Predicting Diabetes and Heart Attack Using a Health Survey

Group 2: Katharine Bloeser,
Emma Perez, and Raphael Sheikh

The National Health and Nutrition Examination Survey (NHANES)

The dataset from the National Health and Nutrition Examination Survey (NHANES), provided by the CDC, contains extensive health data on Americans. It includes information on demographics, dietary habits, medical history, physical activity, and laboratory test results. This data is commonly used to analyze public health trends and evaluate risk factors for various conditions, including diabetes and heart disease. To address the question, "What predicts diabetes and heart attack among Americans?", we can use NHANES data to identify key predictors, such as obesity, high blood pressure, cholesterol levels, smoking, and physical inactivity, among other factors. This data helps inform public health strategies to prevent and manage these chronic conditions.

Reference:
https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

Diabetes Testing Data Confusion Matrix

True Positive: 2,332	False Positive: 28
False Negative: 140	True Negative: 44

Diabetes Testing Data Confusion Matrix: Interpretation

True Positive:
We think they have diabetes, they do have diabetes.
2,332 screened and treated.

False Positive:
We think they have diabetes, they do not have diabetes.
28 screened not treated.

False Negative:
We think they do not have diabetes, they do have diabetes.
140 not screened not treated.

True Negative:
We think they do not have diabetes, they do not have diabetes.
44 not screened not treated.

Testing Data Classification Report

Accuracy: 93% of the time, this model correctly predicted if someone had diabetes or did not have diabetes.

Recall:

The model correctly identifies 99% of the actual population of people who do not have diabetes.
The model correctly identifies just 24% of the actual population of people who do have diabetes.

Precision:

When the model says someone does not have diabetes it is correct 94% of the time. However, 6% of the time, it identifies false negatives, meaning it says someone does not have diabetes, when they in fact, do have diabetes.

When the model says someone has diabetes it is correct 61% of the time. However, 39% of the time, it identifies false positives, meaning it says someone has diabetes, when they do not, in fact, have diabetes.

F1 Score:

When predicting the no diabetes population, the model performs very well (97%).
When predicting the diabetes population, the model does not perform well (34%).

Feature Importance

Age: 0.47

General Health: 0.25

High Blood Pressure: 0.18

Cholesterol Medication Prescription: 0.18

Relative with Diabetes: 0.16

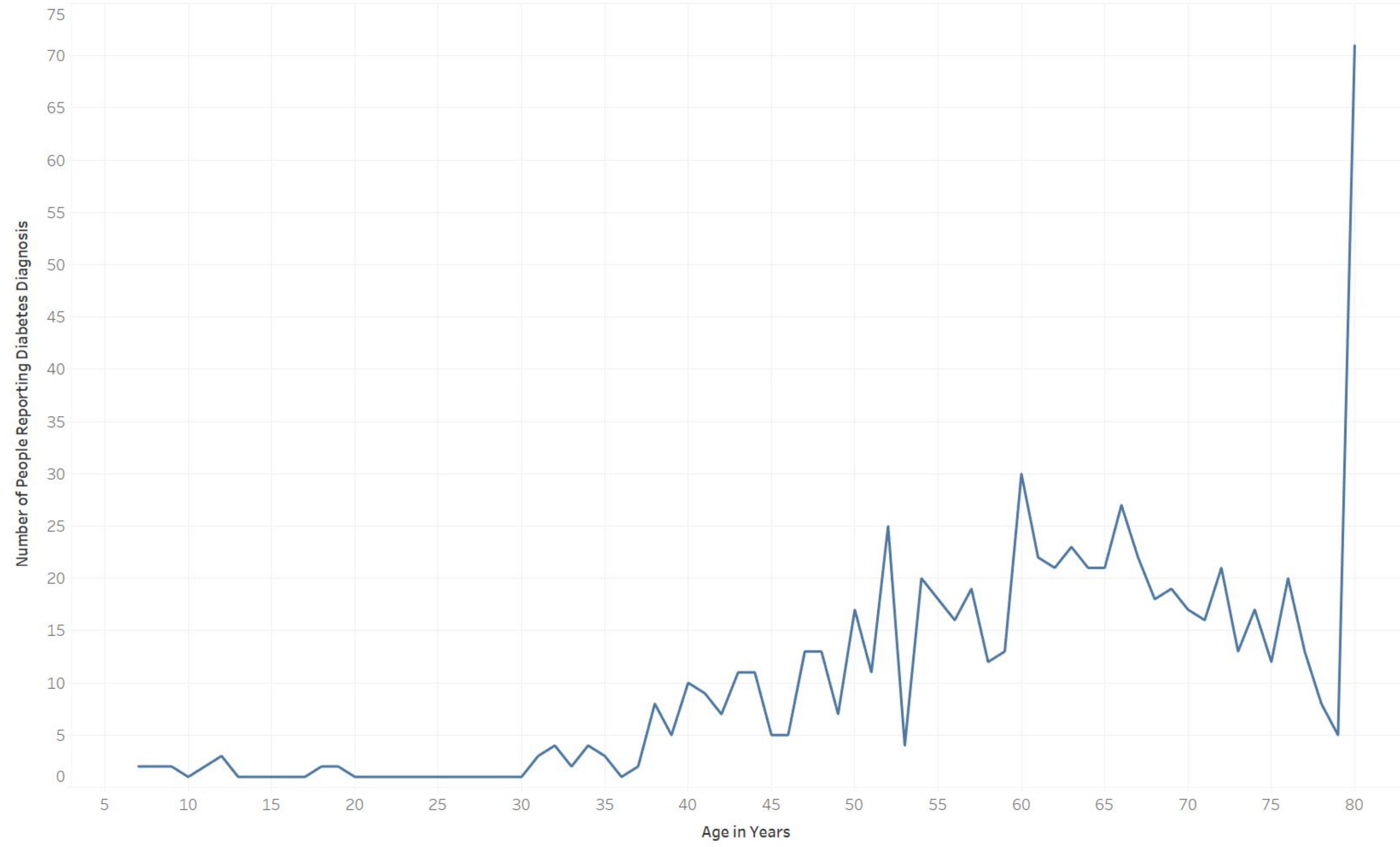
Ever Diagnosed with High Cholesterol: 0.16

Ever told Overweight: 0.15

Marital Status: 0.13

Number of Milk Products Consumed Each Month: 0.12

Ever Diagnosed with Arthritis: 0.09



Number of Diabetes Diagnoses by Age, NHANES 2013-2014

Heart Attack Testing Data Confusion Matrix

True Positive: 1,382	False Positive: 2
False Negative: 58	True Negative: 0

Heart Attack Testing Data Confusion Matrix: Interpretation

True Positive:
We think they have had a heart attack, they have had a heart attack.
1,382 screened and treated.

False Positive:
We think they have had a heart attack, they have not had a heart attack.
2 screened not treated.

False Negative:
We think they have not had a heart attack, they have had a heart attack.
58 not screened not treated

True Negative:
We think they have not had a heart attack, they have not had a heart attack.
0 not screened not treated

Testing Data Classification Report

Accuracy: 96% of the time, this model correctly predicted if someone had a heart attack or did not have a heart attack.

Recall:

The model correctly identifies 100% of the actual population of people who did not have a heart attack.

The model correctly identifies 0% of the actual population of people who did have a heart attack.

Precision:

When the model says someone did not have a heart attack it is correct 96% of the time.

When the model says someone did have a heart attack it is never correct.

F1 Score:

When predicting the no heart attack population, the model performs very well (98%).

When predicting the heart attack population, the model does not perform (0%).

Feature Importance

Age: 0.21

Coronary Heart Disease: 0.13

Gender: 0.10

Cholesterol Prescription: 0.10

High Blood Pressure: 0.09

Congestive Heart Failure: 0.09

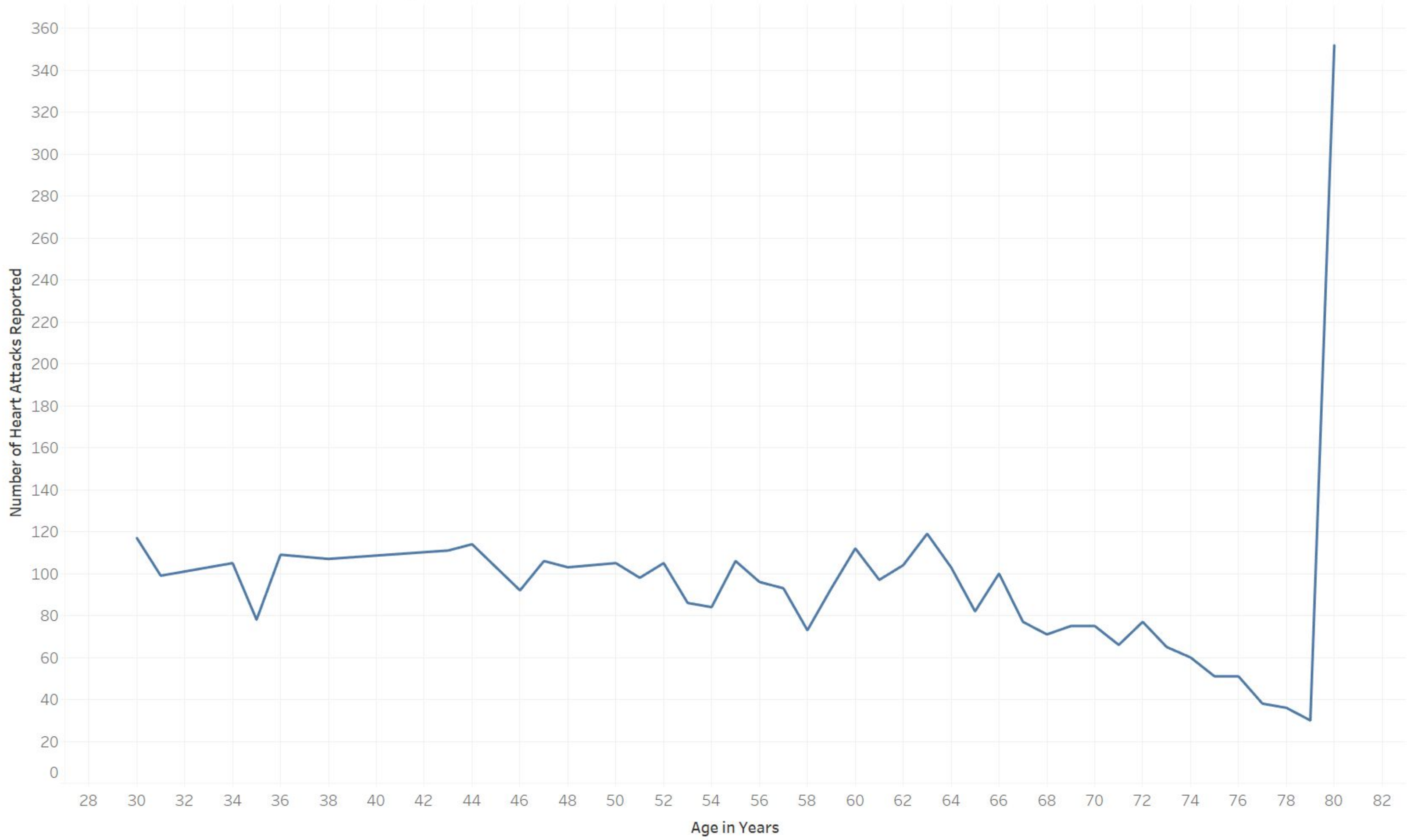
Monthly Income: 0.09

High Cholesterol: 0.08

General Health: 0.07

Blood Transfusion: 0.07

Number of Heart Attacks within Age, NHANES 2013-2014



Summary

1. Our models did a very good job of predicting “healthy” people. They were good at predicting when someone would not have a history of diabetes or a heart attack.
2. Our models were not good at predicting instances of diabetes and heart attack in the population.

Limitations and Implications for Use

Why did our models not fit well to predict the Class 1 (or presence of diabetes/heart attack)? Possibly there were too few cases in the dataset. In the future, we could combine years to get a larger population of people who report the outcome.

We would not recommend these models to predict instances of diabetes and/or heart attack. For both models, there is a preponderance of false negatives, thus the model did not predict the outcome in people who do have the condition. This means we would potentially miss people requiring early intervention.

Final Assessment and Future Considerations

In conclusion, while our models demonstrated strong accuracy in predicting healthy individuals those without a history of diabetes or heart attacks they were less effective in identifying individuals with these conditions. The high incidence of false negatives, particularly in predicting heart attacks, suggests that the models are not well-suited for early intervention purposes. This limitation could be attributed to the dataset's imbalance, with too few instances of diabetes or heart attacks. To improve performance, future efforts could focus on expanding the dataset by combining additional years to capture more instances of the conditions in question.