Predicting Heart Disease Risk from Behavioral and Health Indicators

Yichen Li

Brown University, Data Science Institute

10/23/2025

BROWN
Data Science Institute

github:https://github.com/YichenLi95/DATA1030_HeartDiseaseProject

Introduction

Goal: Predict whether an individual has heart disease using health and lifestyle indicators.

Why it matters:

- Heart disease is one crucial cause of death in the U.S.
- Early intervention can reduce hospitalizations and save lives.

Task Type: Binary Classification

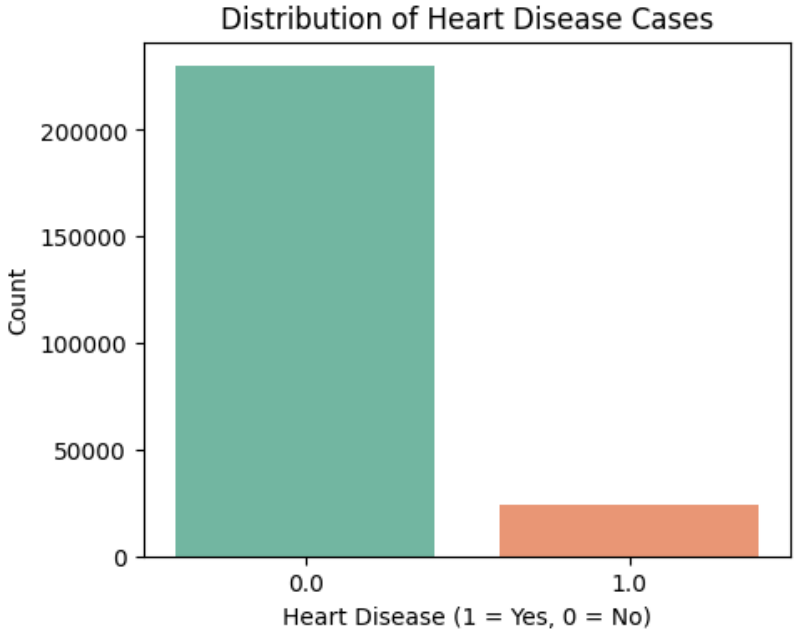
Data Source:

