Predicting Prediabetes Using Health Indicators

-Daniel Ellis Schwartz

Goal

Predict a person's risk of prediabetes using general health survey questions.

Based on responses to 21 general health questions, predict whether a person has:

- No diabetes
- Prediabetes
- Diabetes

with a focus on prediabetes.

Results Summary

- No acceptable model found
  - No models found with sufficiently high performance on target variable
  - Multiple attempts using several methods
- Data quality likely at fault
  - Data collection biases
  - Highly polluted data

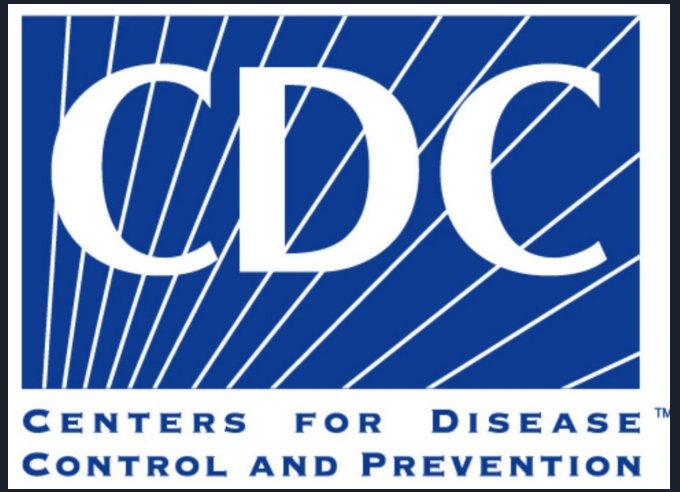
https://www.parkview.com/community/dashboard/could-you-have-prediabetes

Background

- Prediabetes comes before diabetes
- Can be reversed with lifestyle changes
- Early intervention is better

Data Overview

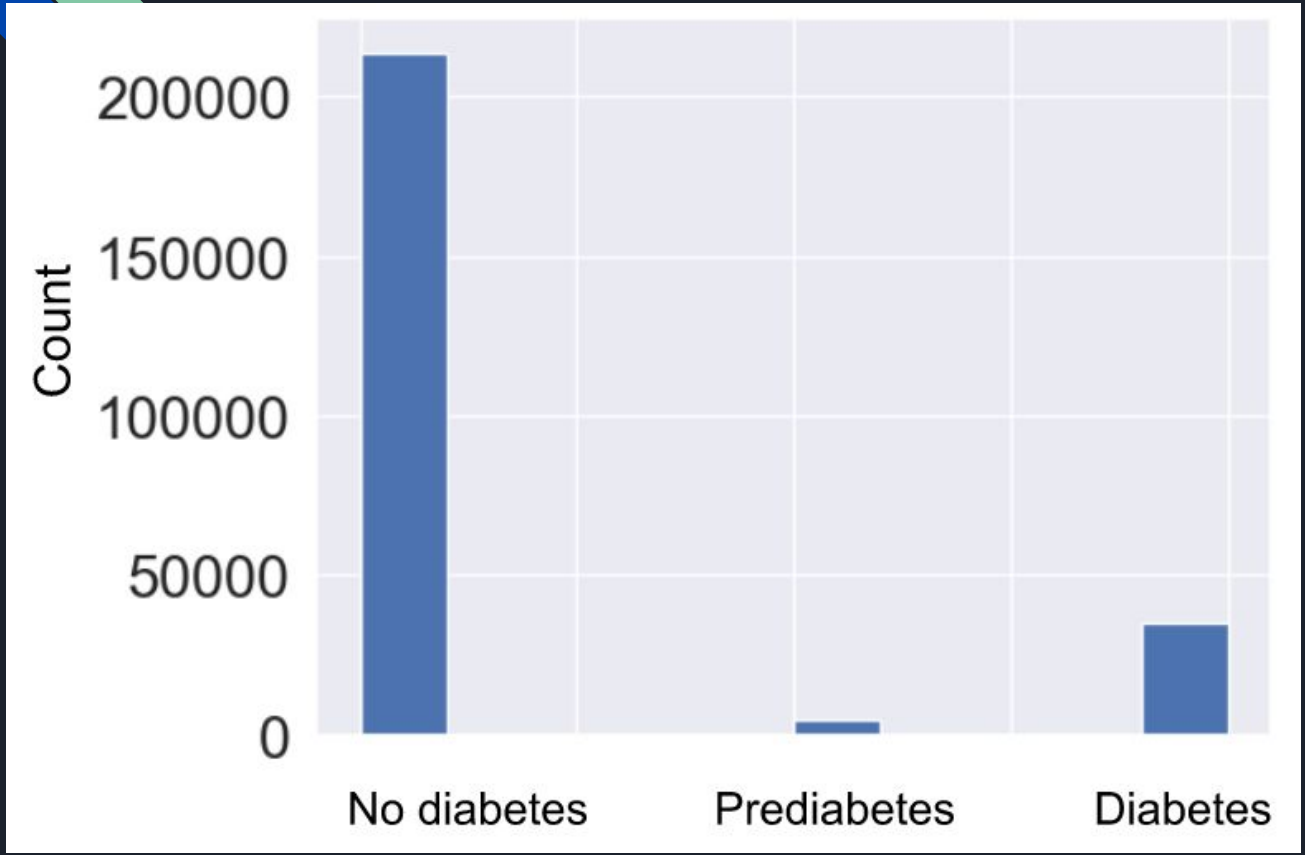
- Subset of general health survey selected by Kaggle user Alex Teboul
- Survey conducted by US Centers for Disease Control and Prevention
  - Between 2011 and 2015
- Subset contains responses from ~400,000 US Americans.
- Responses to 22 survey questions are included



Example Survey Responses

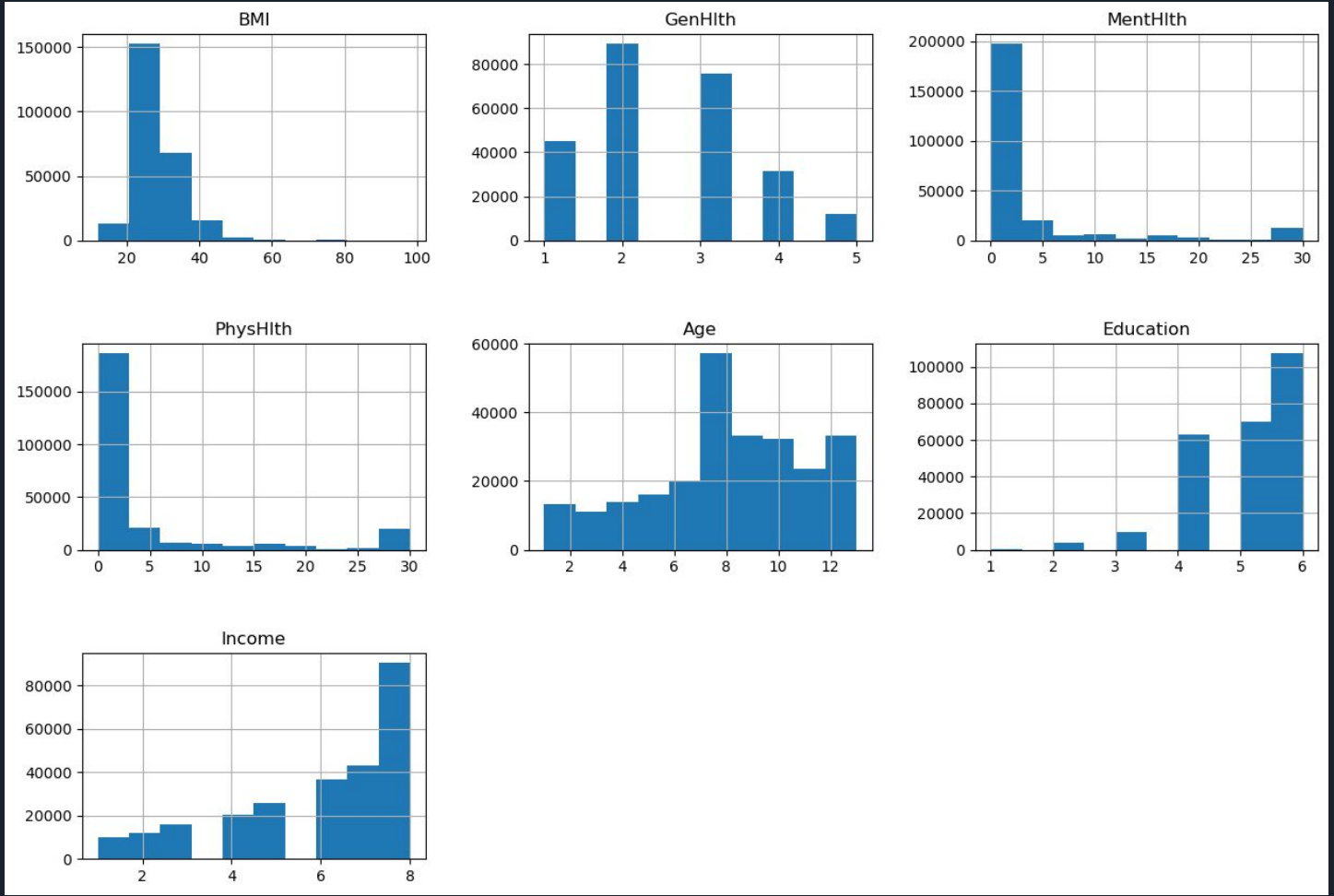
- Diabetes health status:
  - No Diabetes
  - Prediabetes
  - Diabetes
- Have you been diagnosed with high blood pressure? Yes or no.
- Has your cholesterol level been check in the last 5 years? Yes or no.
- Have you done any physical activity in the past 30 days, excluding your work? Yes or no.
- Would you say that in general your health is excellent, very good, good, fair, or poor?
- During the past 30 days how many days was your physical health not good?

