Heart Disease analysis with classification

Authors: Lola Sõukand, Robin Mürk

Aim is to build a binary classifier that predicts a person's risk of a heart attack and to analyse individual attributes to determine what lowers and what increases the risk of heart attack using regression analysis.

Goal 1: train a model to predict whether or not a patient is at high risk or at low risk of a heart attack.
Goal 2: find factors, that lower the risk of heart attack
Goal 3: find factors, that increase the risk of heart attack

Gathering data

Data is from Kaggle (https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset). This dataset is originally from Centers for Disease Control and Prevention (https://www.cdc.gov/brfss/annual_data/annual_2015.html). The data is gathered from USA and the District of Columbia Guam and Puerto Rico by landline and cell phone questionner. The dataset used in this work is already cleaned for the Kaggle (https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook). The dataset is from 2015, we tried to get the newer dataset from CDC, but since they have changed the questions they asked, we decided to use the Kaggle version.

Describing data

Size:

- 22 columns
- 253680 rows

Dataset is CSV (22.74 MB)

The primary dataset is cleaned for heart disease analysis for Kaggle by Alex Teboul.

Columns:

- HeartDiseaseorAttack
  - Label
  - Binary
  - 1=yes
  - 0= no
- HighBP
  - Indicates if the person has been told by a health professional that they have High Blood Pressure
  - Binary
  - 1= yes
  - 0= no
- HighChol

    - Indicates if the person has been told by a health professional that they have High Blood Cholesterol
    - Binary
    - 1= yes
    - 0= no
- CholCheck
    - Cholesterol Check, if the person has their cholesterol levels checked within the last 5 years
    - Binary
    - 1= yes
    - 0= no
- BMI
    - Body Mass Index, calculated by dividing the persons weight (in kilogram) by the square of their height (in meters).
    - Numerical
- Smoker
    - Indicates if the person has smoked at least 100 cigarettes.
    - Binary
    - 1= yes
    - 0= no
- Stroke
    - Indicates if the person has a history of stroke.
    - Binary
    - 1= yes
    - 0= no
- Diabetes
    - Indicates if the person has a history of diabetes, or currently in pre-diabetes, or suffers from either type of diabetes.
    - Numerical
    - 2= type 2 diabetes
    - 1= type 1 diabetes
    - 0= no diabetes
- PhysActivity
    - Indicates if the person has some form of physical activity in their day-to-day routine.
    - Binary
    - 1= yes
    - 0= no
- Fruits
    - Indicates if the person consumes 1 or more fruit(s) daily.
    - Binary
    - 1= yes
    - 0= no
- Veggies
    - Indicates if the person consumes 1 or more vegetable(s) daily.

    - Binary
    - 1= yes
    - 0= no
- HvyAlcoholConsump
    - Indicates if the person has more than 14 drinks per week.
    - Binary
    - 1= yes
    - 0= no
- AnyHealthcare
    - Indicates if the person has any form of health insurance.
    - Binary
    - 1= yes
    - 0= no
- NoDocbcCost
    - Indicates if the person wanted to visit a doctor within the past 1 year but couldn't, due to cost.
    - Binary
    - 1= yes
    - 0= no
- GenHlth
    - Indicates the persons response to how well is their general health, ranging from 1 (excellent) to 5 (poor).
    - Categorical
    - 1= excellent
    - 2= very good
    - 3= good
    - 4= fair
    - 5= poor
- Menthlth
    - Indicates the number of days, within the past 30 days that the person had bad mental health.
    - Numerical
    - 1-30= number of days
- PhysHlth
    - Indicates the number of days, within the past 30 days that the person had bad physical health.
    - Numerical
    - 1-30= number of days
- DiffWalk
    - Indicates if the person has difficulty while walking or climbing stairs.
    - Binary
    - 1= yes
    - 0= no
- Sex

    - Indicates the gender of the person, where 0 is female and 1 is male.
    - Binary
    - 1= male
    - 0= female
- Age
    - Indicates the age class of the person, where 1 is 18 years to 24 years up till 13 which is 80 years or older, each interval between has a 5-year increment.
    - Categorical
    - 1= 18-24
    - 2=25-29
    - 3= 30-34
    - 4= 35-39
    - 5=40-44
    - 6=45-49
    - 7=50-54
    - 8=55-59
    - 9=60-64
    - 10=65-69
    - 11=70-74
    - 12=75-79
    - 13= 80+
- Education
    - Indicates the highest year of school completed, with 1 being never attended or kindergarten only and 6 being, having attended 4 years of college or more.
    - Categorical
    - 1= never attended school or only kindergarten
    - 2= grades 1-8
    - 3= grades 9-11
    - 4= grades 12 or GED
    - 5= collage 1- 3
    - 6= collage 4+
- Income
    - Indicates the total household income, ranging from 1 (at least $10,000) to 6 ($75,000+)
    - Categorical
    - 1= <10 000$
    - 2= <15 000$
    - 3= <20 000$
    - 4= <25 000$
    - 5= <35 000$
    - 6= <50 000$
    - 7= <75 000$
    - 8= >75 000$

Exploring data and verifying data quality

From exploring the data it seems that there are some correlesions with heart diseases and other values in the data, but nothing that is outstanding. There is no missing data.

The data quality seems good. There is no missing data and everything seems okay.