lower performance. However, as a part of the project, I am going to explore methods of combatting imbalance through mechanisms such as data resampling/SMOTE.

I am also relying on the Kaggle authors’ choice of features for diabetes prediction as taken from the original CDC dataset. There may be additional indicators in the original set that may be relevant to prediction that were left out from the subset. Time permitting, I will attempt to explore the original BRFSS codebook + dataset to determine if any relevant indicators were left out.