Trends in the Data: Diabetes Diagnoses

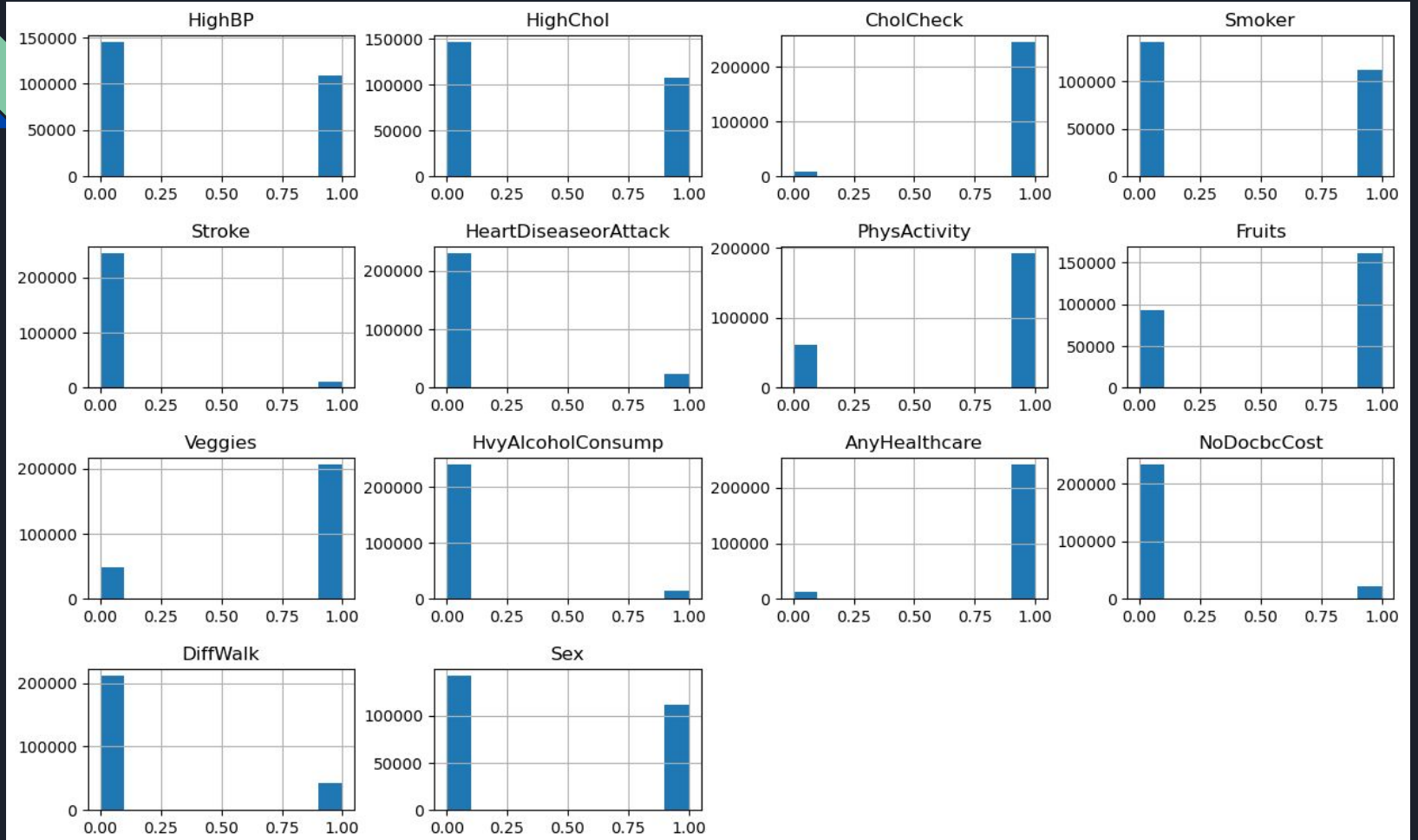


- No diabetes - 84%
- Prediabetes - 2%
- Diabetes - 14%

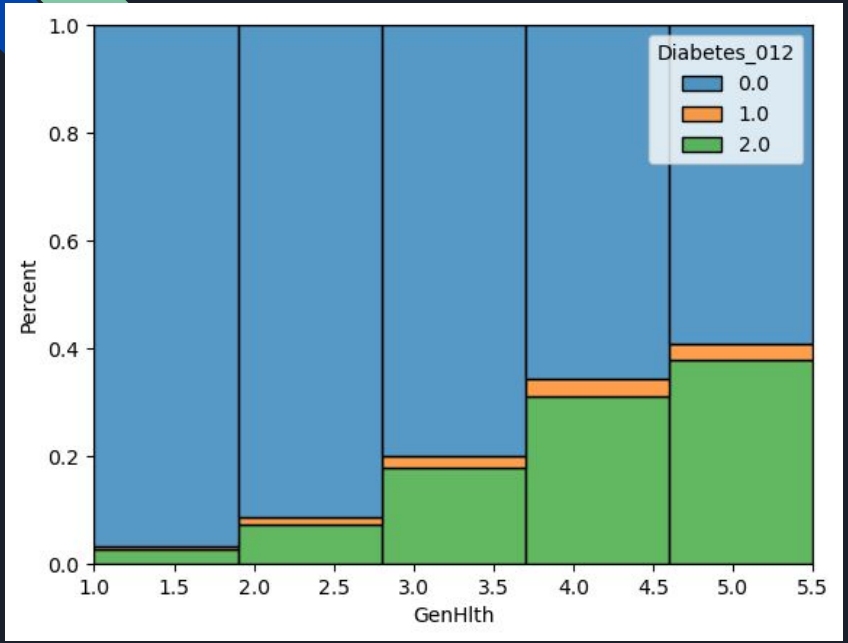
Trends in the Data: Non-Binary Variables

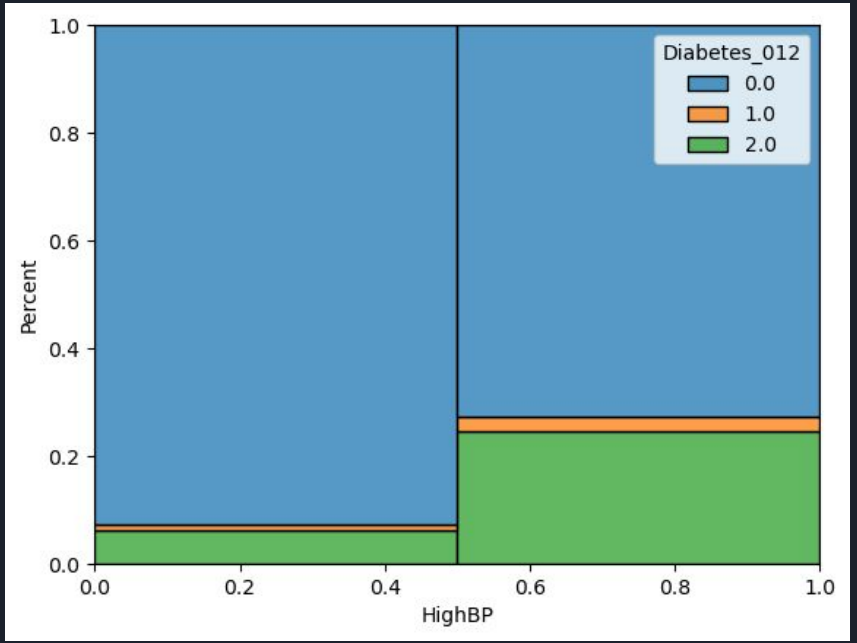


Trends in the Data: Binary Variables



Expected trends are present





Models

Model Type	Resampling Strategy	Precision	Recall
Random Forest	Undersample	No Diabetes: 0.95 Prediabetes: 0.03 Diabetes: 0.35	No Diabetes: 0.62 **Prediabetes: 0.3** Diabetes: 0.59
Random Forest	Undersample & oversample	No Diabetes: 0.95 Prediabetes: 0.02 Diabetes: 0.30	No Diabetes: 0.71 **Prediabetes: 0.01** Diabetes: 0.79
Histogram Gradient Boosting, One vs Rest	Undersample	No Diabetes: 0.96 Prediabetes: 0.03 Diabetes: 0.34	No Diabetes: 0.63 **Prediabetes: 0.31** Diabetes: 0.60
Histogram Gradient Boosting, One vs Rest	Undersample	No Diabetes: 0.95 Prediabetes: 0.03 Diabetes: 0.34	No Diabetes: 0.62 **Prediabetes: 0.30** Diabetes: 0.60

Conclusions & Recommendations

- Unable to make a predictive model for prediabetes using this data
  - Data quality likely at fault

- To predict prediabetes either:
  - Predict diabetes in undiagnosed people or
  - Use better data