Project Proposal Title:

Diabetes Prediction Using 2015 BRFSS Health Survey Data

Group members:

Shreyoshi Ghosh

Topic and Research Question:

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 the provided 2015 BRFSS data to build a model to predict/classify diabetes.

Hypothesis:

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

Data sources:

The original BRFSS data can be found here, along with the 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 narrows 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>

Methods:

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 analysis 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
5. Depending on model performance, either pick one model or create an ensemble, like a voting classifier
6. Repeat process for binary + multiclass classifiers

Expected Results:

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

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

Potential Problems and 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.

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.