Predicting Diabetes

CDC Behavioral Risk Factor Surveillance System (2022)

World Health Organization

2 million

estimated deaths due to diabetes in 2019

Risks

blindness, kidney failure, heart attacks, stroke, lower limb amputation

95%

of people with diabetes have type 2 diabetes

DIABETES

A US REPORT CARD

DIABETES

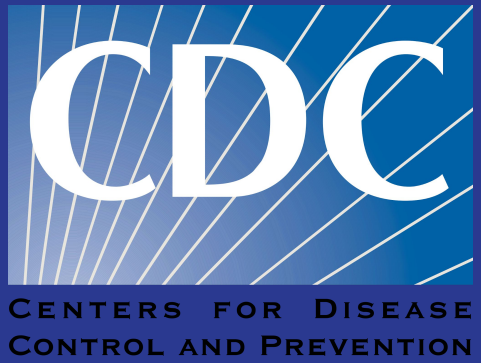
38 Million

About 38 million people **have diabetes**

That's about **1 in every 10** people

1 in 5 people **don't know they have it**

Data Understanding



CDC Behavioral Risk Factor Surveillance System (2022)

300+ columns (smoking habits, cancer diagnoses, demographics)

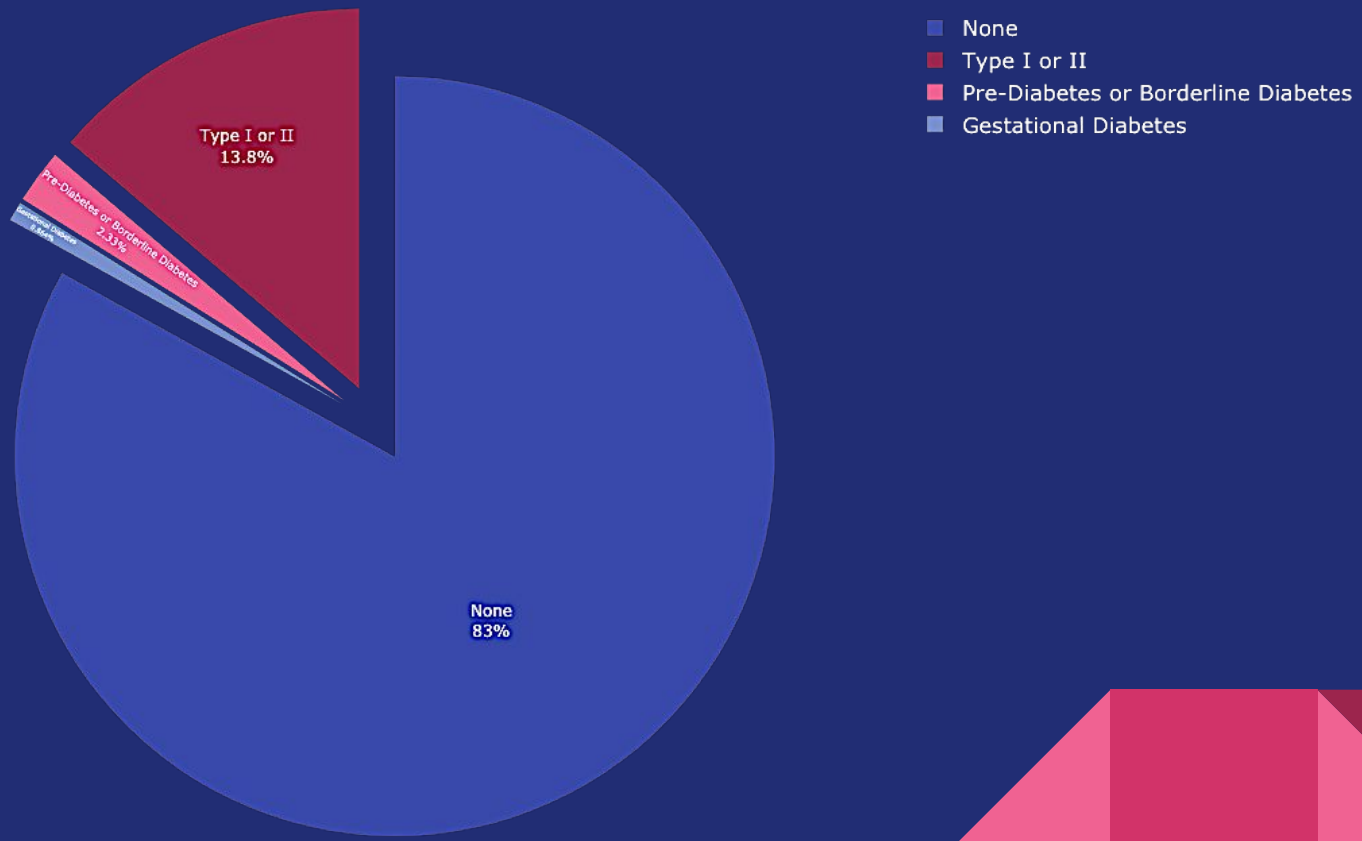
445,132 rows

Survey questions transformed to numeric values

RECALL

FALSE NEGATIVES

Types of Diabetes Diagnosis



Risk Factors

Weight, Exercise, & Activity

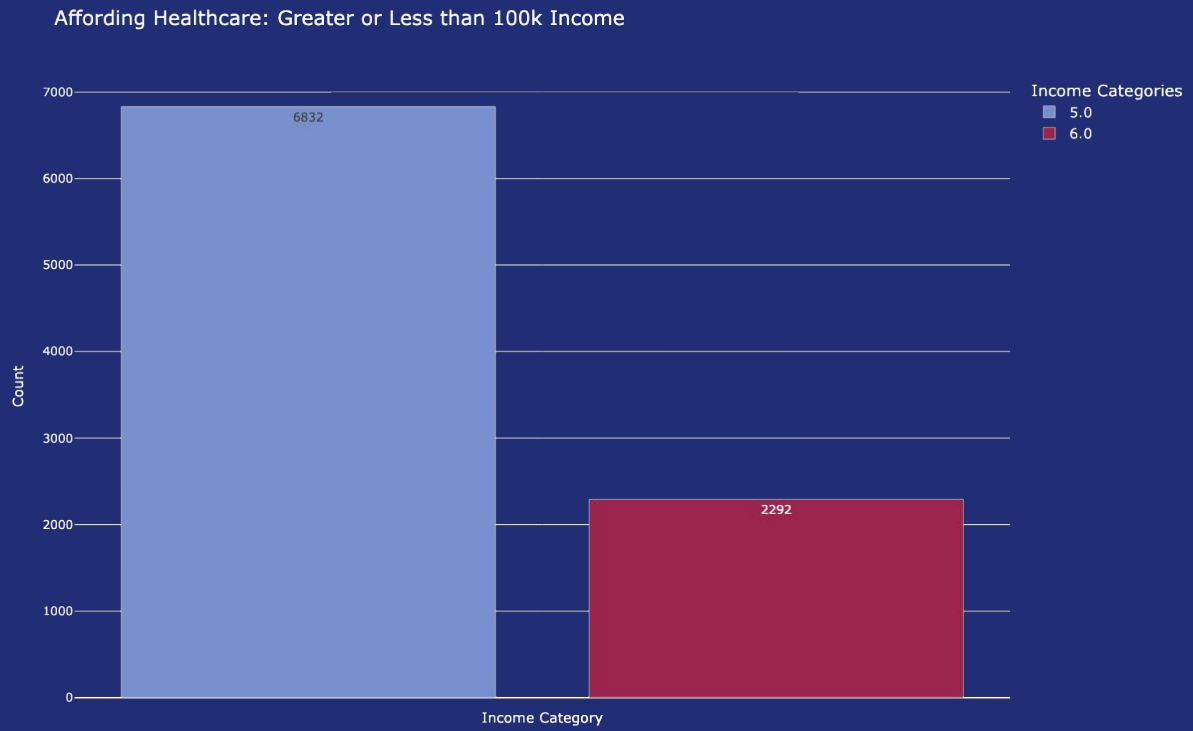
Family History: *parent or sibling*

Age: *risk increases with age*

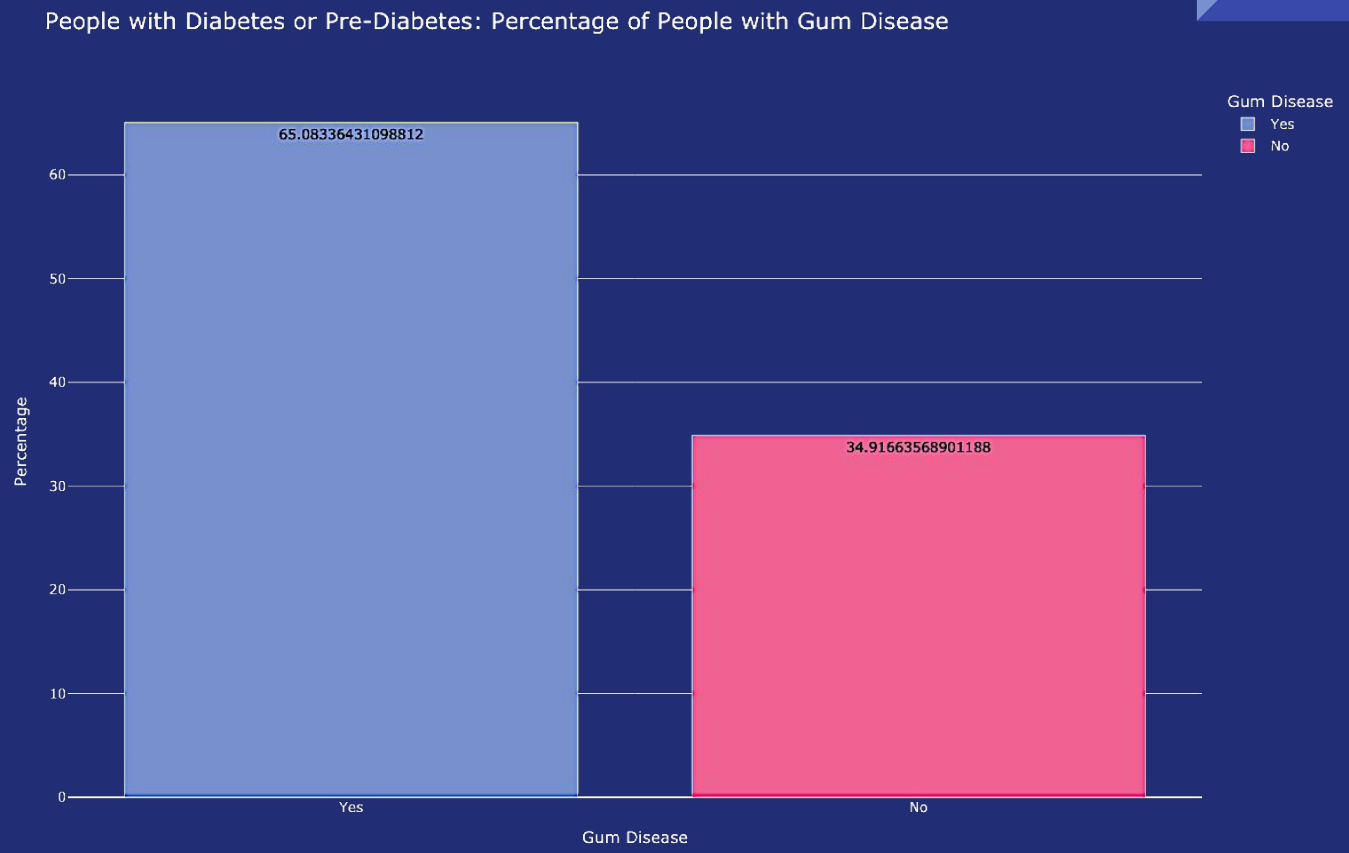
Prediabetes Diagnosis

INCOME_100K

Time in the past year when you needed to see a doctor but could not because you could not afford it?



GUM_DISEASE

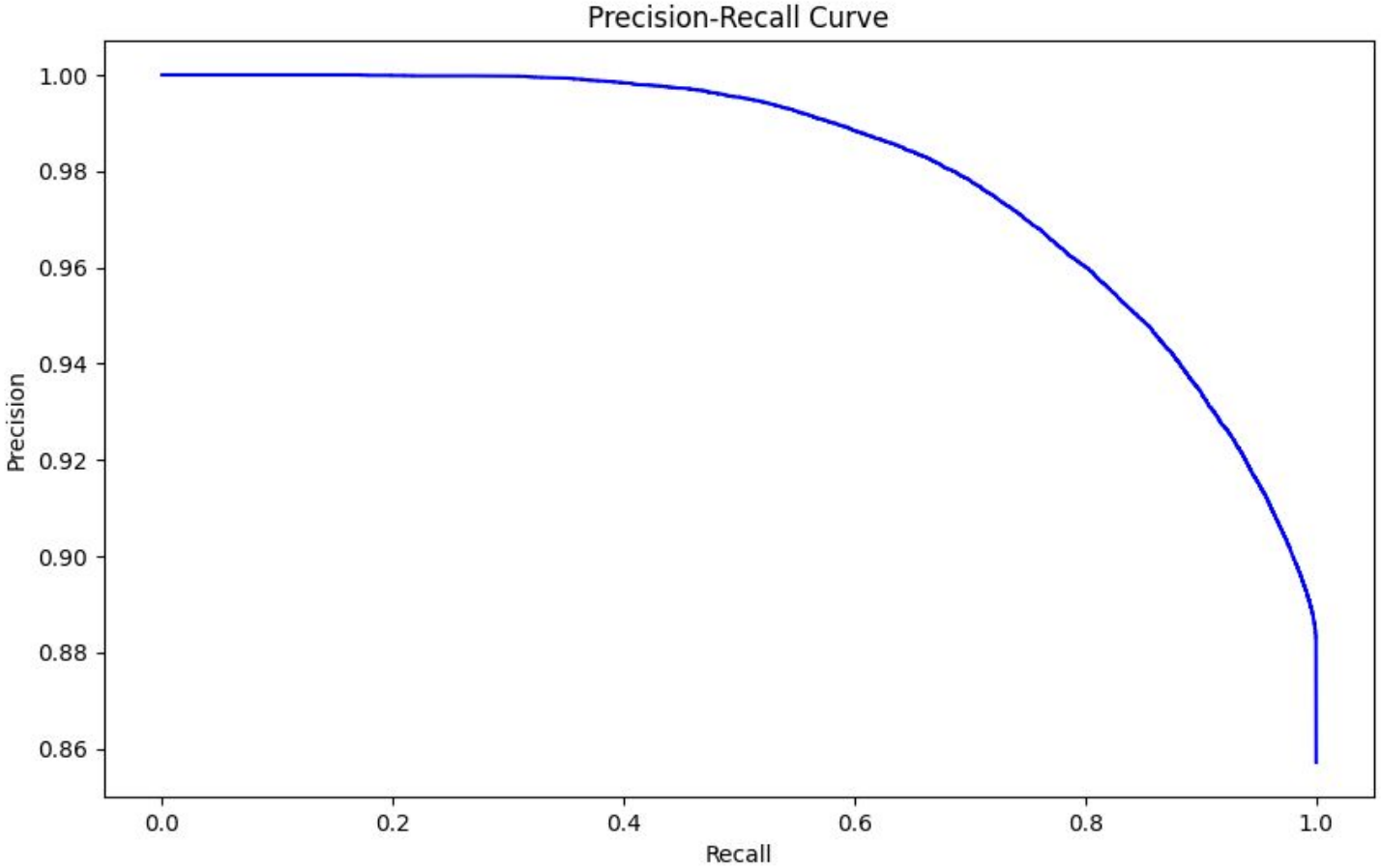

People with Diabetes or Pre-Diabetes: Percentage of People with Gum Disease
Gum Disease
Yes
No
65.08336431098812
34.91663568901188
Percentage
0
10
20
30
40
50
60
Yes
No
Gum Disease

CONCLUSIONS

Best Model: Random Forest

Features: Pre-Diabetes, Age, State, BMI, Sleep Habits

Naive Bayes	Logistic Regression	Random Forest	Random Forest
Recall: 0.6724	**Recall: 0.7549**	**Recall: 0.8234**	**Recall: 0.8643**
Test set score: 0.65	**Test set score: 0.74**	**Test set score: 0.83**	**Test set score: 0.90**
alpha: 0.001	C: 0.01	max_depth: 20	max_depth: 25
fit_prior: True	penalty: L2	min_samples_split: 5	min_samples_split: 3
	solver: liblinear	n_estimators:100	n_estimators: 200


Precision-Recall Curve
Precision
Recall
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.0
0.2
0.4
0.6
0.8
1.0

