CDC Diabetes Health Indicators

Tarun Sudhams

Data Source

Title
CDC Diabetes Health Indicators

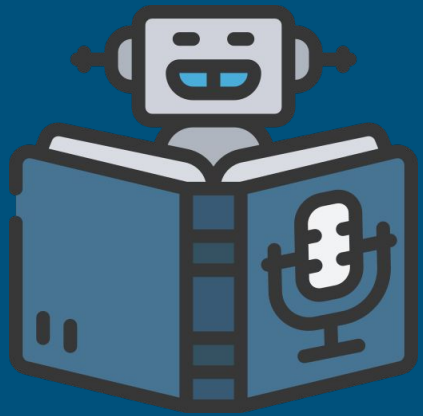
Data Creator
CDC

Data Publisher
UC Irvine Machine Learning Repository

Last Updated
2023

Goals

The primary goal of this project is to **build** and **evaluate** predictive models for diabetes status using the CDC Diabetes Health Indicators dataset, focusing on **Logistic Regression** and **Random Forest** models.



Data Cleaning

Overall Data Quality

Given our data source, we were quite lucky with how clean and organized our data was.

Still needed some cleaning

Despite the nice state of the dataset, we still had to test for missing values and also duplicate rows.

On checking, we found ~22K duplicate rows that we had to get rid of to avoid our models getting skewed.

Exploratory Data Analysis

Correlation Matrix

Allowed us to visualize the impact of features with each other and be able to gauge which ones will play a role in our data analysis

Outliers Assessment

Correlation Matrix helped us get a hint on the outlier data points that could exist in our dataset. We could then apply caps on those values to make sure it doesn't skew our models.

Model Results

Random Forest

```
Test Accuracy: 0.8698
Confusion Matrix:
 [[33935   259]
 [ 4879   385]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93     34194
           1       0.60      0.07      0.13      5264

    accuracy                           0.87     39458
   macro avg       0.74      0.53      0.53     39458
weighted avg       0.84      0.87      0.82     39458
```

Logistic Regression

```
Test Accuracy: 0.8684
Confusion Matrix:
 [[33847   347]
 [ 4847   417]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93     34194
           1       0.55      0.08      0.14      5264

    accuracy                           0.87     39458
   macro avg       0.71      0.53      0.53     39458
weighted avg       0.83      0.87      0.82     39458
```

Jupyter Notebook Demo

Thank You!