Key Indicators in predicting a Heart Attack

Arpit Dang
Data Science Institute at Brown University
October 23rd, 2024
https://github.com/adang66/Data1030-Project

More information about the dataset

Why:

- 1 in 5 deaths in the United States are related to heart diseases [1]
- On average, someone in United States has a heart attack every 40 seconds [2]
- About 47% of Americans have at least 1 of the 3 major risk factors for heart diseases: high blood pressure, high cholesterol or smoking [3]

What: Telephone surveys among 400k+ American adults that collected variables that may contribute to heart attacks

Who: Collected by the Centers for Disease Control and Prevention (CDC)

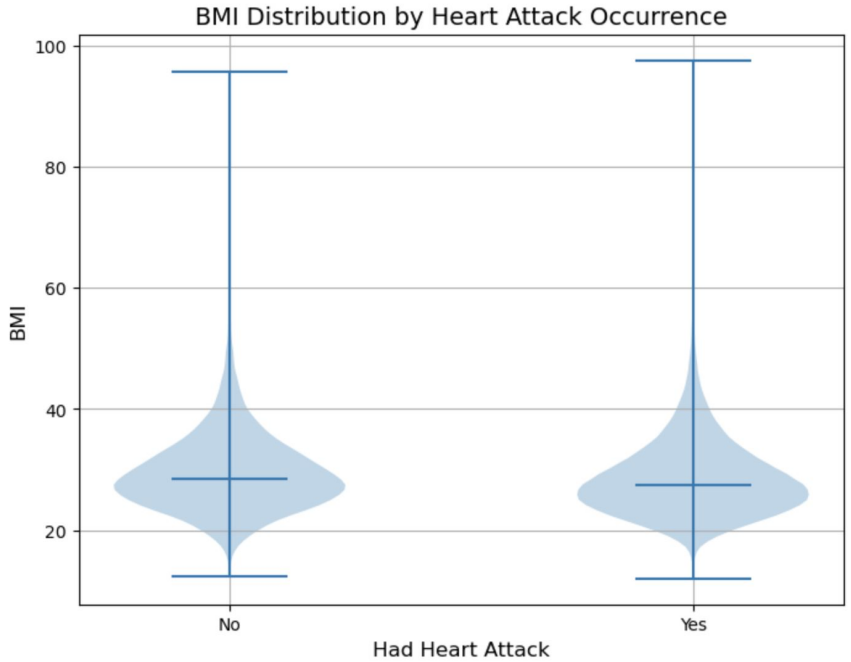
When: It was collected in the year 2022 (relatively recent)

More information about the dataset (cont'd)

- There were initially 400k+ adult data points in the dataset, however the authors removed the data points with missing values leaving **246,022 data points**
- The survey initially included approximately 300 variables but it was reduced to **39 variables** they thought would be the most suitable for this dataset

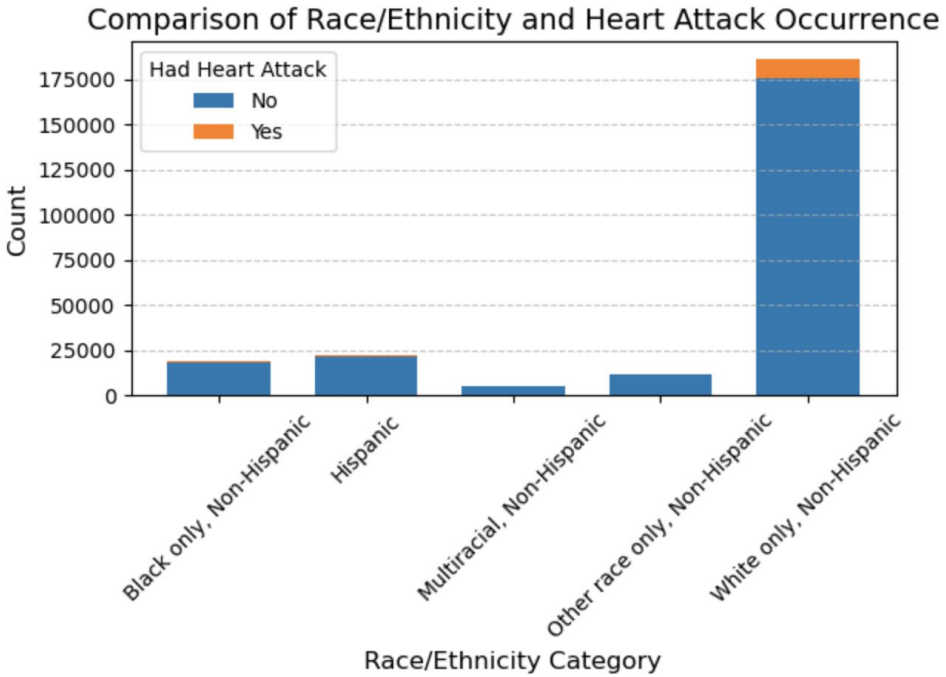
- **Target Variable:** If the participant had a heart attack; (Yes OR No); making this a classification problem

