Diabetes classification

Information about the disease: Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from 253,680 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

Dataset:diabetes_binary_classification_data.csv is a clean dataset of 253,680 survey responses to the Center for Disease Control's (CDC) survey. The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

Answer the following questions based on data preprocessing, visualization, correlation, feature engineering, and using any 3 machine learning models for classification of your choice. (Do not use deep learning models)

Explore some of the following research questions:

1) Can survey questions asked from the CDC provide accurate predictions of whether an individual has diabetes? Do you recommend additional features in the data?
2) What risk factors are most predictive of diabetes risk?
3) Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?
4) What machine learning models are best for classifying the disease? Compare models and explain why a model performed better based on the confusion matrix and minimizing false negatives.

Deliverables:

1) Github code link with the readme and well-commented code.

2) A document providing answers to these questions and how did your collaboration work? There are no completely wrong answers. You need to explore data a lot more to suggest accurate prediction and disease prevention in a population.

Information of the dataset

Diabetes_binary

0 = no diabetes 1 = prediabetes 2 = diabetes

HighBP

0 = no high BP 1 = high BP

HighChol

0 = no high cholesterol 1 = high cholesterol

CholCheck

0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

BMI

Body Mass Index

Smoker

Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes

Stroke

(Ever told) you had a stroke. 0 = no 1 = yes

HeartDiseaseorAttack

coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes

PhysActivity

physical activity in past 30 days - not including job 0 = no 1 = yes

Fruits

Consume Fruit 1 or more times per day 0 = no 1 = yes

Veggies

Consume Vegetables 1 or more times per day 0 = no 1 = yes

HvyAlcoholConsump

(adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes

AnyHealthcare

Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes

NoDocbcCost

Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes

GenHlth

Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor

MentHlth

days of poor mental health scale 1-30 days

PhysHlth

physical illness or injury days in past 30 days scale 1-30

DiffWalk

Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes

Sex

0 = female 1 = male

Age

13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older

_AGEG5YR :

1 Age 18 to 24 Respondents with reported age between 18 and 24 years (18 <= AGE <= 24)
2 Age 25 to 29 Respondents with reported age between 25 and 29 years (25 <= AGE <= 29)
3 Age 30 to 34 Respondents with reported age between 30 and 34 years (30 <= AGE <= 34)
4 Age 35 to 39 Respondents with reported age between 35 and 39 years (35 <= AGE <= 39)
5 Age 40 to 44 Respondents with reported age between 40 and 44 years (40 <= AGE <= 44)
6 Age 45 to 49 Respondents with reported age between 45 and 49 years (45 <= AGE <= 49)
7 Age 50 to 54 Respondents with reported age between 50 and 54 years (50 <= AGE <= 54)
8 Age 55 to 59 Respondents with reported age between 55 and 59 years (55 <= AGE <= 59)
9 Age 60 to 64 Respondents with reported age between 60 and 64 years (60 <= AGE <= 64)
10 Age 65 to 69 Respondents with reported age between 65 and 69 years (65 <= AGE <= 69)
11 Age 70 to 74 Respondents with reported age between 70 and 74 years (70 <= AGE <= 74)
12 Age 75 to 79 Respondents with reported age between 75 and 79 years (75 <= AGE <= 79)
13 Age 80 or older Respondents with reported age between 80 and 99 years (80 <= AGE <= 99)

Education

Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.

EDUCA:

Respondents who reported they did not graduate high school. (EDUCA=1,2,3)
Graduated High School Respondents who reported they graduated high school. (EDUCA=4)
Attended College or Technical School Respondents who reported they attended college or technical school. (EDUCA=5)
Graduated from College or Technical School Respondents who reported they graduated from college or technical school. (EDUCA=6)

Income

Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

INCOME2

Less than $15,000 Respondents whose reported income is less than $15,000. (INCOME2=1,2)
$15,000 to less than $25,000 Respondents whose reported income is $15,000 to less than $25,000. (INCOME2=3,4)

$25,000 to less than $35,000 Respondents whose reported income is $25,000 to less than $35,000. (INCOME2=5)

$35,000 to less than $50,000 Respondents whose reported income is $35,000 to less than

$50,000. (INCOME2=6) 5 $50,000 or more Respondents whose reported income is $50,000 or more. (INCOME2=7,8)