- Via Kaggle Platform
- Collected through nationwide telephone survey

Main Challenge: Large dataset, contain over 250k data points

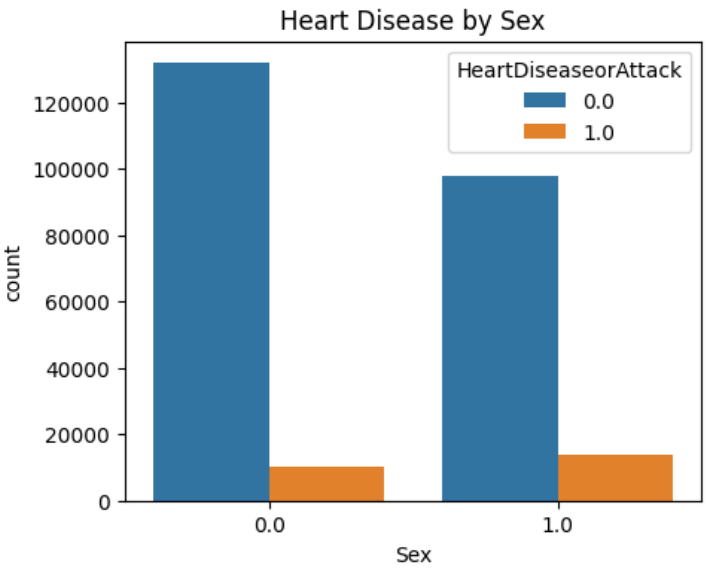
EDA

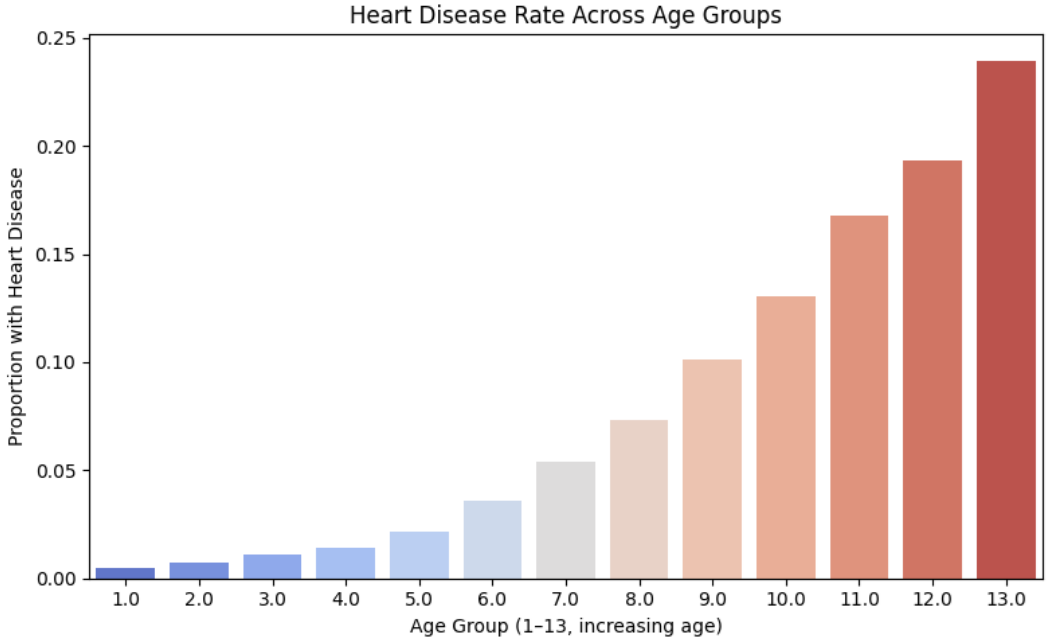
```
Shape: (253680, 22)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```



Imbalance affects model training → need stratified split

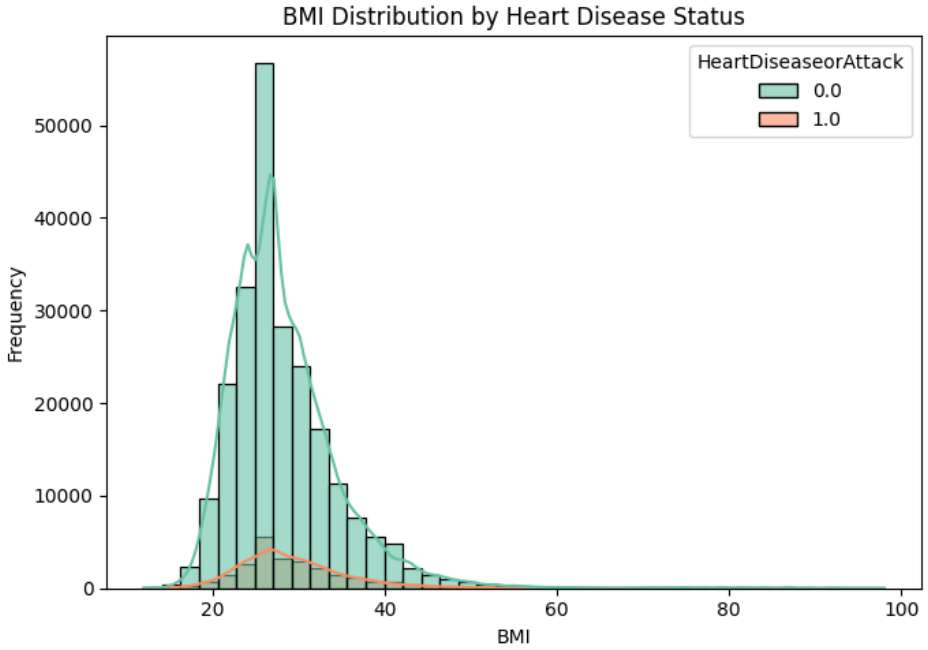
EDA





Highlight features likely to have **high model importance**.

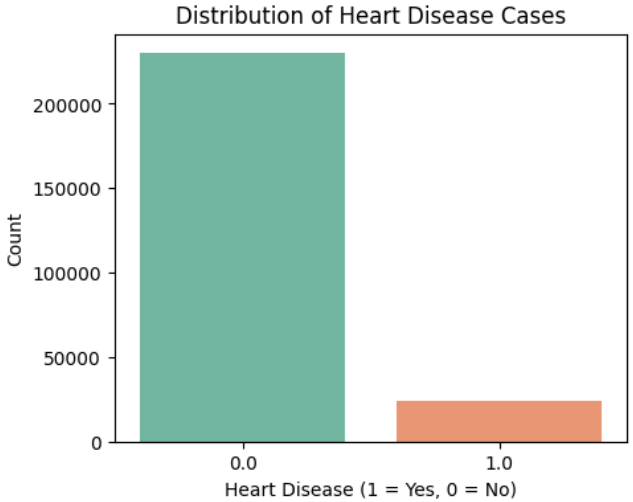
EDA



Boundary considerations

Consider use MinmaxScaler

Data Splitting Strategy



→ **Imbalance** →

Stratified Split

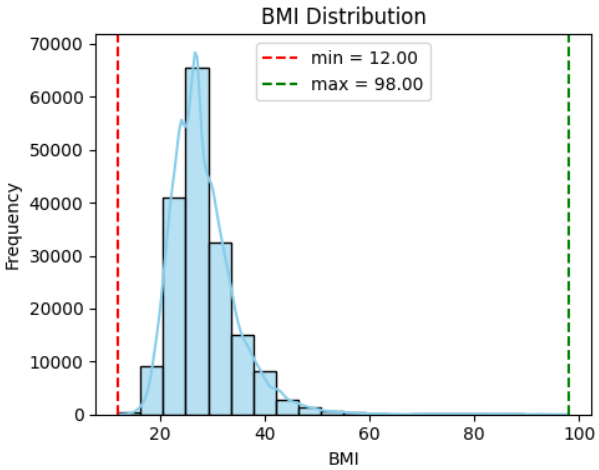
70% Train, 15% Validation, 15% Test

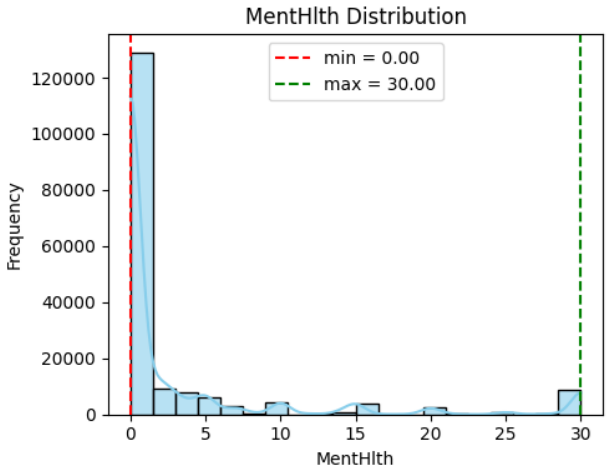
