Heart Attacks & Disease

Predicting heart attacks and diseases based on demographic and biometric data

October 24th

Maxine Baghdadi
Data Science Institute
Github Link

Introduction

Overview

What problem are we solving?

Predicting whether a person will have a heart disease or attack based on general health and demographic data

What type of problem?

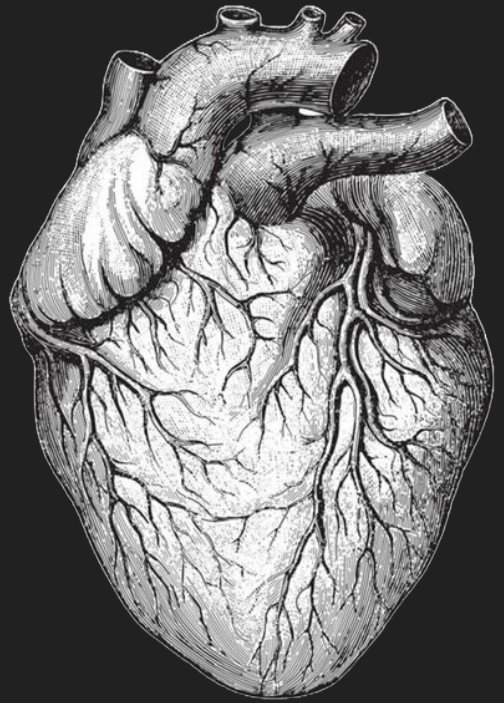
This is a binary classification problem where our target variable is `HeartDiseaseorAttack`

Why is this important?

Improving our ability to detect diseases early and helping inform personalized healthcare recommendations

Where was data from? How was it collected?

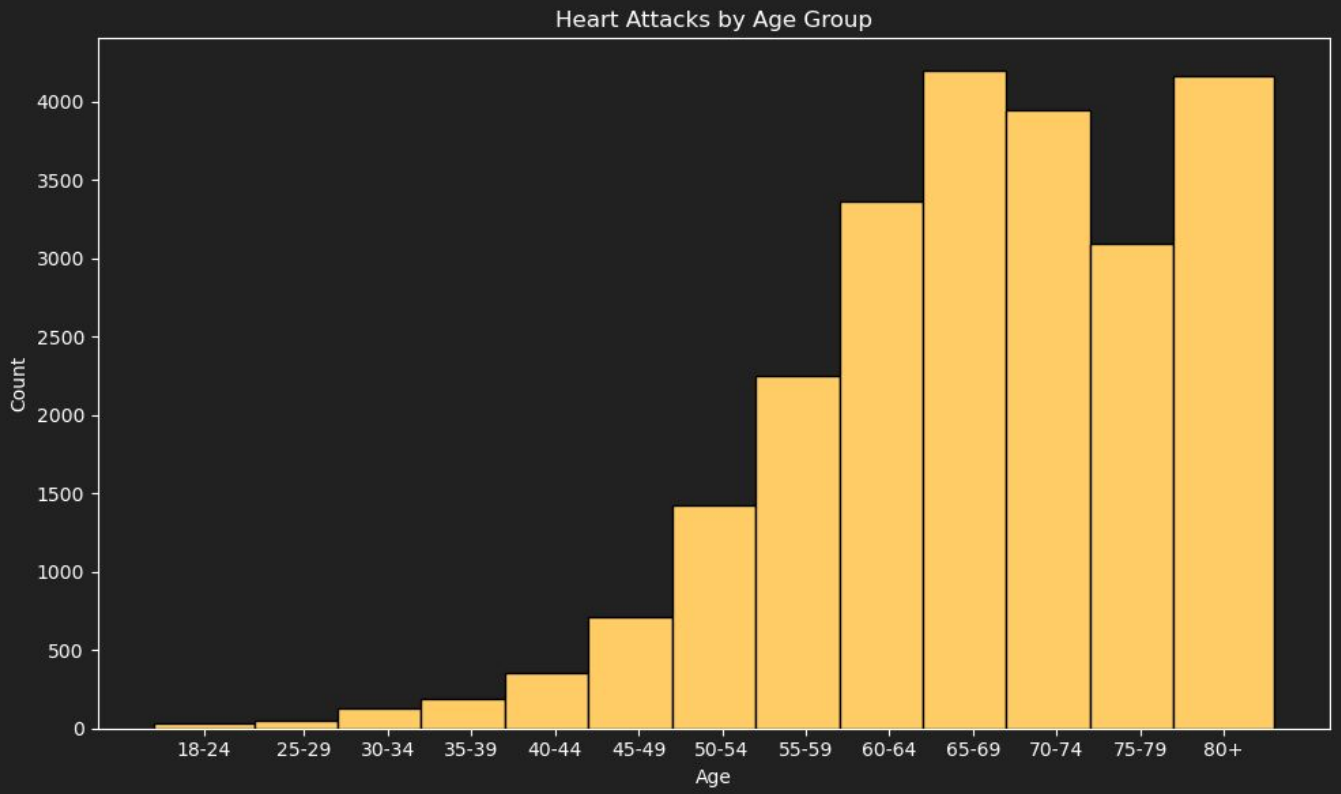
Extracted from survey data collected by the Behavioral Risk Factor Surveillance System (BFRSS) through the Center for Disease Control



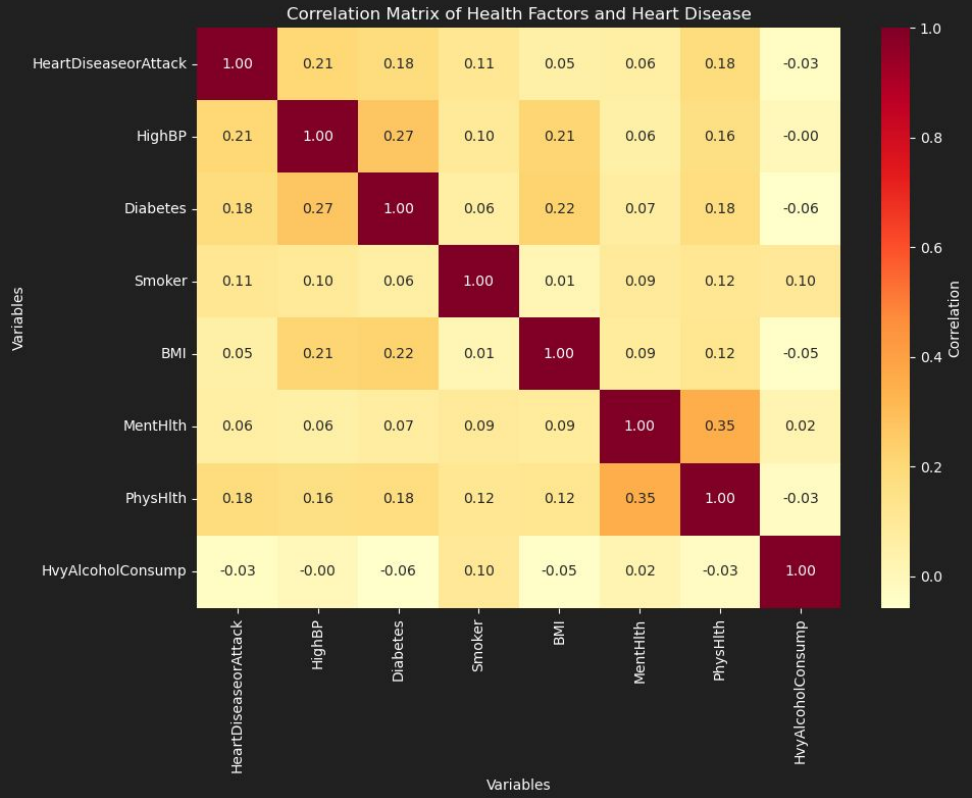
Exploratory Data Analysis

Distribution Across Age

The risk of heart attacks significantly increases as people age, with a **sharp rise** starting after the age of **50**, and peaking between **60 and 69 years old**. The highest counts are seen in individuals aged **65+**, indicating that heart disease prevention should be prioritized for middle-aged and older adults.



