Ying Hong (yh3538), Rohan Sheelvant (rns2167), Ardrian Wong (aaw2179), Liwen Zhu (lz2512)

Diabetes is a chronic health condition that affects how your body turns food into energy. With diabetes, patients don’t make enough insulin or can’t use it well, causing much blood sugar to stay in the bloodstream. This can lead to serious health problems like heart disease, vision loss, and kidney disease.

In 2018, the Centers for Disease Control and Prevention reported that 34.2 million Americans have diabetes and 88 million have prediabetes. What’s more, the CDC estimates that 80% of prediabetics are unaware of their risks. Therefore we plan to build models that would identify whether a person has prediabetes.

**About the Dataset**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey collected annually by the CDC. Our dataset comprises 253,680 survey responses to the CDC's BRFSS2015. The target variable *‘Diabetes’* has 2 classes. 0 is for no diabetes and 1 is for prediabetes. There is a class imbalance in this dataset. This dataset has 21 independent feature variables, including HighBP, HighChol, Smoker, BMI, Stroke, PhysActivity, HvyAlcoholConsump, Sex, Age, Income, etc.

**Machine Learning Models Proposed**

Since we are solving a classification problem, we are considering using the following algorithms and concepts:

● SMOTE

● Precision, Recall, F1 score

● Logistics Regression

● Decision Tree Classifier

● Random Forest

● SVM

● Neural Networks

**Reference**

<https://www.cdc.gov/diabetes/basics/diabetes.html>

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv>