Air University Islamabad

FACULTY OF COMPUTING & ARTIFICIAL INTELLIGENCE

Department of Creative Technologies

BSDS – 4-A

DS-212 Applied Statistics

Project

Statistical Analysis of Heart Patients in Pakistan

Submitted To: Dr. Faridoon Khan & Sir Irfan ul Haq

Submitted By: Ayema Amir (232543)

Malik Usama Arif (233105)

Muhammad Abdullah Younas (232534)

Halima Hanif (232553)

Statistical Analysis of Heart Patients in Pakistan

Table of Contents

Abstract

This project presents a comprehensive statistical analysis of heart disease patients in Pakistan, focusing on identifying key patterns, associations, and risk factors using real-world clinical data. The aim was to evaluate how demographic, clinical, and biochemical markers relate to heart disease severity and mortality risk. Various statistical methods—including chi-square tests, Spearman correlation, Mann-Whitney U, Kruskal-Wallis, and logistic regression—were employed alongside machine learning techniques such as Random Forest and K-Means clustering. The findings reveal that chest pain alone, diabetes, and kidney markers are strongly associated with high severity, and that parameters such as serum creatinine, age, and cholesterol are key indicators for both severity and mortality. Modeling techniques predicted severity levels with reasonable accuracy and highlighted actionable features like thalach (maximum heart rate) and oldpeak (ST depression). Mortality prediction using logistic regression identified five major predictors, and clustering successfully grouped patients into risk categories for targeted intervention. These insights align with prior epidemiological research in Pakistan and global studies on heart disease prediction, underscoring the value of data-driven strategies in cardiovascular care. The results advocate for early screening, better symptom awareness, and multi-disease risk monitoring to reduce the burden of heart disease in Pakistan.

Introduction

Cardiovascular diseases (CVDs), particularly heart disease, remain the leading cause of death worldwide, accounting for nearly 17.9 million lives annually, as reported by the World Health Organization (WHO). The situation is particularly dire in low- and middle-income countries like Pakistan, where a combination of socioeconomic challenges, lifestyle shifts, and inadequate healthcare infrastructure continues to intensify the burden of chronic disease. Despite medical advancements, the detection, prevention, and management of heart disease remain a challenge due to late diagnosis, poor symptom recognition, limited diagnostic infrastructure, and insufficient public health interventions.

Heart disease in Pakistan is exacerbated by factors such as high rates of hypertension, diabetes, tobacco usage, sedentary lifestyles, and dietary habits rich in cholesterol and trans fats. The prevalence of obesity, stress, and physical inactivity further aggravates the situation, contributing to both the onset and severity of cardiac conditions. Notably, due to social and cultural norms, women often underreport symptoms and delay seeking treatment, leading to worse outcomes in many cases.

At the same time, the healthcare infrastructure is often unequipped to perform comprehensive risk assessments or deploy preventive strategies. Patients typically present to clinics and hospitals only after symptoms become severe. There is, therefore, an urgent need for a data-driven approach that can uncover risk patterns earlier, identify vulnerable patient groups, and predict disease severity or mortality using statistical and machine learning tools.

This project was designed with the goal of statistically analyzing heart patient data from Pakistan to uncover the associations between demographic, clinical, and lifestyle factors and heart disease severity. By employing a mixture of classical statistical hypothesis testing and machine learning models, we aim to provide evidence-based insights that can guide clinical decisions and public health strategies.

1. Literature Review

1.1 Heart Disease in Pakistan: A Growing Epidemic

A significant body of research indicates that heart disease in Pakistan is widespread and increasing rapidly. In a cross-sectional study by Kazmi et al. (2022), researchers surveyed 906 adults aged 30 and above in Lahore. Their findings revealed that **17% of the population reported having ischemic heart disease**, while **40.1% had hypertension** and **15.8% suffered from diabetes**. This combination of risk factors—often co-existing—paints a concerning picture of the average Pakistani adult's cardiac health. The study not only highlighted the epidemiological trends but also pointed to a healthcare system that is reactive rather than proactive in its treatment protocols.

What makes the situation even more pressing is the lack of structured follow-up or risk monitoring programs for patients who already exhibit signs of heart disease. In most clinical setups, patients are diagnosed based on acute symptoms, and little is done to monitor or predict future cardiac events or escalating severity levels. The work of Kazmi et al. is particularly relevant to our study because it offers population-level data that aligns with the risk factors and disease outcomes we observed in our own dataset.

Furthermore, their findings support our decision to analyze co-morbid risk clusters, such as those involving diabetes, high blood pressure, and abnormal renal function. The overlap between these conditions and heart disease severity makes it critical to analyze them collectively rather than in isolation.

1.2 Global Studies on Heart Disease Prediction

Internationally, significant strides have been made in using machine learning and statistical models to predict heart disease and its severity. One of the most cited works in this space is by Mohan et al. (2019), who introduced a hybrid model combining Random Forest with Logistic Regression to predict heart disease using the Cleveland dataset. Their method—known as the Hybrid Random Forest with Linear Model (HRFLM)—achieved an **accuracy of 88.7%**, significantly outperforming individual models like decision trees, SVMs, or logistic regression alone.

The strength of their approach lies in its ability to both capture nonlinear interactions (via Random Forest) and maintain interpretability (via Logistic Regression). Features such as maximum heart rate achieved (thalach), cholesterol, ST depression (oldpeak), and exercise-induced angina (exang) were found to be particularly predictive—exactly the type of variables we analyzed in our dataset.

Mohan et al.'s study validated the feasibility and effectiveness of using machine learning models in clinical diagnostics and highlighted the potential for early risk detection. Although their work

is based on a Western dataset, the core methodology is highly transferable to the Pakistani context. Our use of Random Forest, Logistic Regression, and K-Means Clustering is partially inspired by their approach, although our focus is more region-specific and involves unique variables such as renal markers and socio-demographic traits not present in the Cleveland dataset.

1.3 Risk Stratification, Biomarkers, and Gender Disparities

Other studies have emphasized the role of risk stratification and biomarker analysis in predicting cardiac events. Numerous investigations have shown that **serum creatinine, cholesterol, CK-MB, and WBC counts** can signal underlying cardiac issues well before major symptoms arise. Our analysis of these markers aligns with these studies. For instance, increased serum creatinine is linked not just to kidney dysfunction but also to higher cardiac burden due to fluid retention and vascular resistance.

Moreover, recent literature has drawn attention to the **underrepresentation of women** in cardiovascular studies. Gender-based disparities often mean that women present atypical symptoms or receive delayed treatment, leading to worse outcomes. Our dataset revealed that even though male patients were more prevalent in each severity group, female patients—particularly in the 45–59 age group—frequently exhibited high severity despite seemingly normal renal markers. This suggests that gender-aware modeling could improve diagnostic accuracy and early intervention, particularly in rural settings.

2. Objectives

The literature thus provides a clear rationale for the current study. Previous research has shown:

- The **alarming prevalence of heart disease** in Pakistan and its association with lifestyle diseases.
- The **predictive power of machine learning models** in assessing heart disease risk.
- The importance of **biochemical markers and comorbidity clusters**.
- The **need to address gender and age-specific vulnerabilities** in heart patients.

Building on these insights, our study aims to:

1. Analyze real-world clinical data from heart patients in Pakistan to identify key risk factors associated with severity and mortality.
2. Use statistical tests to understand the strength of associations between symptoms, test results, comorbidities, and outcomes.
3. Apply predictive models (Random Forest, Logistic Regression) to classify patients into severity levels and estimate death risk.
4. Cluster patients into risk groups to inform triage strategies and clinical decision-making.

By bridging epidemiological trends with machine learning models, we hope to contribute actionable insights that can be used both at the hospital level and for broader public health planning.

Data Description

The dataset used in this study comprises detailed, anonymized clinical records of heart patients sourced from healthcare settings in Pakistan. It includes a diverse and comprehensive set of variables spanning patient demographics, lifestyle factors, clinical symptoms, laboratory investigations, diagnostic test results, and outcome measures. The richness of this dataset allows for robust statistical and machine learning analysis to identify key predictors of heart disease severity and mortality.

Data Source: https://opendata.com.pk/dataset/heart-patients-in-pakistan/resource/70aec9b8-7674-492f-a390-1bd72666744f

Demographic Information

The dataset captures essential demographic details, including:

- **Age** (as a continuous variable) and **Age Group** (e.g., 41–50, 51–60)
- **Gender** (Male/Female)
- **Locality** (Urban/Rural)
- **Marital Status**

These features provide foundational context for exploring the impact of socio-demographic factors on disease progression.

Lifestyle and Clinical History Variables

Lifestyle-related and comorbidity indicators are well represented, including:

- **Life_style** (coded to reflect physical activity level)
- **Sleep patterns**
- **Smoking status**
- **Presence of Depression**
- **Hyperlipidemia (Hyperlipi)**
- **Family History of heart disease**
- **Diabetes** and **Hypertension (HTN)**
- **Reported Allergies**

These variables are critical for assessing modifiable and hereditary risk factors that may contribute to heart disease severity.

Clinical Observations and Diagnosis

Text-based fields such as **Clinical Observation** and **Diagnosis** contain qualitative information reflecting physician notes, symptom descriptions, and disease classification (e.g., "Chest pain, SOB, cold, sweating", "I/W M.I", "LV dysfunction"). Additionally, variables like **Thrombolysis**, **Thalassemia**, and **Exertional Angina (exang)** are recorded to capture key medical events or conditions related to cardiovascular function.

Biochemical and Hematological Markers

The dataset includes a wide range of lab test results commonly used in cardiac and general health evaluations:

- **Renal Markers**: Serum Creatinine (S.Cr), Blood Urea (B.Urea)
- **Electrolytes**: Sodium (S.Sodium), Potassium (S.Potassium), Chloride (S.Chloride)
- **Cardiac Enzymes**: Creatine Phosphokinase (C.P.K), CK-MB
- **Glucose Profile**: Blood Glucose Random (BGR)
- **Inflammation and Blood Health**:
    - **White Blood Cell Count (WBC)**, **Red Blood Cell Count (RBC)**
    - **Hemoglobin**, **Packed Cell Volume (P.C.V)**, **Mean Corpuscular Volume (M.C.V)**, **Mean Corpuscular Hemoglobin (M.C.H)**, **Mean Corpuscular Hemoglobin Concentration (M.C.H.C)**
    - **Platelet Count**
    - **Differential Counts**: Neutrophil, Lymphocytes (LYMPHO), Monocytes, Eosinophils (EOSIO), Others
    - **Erythrocyte Sedimentation Rate (ESR)**

These biomarkers were selected due to their established roles in diagnosing cardiovascular conditions and related systemic disorders.

Cardiac Diagnostic Parameters

The dataset also includes quantitative results from diagnostic tests routinely used in cardiology, such as:

- **Resting Blood Pressure (trestbps)**
- **Cholesterol (chol)** and **Fasting Blood Sugar (fbs)**
- **Resting ECG Results (restecg)**
- **Maximum Heart Rate Achieved (thalach)**
- **ST Depression Induced by Exercise (oldpeak)**
- **Slope of the Peak Exercise ST Segment (slope)**
- **Number of Major Vessels Colored by Fluoroscopy (ca)**
- **Thalassemia Classification (thal)**

These features are particularly valuable for modeling heart disease severity, especially when interpreted together through multivariate techniques.

Outcome Variables

Two key outcome variables are included:

- **Heart Disease Severity Level (num)**: An ordinal variable indicating the degree of severity on a multi-level scale (e.g., 0–3 or 1–4).
- **Mortality**: A binary variable (0 = survived, 1 = deceased), indicating whether the patient succumbed to the disease.
- **Follow-Up Period**: Recorded in days, this variable reflects the post-treatment observation period and can be used to support survival analysis or time-based outcome evaluation.

Scope and Structure

The dataset contains **several hundred records**, each with **more than 60 variables**, offering a robust structure for both exploratory and predictive modeling. Its diversity allows for detailed sub-analyses, such as gender-specific risk profiling, comorbidity clustering, and blood marker correlation studies.

Overall, the dataset's depth and breadth make it highly suitable for comprehensive statistical evaluation, machine learning modeling, and evidence-based recommendations for early detection and effective management of heart disease in the Pakistani population.

Methodology

This study followed a structured methodology combining statistical analysis and machine learning techniques to examine the severity and mortality risk of heart disease among patients in Pakistan. The process was divided into three key phases: data preparation, statistical analysis, and predictive modeling.

Data Cleaning and Preprocessing

Although the dataset was largely clean, we performed standardization to harmonize similar entries. For instance, symptom combinations such as "chest pain, sneezing and cough" and "sneezing, chest pain and cough" were standardized to a consistent format to ensure uniformity during grouping and analysis. Missing values were handled by either removing records with critical data gaps or imputing values for variables where appropriate. Categorical variables were encoded into numeric form to ensure compatibility with modeling algorithms, while continuous variables were scaled during the clustering and PCA phases to maintain consistent magnitude across features. Data was split into training (80%) and testing (20%) sets to evaluate model performance.

Exploratory Data Analysis (EDA)

We began the analysis with a comprehensive exploratory data analysis to uncover initial trends and distributions. Visualization tools were used to examine how symptoms, disease severity, and comorbidities were distributed across patient records. For example, we found that patients reporting only chest pain were disproportionately present in the highest severity group, indicating potential underestimation of their condition. Similarly, diabetes and kidney-related markers were frequently associated with higher severity levels. Descriptive statistics helped us understand the central tendencies and spread of continuous variables such as age, cholesterol, and biochemical indicators.

