Cardiovascular Health: A Study of Factors Influencing Heart Disease Risk

George Mason University
STAT-515-001 | Prof. Dr. Tokunbo Fadahunsi, PhD.

Team Members:

Lakshmi Durga Teratipally – G01448583
Anusha Goulla – G01452111
Sai Charan Somineni – G01447313

I. Introduction

We have chosen to analyze a Cardiovascular dataset sourced from a medical survey, offering a wide range of health measurements and lifestyle markers for a diverse group of people. It incorporates key attributes like age, gender, height, weight, blood pressure readings (both systolic and diastolic), cholesterol levels, glucose levels, as well as lifestyle habits such as smoking, alcohol consumption, and physical activity. Each entry presents a snapshot of an individual's health profile, offering insights into potential correlations between these factors and the presence or absence of cardiovascular disease. With parameters like cholesterol and glucose levels categorized into multiple levels and lifestyle habits dichotomized as present or absent, this dataset offers a rich ground for exploratory analysis. By scrutinizing this dataset, we can potentially unearth patterns, associations, and risk factors contributing to cardiovascular health, aiding in the formulation of preventive strategies or targeted interventions for managing and mitigating cardiovascular diseases.

Our primary aim in selecting this dataset is to construct a predictive model that effectively anticipates the occurrence of cardiovascular disease. To achieve this, we'll delve into the dataset's diverse factors, analyzing their relevance and impact on predicting the presence or absence of this disease. By identifying the most influential elements among these health metrics and lifestyle indicators, we can build a robust model that accurately forecasts the likelihood of cardiovascular disease in individuals. This model holds significant potential in proactively identifying individuals at risk and implementing targeted interventions or preventive strategies, potentially making a substantial impact on public health initiatives and individual well-being.

II. Dataset

The dataset contains 5 numerical and 7 categorical variables capturing various demographic, health-related, and lifestyle factors that are relevant to cardiovascular health analysis or prediction. This is a multivariate dataset with ‘cardio’ as the target variable/ dependent variable. This dataset offers an opportunity to delve into factors influencing cardiovascular health and potentially derive actionable insights to improve prevention, diagnosis, and management of heart related conditions.

age	gender	height	weight	ap_hi	ap_lo	cholesterc	gluc	smoke	alco	active	cardio
50	2	168	62	110	80	1	1	0	0	1	0
55	1	156	85	140	90	3	1	0	0	1	1
51	1	165	64	130	70	3	1	0	0	0	1
48	2	169	82	150	100	1	1	0	0	1	1
47	1	156	56	100	60	1	1	0	0	0	0
59	1	151	67	120	80	2	2	0	0	0	0
60	1	157	93	130	80	3	1	0	0	1	0
61	2	178	95	130	90	3	3	0	0	1	1
48	1	158	71	110	70	1	1	0	0	1	0
54	1	164	68	110	60	1	1	0	0	0	0
61	1	169	80	120	80	1	1	0	0	1	0
51	2	173	60	120	80	1	1	0	0	1	0
40	2	165	60	120	80	1	1	0	0	0	0
54	1	158	78	110	70	1	1	0	0	1	0
39	2	181	95	130	90	1	1	1	1	1	0
45	2	172	112	120	80	1	1	0	0	0	1
58	1	170	75	130	70	1	1	0	0	0	0
45	1	158	52	110	70	1	3	0	0	1	0
47	1	154	68	100	70	1	1	0	0	0	0

Figure 1: Dataset

Column	Description
age	Age of the individual in years.
gender	Gender of the individual (1 for female, 2 for male).
height	Height of the individual in centimeters.
weight	Weight of the individual in kilograms.
ap_hi	Systolic blood pressure (the higher number in a reading).
ap_lo	Diastolic blood pressure (the lower number in a reading).
cholesterol	Cholesterol level category: 1 for normal, 2 for above normal, 3 for well above normal.
gluc	Glucose level category: 1 for normal, 2 for above normal, 3 for well above normal.
smoke	Indicates smoking status: 0 for non-smoker, 1 for smoker.
alco	Indicates alcohol consumption: 0 for non-drinker, 1 for drinker.
active	Indicates physical activity: 0 for inactive, 1 for active.
cardio	Denotes presence or absence of cardiovascular disease: 0 for absence, 1 for presence.

Table 1: Variable Description

III. Research Questions

Dataset: *Cardiovascular Health*

After studying the data set, we have formulated the below research questions.

- Is there any gender-based differences in the distribution of cardiovascular disease?
- What is the relationship between systolic blood pressure and cardiovascular disease?
- How does diastolic blood pressure relate to the likelihood of having cardiovascular disease?
- Which attributes or factors have the most significant impact on predicting the likelihood of cardiovascular disease?
- How do different health attributes contribute to the likelihood of developing cardiovascular disease?

The above research questions can be answered by studying and analyzing the dataset.

IV. Data Cleaning and Transformation

We have used the Pandas library in Python for cleaning and preprocessing the cardiovascular dataset. Firstly, rows containing missing values were removed to ensure data integrity and reliability. Duplicate rows were eliminated to maintain dataset consistency. The 'age' column was

converted to an integer data type for uniformity and ease of analysis. Additionally, extreme values in the 'ap_hi' (systolic blood pressure) and 'ap_lo' (diastolic blood pressure) columns were filtered out, aiming to mitigate outliers that might adversely impact subsequent analyses. Through these steps, the dataset was refined, ensuring a more robust and accurate foundation for further analysis or modeling related to cardiovascular health.

V. Methodology

In this study our main goal is to thoroughly examine the dataset, utilize visualizations to understand its patterns, and construct predictive models that effectively forecast the occurrence of cardiovascular disease.

Research question 1:

Is there any gender-based differences in the distribution of cardiovascular disease?

To address our research query, we've constructed a grouped bar chart illustrating the frequency of cardiovascular disease occurrence categorized by age groups and gender. The age brackets are segmented into 5-year intervals, and we've employed distinct color codes to represent different genders within the visualization.

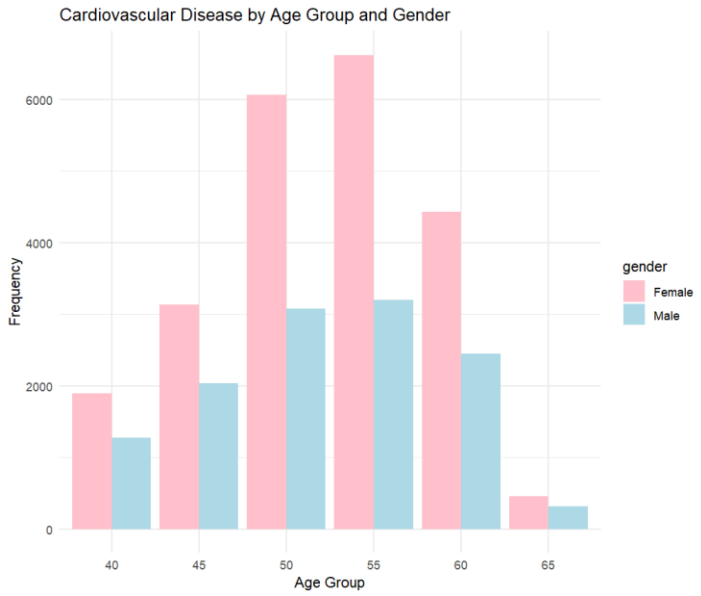

Figure 2: Grouped Bar Chart of Cardiovascular Disease by age group and gender.

In the figure 2 the graph indicates a rise in cardiovascular disease occurrences with age for both men and women. Comparatively, females exhibit a higher likelihood of experiencing cardiovascular diseases across all age groups. Notably, women between the ages of 50 and 55 display particularly elevated chances of developing CVD. The data also suggests a notably lower

incidence of cardiovascular disease among individuals at the age of 65. In conclusion, significant variations based on gender are evident in the occurrence of cardiovascular diseases.

Research question 2:

What is the relationship between systolic blood pressure and cardiovascular disease?

In response to our research query, we've created a box plot focusing on the systolic blood pressure (ap_hi) distribution among individuals diagnosed with CVD. This visualization aids in grasping the typical range, variation, and identification of possible extreme values in systolic blood pressure readings within the group affected by cardiovascular disease.

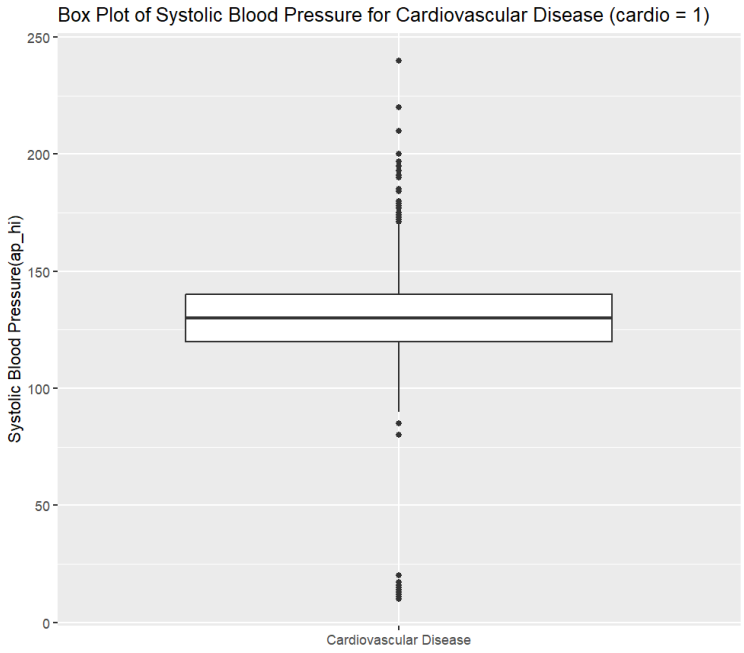

Figure 3: Boxplot of Systolic Blood Pressure for Cardiovascular Disease

The plot illustrates that individuals diagnosed with cardiovascular disease tend to have a higher median systolic blood pressure compared to those without the disease. Additionally, outliers in the plot represent individuals with systolic blood pressure values significantly diverging from the median, possibly indicating distinct medical conditions, or influencing factors. In essence, the plot suggests a positive relationship between Systolic Blood Pressure and Cardiovascular Disease, implying that higher systolic blood pressure values are associated with an increased likelihood of CVD occurrence.

For a more in-depth analysis, we've conducted a comparison between two plots: one that includes outliers and another that excludes them. This comparison aims to explore the impact of outliers on the visualization and gain a deeper understanding of the data's distribution and characteristics.

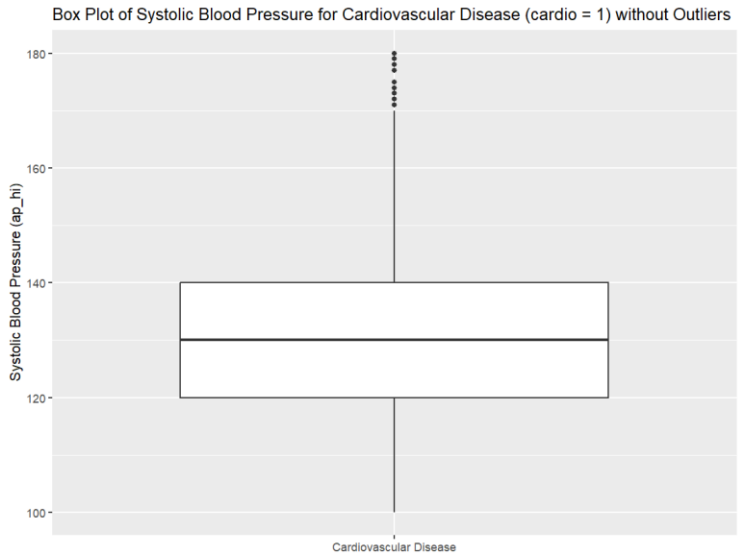

Figure 4: Box plot of Systolic Blood Pressure for Cardiovascular Disease without outliers

The revised graph, post removal of outliers, reveals a notable trend. The observation suggests a specific range of systolic blood pressure, between 120mmHg and 140mmHg, as associated with an increased likelihood of experiencing cardiovascular disease (CVD). This range appears to indicate a critical zone wherein individuals may have a higher susceptibility to CVD occurrence.

Research question 3:

How does diastolic blood pressure relate to the likelihood of having cardiovascular disease?

To answer the question, we have generated a box plot representing the distribution of diastolic blood pressure (ap_lo) among individuals diagnosed with cardiovascular disease (cardio = 1).

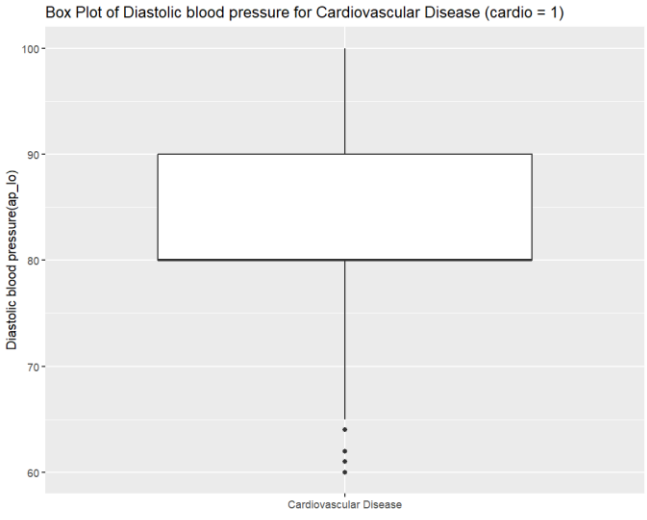

Figure 5: Box plot of Diastolic Blood Pressure for Cardiovascular Disease

From figure 5, it appears that individuals displaying diastolic blood pressure levels falling within the range of 80mmHg to 90mmHg demonstrate an increased likelihood of experiencing cardiovascular disease. The outliers shown in the graph might represent extreme diastolic blood pressure measurements, potentially indicating unusual or exceptional cases. Their presence could hint at unique medical conditions or other factors influencing blood pressure.

Research question 4:

Which attributes or factors have the most significant impact on predicting the likelihood of cardiovascular disease?

We approached this research question by splitting the dataset into distinct training and testing subsets. The training data was used to create a logistic regression model aimed at predicting cardiovascular disease occurrence. Factors such as age, gender, physical characteristics (height, weight), blood pressure metrics (ap_hi and ap_lo), cholesterol, and glucose levels were considered as predictors. Additionally, lifestyle factors like alcohol consumption, smoking habits, and physical activity were included in the model. The summary of the fitted model provided insights into the significance and impact of these factors, aiding in understanding their association with the likelihood of cardiovascular disease.

```
Call:
glm(formula = cardio ~ age + gender + height + weight + ap_hi +
    ap_lo + cholesterol + gluc + smoke + alco + active, family = binomial(link = "logit"),
    data = train_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.122e+01  2.778e-01 -40.386  < 2e-16 ***
age          5.122e-02  1.613e-03  31.745  < 2e-16 ***
gender      -4.829e-03  2.614e-02  -0.185 0.853461
height      -3.807e-03  1.538e-03  -2.476 0.013283 *
weight       1.176e-02  8.262e-04  14.233  < 2e-16 ***
ap_hi        4.836e-02  1.076e-03  44.961  < 2e-16 ***
ap_lo        2.275e-02  1.786e-03  12.737  < 2e-16 ***
cholesterol  5.075e-01  1.872e-02  27.113  < 2e-16 ***
gluc        -1.302e-01  2.107e-02  -6.178 6.50e-10 ***
smoke       -1.544e-01  4.177e-02  -3.697 0.000218 ***
alco        -2.049e-01  5.078e-02  -4.036 5.45e-05 ***
active      -2.223e-01  2.607e-02  -8.528  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65992  on 47614  degrees of freedom
Residual deviance: 53991  on 47603  degrees of freedom
AIC: 54015

Number of Fisher Scoring iterations: 4
```

Figure 6: Summary of Logistic Regression Model

According to the summary in figure 6, Age, weight, blood pressure, cholesterol, and lifestyle factors like smoking, alcohol consumption, and physical activity seem to significantly influence

the likelihood of cardiovascular disease, as indicated by their significant coefficients ($p < 0.05$). However, gender might not play a significant role in predicting cardiovascular disease based on this model (p-value > 0.05).

We have utilized the 'pROC' package in R to generate and plot a Receiver Operating Characteristic (ROC) curve for evaluating the performance of a predictive model. The ROC curve visualizes the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different thresholds of a predictive model.

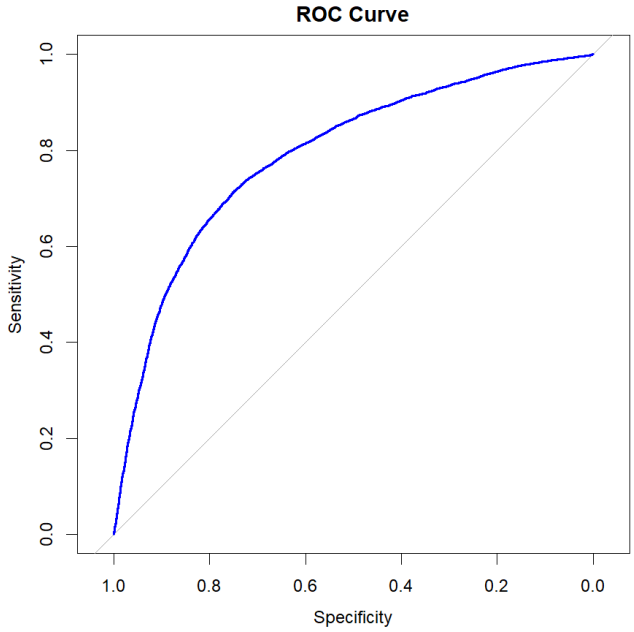

Figure 7: ROC

The higher area under the ROC curve in figure 7 indicates better model performance in distinguishing between classes (Individuals with and without CVD).

Research question 5:

How do different health attributes contribute to the likelihood of developing cardiovascular disease?

We have developed a Shiny web app to predict cardiovascular disease risk by leveraging a logistic regression model. Users can input personal health details, enabling the app to estimate the likelihood of the disease by computing a risk percentage based on the model's predictions as shown in figure 8.

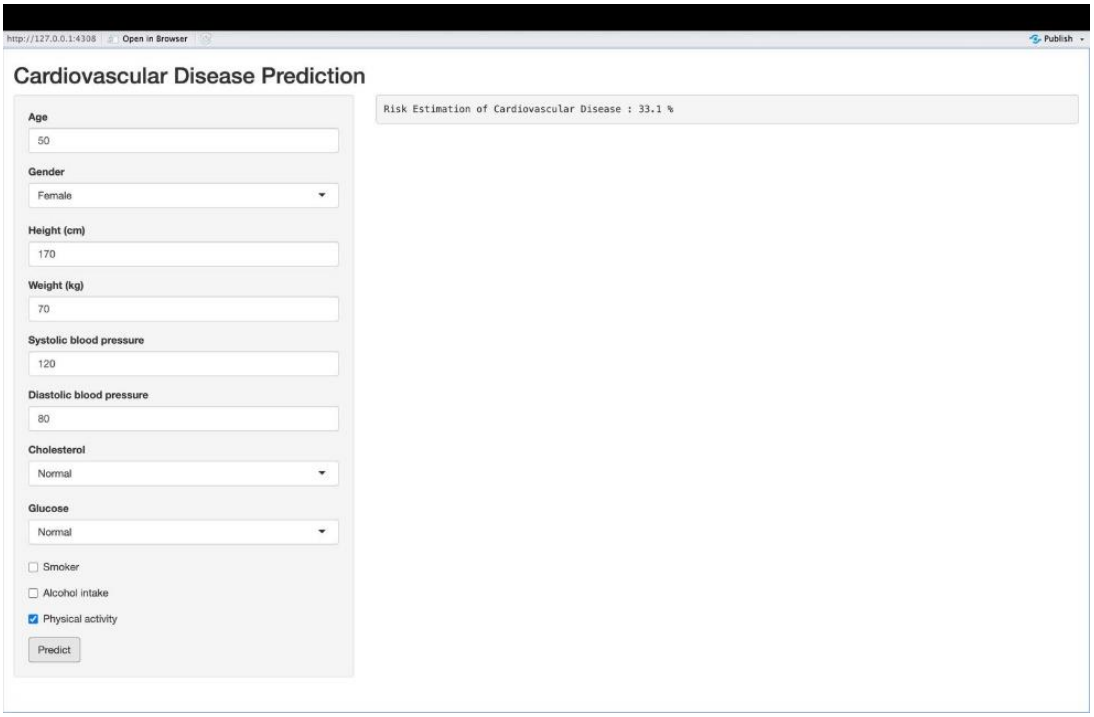

Figure 8: Shiny application for predicting cardiovascular disease.

In summary, the developed application offers an accessible and practical means for individuals to estimate their CVD risk, fostering awareness and encouraging proactive health management.

VI. Conclusion

In conclusion this analysis offers critical insights into cardiovascular disease (CVD) by identifying influential factors and their relationships. Understanding gender disparities, age-related risk changes, and the impact of blood pressure on CVD likelihood provides targeted areas for preventive healthcare interventions. The predictive model's accuracy in assessing CVD risk enables proactive identification of high-risk individuals. Moreover, the user-friendly Shiny app empowers individuals to estimate their CVD risk, fostering health awareness and encouraging proactive health management. Overall, these findings contribute significantly to public health strategies, enabling tailored interventions for at-risk populations and promoting individual well-being through early detection and preventive measures.

References

Cardiovascular Disease dataset. (2019, January 20).
Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset