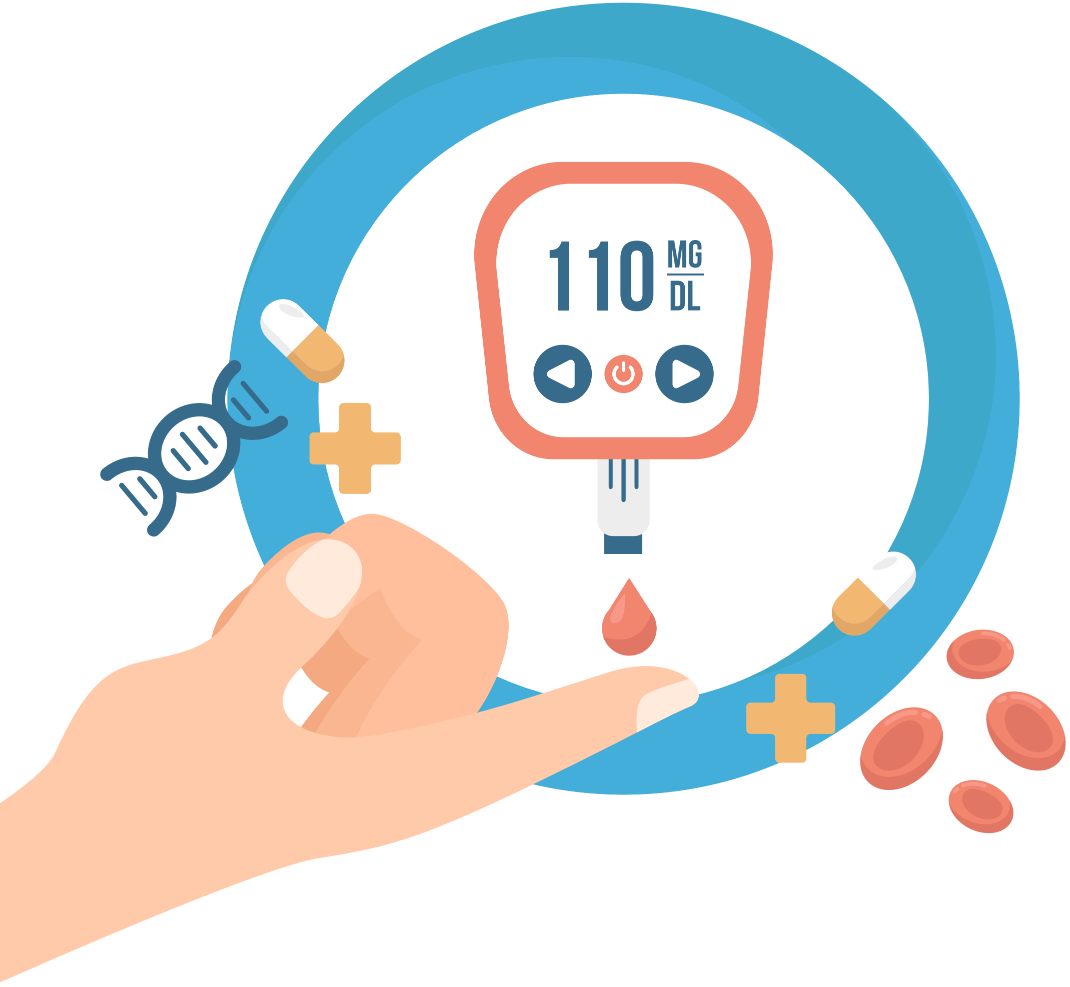


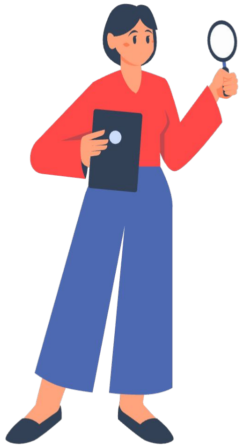
Predicting Diabetes: Data-Driven Insights

Yun Ma
Ziyin Zheng

Introduction

According to CDC, more than 11% of the US population suffer from diabetes

Approximately 1 in 5 in this population is undiagnosed



Why interested

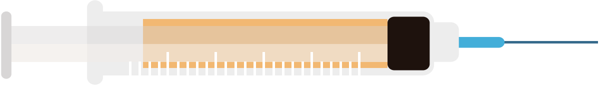
While Diabetes is hard to cure, it is reversible if discovered early

Our study aims to use various health and lifestyle indicators to distinguish patients that are health, diabetic, or with diabetes onset.

Our Research Question

- Can we **train and evaluate a model** that effectively distinguish patients that are **healthy, diabetic, and pre-diabetic**?

Diabetes Dataset



Data

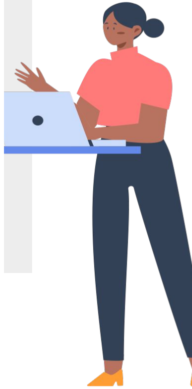
Original Source

Our data is collect by the **CDC** designed to understand the relationship between lifestyle and diabetes in the U.S.
Aim for **Public health patterns and risk behavior monitoring.**

Brief information

The data set surveyed 253680 individuals on 21 lifestyle related questions.
Each individual is categorized as either **diabetic, prediabetic, or healthy**

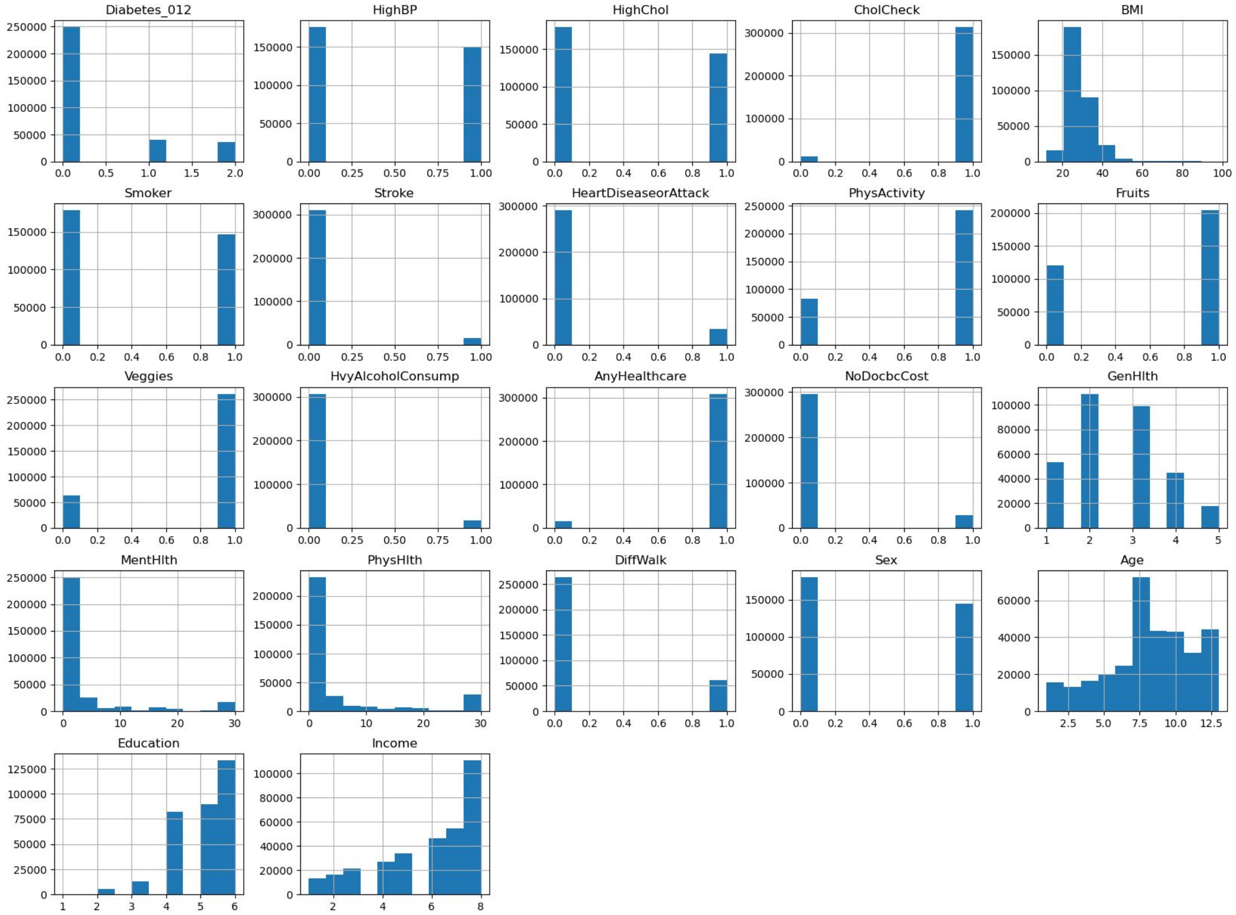
```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```



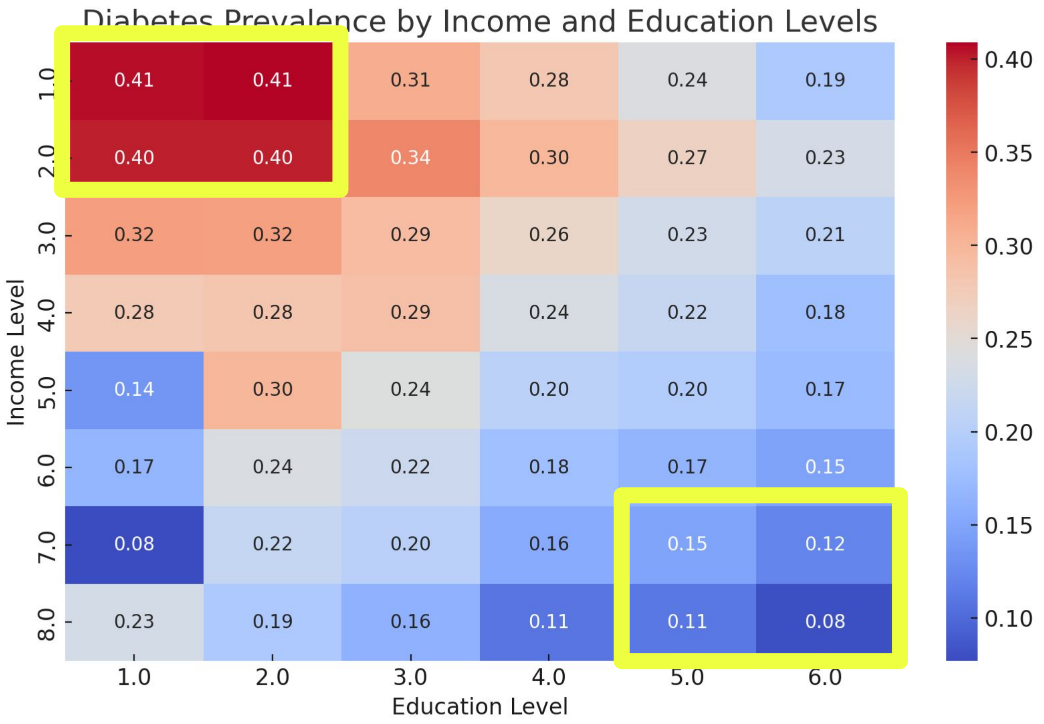
Data

Distribution of 21 variables
Can be self-reported

No missing data/ Pre-cleaned



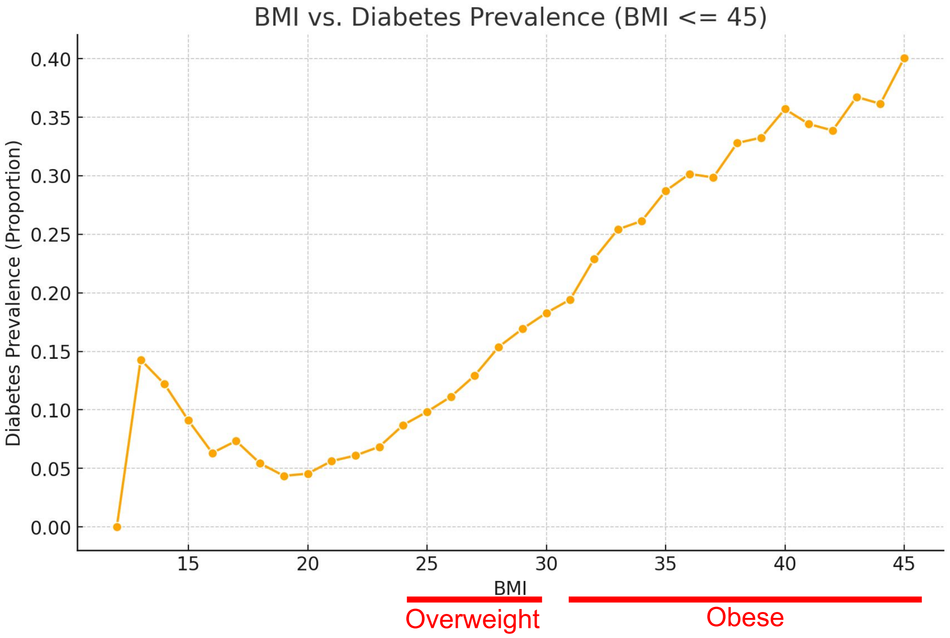
EDA – Diabetes vs Income + Education level

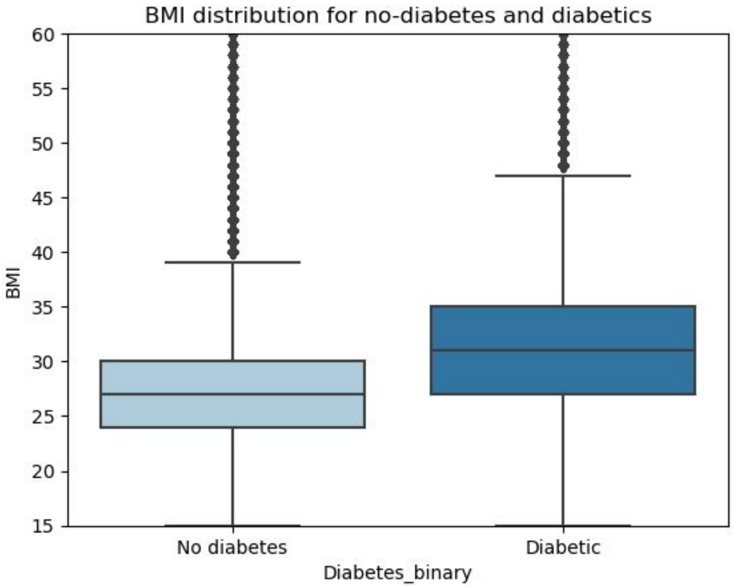


Education level
1 = Never school or only kindergarten
2 = Grades 1 through 8 (Elementary)
3 = Grades 9 through 11 (Some high school)
4 = Grade 12 or GED (High school graduate)