Statistical Testing

To quantify relationships between variables, we applied multiple statistical tests tailored to the nature of the data. Normality of numerical columns was assessed using the Shapiro-Wilk test, which guided our choice of non-parametric tests due to widespread non-normality. For categorical data, we used the Chi-Square test to assess associations between variables such as symptom type, lifestyle habits, and disease severity. The Mann-Whitney U test was employed to compare distributions of continuous variables between gender groups, while the Kruskal-Wallis test was used to analyze how age varied across different severity levels. To examine relationships between continuous or ordinal variables, particularly among blood biomarkers and test results, we used Spearman's Rank Correlation, which is well-suited for non-linear, non-parametric data.

Model Development

We developed several predictive models to assess heart disease severity and mortality risk. For predicting severity levels (categorized into four classes), we implemented two models: Multinomial Logistic Regression and Random Forest Classifier. The logistic regression model was selected for its interpretability, while the random forest was used for its ability to model complex, non-linear relationships and to rank feature importance. These models were trained on the 80% training set and evaluated on the 20% test set using accuracy and confusion matrices. Feature importance scores from the Random Forest revealed key predictors such as thalach (maximum heart rate), oldpeak (ST depression), age, and several blood parameters like WBC and neutrophils.

To predict mortality, we used Binary Logistic Regression. The modeling process included univariate screening of all available variables to identify those significantly associated with death ($p < 0.1$). To reduce multicollinearity, we performed VIF analysis and removed redundant variables. Model optimization was done using stepwise selection via the stepAIC function, yielding a parsimonious model with strong predictive power. A patient risk score was calculated using the final model's coefficients, enabling the stratification of patients based on death risk. The risk score was visualized to show clear separation between survivors and non-survivors.

Clustering and Dimensionality Reduction

To explore patient subtypes, we applied K-Means clustering using six key clinical features: age, cholesterol, CK-MB, serum creatinine, resting blood pressure, and thalach. The data was standardized before clustering, and three distinct patient clusters were identified, corresponding to low-risk, moderate-risk, and high-risk groups. Principal Component Analysis (PCA) was then performed to reduce dimensionality and aid in visualizing the clusters in two-dimensional space. PCA plots provided an intuitive way to observe how patients group together based on shared characteristics, offering practical insights for targeted intervention strategies.

Software and Tools

The analysis and modeling for this project were conducted primarily using **R programming**, complemented by **Microsoft Excel** for preliminary data inspection and formatting. Excel was utilized in the initial stages to perform basic data cleaning, structure tabular information, and ensure consistent formatting of categorical variables, especially symptoms.

For the core statistical analysis, modeling, and data visualization, the **R environment** was extensively used due to its powerful statistical computing capabilities and a wide ecosystem of packages. In addition to traditional scripting and plotting, the project leveraged **Shiny** and **Shiny Dashboard** to build an **interactive web-based dashboard**, which allowed users (instructors, evaluators, or healthcare professionals) to **explore key visualizations, filter patient groups, and interact with severity and risk plots in real time**. This enhanced the interpretability and accessibility of the insights generated.

A wide array of R packages was employed to support various phases of the analysis:

- **Data Cleaning and Manipulation**: `dplyr`, `tidyr`, `readr`, `stringr`
- **Statistical Testing**: `stats`, `car`
- **Visualization**: `ggplot2`, `plotly`
- **Modeling**: `randomForest`, `nnet`, `caret`, `MASS`
- **Clustering and Dimensionality Reduction**: `cluster`
- **Shiny Interface**: `shiny`, `shinydashboard`

These tools collectively enabled comprehensive exploratory data analysis, hypothesis testing, predictive modeling, and real-time interactive visual reporting. The integration of Shiny into the project added a practical layer of usability, simulating a real-world healthcare analytics tool that could assist clinicians and policymakers in making informed, data-backed decisions about heart disease risk and management.

Ethical Considerations

All data used in this study was anonymized and handled strictly for academic and educational purposes. No personally identifiable information (PII) was used at any stage of the analysis, ensuring compliance with ethical standards for data use and privacy.

Analysis

This section include the detailed explanation of the process and the visuals

Data Cleaning & Preprocessing

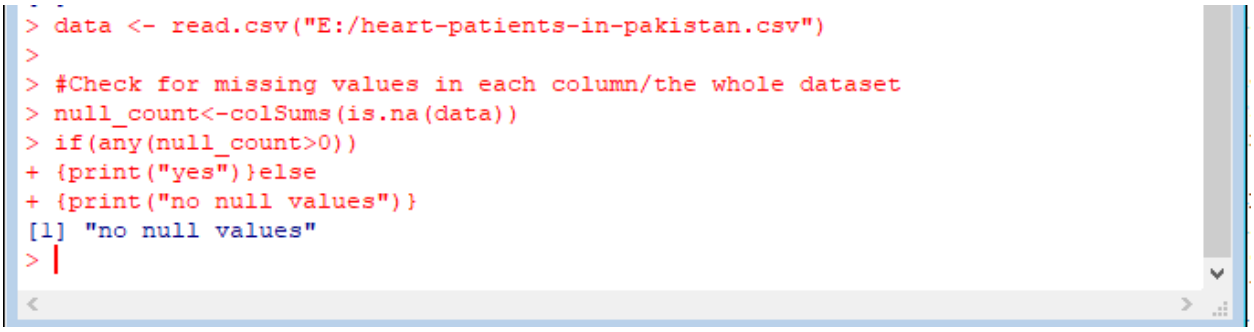
Although the dataset was relatively well-organized upon acquisition, preliminary inspection revealed minor inconsistencies in the formatting of categorical variables—particularly in the symptom descriptions recorded as free-text fields. To ensure uniformity and prevent analytical discrepancies, we performed data standardization using Microsoft Excel during the initial cleaning phase.

One specific challenge involved entries that described multiple symptoms in varying word orders. For example, phrases such as **"chest pain, sneezing and cough"** and **"sneezing, chest pain and cough"**—although semantically identical—were treated as distinct categories by the software. To address this, we developed a consistent formatting rule: symptoms were tokenized, alphabetized, and rewritten in a standardized order. This ensured that all patients presenting with the same combination of symptoms were grouped together appropriately in the analysis, avoiding artificial fragmentation of symptom categories.

In addition to symptom normalization, we also conducted general data validation, including:

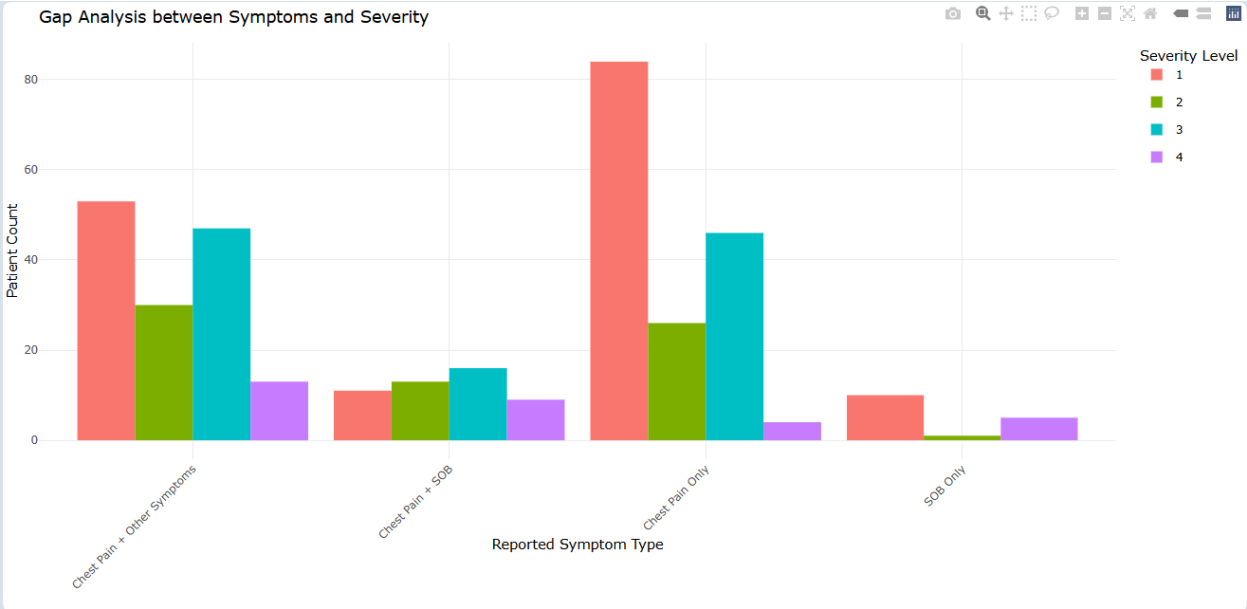
- Converting categorical text fields into lower case for consistency,
- Removing trailing or embedded spaces,
- Ensuring binary variables (e.g., smoking, diabetes, HTN) were properly encoded (e.g., 0 for absence, 1 for presence),
- Verifying numeric fields for outliers or invalid entries (e.g., blood pressure values outside human physiological limits).

This standardization process was essential for reliable group-wise comparisons, accurate model training, and effective feature aggregation in downstream statistical and machine learning analyses.



Exploratory Data Analysis

GAP ANALYSIS BETWEEN SYMPTOMS AND SEVERITY LEVEL

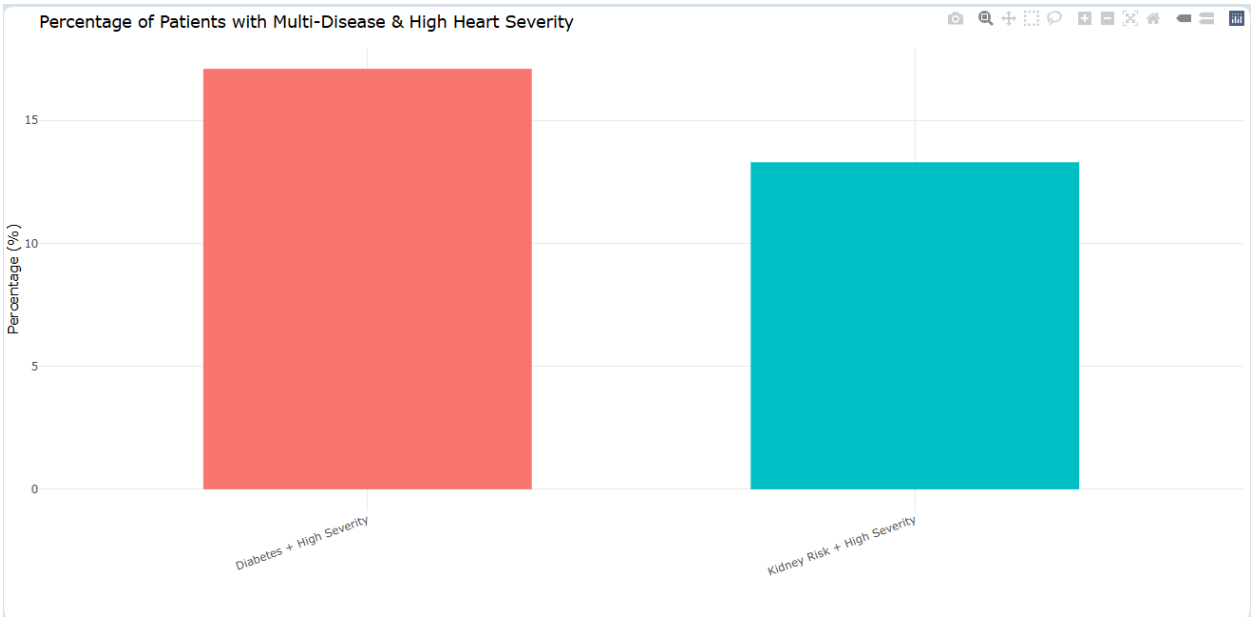


Insight:

- **Chest Pain Only** is the most common symptom type, but it's also associated with the highest proportion of patients in Severity Level 1 (highest severity)**.** This suggests many patients with just chest pain may have severe conditions, indicating underestimation of risk by patients reporting only chest pain**.**
- **Chest Pain + Other Symptoms** (e.g., nausea, dizziness) also shows a **large patient count**, but with a **more balanced severity distribution**, suggesting these patients may seek help earlier or have more recognizable symptoms.
- **SOB Only** (shortness of breath) is much less frequent, but patients in this category have a mix of severity levels, suggesting that even isolated SOB shouldn't be ignored.
- **Statistical test (Chi-Square)** shows that there's a significant association between symptom type and severity level—meaning the reported symptoms **don't always align with the actual severity**. This indicates a **gap in symptom recognition or reporting**.

```
> print(symptom_chi)

        Pearson's Chi-squared test

data:  symptom_table
X-squared = 42.362, df = 9, p-value = 2.821e-06
```

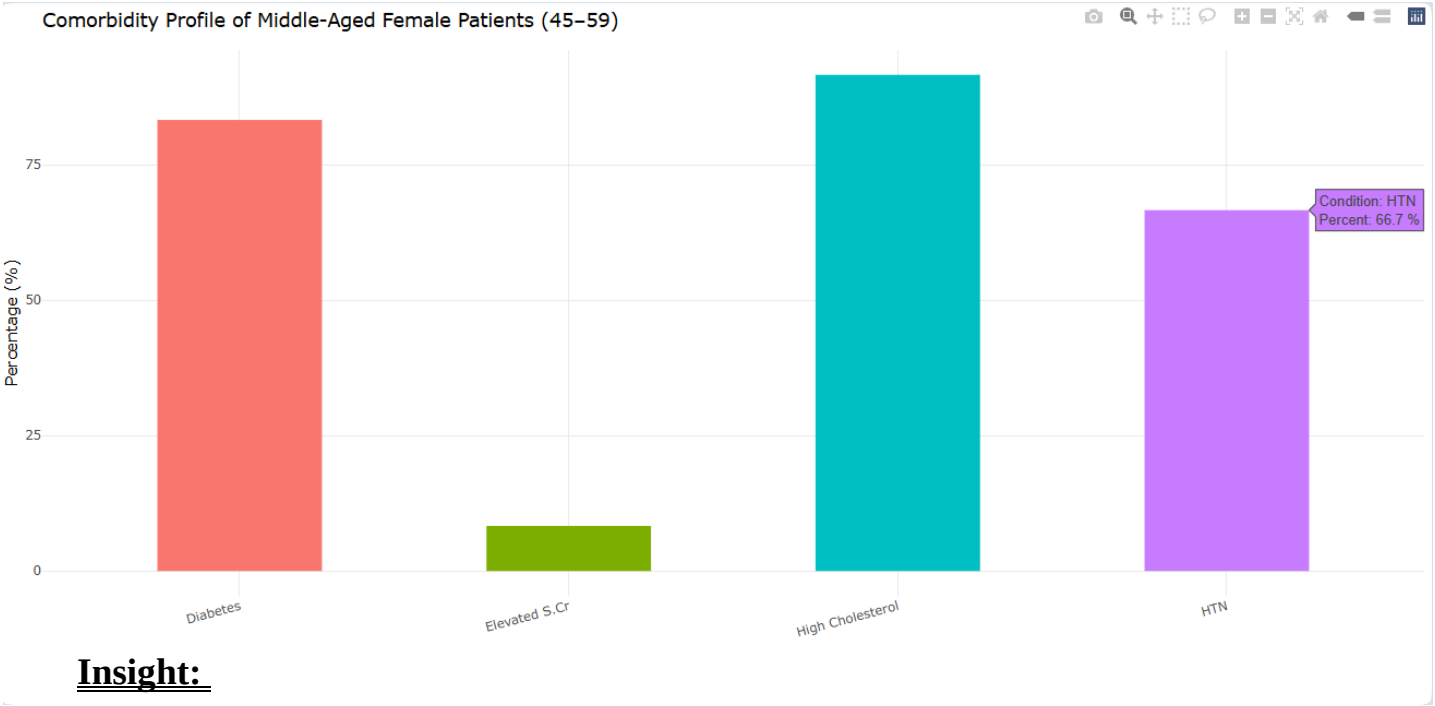
MULTI-DISEASE RISK DETECTION



Insight:

- **Diabetes + High Severity** patients represent the **largest proportion (~17%)** of the population, indicating that diabetes may significantly compound heart disease severity. This suggests a strong need for **integrated diabetic–cardiac care** programs.
- **Kidney Risk + High Severity** affects a **substantial group (~13%)** as well. Patients with elevated creatinine or urea levels are also prone to severe heart issues, reinforcing the link between **renal dysfunction and cardiac burden**.
- This comparison highlights a **clear overlap** between chronic conditions (like diabetes and kidney risk) and heart disease severity. These overlapping cases are **not minor**, making up **a notable percentage of the overall population**.
- The insight supports **policy action**: incorporating **routine screenings for kidney function and blood sugar** in heart disease management could prevent complications and reduce long-term healthcare costs.

HIGH ATTENTION TO MIDDLE-AGED FEMALE PATIENTS



Insight:

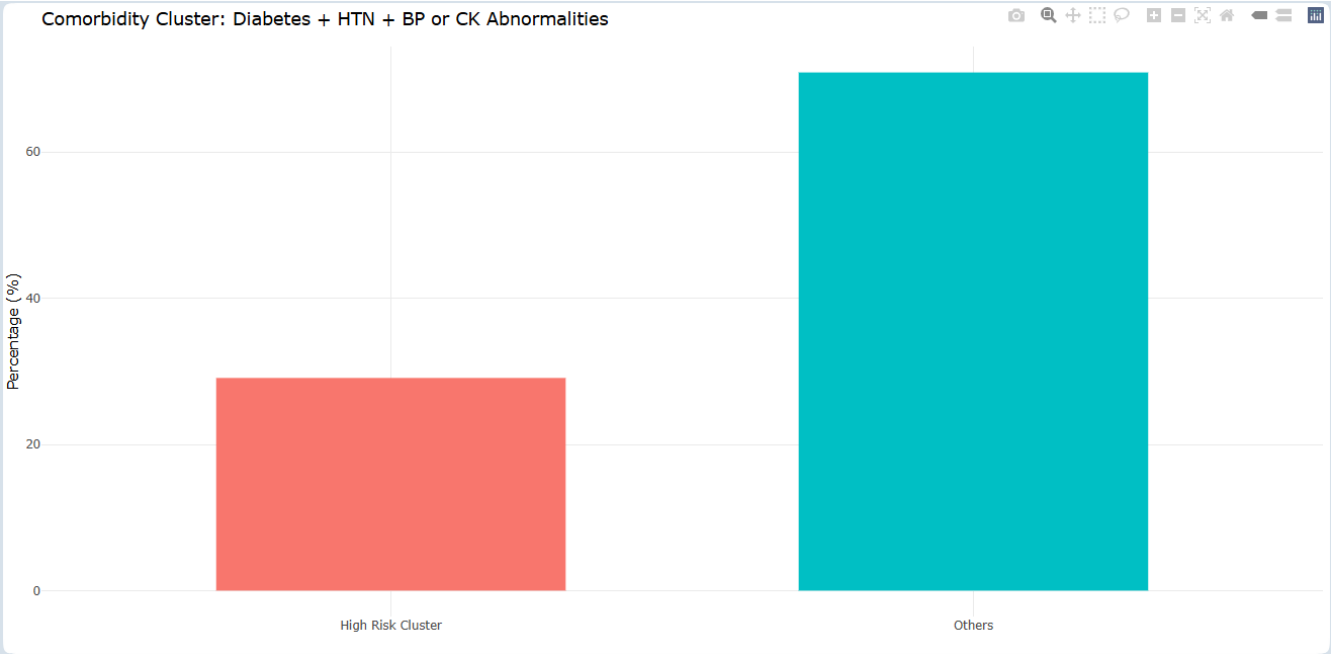
Middle-aged female heart patients (aged 45–59) have:

- **Very high rates** of **diabetes** and **high cholesterol** (above 85%)
- **66.7% have hypertension (HTN)**
- **Only a small number** have **elevated serum creatinine (S.Cr)**

This group is **at high heart risk** due to **multiple lifestyle-related diseases**, even if kidney markers look normal.

Prioritize this group for **early screening** and **preventive care** programs.

COMBORBIDITY CLUSTERS NEED SPECIAL ATTENTION



Insight:

1. **High-Risk Patient Concentration**

- Approximately **28-30%** of your patients fall into the "High Risk Cluster"
- This means nearly **1 in 3 patients** have multiple serious comorbidities simultaneously

2. **Comorbidity Clustering Pattern**

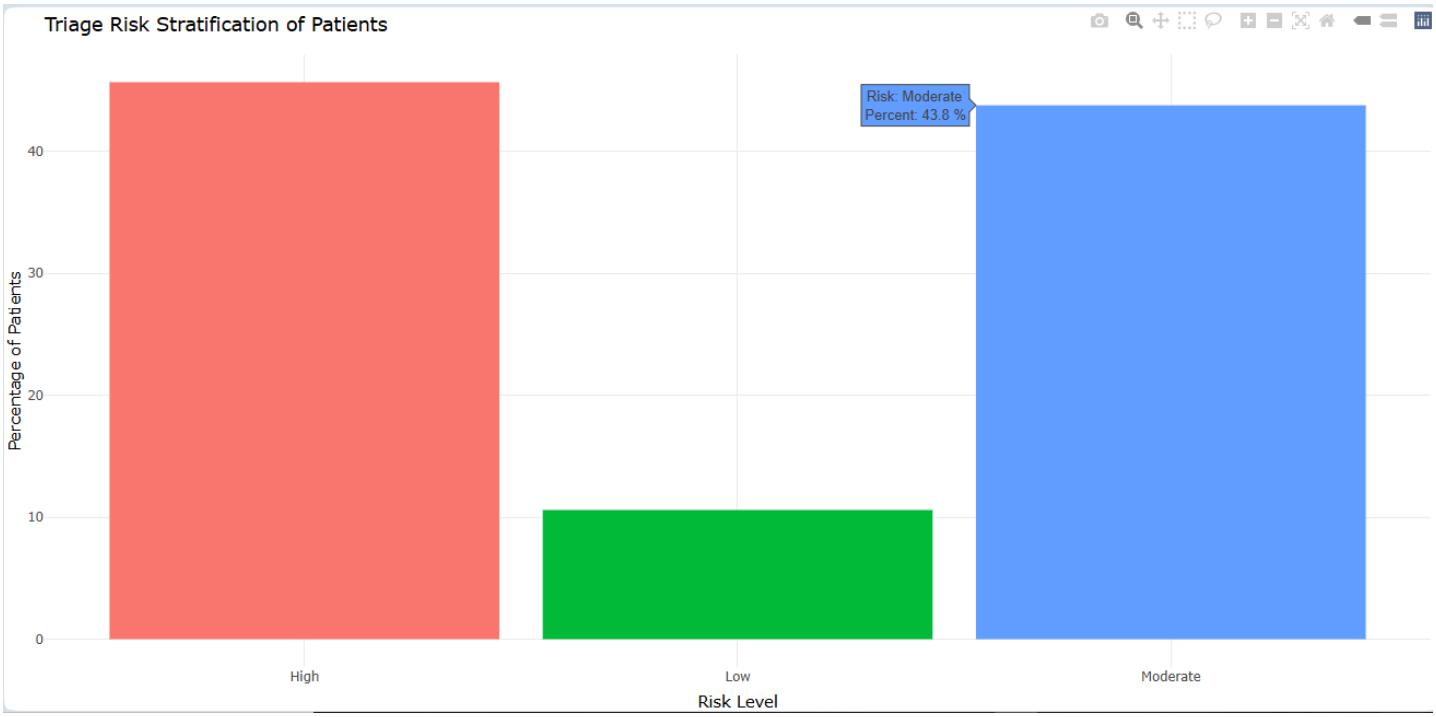
- The "High Risk Cluster" represents patients with **all three conditions**:
    - Diabetes
    - Hypertension (HTN)
    - Either high blood pressure (≥140) OR abnormal cardiac enzymes (CK-MB/CPK)

3. **Clinical Significance**

- **28-30% is a substantial proportion** - this isn't a rare combination
- These patients likely require **intensive monitoring and coordinated care**
- The clustering suggests these conditions often occur together, indicating shared risk factors

This is highlighting the need for comprehensive, multi-condition management strategies rather than treating diseases in isolation.

RISK STRATIFICATION AT TRIAGE



Insight:

- **Resource allocation:**

Tells us how many patients may need urgent resources (doctors, beds, monitoring) right away.

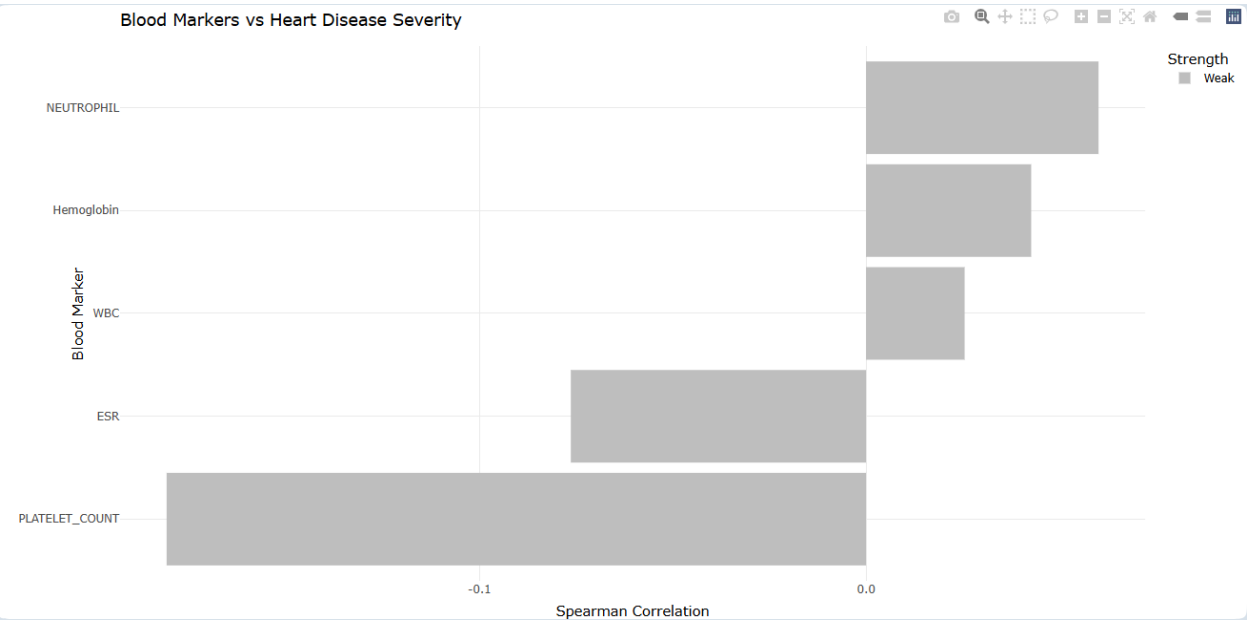
- **Clinical workflow:**

Helps prioritize which patients to see first.

- **Policy and planning:**

Useful for hospital managers and policymakers to understand the burden of moderate/high-risk patients in your setting.
It summarizes the urgency and risk profile of your incoming patient population, which is crucial for both clinical and administrative decision-making.

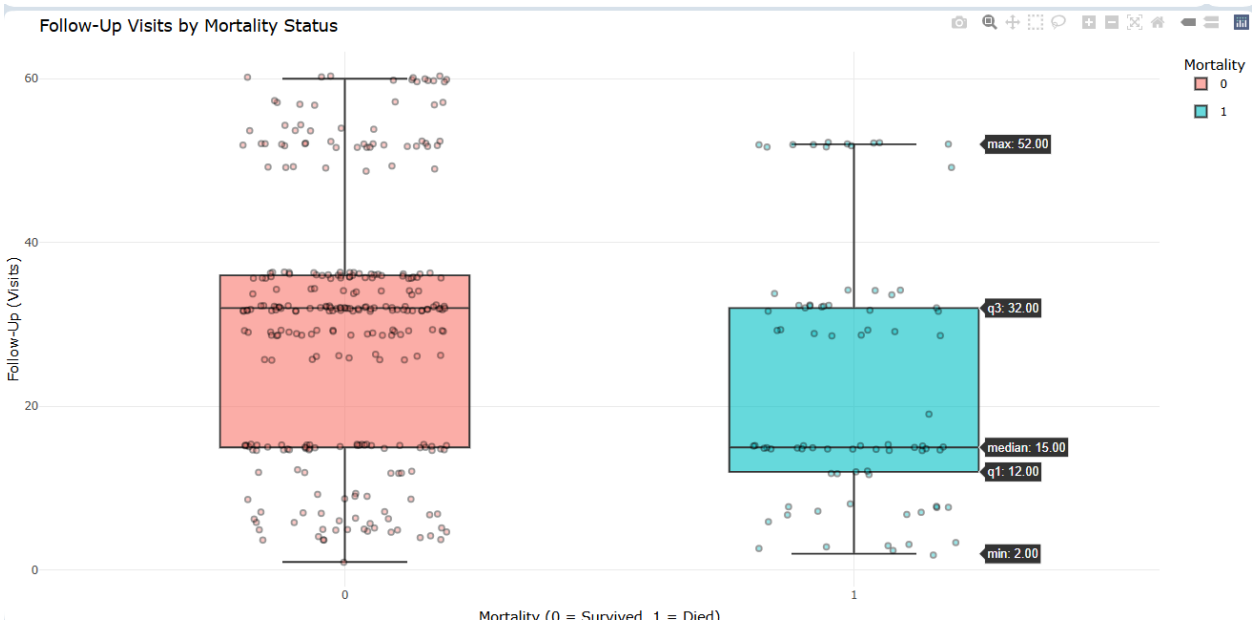
BLOOD MAKERS VS HEART DISEASE SEVERITY



Insight:

- Basic blood markers alone are poor predictors of heart disease severity, suggesting the need for more specific cardiac biomarkers or multi-parameter approaches for accurate severity assessment. Thus, **don't rely solely on basic blood counts** for severity assessment

FOLLOW-UP ANALYSIS BY SEVERITY LEVEL AND MORTALITY



Insight:

Median Follow-Up Visits:

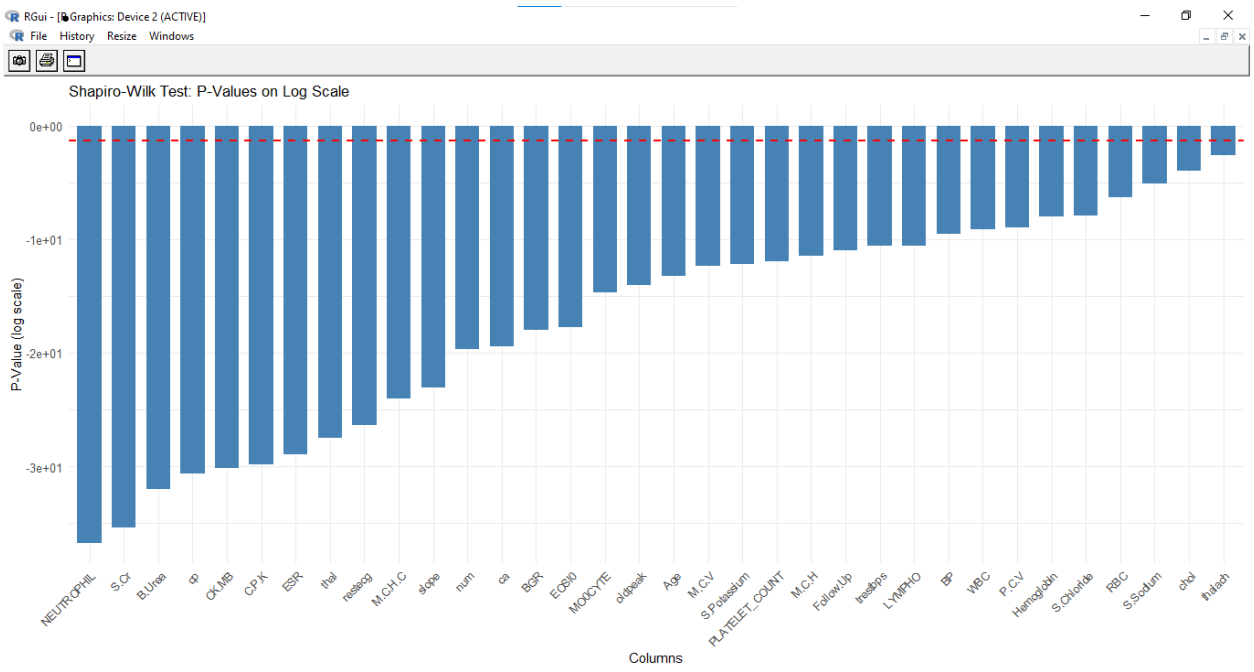
- **Survived (Mortality = 0)**: Median ≈ **31**
- **Died (Mortality = 1)**: Median = **15**

Patients who **survived** had **more follow-up visits** on average than those who died.

Hypothesis Testing

NORMALIRY TEST (Shapiro-Wilk Test)

It checks if data comes from a normal distribution. It's more powerful and accurate for small to medium datasets and our data is medium.



Insight:

The **most non-normal column is NEUTROPHIL**, followed by S.Cr, B.Urea, and P.

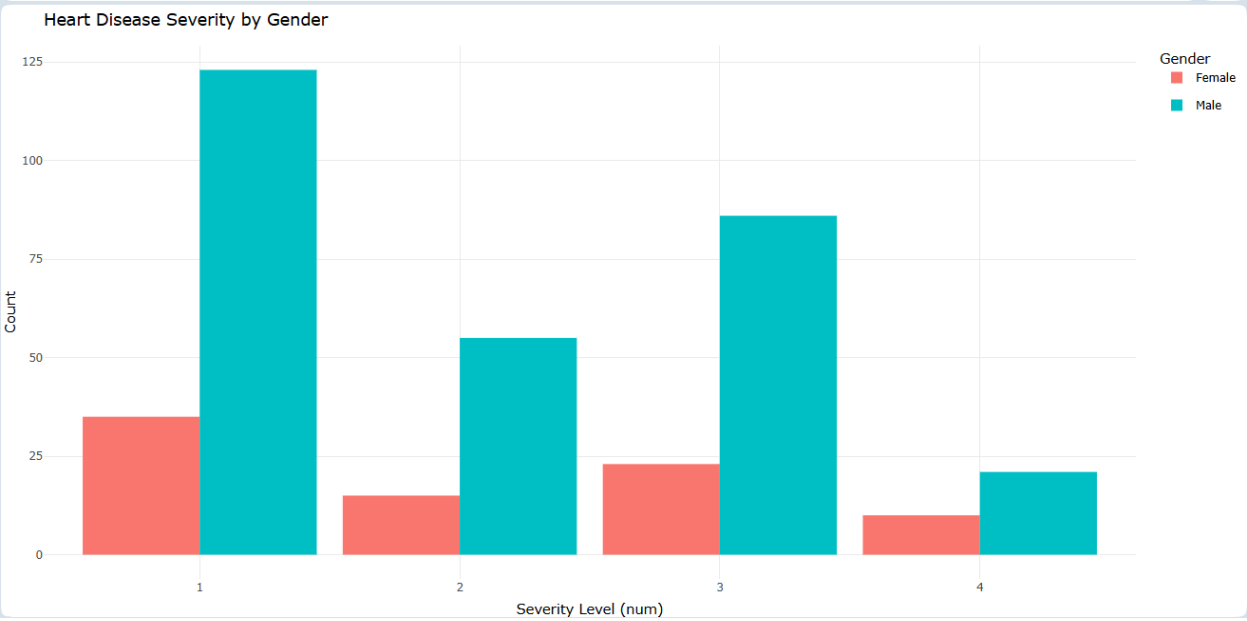
HEART DISEASE SEVERITY BY GENDER

Statistical Tests:

- **Mann-Whitney U Test (Wilcoxon Rank-Sum Test)**

Why it's used: Variable num (severity level) is ordinal and not normally distributed and groups are independent and this test works well for comparing two groups in such cases.

```
> print(wilcox_result)

        Wilcoxon rank sum test with continuity correction

data:  num by Gender
W = 12238, p-value = 0.6098
alternative hypothesis: true location shift is not equal to 0
```



Insight:

The severity level distribution between males and females is not significantly different, although more males appear in every severity category.

ECG AND STRESS TEST INDICATORS

Statistical Tests:

- **Chi-Square Test**

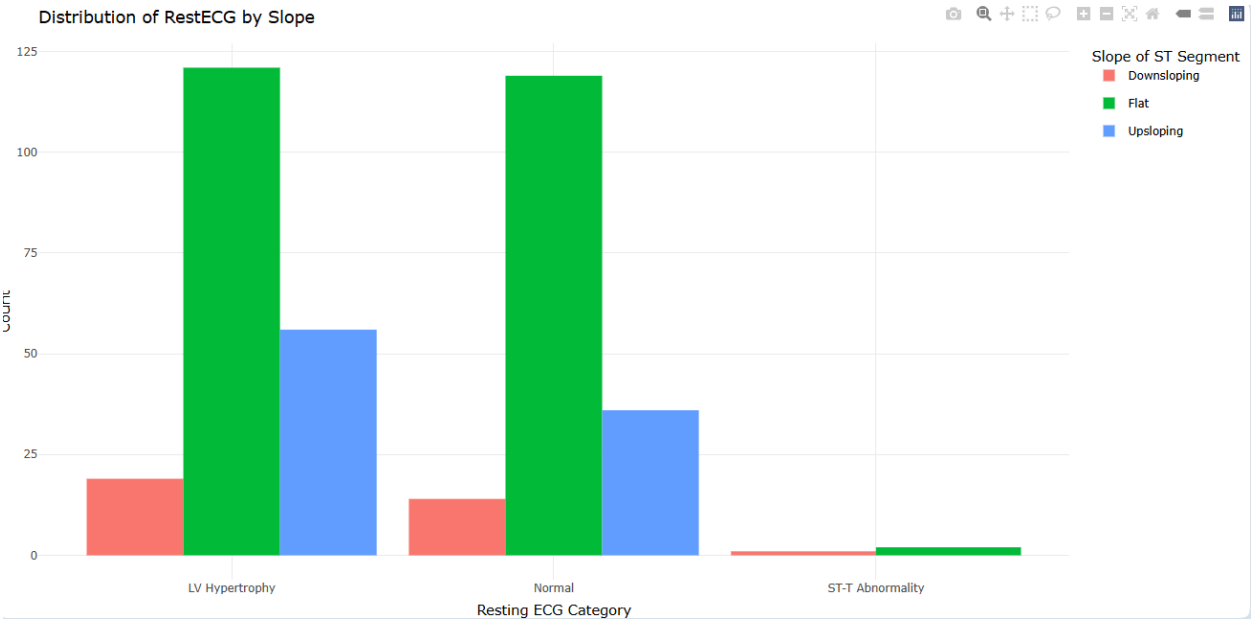
Why it's used: Both variables Resting ECG results (restecg) and Slope of the ST segment (slope) are categorical. This test tells us if there's a relationship between them.

```
> print(chi_result)

        Pearson's Chi-squared test

data:  table_restecg_slope
X-squared = 5.8103, df = 4, p-value = 0.2138
```

- **Spearman's Rank Correlation**

Why it's used: It checks if there's a monotonic relationship between maximum heart rate achieved (thalach) and ST depression (oldpeak). The data is not normally distributed and both columns are numeric, so Spearman is a safer choice for non-parametric data.

```
  cannot compute exact p-value with ties
> print(cor_result)

        Spearman's rank correlation rho

data:  data$thalach and data$oldpeak
S = 11133174, p-value = 1.96le-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.3403863
```



Axis Explanation:

X-axis (Resting ECG Category):
Represents the type of resting ECG result, which includes:

- Normal
- ST-T Abnormality (suggests possible heart strain)
- LV Hypertrophy (thickened heart muscle)

•**Y-axis(Count)**:
Shows the number of people (patients) who fall into each combination of RestECG and Slope category.

• **Bar colors (Slope of ST Segment)**:

- **Red**: Down sloping (often associated with higher risk)
- **Green**: Flat (possible sign of heart issues during stress)
- **Blue**: Upsloping (usually normal response)

Insight:

A flat ST slope during a stress test is commonly seen in this data, which may suggest these individuals are at higher risk of heart disease, even if their resting ECG appears normal.

ENABLE EARLY DETECTION THROUGH BLOOD TESTS

Statistical Tests:

- **Mann-Whitney U Test**

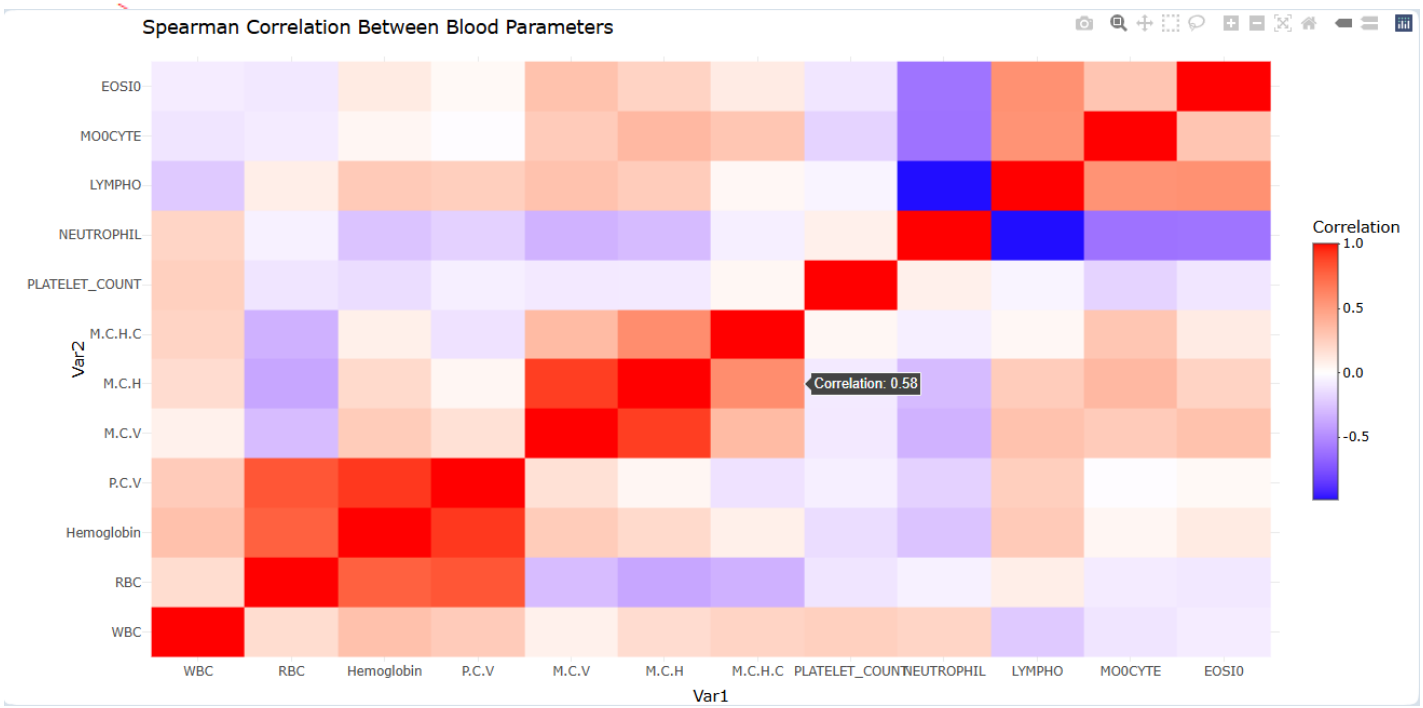
Why it's used: Data is not normally distributed and Gender has two groups (Male/Female). Identifies if blood parameters differ significantly between genders.

```
$WBC
[1] 2.712415e-06

$RBC
[1] 3.708606e-16

$Hemoglobin
[1] 3.367066e-16

$P.C.V
[1] 4.66729e-15

$M.C.V
[1] 0.3241417

$M.C.H
[1] 0.6875542

$M.C.H.C
[1] 0.6920477

$PLATELET_COUNT
[1] 0.1094726

$NEUTROPHIL
[1] 0.109699

$LYMPHO
[1] 0.124944

$MOOCYTE
[1] 0.6395941

$EOSIO
[1] 1.458088e-05
```

- **Spearman Correlation**

Why it's used: Blood data is not normal; this test detects monotonic trends. Reveals how blood values rise/fall together, helping spot key related indicators.

```
> print(cor_matrix)
          BGR       B.Urea          S.Cr        CK.MB        C.P.K         chol
-0.007290143 -0.067179447  0.039492777 -0.120127269 -0.063735614  0.006416735
     S.Sodium  S.Potassium   S.Chloride          ESR
-0.012608931 -0.118105777 -0.105384204 -0.076355108
>
```



Insight:

Strongly linked blood parameters are WBC, RBC, Hemoglobin, P.C.V, Neutrophil, Platelet_Count, Lymphocytes, which may help spot early signs of heart or blood health problems.

AGE VS HEART DISEASE SEVERITY LEVEL

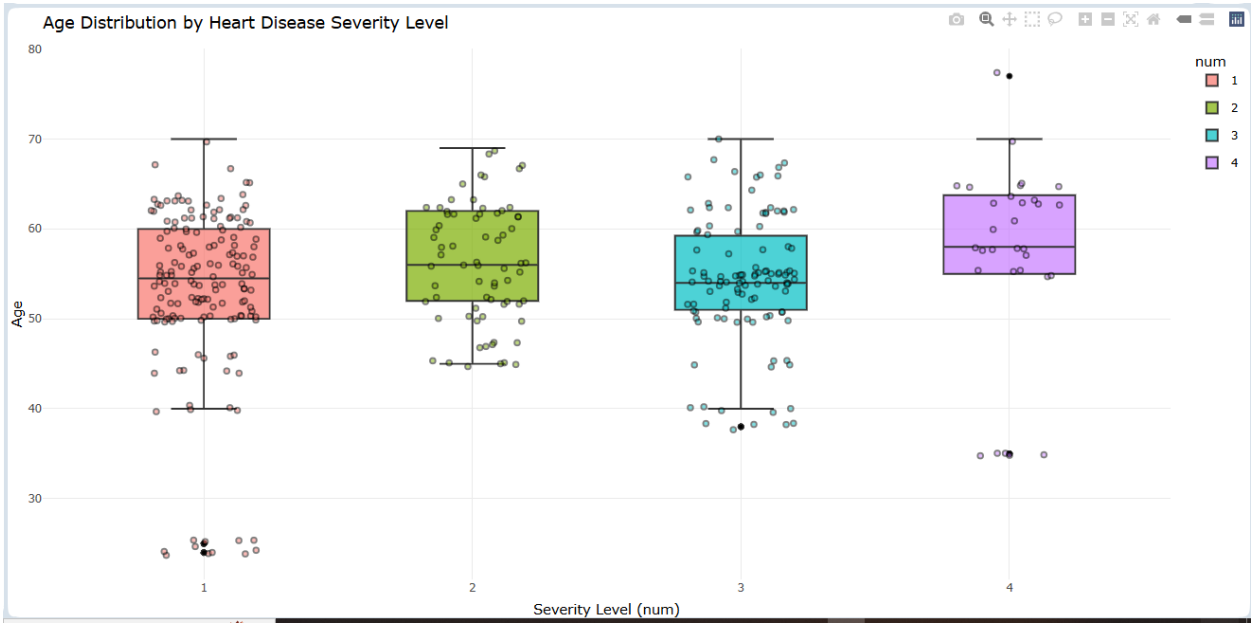
Statistical Tests:

- **Kruskal-Wallis Test**

Why it's used: Age is not normally distributed, and severity level has more than two groups. Kruskal-Wallis is a non-parametric alternative to ANOVA when data doesn't meet

assumptions. Identifies if there is a significant difference in Age distributions across severity levels.

```
> print(kruskal_result)

        Kruskal-Wallis rank sum test

data:  Age by num
Kruskal-Wallis chi-squared = 12.332, df = 3, p-value = 0.006328

>
```



Insight:

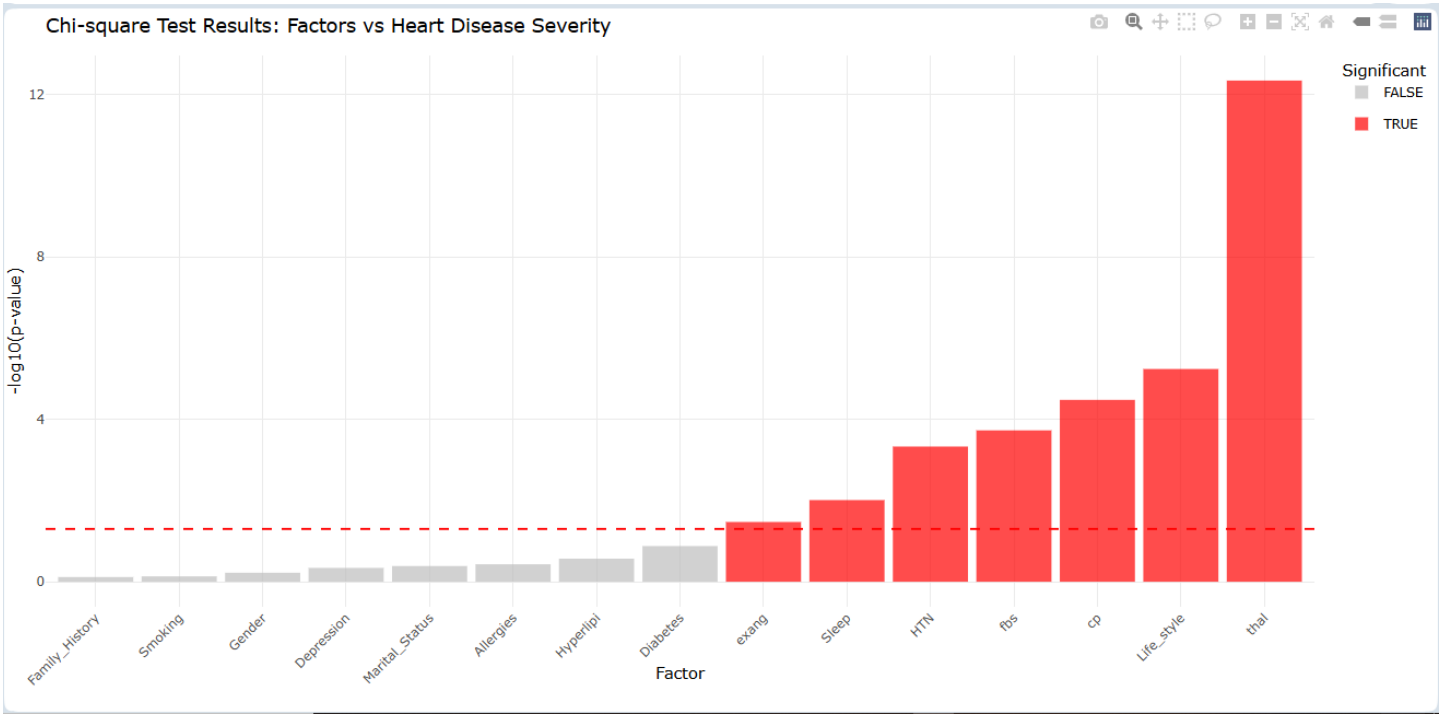
Older patients tend to have more severe heart disease, and the difference in age by severity level is statistically meaningful.

LIFE STYLE AND OTHER FACTORS ASSOCIATION

Statistical Tests:

- **Chi-Square Test**
  Why it's used: Both variables are categorical (or ordinal) and the test detects **associations, not causation**, especially for non-parametric data. It identifies which lifestyle/clinical factors **significantly differ** across heart disease levels. This helps target **risk factors** for prevention strategies.

```
> 
> # Print summary table
> print(chi_sq_summary)
                     Factor Chi_sq_statistic      p_value
X-squared...1         Smoking         1.265276 7.373947e-01
X-squared...2        Diabetes         5.611025 1.321469e-01
X-squared...3             HTN        17.876007 4.665223e-04
X-squared...4      Life_style        27.056107 5.730038e-06
X-squared...5  Marital_Status         2.882239 4.101402e-01
X-squared...6      Depression         2.604313 4.567339e-01
X-squared...7  Family_History         1.138788 7.677197e-01
X-squared...8       Allergies         3.139525 3.706094e-01
X-squared...9           Sleep        11.402575 9.736767e-03
X-squared...10      Hyperlipi         3.917841 2.704736e-01
X-squared...11         Gender         1.868393 6.001660e-01
X-squared...12          exang         8.689619 3.371530e-02
X-squared...13             cp        36.445182 3.304536e-05
X-squared...14            fbs        19.803043 1.864658e-04
X-squared...15           thal        69.805358 4.482494e-13
```



Insight:

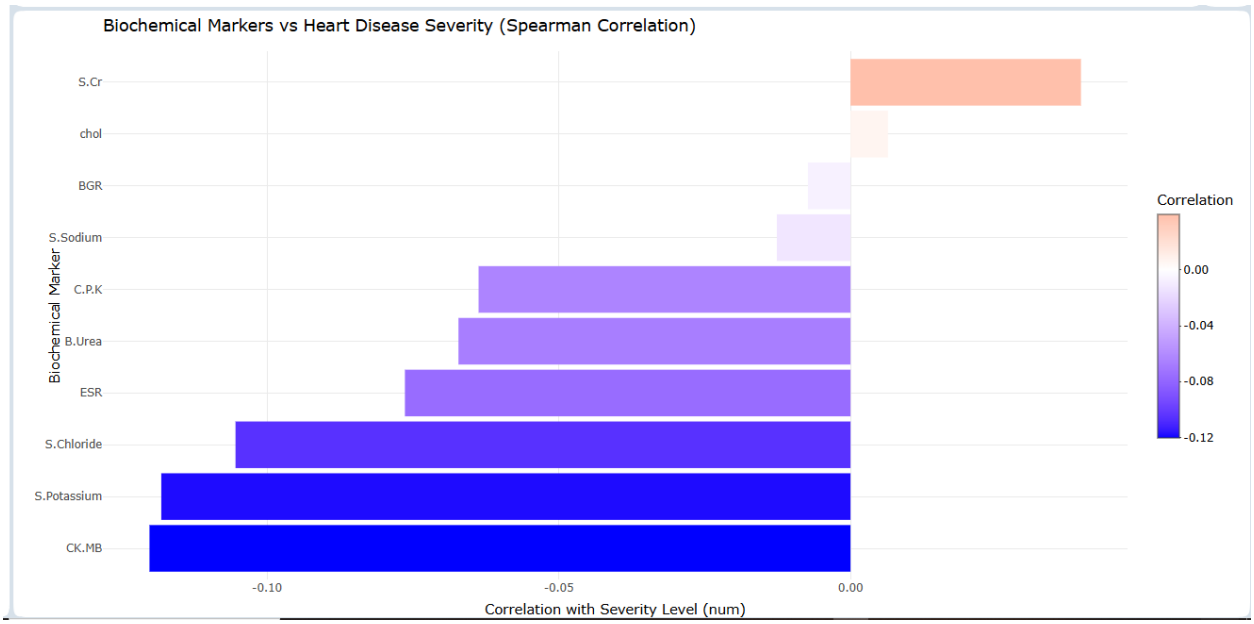
Thalassemia has the strongest association, suggesting it's the most influential factor.

BIOCHEMICAL MAKERS VS HEART DISEASE SEVERITY LEVEL

Statistical Tests:

- **Spearman Correlation**

Why it's used: It measures the strength and direction of the relationship between continuous or ordinal variables. Since both severity level and biomarkers are continuous or ordinal, this test identifies which markers have strong, moderate, or weak correlations with heart disease severity.



Insight:

- **S.Cr (Serum Creatinine)**: Positively correlated with heart disease severity. This suggests **kidney dysfunction is linked to worsening heart disease**.
- **CK-MB (Creatine Kinase-MB)**: Negatively correlated with severity. High levels generally indicate **acute cardiac injury**; a drop with worsening severity may imply **chronic damage or adaptation**.

So, the health department could focus on **improving kidney health monitoring** and **early cardiac injury detection** to manage heart disease severity.

Modeling

SEVERITY PREDICTION MODEL

Multinomial Logistic Regression and Random Forest:

We want to predict heart severity levels based on patient data (age, tests, lifestyle) to help doctors identify high-risk patients early.

Models:

- **Logistic Regression (Multinomial)**: Good for predicting categories (e.g., severity levels 1-4) when the relationship is somewhat linear.
- **Random Forest**: A tree-based method that handles non-linear relationships and interactions between features. It's robust and gives feature importance.

Combining simple (logistic regression) and complex (random forest) models gives us both **interpretability** and **accuracy**.

Prepare Data:

- Convert variables to the correct type (categorical or numeric) so models can process them.
- Remove missing values to ensure clean input.

Split Data (80/20):

- Training data helps the model learn patterns.
- Test data checks how well the model works on new data.

Train Models:

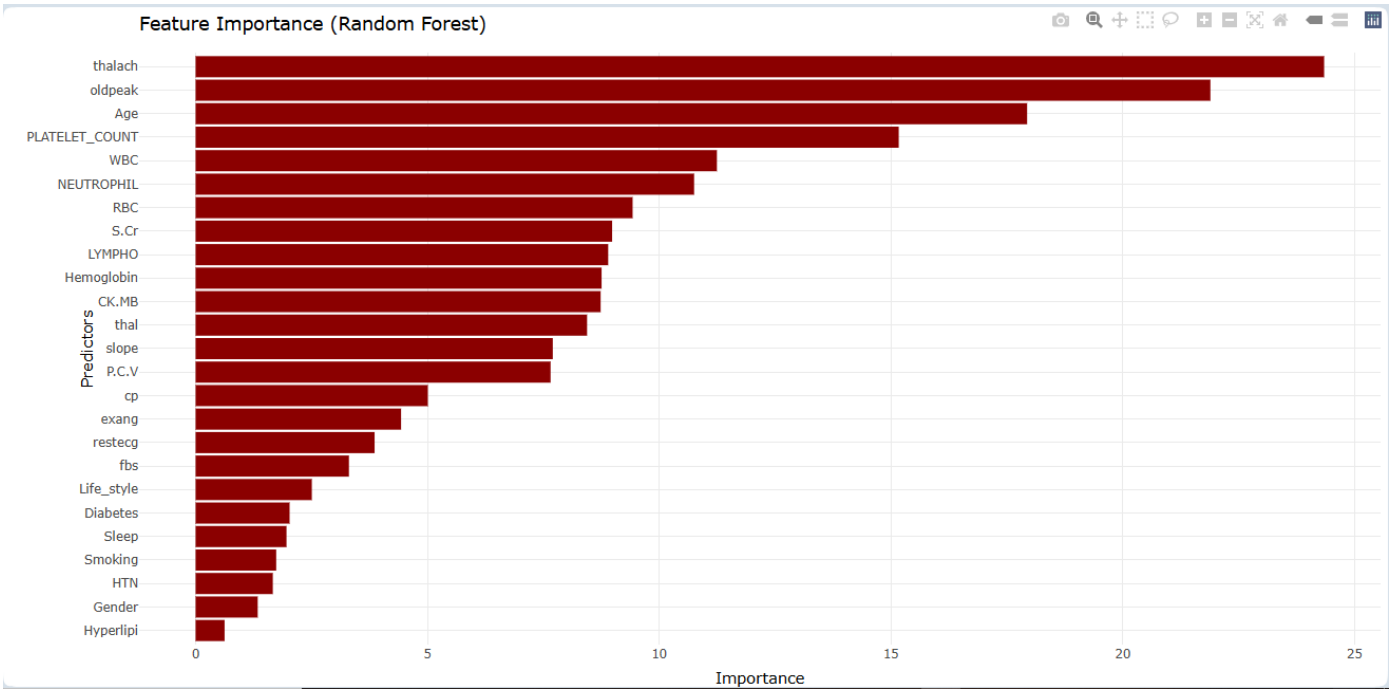
- Fit logistic regression and random forest models on training data.

Evaluate:

- Used **accuracy** and **confusion matrices** to check how many predictions were correct and where the model made mistakes.

Feature Importance (from Random Forest):

Which factors (like cholesterol, blood pressure) are most important for predicting severity.



Insight:

Feature	Meaning	Insight from Plot
thalach	Max heart rate achieved	Most important predictor for severity
oldpeak	ST depression	Strong link to stress-related severity changes
Age	Age of patient	Older age increases severity risk
PLATELET_COUNT, WBC, NEUTROPHIL	Blood markers	Blood values play a **key role** in predicting severity
Gender, HTN, Smoking	Lower bars	Less predictive, but still informative

PREDICTING DEATH USING LOGISTIC REGRESSION

Logistic Regression

Model:

• **Logistic Regression (Binary)**: A simple, interpretable model used to predict whether a patient survived or died (Mortality = 0 or 1). It helps identify how each factor influences death risk.

Prepare Data:

• Converted text columns to categorical (factor) format so the model understands the data.
• Recoded some variables (like thal) into meaningful categories (e.g., normal, fixed defect).

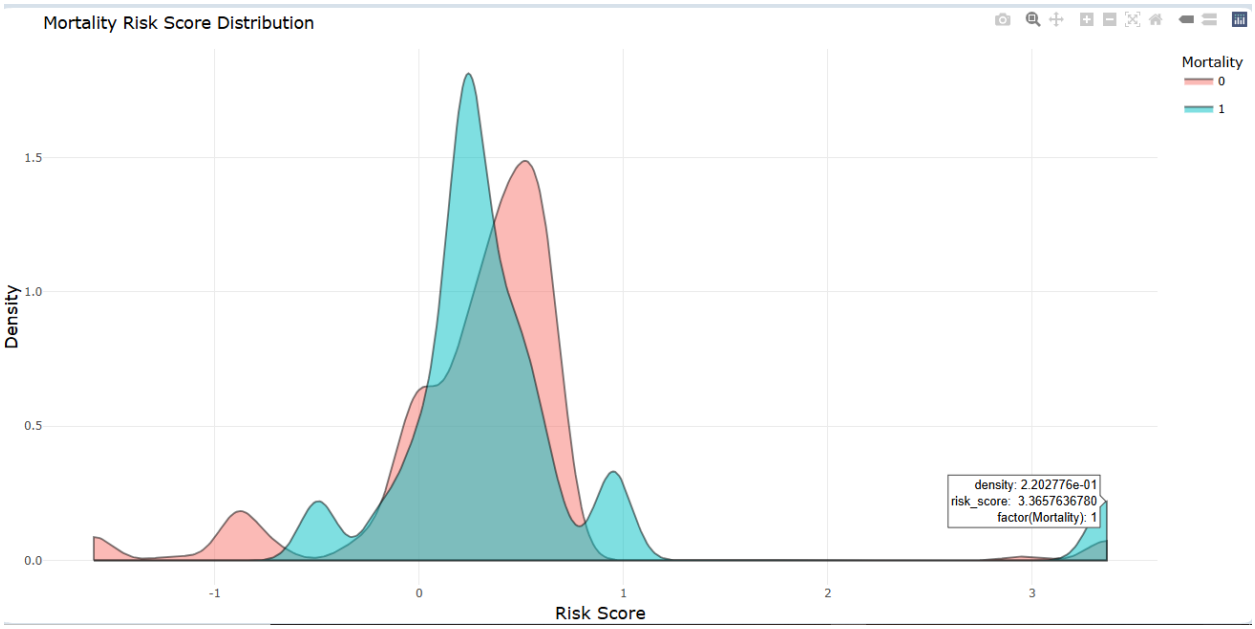
Univariate Screening:
- Tested each predictor one at a time to see if it was related to death.
- Kept predictors with a p-value less than 0.1 for further analysis.

Check Multicollinearity:
- Removed variables that were strongly related to each other (using VIF and alias detection).
- Kept only independent predictors to avoid confusion in the model.

Build Final Model:
- Used stepAIC to select the best combination of predictors for logistic regression.
- This reduced the model to the most important and non-redundant variables.

Create Risk Score:
- Combined top predictors into a simple formula (weighted sum) to create a patient risk score.
- Higher scores indicate greater risk of death.

Evaluate:
- Used logistic regression accuracy to check model performance.
- Created a density plot showing how risk scores differ between survivors and non-survivors.



Insight:

Higher risk scores are more likely in patients who died**,** so the model helps doctors spot high-risk patients before it's too late.

This is the output of the top 5 predictors selected by your logistic regression model (with stepAIC()), along with a **recommendation** for each one.

```
>
> print(policy_table)
   Predictor                                  Recommendation
1        Age Monitor serum creatinine in routine checkups
2 GenderMale              Prioritize thalassemia screening
3      exang      Implement hypertension control programs
4    oldpeak               Mandate exercise stress testing
5       chol            Track CK-MB levels post-treatment
>
```

Predictor	Meaning	What Policy Makers Should Do
Age	Older people are at higher risk	Start early screening programs for senior citizens
Gender Male	Males had higher death risk	Create gender-focused heart disease awareness
exang	Chest pain during exercise is risky	Mandate treadmill stress testing
oldpeak	Measures ST depression on ECG	Train staff to interpret ECG abnormalities more critically
chol	Cholesterol level	Regular lipid testing and dietary advice programs

These 5 variables were the **most powerful predictors of death**, and the recommendations tell **hospitals or health departments** what action to take

CLUSTERING PATIENTS INTO RISK GROUPS

K-Means Clustering and PCA:

Model:
This analysis groups patients into subtypes based on key health markers using clustering and PCA (Principal Component Analysis). It helps identify high-risk vs. low-risk groups for targeted healthcare interventions.

Prepare Data:

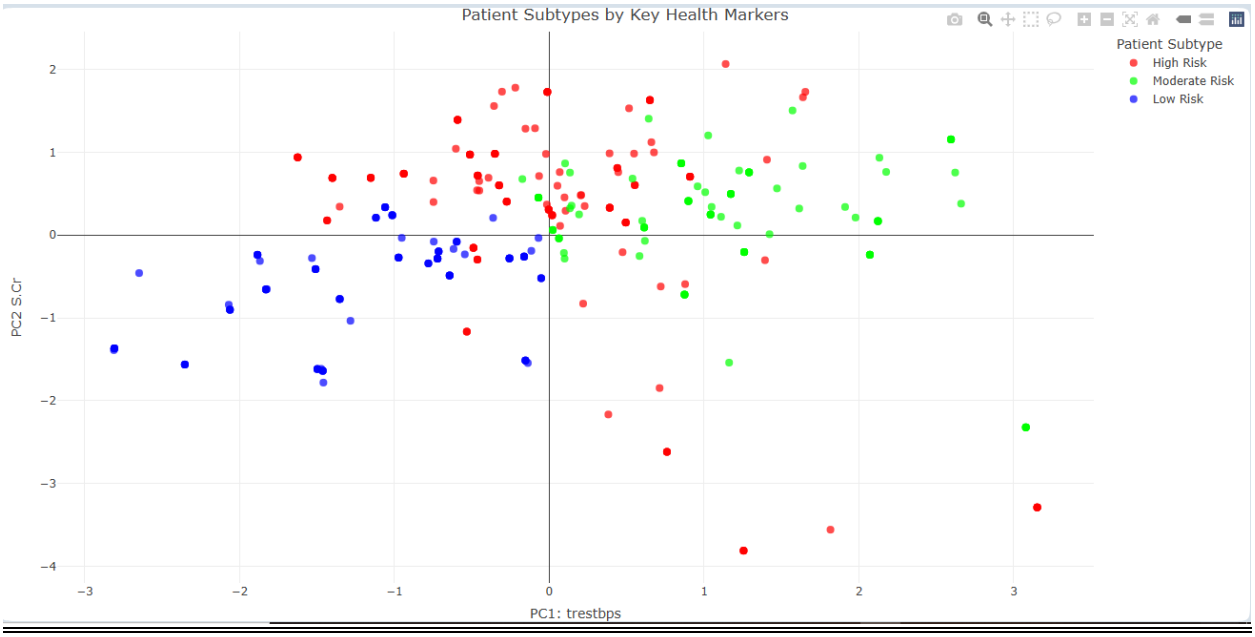
- Selected key health variables: Age, Serum Creatinine (S.Cr), CK-MB, Cholesterol, Resting BP (trestbps), and Max Heart Rate (thalach).
- Scaled data to standardize values for clustering and PCA.

Clustering Approach:

- Used **K-Means** clustering with 3 clusters (High Risk, Moderate Risk, Low Risk).
- K-Means groups patients by similarity in their health markers.
- Summarized clusters based on median health marker values to interpret the risk levels.

PCA for Visualization:

- PCA reduced the data to two principal components (PC1 and PC2), capturing the most variation in the data.
- Identified key contributors to PC1 and PC2 for better interpretation (e.g., PC1 is influenced most by Age or CK-MB).
- Created an interactive scatter plot where each point is a patient colored by cluster (High, Moderate, Low Risk).




```
> print(cluster_summary)
# A tibble: 3 × 6
  patient_type median_age median_SCr median_CKMB median_chol count
         <int>      <dbl>      <dbl>       <dbl>        <dbl> <int>
1            1         56          1          43          246   159
2            2         60          1          30          288   102
3            3         50        0.9          32          219   107
>
```

Insight:

Three patient subtypes were identified:

- **High Risk** (Cluster 1): Older patients (median age 56), higher CK-MB (43), higher cholesterol (246), and moderate serum creatinine (1).

- **Moderate Risk** (Cluster 2): Older (median age 60), highest cholesterol (288), moderate CK-MB (30), and serum creatinine (1).
- **Low Risk** (Cluster 3): Younger (median age 50), lowest serum creatinine (0.9), moderate CK-MB (32), and lowest cholesterol (219).

Visual Insight from PCA Plots:

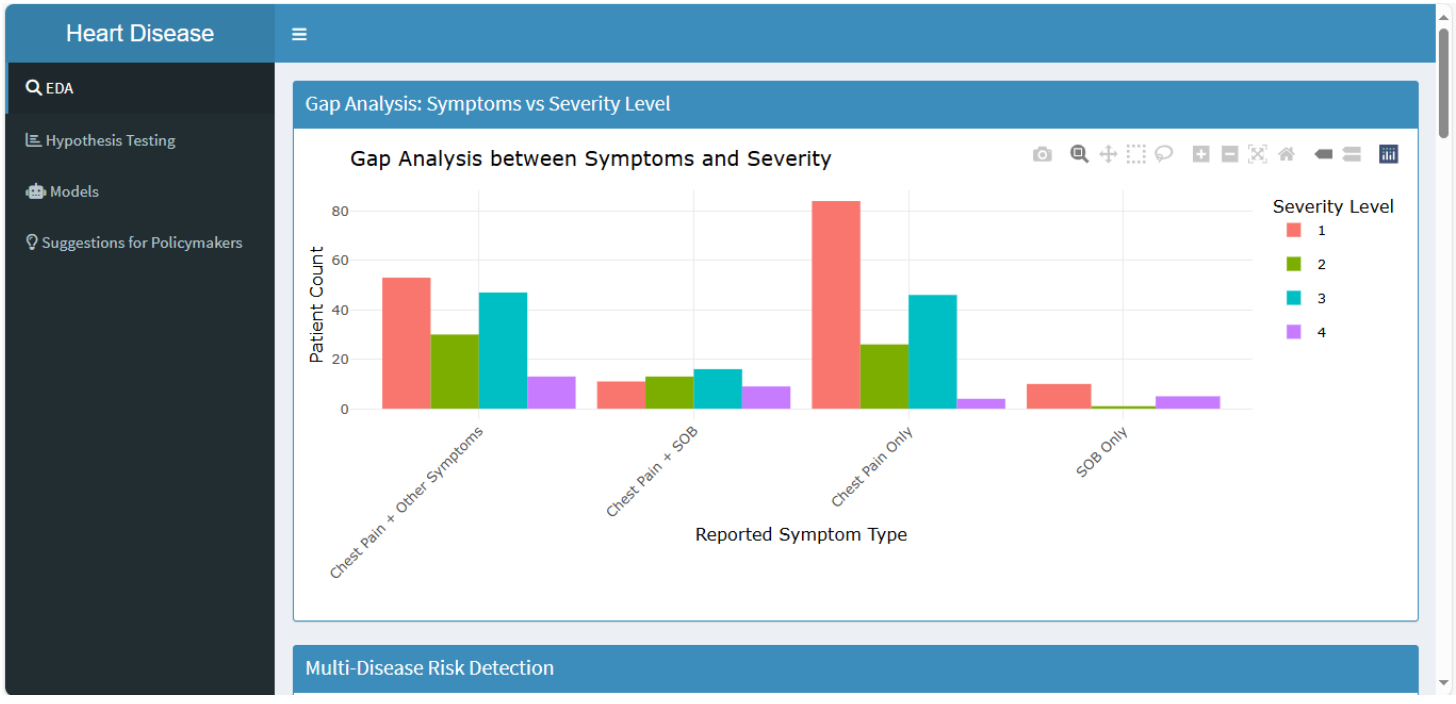
- Patients are grouped along PC1 and PC2 axes, driven by markers like resting BP (trestbps) and serum creatinine (S.Cr).
- High-risk patients cluster together, showing distinct profiles in terms of health markers compared to lower-risk groups.

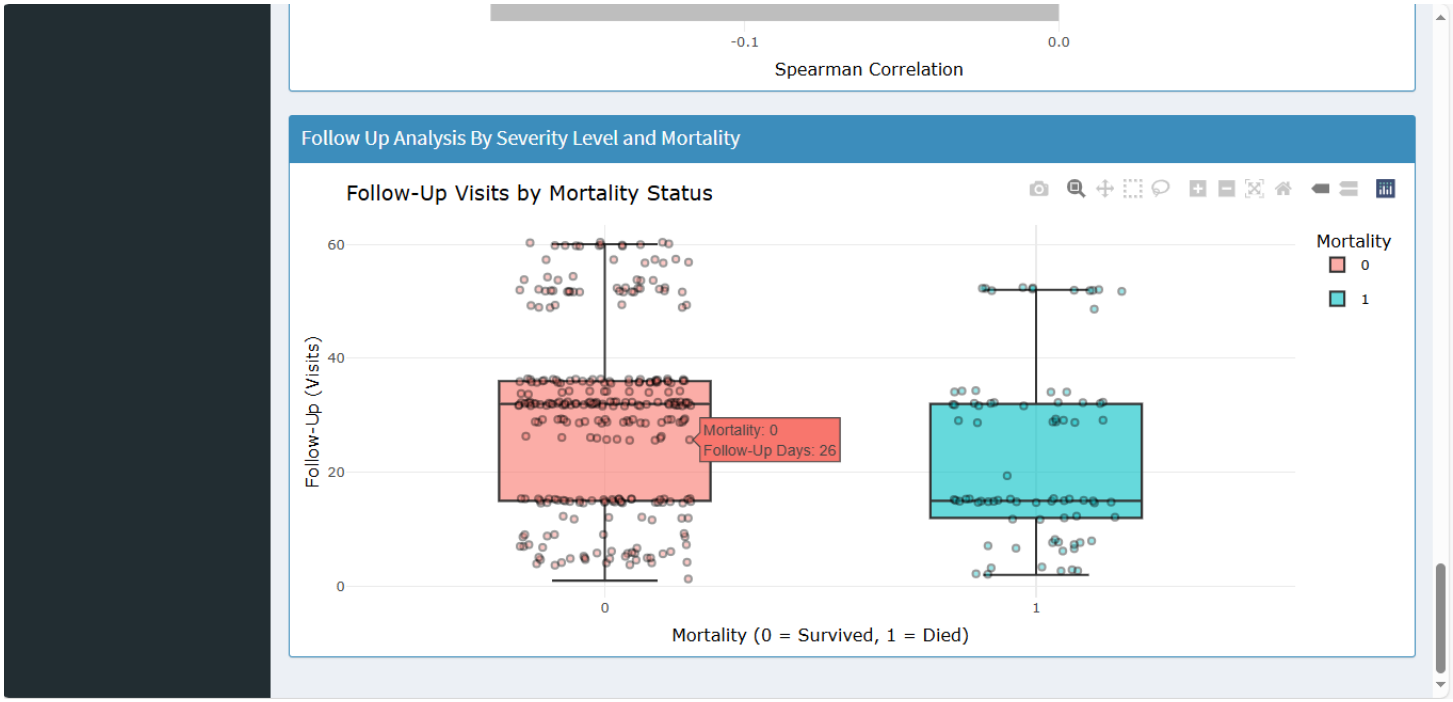
Overall Conclusion:

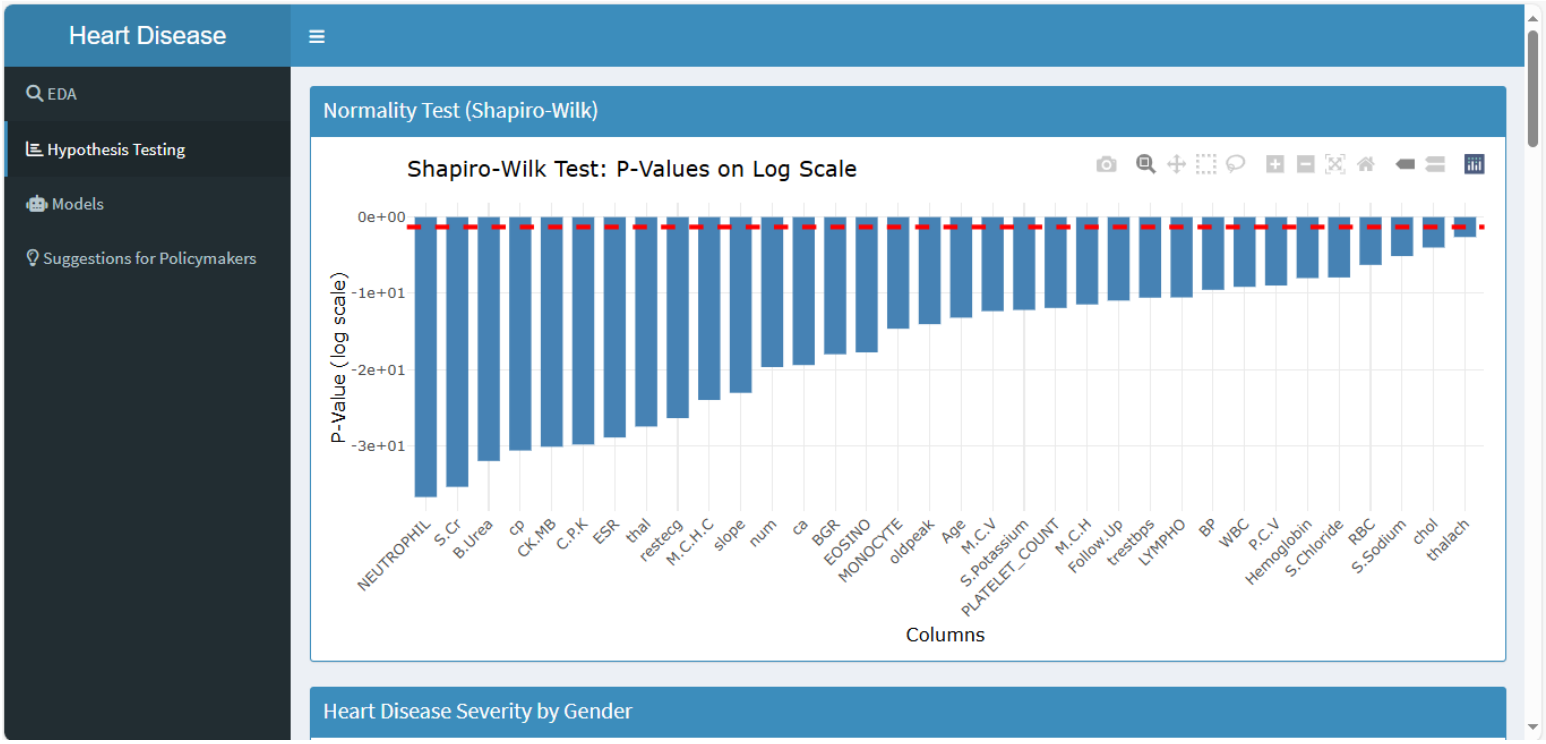
- Older age, higher CK-MB, and cholesterol levels are associated with higher risk.
- Younger patients with lower cholesterol and creatinine levels tend to fall into the low-risk group.

Dashboard

In this dashboard, we've integrated dynamic tooltips and interactive visuals that empower users to effortlessly explore key insights, uncover hidden patterns, and make informed, data-driven decisions with confidence and ease.




-0.1
0.0
Spearman Correlation
Follow Up Analysis By Severity Level and Mortality
Follow-Up Visits by Mortality Status
Mortality
0
1
60
40
20
0
Follow-Up (Visits)
Mortality: 0
Follow-Up Days: 26
0
1
Mortality (0 = Survived, 1 = Died)


Heart Disease
EDA
Hypothesis Testing
Models
Suggestions for Policymakers
Normality Test (Shapiro-Wilk)
Shapiro-Wilk Test: P-Values on Log Scale
0e+00
-1e+01
-2e+01
-3e+01
P-Value (log scale)
NEUTROPHIL
S.Cr
B.Urea
cp
CK.MB
C.P.K
ESR
thal
restecg
M.C.H.C
slope
num
ca
BGR
EOSINO
MONOCYTE
oldpeak
Age
M.C.V
S.Potassium
PLATELET_COUNT
M.C.H
Follow.Up
trestbps
LYMPHO
BP
WBC
P.C.V
Hemoglobin
S.Chloride
RBC
S.Sodium
chol
thalach
Columns
Heart Disease Severity by Gender

Columns

Heart Disease Severity by Gender

ECG & Stress Test Indicators

Heart Disease

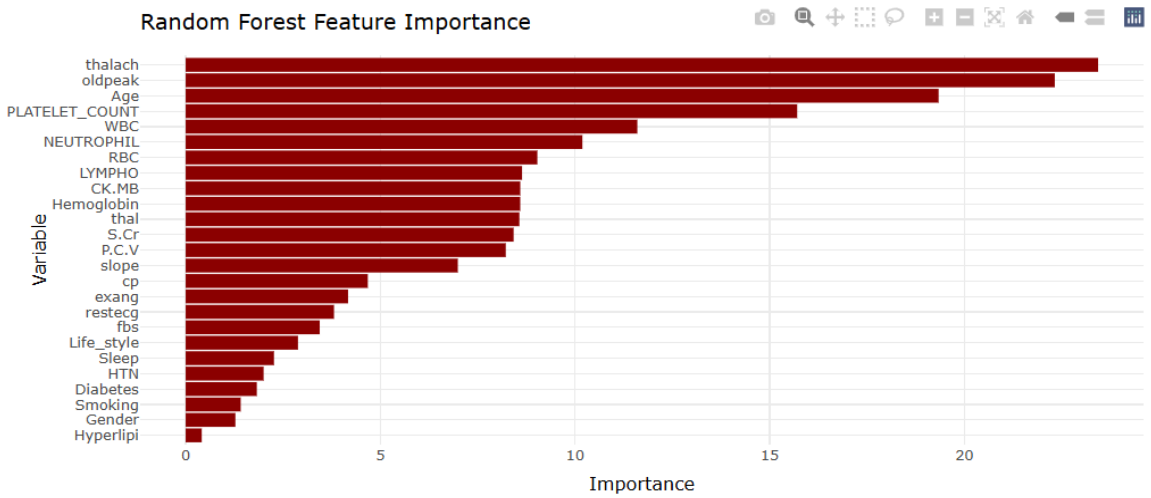
EDA

Hypothesis Testing

Models

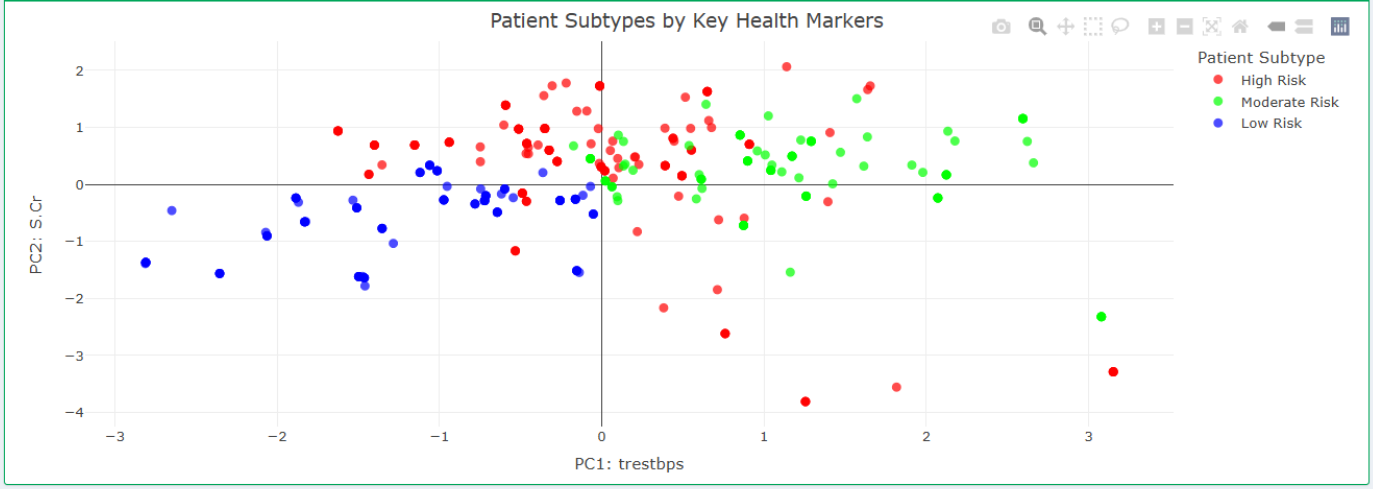
Suggestions for Policymakers

Severity Prediction Model



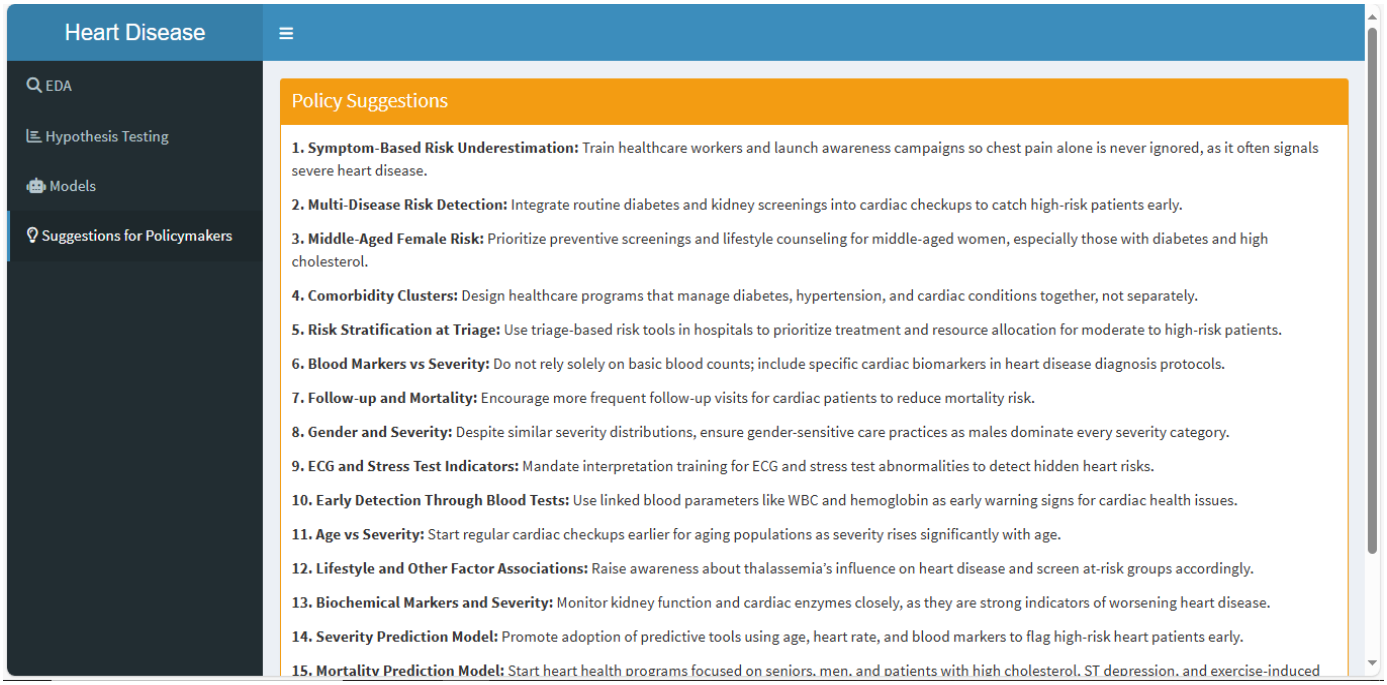
Predicting Death Using Logistic Regression

Clustering Patients into Risk Groups



Predicting Death Using Logistic Regression

Clustering Patients into Risk Groups



Conclusion

This study presents a comprehensive statistical evaluation of heart disease in Pakistan using real-world clinical data and advanced analytics. The findings emphasize several critical insights:

1. **Chest Pain as a Severe Predictor**: Patients presenting with chest pain only were disproportionately represented in the highest severity category. This suggests that even isolated chest pain should not be underestimated, as it may indicate serious underlying conditions.
2. **Comorbid Conditions Exacerbate Severity**: Diabetes and renal dysfunction significantly increase heart disease severity. Patients with both diabetes and kidney risk factors accounted for over 30% of high-severity cases, reinforcing the interconnected nature of chronic diseases.
3. **Gender-Specific Risk Trends**: Middle-aged women (aged 45–59) exhibited high rates of diabetes, cholesterol, and hypertension. Despite normal renal markers, these patients were often classified as high severity, suggesting possible underdiagnosis or delayed care-seeking behavior.
4. **Predictive Modeling Validity**: The Multinomial Logistic Regression and Random Forest models were effective in predicting heart disease severity. Key predictors included thalach (maximum heart rate), oldpeak (ST depression), age, and blood markers (WBC, platelet count, and neutrophils).
5. **Mortality Risk Modeling**: Binary Logistic Regression identified age, gender, exercise-induced angina, oldpeak, and cholesterol as the top predictors of mortality. A weighted risk score derived from these variables demonstrated strong discrimination between survivors and non-survivors.

6. **Clustering and Risk Segmentation**: K-Means clustering grouped patients into high, moderate, and low-risk categories based on biochemical and physiological markers. This stratification offers actionable insights for triage and targeted intervention.
7. **Follow-up Importance**: Patients with higher follow-up visit frequencies showed better survival outcomes, highlighting the value of consistent monitoring in post-diagnosis care.
8. **Blood Markers and Lifestyle Indicators**: Basic hematological markers alone were insufficient for severity prediction, advocating for a multi-marker and lifestyle-integrated assessment approach.

Overall, the study confirms that heart disease in Pakistan is a multifactorial problem requiring data-driven diagnosis, gender-sensitive evaluation, and integrated chronic disease management.

Recommendations

Based on the conclusions drawn, the following recommendations are proposed:

1. **Establish Early Screening Protocols**:
    - Integrate chest pain-only screening into primary care triage to prevent underdiagnosis.
    - Initiate early diagnostic testing for patients reporting even isolated symptoms.
2. **Integrate Comorbidity Management Programs**:
    - Establish combined care pathways for patients with diabetes, hypertension, and renal dysfunction.
    - Prioritize these high-risk clusters for preventive treatment and regular monitoring.
3. **Launch Gender-Specific Interventions**:
    - Create heart health awareness campaigns focused on middle-aged women.
    - Encourage regular lipid and hypertension screening for women, especially in underrepresented communities.
4. **Deploy Predictive Tools in Clinical Practice**:
    - Utilize logistic regression models and risk scores in hospitals to flag high-risk patients.
    - Embed Random Forest-based severity prediction into electronic health systems for real-time risk alerts.
5. **Promote Routine Follow-Up Protocols**:
    - Design standardized follow-up schedules for all discharged heart patients.
    - Use mobile health technologies or dashboards (e.g., Shiny apps) to track and engage patients post-treatment.
6. **Enhance Diagnostic Testing Infrastructure**:
    - Equip rural and urban healthcare centers with tools to measure critical markers like thalach, oldpeak, CK-MB, and serum creatinine.
    - Train frontline health workers to interpret ECG and stress test results effectively.
7. **Adopt Risk Clustering for Triage**:

- Apply cluster-based patient categorization in emergency departments for better resource allocation.
- Use PCA insights to develop visual risk dashboards for quick clinical decisions.

8. **Encourage Further Research and Data Collection**:
    - Conduct longitudinal studies to monitor how risk scores translate into outcomes over time.
    - Expand datasets to include additional variables such as socioeconomic status, family history, and medication adherence.

By implementing these recommendations, healthcare policymakers and practitioners in Pakistan can substantially improve early detection, triage accuracy, and long-term outcomes for heart patients, ultimately reducing the national burden of cardiovascular disease.