Lasering into health factors

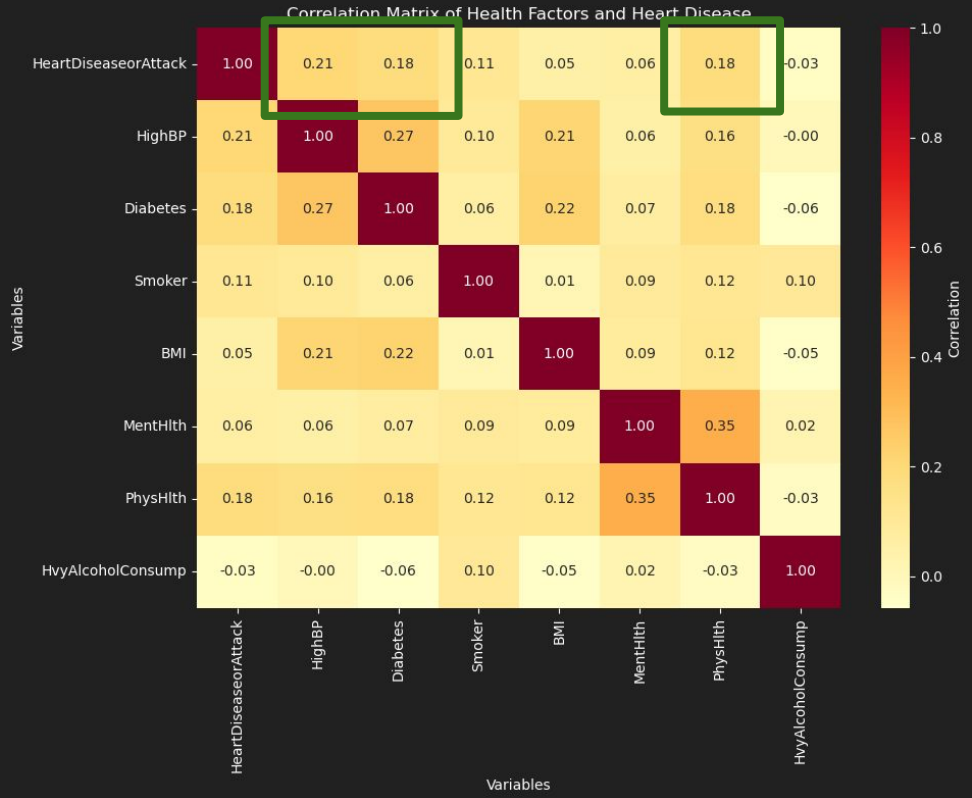


Analyzed potential **health specific variables** and their correlation with having a heart attack or disease

Takeaways:

1. BMI and Smoker have weak correlations with Heart Disease
2. Heavy Alcohol Consumption shows a negative correlation
3. Strongest correlations are high blood pressure, diabetes and physical health

Lasering into health factors



Analyzed potential **health specific variables** and their correlation with having a heart attack or disease

Takeaways:

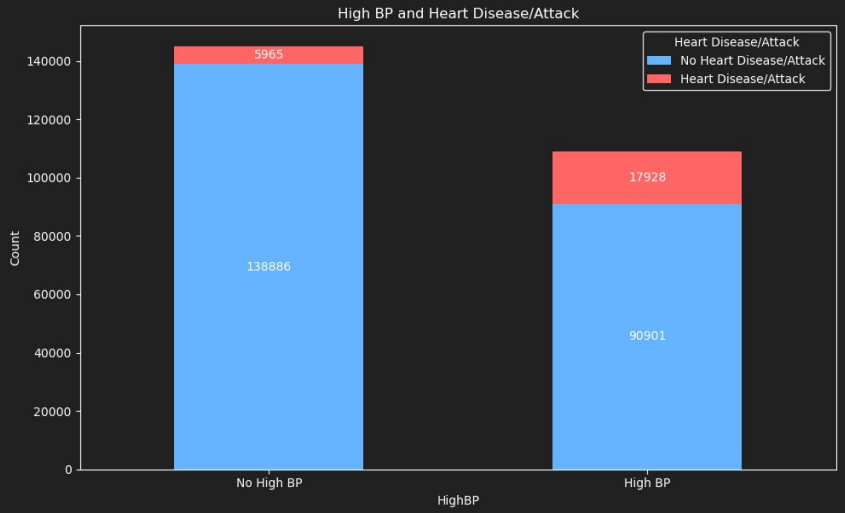
1. BMI and Smoker have weak correlations with Heart Disease
2. Heavy Alcohol Consumption shows a negative correlation
3. Strongest correlations are high blood pressure, diabetes and physical health

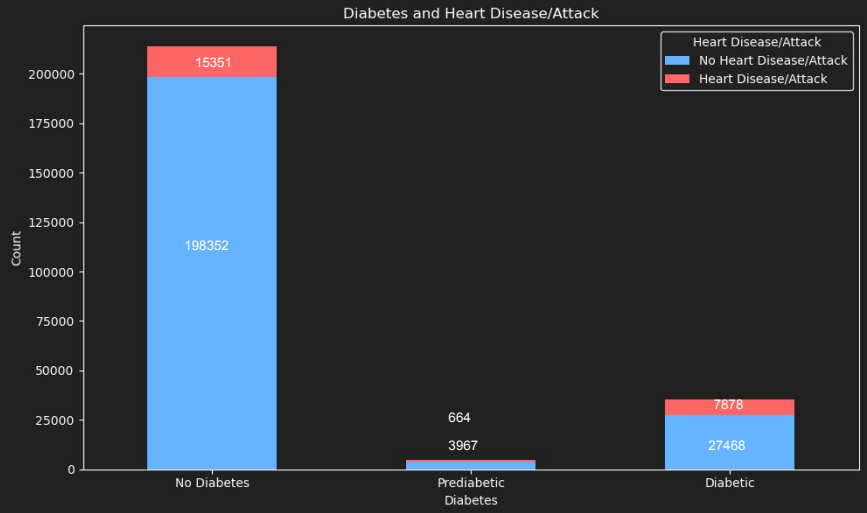
Let’s dive deeper into #3!

Digging deeper into BP & Diabetes

There is a **consistent trend**: A higher proportion of those who have experienced health conditions have experienced a heart attack

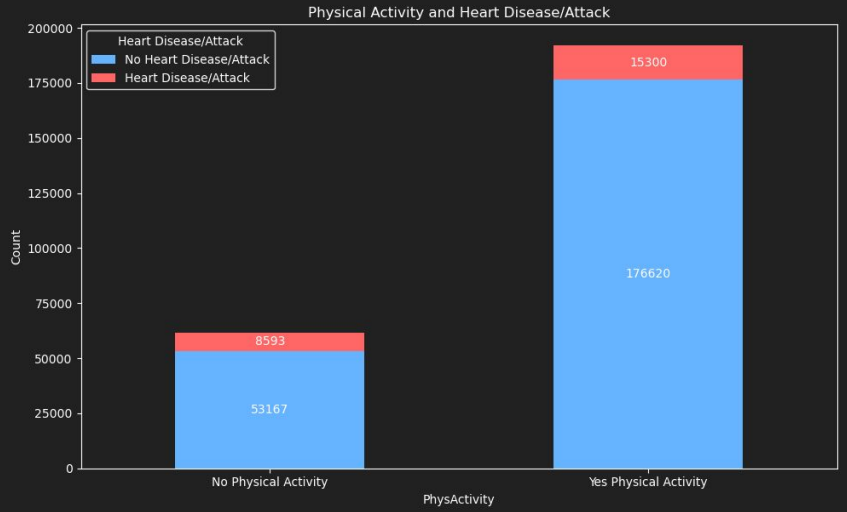
- High BP: 4% versus 16%
- Diabetes: 7% versus 21%

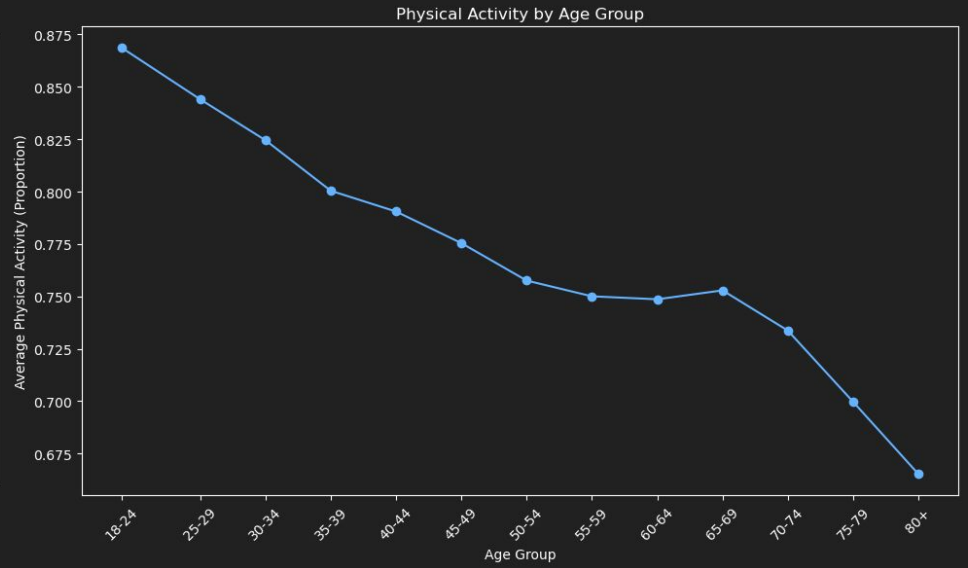




And what about physical activity?

When it came to physical activity, you can see a higher percentage of people who didn't complete physical activity have diabetes. You can also see that those with little activity are for older age groups.





Splitting & Preprocessing

Splitting the data

Considerations:

- Only 9% of the data is in Class 1
- The dataset contains over 250k rows

Approach: Stratified Split

- 60% in train
- 20% in test
- 20% in validation

Class	Count	Percentage
Class 0	229,787	90.6%
Class 1	23,893	9.4%

```
**balance with stratification:**
(array([0., 1.]), array([137872,  14336]))
(array([0., 1.]), array([45958,  4778]))
(array([0., 1.]), array([45957,  4779]))
```

Preprocessing the data

- ☑ No missing values
- ☑ Strings converted to numbers
- ☑ Features scaled

	Mean	Standard Deviation
HeartDiseaseorAttack	0.094186	0.292087
HighBP	0.429001	0.494934
HighChol	0.424121	0.494210
CholCheck	0.962670	0.189571
BMI	28.382364	6.608694
Smoker	0.443169	0.496761
Stroke	0.040571	0.197294
Diabetes	0.296921	0.698160
PhysActivity	0.756544	0.429169
Fruits	0.634256	0.481639
Veggies	0.811420	0.391175
HvyAlcoholConsump	0.056197	0.230302
AnyHealthcare	0.951053	0.215759
NoDocbcCost	0.084177	0.277654
GenHlth	2.511392	1.068477
MentHlth	3.184772	7.412847
PhysHlth	4.242081	8.717951
DiffWalk	0.168224	0.374066
Sex	0.440342	0.496429
Age	8.032119	3.054220
Education	5.050434	0.985774
Income	6.053875	2.071148

	Mean	Standard Deviation
HeartDiseaseorAttack	9.418559e-02	0.292087
HighBP	4.290011e-01	0.494934
HighChol	4.241209e-01	0.494210
CholCheck	9.626695e-01	0.189571
BMI	-2.505162e-16	1.000002
Smoker	4.431686e-01	0.496761
Stroke	4.057080e-02	0.197294
Diabetes	2.969213e-01	0.698160
PhysActivity	7.565437e-01	0.429169
Fruits	6.342558e-01	0.481639
Veggies	8.114199e-01	0.391175
HvyAlcoholConsump	5.619678e-02	0.230302
AnyHealthcare	9.510525e-01	0.215759
NoDocbcCost	-1.792602e-17	1.000002
GenHlth	2.511392e+00	1.068477
MentHlth	8.963011e-18	1.000002
PhysHlth	3.450759e-17	1.000002
DiffWalk	1.682237e-01	0.374066
Sex	4.403422e-01	0.496429
Age	8.032119e+00	3.054220
Education	5.050434e+00	0.985774
Income	6.053875e+00	2.071148

Next Steps

Questions?