Predictive Analytics for Heart Disease Risk Based on Key Indicators

Ruihan Dou (rd593), Anqi Li (al2555)

1 Leading Question

Background: In 2023, heart disease ranked as the leading cause of death in the United States, according to CDC. The suddenness of heart disease and the potential for rapid death without timely treatment underscore the importance of early risk assessment. The CDC has identified several major contributors to heart disease, including high blood pressure, smoking, diabetes, alcohol drinking, high BMI, and various other factors. This project aims to quantify the impact of these factors on an individual's susceptibility to heart disease. Python will be utilized throughout the project for model development and data visualization.

Goal: This project aims to identify the key indicators contributing to heart disease risk and develop predictive models to determine whether an individual is vulnerable to a heart disease. In this project, we will conduct exploratory data analytics and feature engineering to investigate various indicators. Additionally, we will build classification models, including K-Nearest Neighbors (KNN), logistic regression, Support Vector Machine (SVM), and Neural Networks, as taught in ORIE 5741.

Significance: The significance of this project lies in its potential to address a critical public health issue and drive healthcare innovation. By developing accurate predictive models for measuring heart disease, this project contributes to the early detection and prevention of the leading cause of death in the United States. The actionable insights derived from the project will empower individuals to take proactive measures in managing their heart health, potentially reducing the burden on healthcare systems and saving lives. Moreover, the project's findings can be leveraged to develop innovative healthcare solutions, such as personalized health management insurances or targeted interventions, revolutionizing the approach to heart disease prevention.

2 Datasets and Variables (sustainability of the dataset)

The dataset originates from the CDC and forms a significant component of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts yearly telephone surveys to gather information on the health conditions of individuals residing in the United States. According to the CDC, BRFSS was established in 1984 with participation from 15 states, and it currently gathers data from all 50 states, the District of Columbia, and three U.S. territories. BRFSS conducts over 400,000 interviews with adults annually, making it the most extensive ongoing health survey system globally. The dataset includes more than 400,000 rows and 40 variables(columns). The variables are:

- State : The U.S. state where the individual resides.
- Sex : Gender of the individual (Male or Female).
- GeneralHealth : Self-reported general health status of the individual.
- PhysicalHealthDays : Number of days in the past 30 days that physical health was not good.
- MentalHealthDays : Number of days in the past 30 days that mental health was not good.
- LastCheckupTime : Time since the last routine checkup or health examination.
- PhysicalActivities : Frequency of engaging in physical activities or exercises.
- SleepHours : Average number of hours of sleep per night.
- RemovedTeeth : Number of permanent teeth removed due to dental issues.
- HadHeartAttack : Whether the individual has had a heart attack.
- HadAngina : Whether the individual has experienced angina (chest pain or discomfort).
- HadStroke : Whether the individual has had a stroke.
- HadAsthma : Whether the individual has had asthma.
- HadSkinCancer : Whether the individual has had skin cancer.
- HadCOPD : Whether the individual has had Chronic Obstructive Pulmonary Disease (COPD).
- HadDepressiveDisorder : Whether the individual has had a depressive disorder.
- HadKidneyDisease : Whether the individual has had kidney disease.
- HadArthritis : Whether the individual has had arthritis.
- HadDiabetes : Whether the individual has had diabetes.
- DeafOrHardOfHearing : Whether the individual is deaf or hard of hearing.
- BlindOrVisionDifficulty : Whether the individual has blindness or vision difficulty.
- DifficultyConcentrating : Self-reported difficulty in concentrating.
- DifficultyWalking : Self-reported difficulty in walking.
- DifficultyDressingBathing : Self-reported difficulty in dressing or bathing.
- DifficultyErrands : Self-reported difficulty in running errands.
- SmokerStatus : Current smoking status (smoker, former smoker, non-smoker).
- ECigaretteUsage : Whether the individual uses e-cigarettes.
- ChestScan : Whether the individual has had a chest scan.
- RaceEthnicityCategory : Categorized race or ethnicity of the individual.
- AgeCategory : Categorized age group of the individual.
- HeightInMeters : Height of the individual in meters.
- WeightInKilograms : Weight of the individual in kilograms.
- BMI : Body Mass Index calculated from height and weight.
- AlcoholDrinkers : Whether the individual consumes alcohol.
- HIVTesting : Whether the individual has undergone HIV testing.
- FluVaxLast12 : Whether the individual received a flu vaccine in the last 12 months.
- PneumoVaxEver : Whether the individual has ever received a pneumonia vaccine.
- TetanusLast10Tdap : Time since the last tetanus vaccination (in the last 10 years, received Tdap).
- HighRiskLastYear : Whether the individual has been considered at high risk for the past year.
- CovidPos : Whether the individual tested positive for COVID-19.

