Predicting Heart Disease using Machine Learning

Tai Nguyen, Daren Saenz, Rakshita Dayal

Table of contents

INTRODUCTION

Problem Statement

- Heart disease is the top cause of death across most demographics in the United States.
- Every 33 seconds, cardiovascular disease claims a life in the U.S.
- In 2021, heart disease was responsible for one in every five U.S. deaths, totaling about 695,000.
- Early diagnosis is crucial to prevent sudden deaths from heart disease.

According to the CDC

INTRODUCTION

Proposed Solution

Create a prediction method for early detection

- Examine various factors that increase risk.
- Identify the core factors and correlations.
- Create a machine learning model to detect heart disease.

Goals

- Be able to detect heart attack risk.
- Identify main factors that cause heart attack.
- Developing preventative strategies.
- Increase public awareness on heart attack.

DATASETS

Data collection

- 2022 annual CDC survey data
- Collected from Kaggle
- Include more than 400,000 adults

Data Variables

- There are 24 variables
  - 4 numerical variables
  - 20 categorical variables
- Contains 246,022 rows
- 3,390 rows of data were eliminated due to duplicates
- There is no missing values in the data

DESCRIPTIVE ANALYSIS

Sample Size

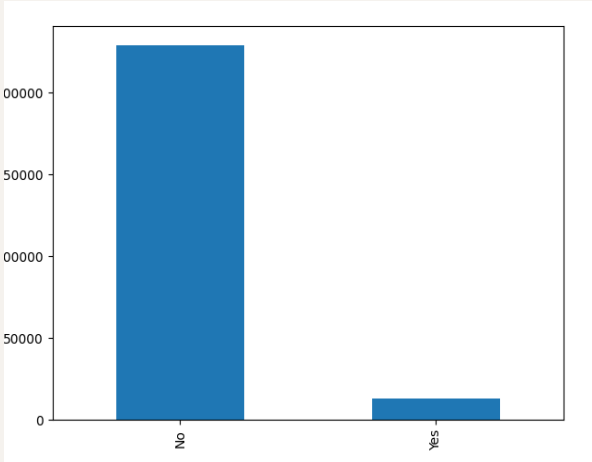
- The sample contains 240,649 rows after removing duplicates and outliers.
- 24 variable remains with 20 categorical and 4 numerical

Numerical Variables

Statistic	Physical Days	Mental Days	Sleep Hours	BMI
Mean	4.17	4.22	7.01	28.70
Standard Dev	8.4	8.14	1.44	6.54
Minimum	0	0	1	12.02
Maximum	30	30	24	97.65

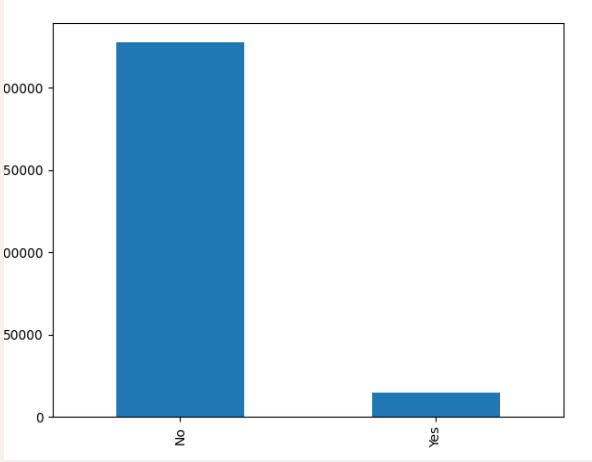
DESCRIPTIVE ANALYSIS

Important histograms



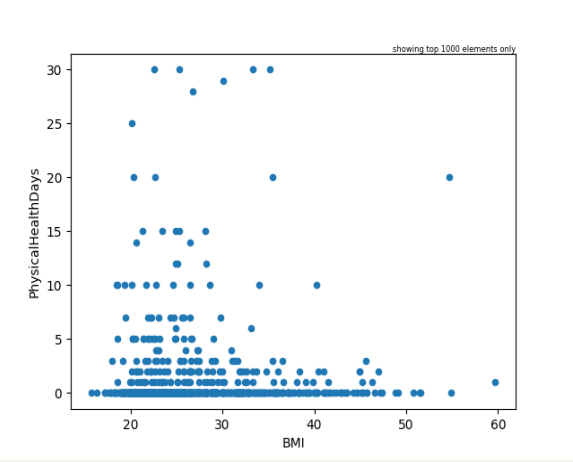
Heart Attack (dependant variable)

Data is imbalanced. 94.46% have no heart attack history



Angina

93.84% have not experienced angina, a major risk factor for heart attacks



BMI

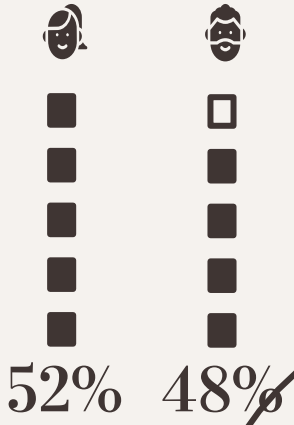
Patients with more physical health have a BMI between 20-40

DESCRIPTIVE ANALYSIS

Categorical Variables

Here is some other relevant statistic:

- Females vs males, 52.03% to 47.97%.
- Predominantly white 75.42%.
- Non-smokers make up 59.71%.
- Over half (54.71%) consume alcohol.
- Majority (77.48%) are physically active.
- COPD affects 7.83%, significant but less common.
- Age distribution is fairly spread out, with a slight increase towards older people



DESCRIPTIVE ANALYSIS

Numerical Variable Correlations
Using the correlation matrix for numerical variables

1.00	0.17	-0.03	0.07
0.17	1.00	-0.07	0.05
-0.03	-0.07	1.00	-0.05
0.07	0.05	-0.05	1.00

Exam The correlation matrix:

- Physical Health Days and Mental Health Days correlate moderately (0.17)
- Mental Health Days and Sleep Hours have a weak negative correlation (-0.7).
- Physical Health Days vs BMI: Weak positive correlation (0.07)

DESCRIPTIVE ANALYSIS

Categorical Variable Correlations
Using the Chi-square test, here is top 5 correlations:

- HadAngina - Extremely strong correlation, indicating angina-related conditions have significant predictive or associative power.
- GeneralHealth - Strong correlation, suggesting overall health significantly influences the outcome.
- AgeCategory - Strong correlation, showing age as a critical determinant.
- HadStroke - Strong correlation, highlighting stroke history as a major factor.
- DifficultyWalking - Substantial correlation, suggesting mobility issues significantly impact health outcomes.

Variable	Value
HadAngina	47795
GeneralHealth	9591
AgeCategory	7927
HadStroke	7517
DifficultyWalking	6070
HadDiabetes	5159
HadCOPD	4224
HadKidneyDisease	2838
SmokerStatus	2159
DifficultyDressingBathing	1630
PhysicalActivities	1615
Sex	1306
AlcoholDrinkers	1299
CovidPos	226
RaceEthnicityCategory	213
ECigaretteUsage	122
HadAsthma	122
HadDepressiveDisorder	120
HighRiskLastYear	117

DATA ANALYSIS

Method of Analysis

Since the dependent variable is classification (yes/no),
we will apply two methods:

- Logistic Regression
  - Draws a straight line to predict heart disease based on input features.
- Random Forest
  - Combines multiple decision trees to predict heart disease by taking a vote among them.

DATA ANALYSIS

Data preparation
Before applying the models, the data need to be ready for training steps

- Converted all categorical columns into numerical.
- All the numerical columns need to be scaled.
- All independent variables have to be converted into a single vector using VectorAssembler.

Data training

- Splitting the data to 70-30
- Building models
- Create a pipeline for modeling
- Model turning using Parameter Grid and Cross-Validation for better accuracy.

DATA ANALYSIS

Examine the result

- F1 Score (0.93): Indicates very high accuracy in identifying both actual cases of heart disease and those without the disease.
- Weighted Precision (0.93): Suggests that 93% of the model's predictions of heart disease are correct, minimizing false positives.
- Weighted Recall (0.95): Reflects that the model successfully identifies 95% of all true cases of heart disease, reducing the risk of missed diagnoses.
- Accuracy (0.95): Shows that the model correctly predicts the presence or absence of heart disease 95% of the time, demonstrating its reliability for clinical use.

F1 Score	93%
Precision	93%
Recall	95%
Accuracy	95%

SUMMARY OF FINDINGS

After applying the models, we have these findings:

- Using the models, we can accuracy predict heart attack risk more than 90%.
- There are strong correlations of heart attack with following variables:
  - People with Angina have the strongest correlation to heart attack.
  - One needs to maintain general health to reduce the risk.
  - People with old age tend to have more risk.
  - People who had stroke also can be in the risk zone.
  - Having difficulty in basic daily activities could be the indicate of heart attack.
  - Increase both physical and mental health will reduce risk of heart attack.

IMPLICATIONS

Application of models:

- Supporting Medical Institutions: Assist healthcare providers in early detection, leading to proactive treatment, increase productivity.
- Increasing Accessibility: Extend healthcare access to regions lacking adequate medical infrastructure.
- Identifying Correlations: Highlights factors like angina, age, and stroke history, aiding individuals in healthier living.
- Promoting Awareness: Educate individuals about heart disease risk factors, empowering them to take preventive action.
- Empowering Individuals: Encourages lifestyle changes and timely medical intervention. Physical therapy is vital for those with limited mobility.

LIMITATIONS

There are some limitation need to be considered:

- Data Privacy: as patient need to input data for risk calculation, data privacy need to be taken seriously.
- Data accuracy: training data need to be collected from reliable source to increase model accuracy.
- Model could be overfitting and need to be tested thoroughly.
- Business aspect: There are competitors in medical application such as Apple health, and Myfitnesspal.

Work Cited

Data Source: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Centers for Disease Control and Prevention. (2023, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/

Thank you for your attention!