1
18 to 24 yrs

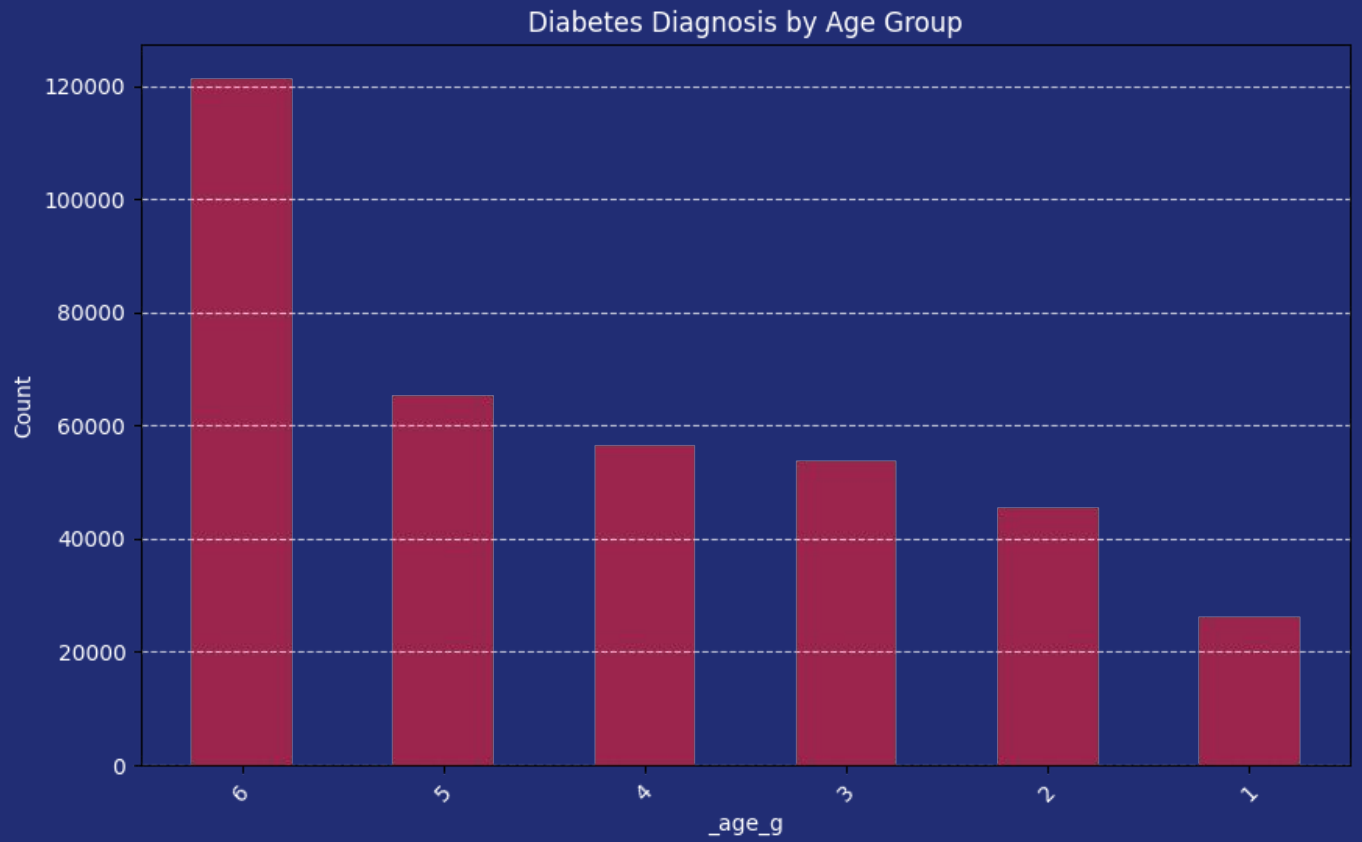
2
25 to 34 yrs

3
35 to 44 yrs

4
45 to 54 yrs

5
55 to 64 yrs

6
65+ yrs



Top 5 States

Vermont

Colorado

District of Columbia

New Jersey

Connecticut

Next Steps

- **Genetic VS. Behavioral Distinction**
  - Build a model based solely on genetic factors that can help screen newborn babies at birth for whether they're higher-risk
- **Explore Feature Importance Findings**
  - Examine why belong to certain states, age groups, etc. makes a person more prone to diabetes diagnosis
- **Application development**
  - Interactive application for mobile-phones allowing insurance-users to track behavioral and medical lifestyle fluctuations, & determine whether they should be medically cautious about developing certain diseases

Questions?