Variable "HadHeartAttack" is our response variable, and we will treat it as binary ("Yes" - respondent had heart disease; "No" - respondent did not have heart disease). Other variables are predictor variables, utilized to predict the outcome of the response variable.

3 Conclusion (overall why it would succeed)

The dataset originates from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect data on the health status of U.S. residents. This is a reliable and established source of data collection, ensuring continuity and reliability over time. Meanwhile, The dataset has undergone treatments to select the most relevant variables related to heart disease, reducing the original set of nearly 300 variables to 40.

The dataset is derived from the 2022 annual CDC survey data of over 400k adults. The regularity of data collection on an annual basis allows for ongoing updates and ensures the dataset remains current and relevant to evolving health trends and patterns.

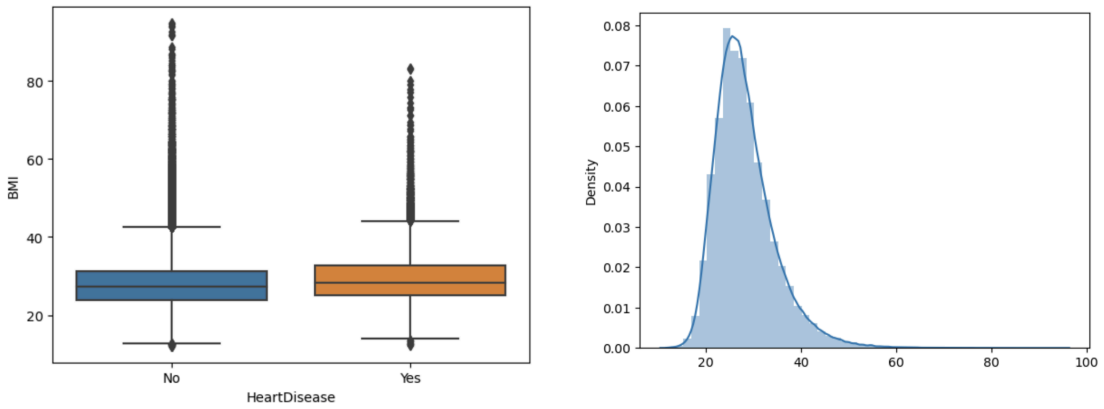
The dataset offers opportunities for various analyses and applications, including exploratory data analysis (EDA) and the application of machine learning methods for predictive modeling, such as logistic regression, SVM, and random forest. The availability of both balanced and imbalanced versions of the dataset allows for flexibility in modeling approaches.

4 Overview of the Dataset

A thorough examination of the dataset confirms its appropriateness and utility for our project. The dataset comprises four columns with floating-point data types: BMI, PhysicalHealth, MentalHealth, and SleepTime. An initial statistical analysis of these variables is presented below.

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319795.000000	319795.00000	319795.000000	319795.000000
mean	28.325399	3.37171	3.898366	7.097075
std	6.356100	7.95085	7.955235	1.436007
min	12.020000	0.00000	0.000000	1.000000
25%	24.030000	0.00000	0.000000	6.000000
50%	27.340000	0.00000	0.000000	7.000000
75%	31.420000	2.00000	3.000000	8.000000
max	94.850000	30.00000	30.000000	24.000000

The left graph delineates the distribution of BMI across individuals, segmented by the presence of heart disease. It employs box-and-whisker plots to convey the dispersion and skewness of BMI values within each group. Adjacent to this, the right graph features a density plot, providing a visual interpretation of the distribution's shape and spread. The curve's peak and tails offer insights into the central tendency and variance in the dataset.



This preliminary exploration lays the groundwork for further analysis. The dataset contains enough dimension of features we need for the classification, and also, data in each feature are well distributed and informative, which is supportive and sustainable for our research.