Diabetes Classification Model

Snizhana Kurylyuk

Type 2 Diabetes

An estimated 34.2 million people have diabetes (10.5 percent of the U.S. population).

- An estimated 7.3 million adults ages 18 years or older have diabetes but are undiagnosed (21.4 percent of adults with diabetes).

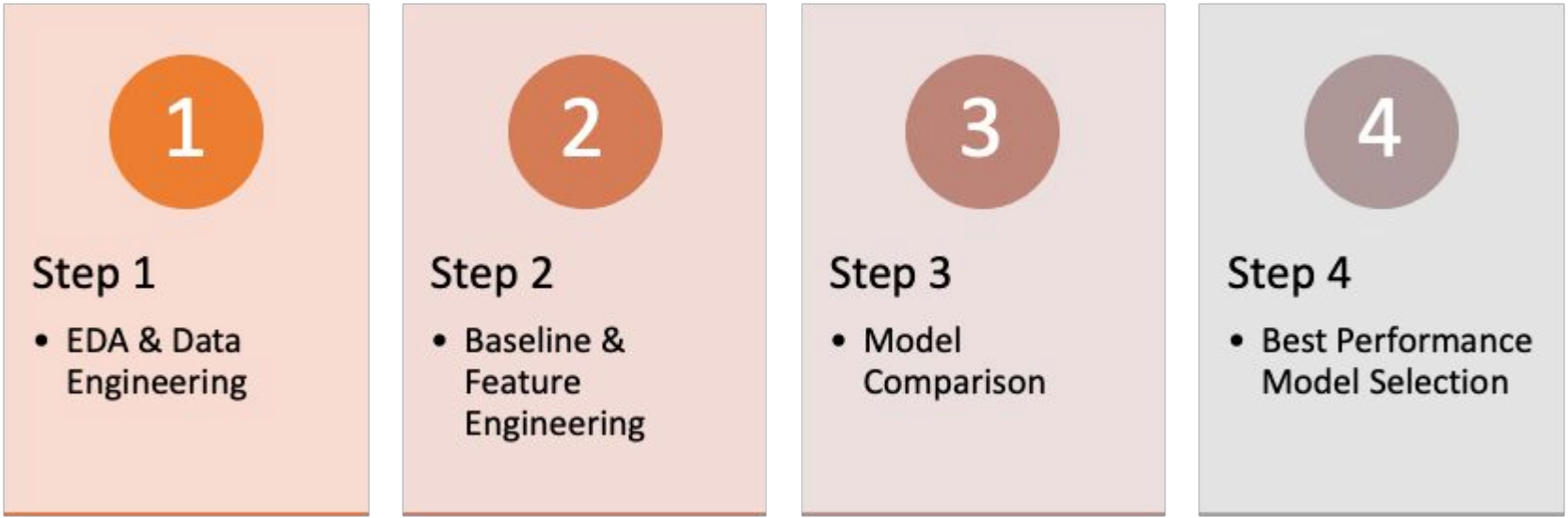
Data: 250K Data points obtained from the Behavioral Risk Factor Surveillance System (BRFSS) a health-related telephone survey that is collected annually by the CDC.

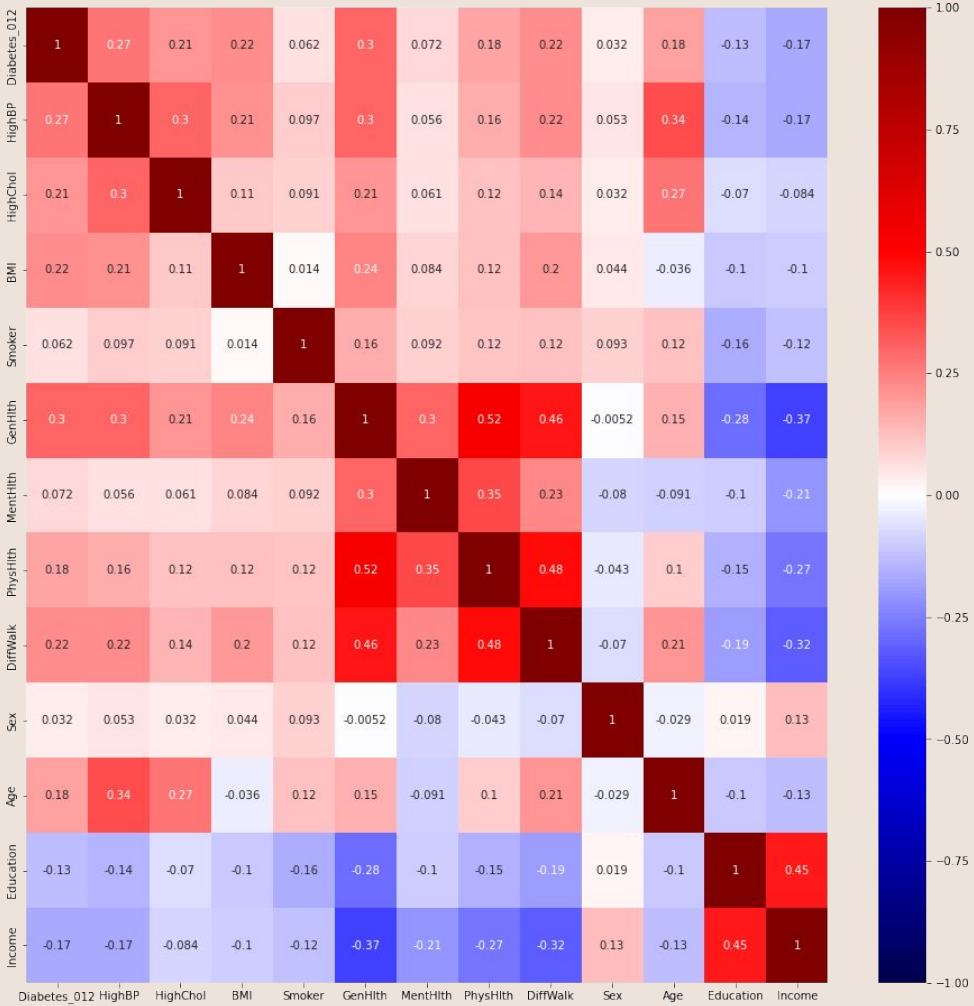
Target: Build a classification model that can diagnose whether or not an individual has diabetes.

