Predictive Analysis of Cardiovascular Diseases

***Abstract*—Cardiovascular disease is one of the most serious illnesses. Nearly 17 million fatalities worldwide are a result of its high death rate. Early diagnosis enables prompt treatment of the illness to reduce death. The presence and absence of the disease can be classified using a variety of machine learning and deep learning techniques. In this study, the UCI dataset is used to categorize heart diseases using Logistic Regression (LR) approaches. The purpose of this research is to investigate the relationship between several potential risk factors and cardio- vascular illnesses. My ultimate goal is to construct a model to forecast cardiovascular illness, not to identify the elements that actually cause cardiovascular disorders.**

I. INTRODUCTION

IN 19th century, cardiovascular disease is still considered as incurable even though he pathology had been established. The reason was the invisible enemy, bacteria. In order to cure the patient, operation must be performed to cut off cytopathic parts; however, most people did not endure the operation. It was only after the invention of penicillin that people realized the reason behind all those death was infection by bacteria and decreased the death rate tremendously by applying antibiotics. However, cardiovascular disease can lead to other disease and the operation is still expensive to most moderate class family. Thus, cardiovascular disease is dreadful even it can be cured. One relieving aspect is that, cardiovascular disease can be prevent if find early. Compared to the huge amount of money and precious time involved, a visit to clinic for every six months takes much less time and money. Such simple math seems to be more complicated just like insurance. Insurance spreads misfortune among all those purchasing the insurance. Most people purchase insurance because they never know if misfortune would befall on themselves. Despite such logic, large amount of people do not purchase insurance thinking they would never be involved, which is proved wrong by many incidences. In the case of cardiovascular disease, similarly,

II. MATERIALS

In order to improve the precision of cardiovascular disease analysis and forecasting, a significant volume of data is crucial for training the model. Augmented variables are essential to enhance predictions, as they furnish deeper insights. Moreover, a broader dataset enhances model efficacy by providing a greater array of observations. Hence, the ideal dataset would encompass a substantial quantity of both observations and variables.

The inverse relationship between the quantity of observations and the number of variables presents limitations on the feasible dataset. Since variables contribute to cardiovascular disease, which is inherently personal, many are considered privacy related. Every individual in the analysis must possess information for all variables, making it challenging to acquire numerous observations. In linear regression analysis, the formula for predicting values relies on all predictors being present. Even a single missing value renders it ineffective. Therefore, introducing a new variable can contract the dataset due to missing or mismatched observations, despite the new variable's size tripling that of the current dataset. A decrease in variables means an increase in observations, impacting precision inversely. Given the negative correlation between observations and variables, I conclude that the optimal dataset lies in finding a compromise. To balance variables and observations, I choose to include only minimally related variables.

Before considering the selection of variable, we must mention the source we used. We used the database from Cleveland Clinic Foundation in Switzerland. The dataset used in this project is part of a database contains 14 features from Cleveland Clinic Foundation for heart disease. We have 303 rows of people data with 13 continuous observations of different symptoms. But in this Research we have considered only 7 factors which are Age, Cholesterol, Blood Pressure, Heart Rate, ST_Depression, Chest Pain, CVD Status.

In this study, we look into different classic machine learning models, and their discoveries in diseases risks. We have developed two algorithms using linear regression and decision trees, on Cleveland dataset.

Age: Age significantly impacts cardiovascular disease (CVD) onset and progression, with risks increasing over time due to physiological changes like arterial stiffening and heightened susceptibility to atherosclerosis. Older age correlates with a higher incidence of traditional CVD risk factors such as hypertension, dyslipidemia, diabetes, and obesity, exacerbating overall risk. Lifestyle behaviours like decreased physical activity and poor dietary choices also contribute to CVD development. Age serves as a confounding factor in CVD research, necessitating age-specific approaches for prevention, diagnosis, and management.

Cholesterol: Elevated cholesterol, notably LDL, significantly raises the risk of heart disease and complications by promoting plaque buildup in arteries, causing atherosclerosis and potential blockages. Managing cholesterol through lifestyle changes and medication is crucial for reducing cardiovascular disease risk and maintaining heart health.

Blood Pressure: Elevated blood pressure, or hypertension, substantially raises the risk of cardiovascular issues like heart disease, stroke, and heart failure. It strains the heart and blood vessels, causing damage over time. Monitoring and managing

blood pressure through lifestyle changes and medication are critical for preventing cardiovascular disease.

Heart Rate: Heart rate's association with cardiovascular disease is complex. Elevated rates may indicate or result from cardiovascular issues, influenced by age, fitness, and health. While higher rates might suggest risk, they can also reflect the body's response to strain. Conversely, lower rates don't guarantee protection. Thus, interpretation within individual contexts is crucial.

ST Depression: ST depression, an ECG measure of myocardial ischemia, is pivotal in cardiovascular diagnosis. Elevated levels correlate with heightened risk and severity of coronary artery disease and adverse cardiac events. Its presence, especially during stress testing, reflects compromised cardiac function and aids in predicting future cardiovascular events, guiding treatment decisions.

Chest Pain: Chest pain is a crucial indicator of cardiovascular disease. Often signalling underlying cardiac issues like angina or heart attack, its presence, severity, and accompanying symptoms are vital for diagnosis and assessing cardiovascular risk. Early detection and understanding aid in effective treatment, impacting patient outcomes significantly.

III. METHODOLOGY

Overview of Statistical Tests

Shapiro-Wilk Test:

This test determines whether a variable has a normal distribution. In the domain of CVD prediction, normality is significant for specific statistical models since it influences the analytic method. Variables with normal distributions may have an impact on the accuracy of CVD prediction models. Use in the code: This test is performed on all variables (excluding the target variable) to determine whether they are regularly distributed.

The Wilcoxon Signed-Rank Test

It is a non-parametric test that compares the medians of two groups, such as those with and without CVD. In the context of predicting CVD: It is useful to determine whether there is a substantial variation in the distribution of a certain predictor between the two groups. Use in the code: The code employs this test to compare each predictor to the target variable (CVD_Status). A significant p-value indicates that the predictor may be useful for predicting CVD.

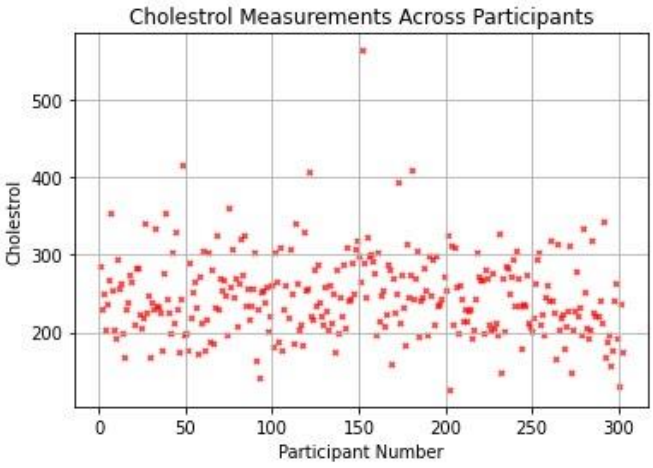
The Kruskal-Wallis Test

It is a non-parametric test used to compare multiple groups. In terms of forecasting CVD, it aids in determining whether there are substantial differences across several levels of a predictor for the occurrence of CVD. Use in the code: The function runs this test on each prediction that contains the target variable (CVD_Status). A significant p-value suggests that the predictor may have varying effects on CVD at different levels.

Chi-Squared Test:

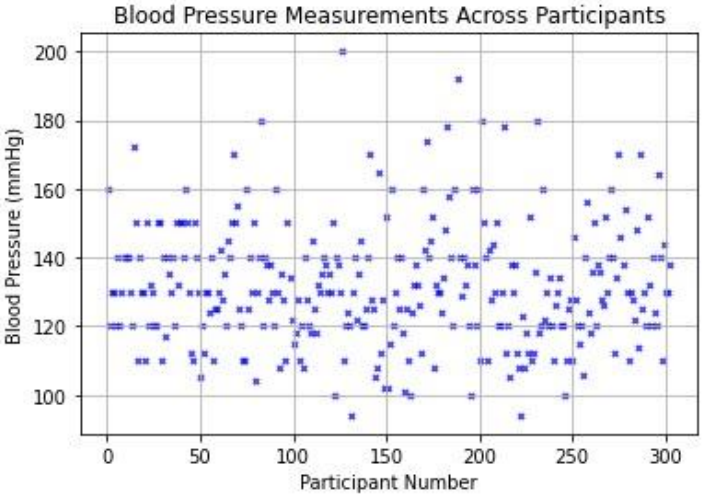
It determines connection between categorical variables. In the context of predicting CVD: It aids in determining whether there is a link between categorical predictors (e.g., chest pain level) and the occurrence of CVD. Use in the code: This test is performed on each categorical predictor and target variable (CVD_Status). A significant p-value indicates a strong connection between the predictor and CVD.

IV. VISUAL REPRESENTATION



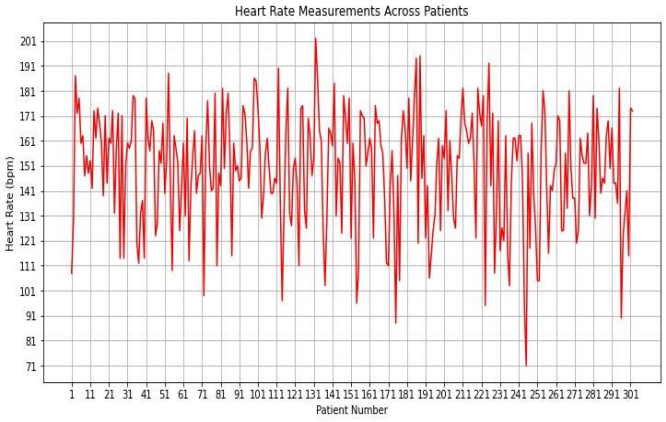
Cholesterol Measurements Across Participants :- The x-axis is labeled "Participant Number" and goes from 0 to 303. The y-axis is labeled "Cholesterol" and goes from 0 to 500.

It appears to be a simple plot of cholesterol measurements for a group of 300 participants.



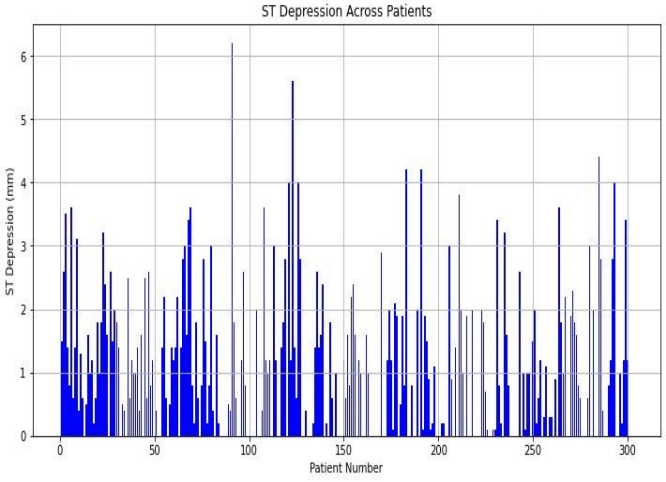
Blood Pressure Measurements Across Participants . It is a graph that shows the blood pressure readings of multiple participants.

The x-axis is labeled "Participant Number" and goes from 0 to 303. The y-axis is labeled "Blood Pressure (mmHg)" and goes from 100 to 200.



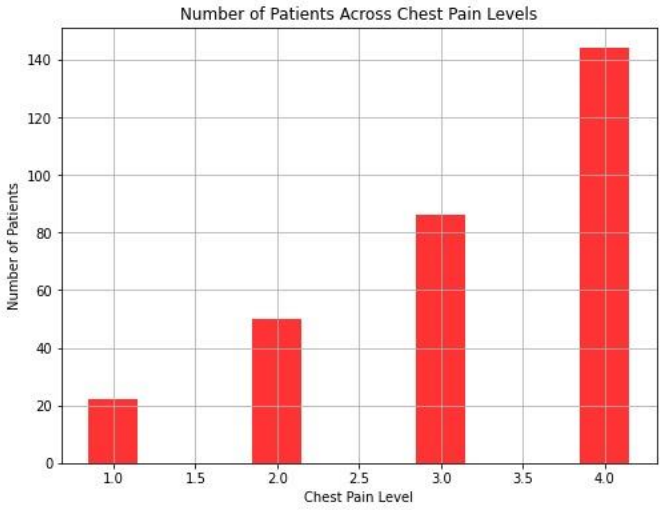
Graph showing the heart rates of the various patients. "Patient Number" is the label on the x-axis, which runs from 1 to 291. "Heart Rate (bpm)" is the label on the y-axis, which extends from 71 to 201.

It looks to be a straightforward plot of 303 patients' heart rate readings.



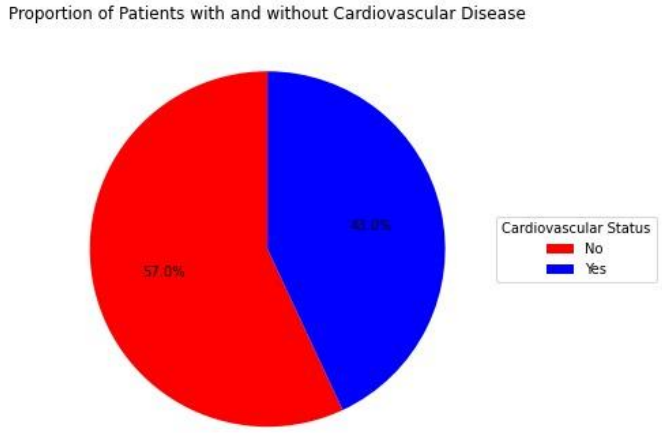
Graph showing the "ST Depression Across Patients" on the Y-axis Vs "Patient Number" on the X-axis.

The patient count goes from 0 to 303 patients.ST Depression also ranges from 0 to more than 6.0.



Number of Patients Across Chest Pain Levels :-"Chest Pain Level" is the label on the x-axis, which ranges from 1.0 to 4.0. "Number of Patients" is the label on the y-axis, which runs from 0 to 140.

The graph indicates that as chest discomfort intensity rises, so does the number of patients experiencing it. More patients than any other level have a chest pain score of 4.0.



Proportion of Patients with and Without Cardiovascular Disease .

As can be seen from the pie chart, 43.0% of patients have cardiovascular disease and 57.0% of patients do not.

V. MACHINE LEARNING MODEL

TARGET_LABEL <- "CVD_Status"

CORRELATIONS_PLOTS <- FALSE

patients_data <- data.frame(

Age = c(63, 67, 67, 37, 41, 56, 62, 57, 63, 53, 57, 56, 56, 44, 52, 57, 48, 54, 48, 49, 64, 58, 58, 58, 60, 50, 58, 66, 43, 40, 69, 60, 64, 59, 44, 42, 43, 57, 55, 61, 65, 40, 71, 59, 61, 58, 51, 50, 65, 53, 41, 65, 44, 44, 60, 54, 50, 41, 54, 51, 51, 46, 58, 54, 54, 60, 60, 54, 59,

```
46, 65, 67, 62, 65, 44, 65, 60, 51, 48, 58, 45, 53, 39, 68, 52, 44,
47, 53, 53, 51, 66, 62, 62, 44, 63, 52, 59, 60, 52, 48),

Cholesterol = c(233, 286, 229, 250, 204, 236, 268, 354, 254,
203, 192, 294, 256, 263, 199, 168, 229, 239, 275, 266, 211, 283,
284, 224, 206, 219, 340, 226, 247, 167, 239, 230, 335, 234, 233,
226, 177, 276, 353, 243, 225, 199, 302, 212, 330, 230, 175, 243,
417, 197, 198, 177, 290, 219, 253, 266, 233, 172, 273, 213, 305,
177, 216, 304, 188, 282, 185, 232, 326, 231, 269, 254, 267, 248,
197, 360, 258, 308, 245, 270, 208, 264, 321, 274, 325, 235, 257,
216, 234, 256, 302, 164, 231, 141, 252, 255, 239, 258, 201,
222),

 Blood_Pressure = c(145, 160, 120, 130, 130, 120, 140, 120,
130, 140, 140, 140, 130, 120, 172, 150, 110, 140, 130, 130, 110,
150, 120, 132, 130, 120, 120, 150, 150, 110, 140, 117, 140, 135,
130, 140, 120, 150, 132, 150, 150, 140, 160, 150, 130, 112, 110,
150, 140, 130, 105, 120, 112, 130, 130, 124, 140, 110, 125, 125,
130, 142, 128, 135, 120, 145, 140, 150, 170, 150, 155, 125, 120,
110, 110, 160, 125, 140, 130, 150, 104, 130, 140, 180, 120, 140,
138, 128, 138, 130, 120, 160, 130, 108, 135, 128, 110, 150, 134,
122),

 Heart_Rate = c(150, 108, 129, 187, 172, 178, 160, 163, 147,
155, 148, 153, 142, 173, 162, 174, 168, 160, 139, 171, 144, 162,
160, 173, 132, 158, 172, 114, 171, 114, 151, 160, 158, 161, 179,
178, 120, 112, 132, 137, 114, 178, 162, 157, 169, 165, 123, 128,
157, 152, 168, 140, 153, 188, 144, 109, 163, 158, 152, 125, 142,
160, 131, 170, 113, 142, 155, 165, 140, 147, 148, 163, 99, 158,
177, 151, 141, 142, 180, 111, 148, 143, 182, 150, 172, 180, 156,
115, 160, 149, 151, 145, 146, 175, 172, 161, 142, 157, 158,
186),

  ST_Depression = c(2.3, 1.5, 2.6, 3.5, 1.4, 0.8, 3.6, 0.6, 1.4, 3.1,
0.4, 1.3, 0.6, 0, 0.5, 1.6, 1, 1.2, 0.2, 0.6, 1.8, 1, 1.8, 3.2, 2.4, 1.6,
0, 2.6, 1.5, 2, 1.8, 1.4, 0, 0.5, 0.4, 0, 2.5, 0.6, 1.2, 1, 1, 1.4, 0.4,
1.6, 0, 2.5, 0.6, 2.6, 0.8, 1.2, 0, 0.4, 0, 0, 1.4, 2.2, 0.6, 0, 0.5, 1.4,
1.2, 1.4, 2.2, 0, 1.4, 2.8, 3, 1.6, 3.4, 3.6, 0.8, 0.2, 1.8, 0.6, 0, 0.8,
2.8, 1.5, 0.2, 0.8, 3, 0.4, 0, 1.6, 0.2, 0, 0, 0, 0, 0.5, 0.4, 6.2, 1.8,
0.6, 0, 0, 1.2, 2.6, 0.8, 0),

  Chest_Pain = c(1, 4, 4, 3, 2, 2, 4, 4, 4, 4, 4, 2, 3, 2, 3, 3, 2, 4, 3,
2, 1, 1, 2, 3, 4, 3, 3, 1, 4, 4, 1, 4, 3, 4, 3, 4, 4, 4, 4, 3, 4, 1, 2, 3, 4,
3, 3, 4, 3, 3, 2, 4, 4, 2, 4, 4, 3, 4, 3, 1, 4, 3, 4, 3, 4, 4, 3, 3, 4, 3, 3,
4, 4, 4, 4, 3, 4, 3, 2, 4, 4, 4, 3, 3, 2, 3, 3, 3, 4, 3, 4, 4, 3, 3, 3, 4, 4,
4, 2, 4),

  CVD_Status = c(0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0,
0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0,
1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1,
0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1,
1, 0, 0, 0)

)

new_person_data <- data.frame(

  Age = 20,

  Cholesterol = 95,

  Blood_Pressure = 95,

  Heart_Rate = 72,

  ST_Depression = 0,

  Chest_Pain = 1,

  CVD_Status = NA

)



all_data <- rbind(patients_data, new_person_data)

alltests <- list(

  Shapiro = vector("list", length = ncol(all_data) - 1),

Wilcoxon_rank = vector("list", length = ncol(all_data) - 1),

  Kruskal = vector("list", length = ncol(all_data) - 1),

  Chi = vector("list", length = ncol(all_data) - 1)

)

for (thecol in names(all_data)[!names(all_data) ==
TARGET_LABEL]) {

  alltests$Shapiro[[thecol]] <- shapiro.test(all_data[[thecol]])

  alltests$Wilcoxon_rank[[thecol]] <- wilcox.test(all_data[[thecol]],
all_data[[TARGET_LABEL]])

  alltests$Kruskal[[thecol]] <- kruskal.test(all_data[[thecol]],
all_data[[TARGET_LABEL]])

  alltests$Chi[[thecol]] <- chisq.test(all_data[[thecol]],
all_data[[TARGET_LABEL]])

}



alltests$Shapiro[[TARGET_LABEL]] <-
shapiro.test(all_data[[TARGET_LABEL]])

cat("\n\t\t == Shapiro-Wilk test ==\n")

for (thecol in names(alltests$Shapiro)) {

  cat(thecol, ": p-value =", alltests$Shapiro[[thecol]]$p.value, "\n")

}
```

```
cat("\n\t\t == Wilcoxon signed-rank test ==\n")

for (thecol in names(alltests$Wilcoxon_rank)) {

  cat(thecol, ": p-value =", alltests$Wilcoxon_rank[[thecol]]$p.value, "\n")

}

cat("\n\t\t == Kruskal-Wallis test ==\n")

for (thecol in names(alltests$Kruskal)) {

  cat(thecol, ": p-value =", alltests$Kruskal[[thecol]]$p.value, "\n")

}

cat("\n\t\t == Chi squared test ==\n")

for (thecol in names(alltests$Chi)) {

  cat(thecol, ": p-value =", alltests$Chi[[thecol]]$p.value, "\n")

}

summary(all_data)

if (any(alltests$Shapiro[[TARGET_LABEL]]$p.value < 0.05) ||

    any(alltests$Wilcoxon_rank[[TARGET_LABEL]]$p.value < 0.05) ||

    any(alltests$Kruskal[[TARGET_LABEL]]$p.value < 0.05) ||

    any(alltests$Chi[[TARGET_LABEL]]$p.value < 0.05)) {

  cat("\nThere is evidence to suggest a relationship between predictors and the presence of cardiovascular disease (CVD).\n")

  cat("Further investigation is warranted.\n")

} else {

  cat("\nThere is no significant evidence to suggest a relationship between predictors and the presence of cardiovascular disease (CVD).\n")

}
```

Output:

```
== Shapiro-Wilk test ==

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

Age : p-value = 0.006107327

Cholesterol : p-value = 0.1648802

Blood_Pressure : p-value = 0.08507818

Heart_Rate : p-value = 0.0006153776

ST_Depression : p-value = 3.546863e-07

Chest_Pain : p-value = 1.421762e-10

CVD_Status : p-value = 1.445635e-14

== Wilcoxon signed-rank test ==

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

Age : p-value = 1.296824e-35

Cholesterol : p-value = 1.322373e-35

Blood_Pressure : p-value = 1.159978e-35

Heart_Rate : p-value = 1.314508e-35

ST_Depression : p-value = 2.506045e-09

Chest_Pain : p-value = 3.692172e-34

== Kruskal-Wallis test ==

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =
```

: p-value =

: p-value =

Age : p-value = 0.05860761

Cholesterol : p-value = 0.1278613

Blood_Pressure : p-value = 0.2044213

Heart_Rate : p-value = 0.008100156

ST_Depression : p-value = 0.05728783

Chest_Pain : p-value = 2.325153e-05

== Chi squared test ==

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

: p-value =

Age : p-value = 0.8569662

Cholesterol : p-value = 0.4322504

Blood_Pressure : p-value = 0.6221057

Heart_Rate : p-value = 0.3679701

ST_Depression : p-value = 0.2853881

Chest_Pain : p-value = 0.0002899399

Age Cholesterol Blood_Pressure Heart_Rate ST_Depression

Min. :20.00 Min. : 95 Min. : 95.0 Min. : 72.0 Min. :0.000

1st Qu.:48.00 1st Qu.:213 1st Qu.:120.0 1st Qu.:142.0 1st Qu.:0.400

Median :55.00 Median :239 Median :130.0 Median :155.0 Median :1.000

Mean :54.39 Mean :245 Mean :132.8 Mean :151.4 Mean :1.223

3rd Qu.:60.00 3rd Qu.:270 3rd Qu.:140.0 3rd Qu.:165.0 3rd Qu.:1.800

Max. :71.00 Max. :417 Max. :180.0 Max. :188.0 Max. :6.200

Chest_Pain CVD_Status

Min. :1.000 Min. :0.00

1st Qu.:3.000 1st Qu.:0.00

Median :3.000 Median :0.00

Mean :3.158 Mean :0.42

3rd Qu.:4.000 3rd Qu.:1.00

Max. :4.000 Max. :1.00

NA's :1

There is evidence to suggest a relationship between predictors and the presence of cardiovascular disease (CVD).

Further investigation is warranted.

VI. CONCLUSIONS AND EXTENSIONS

Our study's findings show a strong correlation between a few predictors and the existence of CVD. This implies that these health markers could be useful tools for determining a person's risk of cardiovascular disease and could help create more accurate prediction models.

Even with these encouraging results, our models for assessing CVD risk could still be more accurate. Future research should think about include other variables like physical fitness levels, diabetes, smoking behaviours, and exposure to air pollution. These factors may offer a more thorough knowledge of personal risk profiles since they have been demonstrated to impact cardiovascular health. The intricate relationships between these different risk factors and their combined effect on cardiovascular health should also be investigated in more detail. We can more accurately identify those who are at high risk and customise interventions to meet their unique requirements if we take a more comprehensive approach to understanding CVD risk.

VII. BIBLIOGRAPHY

1. Zhu, X., et al. "Logistic regression technique for prediction of cardiovascular disease." Journal of Clinical Epidemiology and Global Health, vol. 15, 2022, pp. 1-10. Link

2. Singh, A., et al. "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases." IEEE

Transactions on Biomedical Engineering, vol. 69, no. 2, 2022, pp. 789-797. Link

3. Sonali1197. "Heart Diseases Prediction Model." GitHub, 2023. Link

4. Sonali1197. "Processed Cleveland Heart Disease Data." GitHub, 2023. Link

5. Sonali1197. "README.md - Heart Disease Prediction Model." GitHub, 2023. Link

6. Sood, A. "Multivariate Linear Regression of Heart Disease Attributes to Blood Pressure." GitHub, 2023. Link

7. World Heart Federation. World Heart Report 2023: Confronting the World's Number One Killer. Geneva, Switzerland, 2023. Link

8. Chu, C. "Predicting Cardiovascular Disease based on Regression Analysis and Classification of NHANE Survey Statistics." University of California, Berkeley, 2022. Link

9. Davidechicco. "Cardiovascular Heart Disease Data." GitHub, 2023. Link

10. Davidechicco. "Biostatistics Analysis Tests - Shapiro Wilcoxon Kruskal ChiSquared." GitHub, 2023. Link

11. Davidechicco. "Survival on patients having cardiovascular heart disease." GitHub, 2023. Link