BMI Distribution by Heart Attack Occurrence



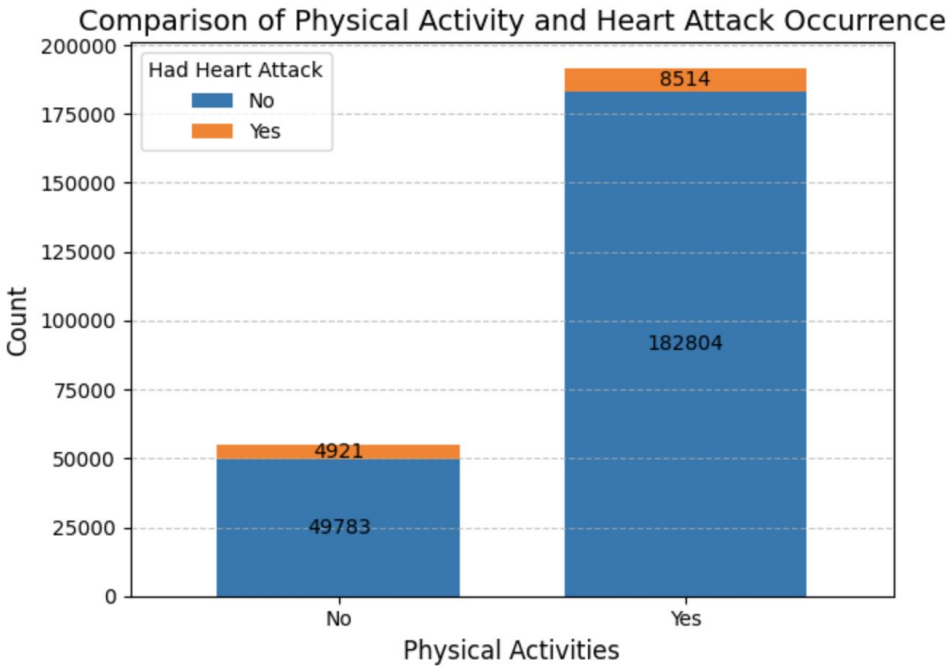
- Expected that people who have had a heart attack would have a higher BMI [4]
- But the BMI of people with and without heart attack was relatively the same
- Odd because you would expect higher median BMI in people who have has a heart attack occurrence

Heart Attack Occurrence by Ethnicity & Race



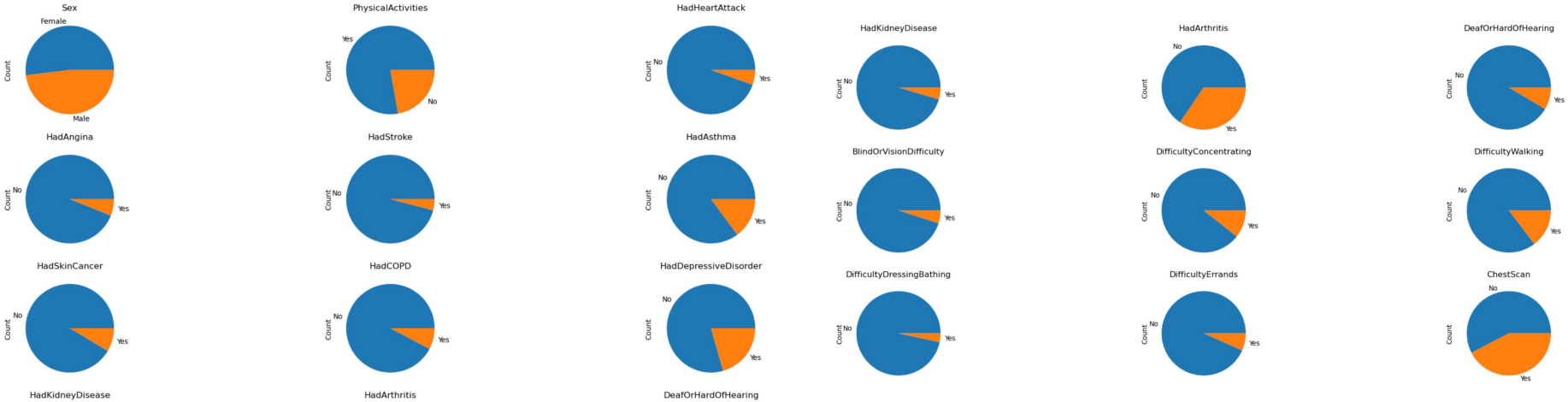
- There is a lack of representation in data for most non-white races in this dataset
- It does somewhat correlate with the distribution of the American population [5]
- Here are percentages for people who suffered a heart attack by ethnicity/race:
  - 6.5% in Multiracial
  - 6.1% in White
  - 5.1% in Other Races
  - 4.8% in Black
  - 4.0% in Hispanic

Comparison of Physical Activity & Heart Attack Occurrence



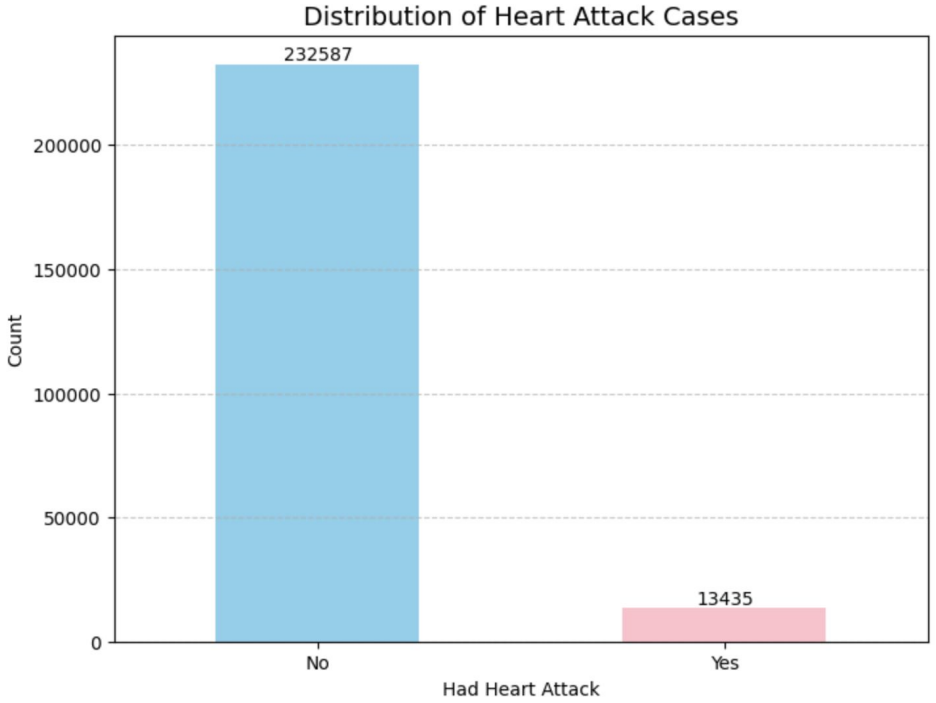
- 9.8% of people had a heart attack occurrence that did not participate in physical activity
- 4.7% of people had a heart attack occurrence that participated in physical activity
- Limitation: not descriptive enough. 23/39 variables in the dataset are binary (Yes OR No)

Binary Features



23/39 features are binary features.
Limitation: limited information representation (ordinal might be better), lack of qualitative data, it can lead to reduced model accuracy or interpretability

Distribution of Heart Attack Cases (target variable)



- This is an independently and identically distributed (i.i.d) dataset
- No missing values
- 5.5% of the participants has an incident of a heart attack, making it an imbalanced dataset
- Solution: stratified split

Stratified split

- 60% training set, 20% validation set, 20% test set

Type of set	Total data points	‘No’ to HadHeartAttack	‘Yes’ to HadHeartAttack
Training set	147 613	139 552	8 061
Validation set	49 204	46 517	2 687
Test set	49 205	46 518	2 687

Preprocessing

- OneHotCoder: 27 features (categorical features)
- OrdinalEncoder: 6 features (ordinal features)
- MinMaxScalar: 6 features (continuous features)

X_train.shape = (147 613, 39)

X_train_prep.shape = (147 613, 126)

Thank you for listening!

Questions?

References

1. Martin SS, Aday AW, Almarzooq ZI, et al.; American Heart Association Council on Epidemiology and Prevention Statistics Committee; Stroke Statistics Subcommittee. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation. 2024;149:e347–913.
2. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. Circulation. 2023;147:e93–e621.
3. Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS Data Brief. 2012;(103):1–8.
4. Adams, B., Jacocks, L., & Guo, H. (2020). Higher BMI is linked to an increased risk of heart attacks in European adults: A Mendelian Randomisation study. BMC Cardiovascular Disorders, 20(1). https://doi.org/10.1186/s12872-020-01542-w