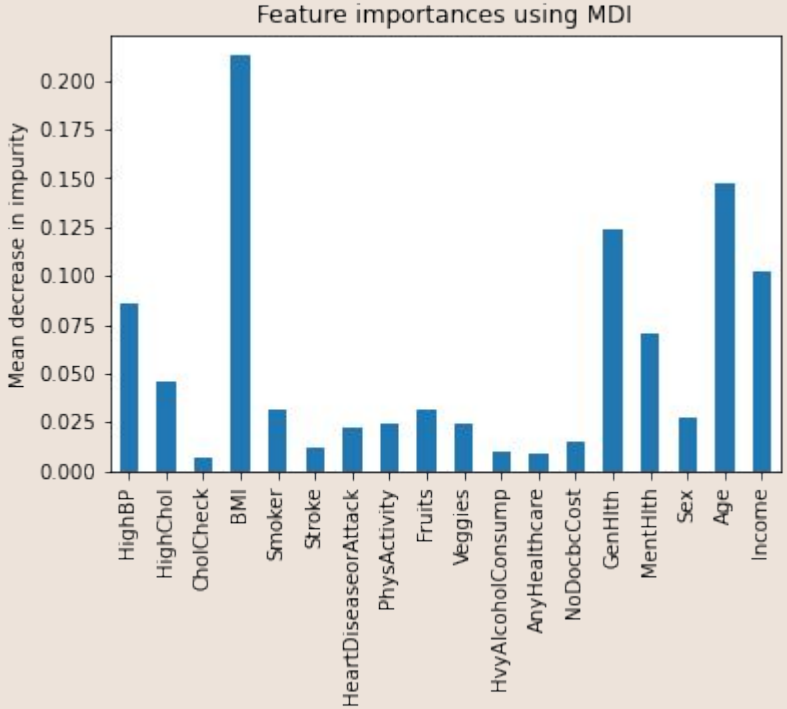
Features

- High Blood Pressure
- High Cholesterol
- Cholesterol Check
- Body Mass Index
- Smoker
- Stroke
- Heart Disease or Attack
- Physical Activity
- Fruits
- Veggies
- Heavy Alcohol Consumption
- Any Healthcare
- No Doctor because of cost
- General Health
- Mental Health
- Physical Health
- Difficulty Walking
- Sex
- Age
- Education
- Income

Can we predict if someone has Diabetes?




Diabetes_012 1 0.27 0.21 0.22 0.062 0.3 0.072 0.18 0.22 0.032 0.18 -0.13 -0.17
HighBP 0.27 1 0.3 0.21 0.097 0.3 0.056 0.16 0.22 0.053 0.34 -0.14 -0.17
HighChol 0.21 0.3 1 0.11 0.091 0.21 0.061 0.12 0.14 0.032 0.27 -0.07 -0.084
BMI 0.22 0.21 0.11 1 0.014 0.24 0.084 0.12 0.2 0.044 -0.036 -0.1 -0.1
Smoker 0.062 0.097 0.091 0.014 1 0.16 0.092 0.12 0.12 0.093 0.12 -0.16 -0.12
GenHlth 0.3 0.3 0.21 0.24 0.16 1 0.3 0.52 0.46 -0.0052 0.15 -0.28 -0.37
MentHlth 0.072 0.056 0.061 0.084 0.092 0.3 1 0.35 0.23 -0.08 -0.091 -0.1 -0.21
PhysHlth 0.18 0.16 0.12 0.12 0.12 0.52 0.35 1 0.48 -0.043 0.1 -0.15 -0.27
DiffWalk 0.22 0.22 0.14 0.2 0.12 0.46 0.23 0.48 1 -0.07 0.21 -0.19 -0.32
Sex 0.032 0.053 0.032 0.044 0.093 -0.0052 -0.08 -0.043 -0.07 1 -0.029 0.019 0.13
Age 0.18 0.34 0.27 -0.036 0.12 0.15 -0.091 0.1 0.21 -0.029 1 -0.1 -0.13
Education -0.13 -0.14 -0.07 -0.1 -0.16 -0.28 -0.1 -0.15 -0.19 0.019 -0.1 1 0.45
Income -0.17 -0.17 -0.084 -0.1 -0.12 -0.37 -0.21 -0.27 -0.32 0.13 -0.13 0.45 1
Diabetes_012 HighBP HighChol BMI Smoker GenHlth MentHlth PhysHlth DiffWalk Sex Age Education Income
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00


Feature importances using MDI
Mean decrease in impurity
0.200 0.175 0.150 0.125 0.100 0.075 0.050 0.025 0.000
HighBP HighChol CholCheck BMI Smoker Stroke HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost GenHlth MentHlth Sex Age Income

Classification Algorithms Validation Metrics

K Nearest Neighbors

- Accuracy: 0.7673
- Precision: 0.3128
- Recall: 0.5360
- F1: 0.3951
- Cross Validation Accuracy Score : 0.77854

Random Forest:

- Accuracy: 0.7698
- Precision: 0.3488
- Recall: 0.7195
- F1: 0.4698
- Cross Validation Accuracy Score: 0.80256

Logistic Regression:

- Accuracy: 0.7896
- Precision: 0.3639
- Recall: 0.6476
- F1: 0.4659
- Cross Validation Accuracy Score : 0.74455

Stacking Ensembling:

- Accuracy: 0.8623
- Precision: 0.5388
- Recall: 0.1988
- F1: 0.2905
- Cross Validation Accuracy Score: 0.92498

Decision Tree:

- Accuracy: 0.7699
- Precision: 0.3330
- Recall: 0.6218
- F1: 0.4337
- Cross Validation Accuracy Score: 0.71748

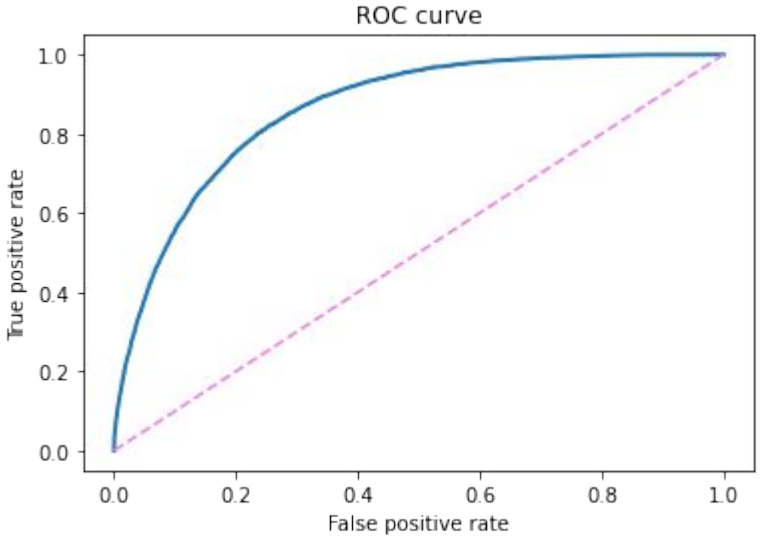
Voting Ensembling:

- Accuracy: 0.7974
- Precision: 0.3560
- Recall: 0.5296
- F1: 0.4258
- Cross Validation Accuracy Score: 0.8623

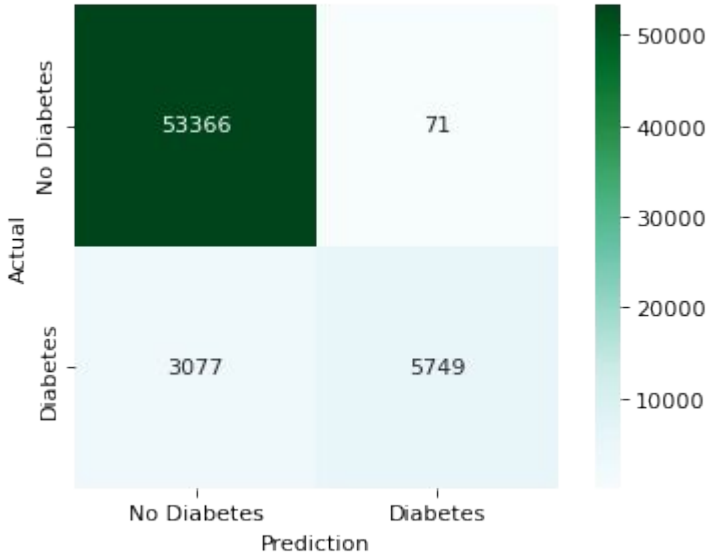
Hyperparameter Tuning

First Model	Second Model	Third Model	Best Model
Accuracy: 0.7698	**Accuracy:** 0.8201	**Accuracy:** 0.8312	**Accuracy:** 0.8263
Precision: 0.3488	**Precision:** 0.3440	**Precision:** 0.3881	**Precision:** 0.4175
Recall: 0.7195	**Recall:** 0.2969	**Recall:** 0.3310	**Recall:** 0.5710
F1: 0.4698	**F1:** 0.3187	**F1:** 0.3572	**F1:** 0.4824
Cross Validation Accuracy Score: 0.80256	**Cross Validation Accuracy Score:** 0.92541	**Cross Validation Accuracy Score:** 0.91561	**Cross Validation Accuracy Score:** 0.82484

Best Performance Model: Random Forest



ROC AUC Score = 0.8625



Questions