5 = College 1 year to 3 years (Some college or technical school)
6 = College 4 years or more (College graduate)

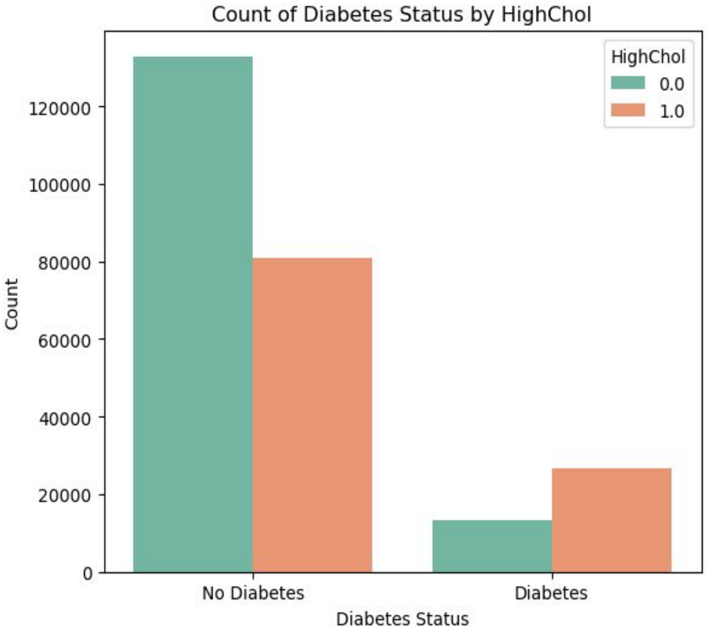
Income scale
1 = less than $10,000
5 = less than $35,000
8 = $75,000 or more

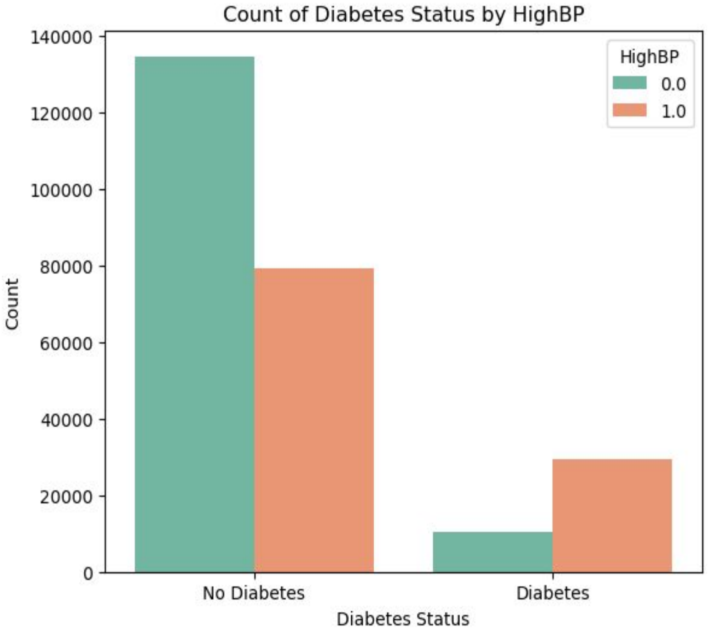
EDA –Diabetes vs BMI





EDA –Diabetes vs highBP and HighChol



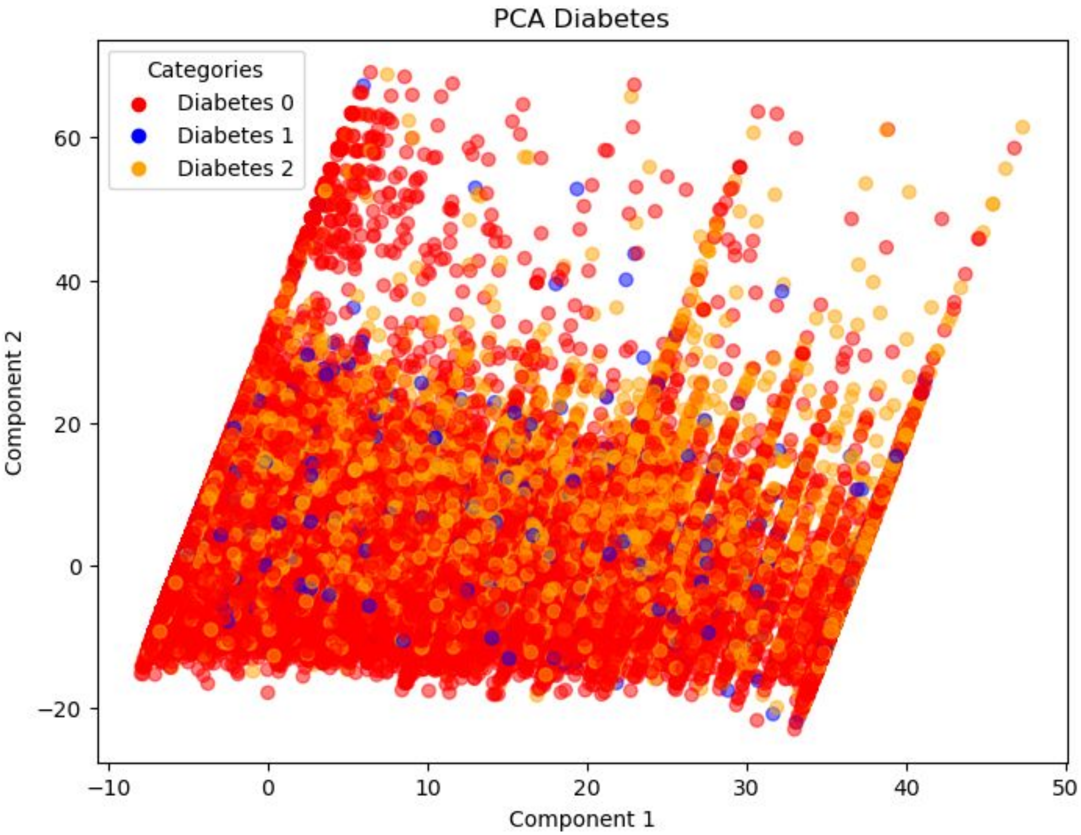


EDA– PCA

Challenges:

- The 3 classes are embedded in each other

- The class size are extremely imbalanced



Class Imbalance

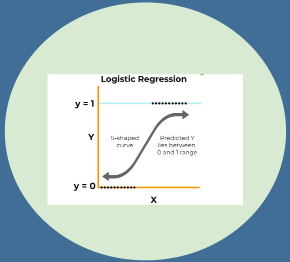
Original Class Size

	0	1	2
Original	213703	35346	4631
Train, Test (0.8:0.2)	(170908, 42795)	(28349, 6997)	(3687, 944)

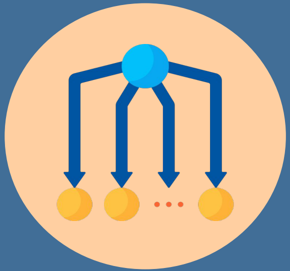
	SMOTE	NearMiss
Approach	Oversampling minority class	Downsampling majority class
Considerations	May amplify noise	May cause data lost
Training Size	(170908, 170908, 170908)	(3687, 3687, 3687)

Model selection

Random Forest

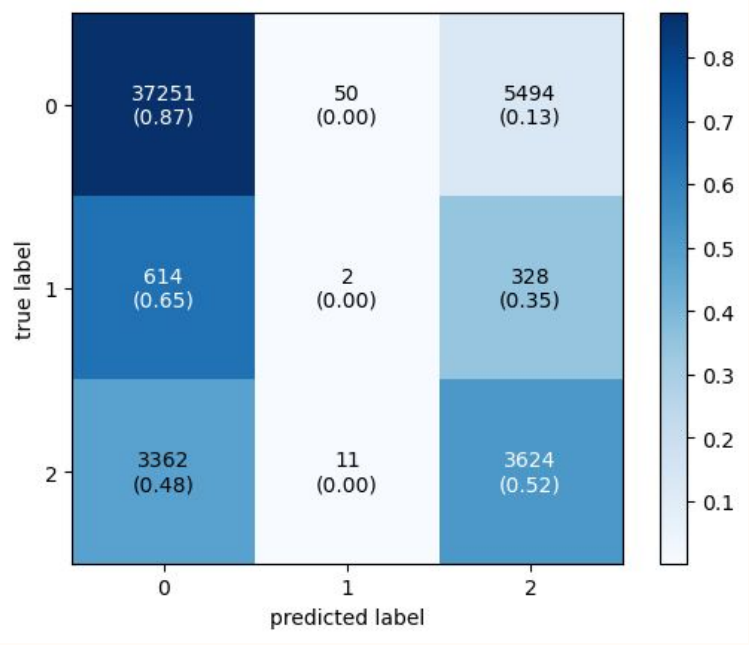


Logistic

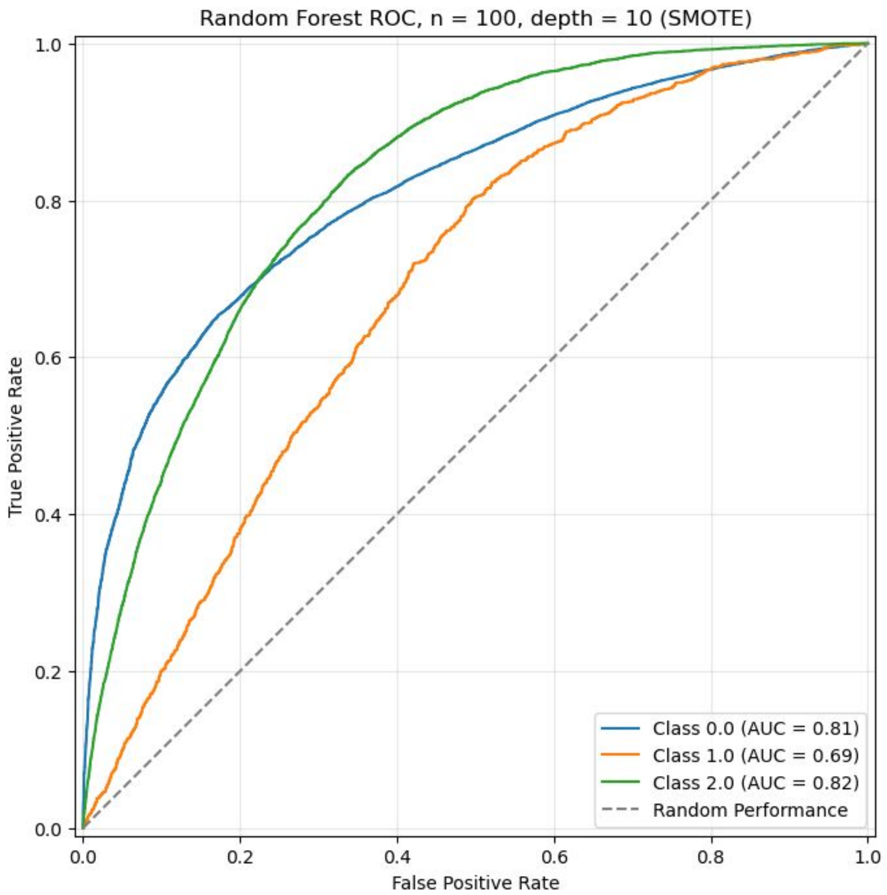


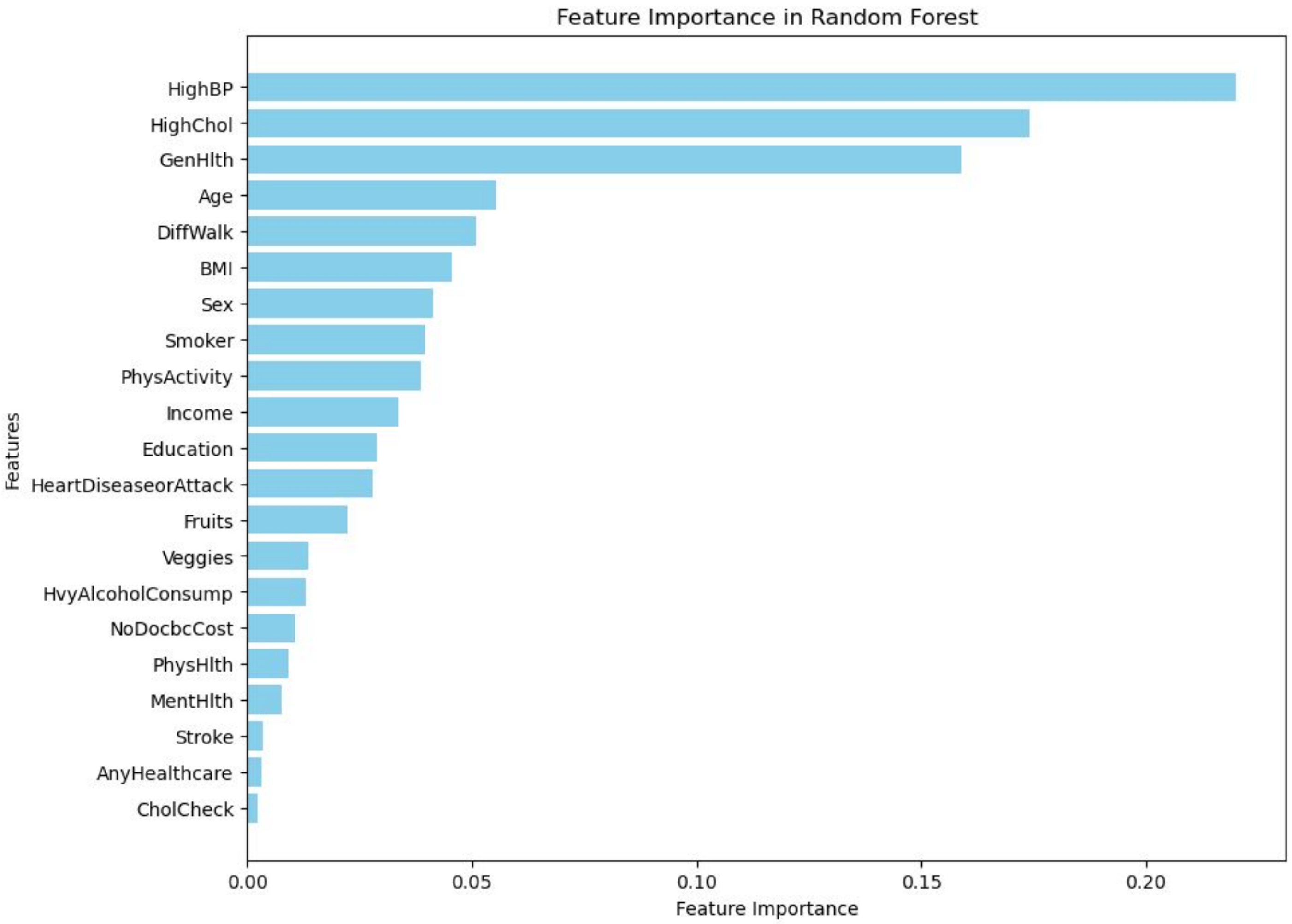
Naive Bayes

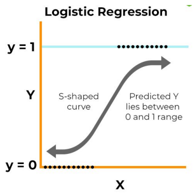
Random Forest



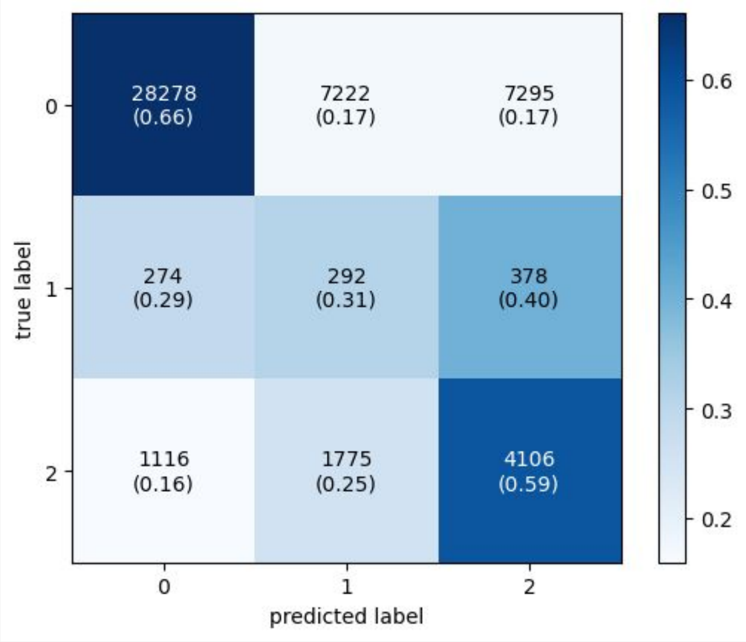
Training set score: 0.73
Test set score: 0.81




Feature Importance in Random Forest
HighBP
HighChol
GenHlth
Age
DiffWalk
BMI
Sex
Smoker
PhysActivity
Income
Education
HeartDiseaseorAttack
Fruits
Veggies
HvyAlcoholConsump
NoDocbcCost
PhysHlth
MentHlth
Stroke
AnyHealthcare
CholCheck
Features
0.00
0.05
0.10
0.15
0.20
Feature Importance

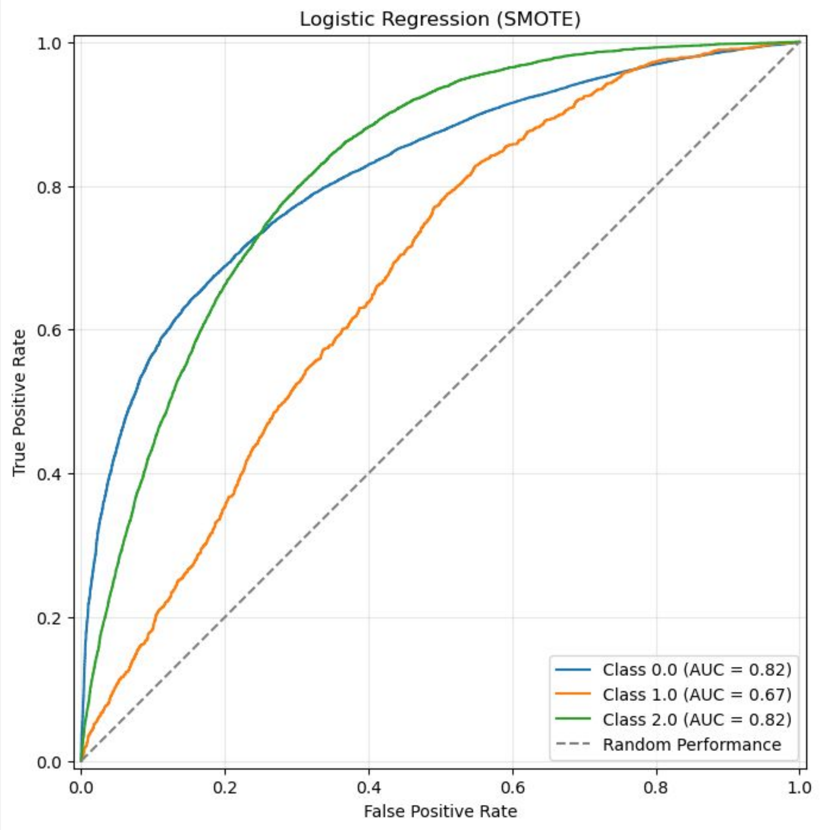


Logistic Regression



Train Accuracy: 0.53

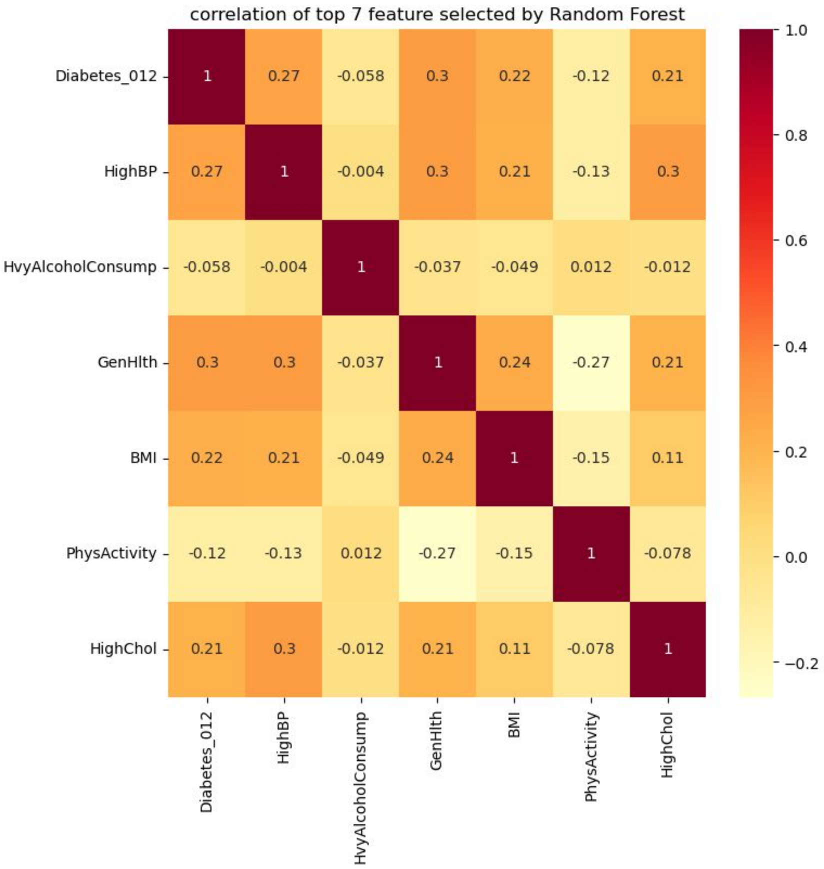
Test Accuracy: 0.64



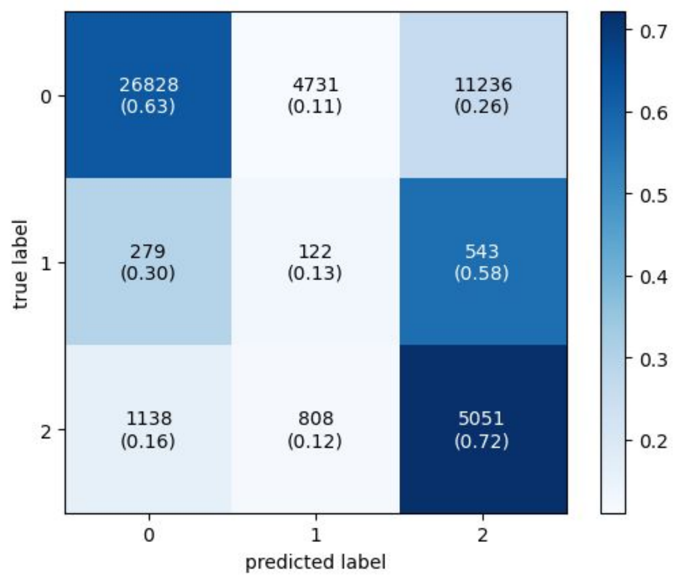
Naive Bayes: Motivation

The correlation between predictor variables seems to be low.

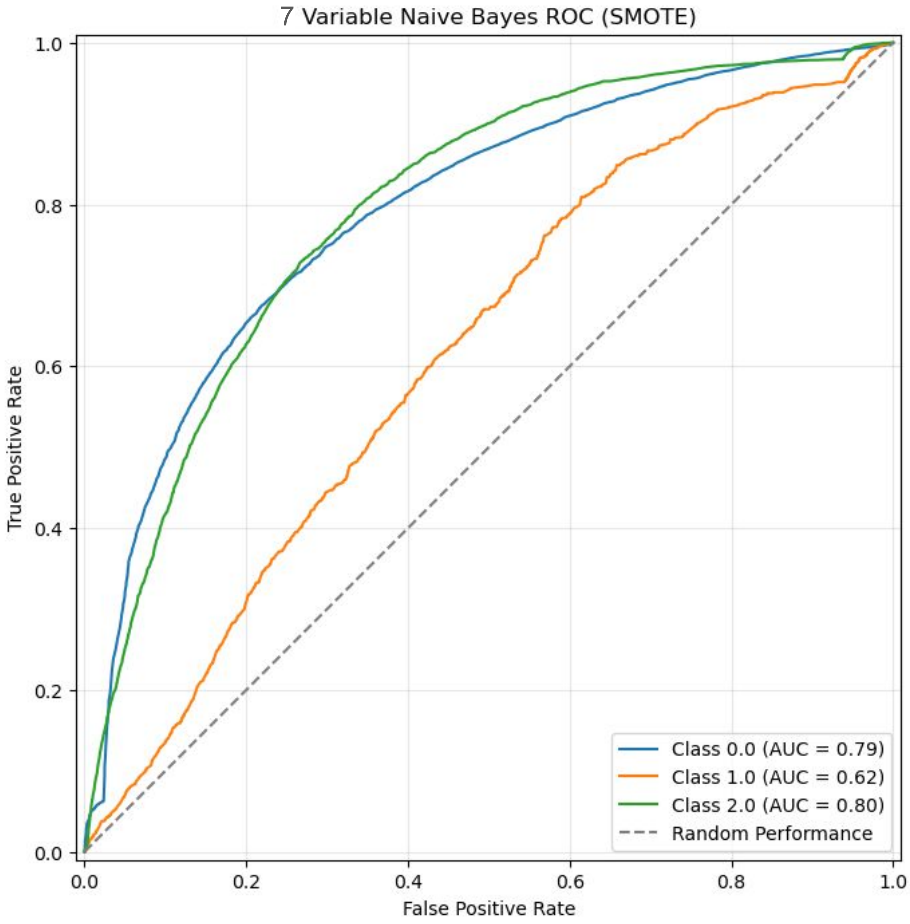
This indicates the assumption for Naive Bayes may be satisfied (independence)



Naive Bayes



Train Accuracy: 0.51
Test Accuracy: 0.63



Summary

- We solved the class imbalance issue with over-sampling the minority class with SMOTE
- Random Forest Model is most accurate, but it is heavily biased for making healthy prediction, which is unfavorable in-practice
- Naive Bayes and Logistic Regression have similar performance
    - Naive Bayes is better at predicting Diabetes
    - Logistic Regression is more balanced at making prediction
- PCA revealed better predictors may be needed to make more robust prediction

Thank You !

Q&A