Ensures same heart disease ratio across all splits

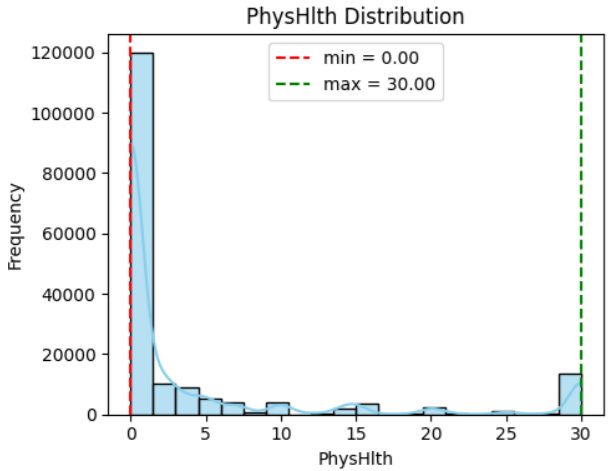
Preprocessing

Feature Type	Example Columns	Transformer
Binary	Other 13 Columns	None
Ordinal	GENHLTH, DIABETE, Age, Education, Income	Already ordinal encoded in original dataset
Continuous numeric	BMI, MentHlth, PhysHlth	MinMaxScaler

Preprocessing







Reasonable Boundary: [12,98] ➡ MinMaxScaler ⬅ MentHlth & PhysHlth lies between [0,30]

Thank you!

BROWN
Data Science Institute

References

Heart Disease Facts. (2024, October 24). Heart Disease. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html?utm_source=chatgpt.com

Heart Disease Health Indicators Dataset. (2022, March 10). Kaggle. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset