MULTIVARIATE STATISTICAL ANALYSIS ON HEART FAILURE PREDICTION DATASET

TUSHITA PANDEY
IIT DELHI

DINTYALA RAHUL BHARDWAJ
IIT DELHI

***Abstract*—Heart disease is a leading cause of global mortality, making early prediction crucial. This study analyzes the Heart Failure Prediction dataset[Sor21], containing patient demographic, clinical, and lifestyle data related to heart failure risk. Through exploratory and inferential statistics, we examined associations between predictor variables—age, cholesterol, blood pressure, smoking status, and others—and heart failure. Key findings indicate that age, cholesterol, and blood pressure strongly correlate with heart failure, with lifestyle factors like smoking and diabetes also significant. We use Logistic regression models to predict high-risk patients, providing insights that may aid healthcare professionals in identifying critical risk factors and preventative strategies.**

I Introduction

The Heart Failure Prediction dataset[Sor21] provides a comprehensive framework for examining the complex interplay of demographic, clinical, and behavioral factors associated with heart disease risk. This dataset includes essential variables such as age, gender, serum cholesterol levels, resting blood pressure, and maximum heart rate, which collectively capture baseline and stress-related cardiovascular indicators. Additional clinical metrics, including chest pain typology, fasting blood glucose, and electrocardiogram (ECG) results, provide valuable insights into underlying physiological and symptomatic variations among patients. These features contribute to a binary target variable that denotes the presence or absence of heart disease, facilitating an investigation into predictive patterns within a clinical context.

Through rigorous analysis of these variables, we seek to quantify the significance of individual and combined risk factors in predicting heart disease. This dataset thus offers a critical foundation for developing predictive models aimed at identifying high-risk individuals, as well as for elucidating broader patterns that may inform preventative strategies and risk stratification in cardiovascular healthcare.

II Pre Processing

The preprocessing steps performed here focus on identifying and encoding categorical variables to make the dataset suitable for machine learning models. First, categorical and numerical features are separated based on the number of unique values in each column. Features with more than 6 unique values are classified as numerical, while those with 6 or fewer unique values are treated as categorical. This distinction is important because categorical variables typically require encoding before being fed into a machine learning algorithm and because because numerical features generally require different preprocessing techniques (e.g., scaling or normalization), while categorical features need to be encoded (e.g., using one-hot encoding or label encoding) before being input into machine learning models.

After identifying the categorical features, Label Encoding is applied to transform these categorical variables into numerical representations. Label Encoding assigns a unique integer to each category, making the data compatible with models that require numerical input. This step ensures that the categorical features, such as Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope, are appropriately transformed for subsequent analysis. By encoding these variables, we prepare the dataset for efficient processing and model training.

III Exploratory Data Analysis

III-A Attribute Analysis

The dataset comprises a range of features, including categorical variables such as `Sex`, `ChestPainType`, `FastingBS`, `RestingECG`, `ExerciseAngina`, `ST_Slope`, and `HeartDisease`. In addition, it includes numerical variables such as `Age`, `RestingBP`, `Cholesterol`, `MaxHR`, and `Oldpeak`.

TABLE I: Attribute Information

Attribute	Description	Values
Age	Age of the patient	Years
Sex	Sex of the patient	M, F
ChestPainType	Chest pain type	TA, ATA, NAP, ASY
RestingBP	Resting blood pressure	mm Hg
Cholesterol	Serum cholesterol	mm/dl
FastingBS	Fasting blood sugar	1, 0
RestingECG	Resting ECG results	Normal, ST, LVH
MaxHR	Maximum heart rate	Beats per minute
ExerciseAngina	Exercise-induced angina	Y, N
Oldpeak	Oldpeak = ST	Numeric value
ST_Slope	Slope of peak exercise ST segment	Up, Flat, Down
HeartDisease	Output class	1, 0

In Table II, we present the sample mean, sample standard deviation, minimum, maximum, and the three quartile values.

TABLE II: Attribute Analysis

	COUNT	MEAN	STD	MIN	25%	50 %	75 %	MAX
AGE	918.00	53.51	9.43	28.00	47.00	54.00	60.00	77.00
RESTINGBP	918.00	132.40	18.51	0.00	120.00	130.00	140.00	200.00
CHOLESTEROL	918.00	198.80	109.38	0.00	173.25	223.00	267.00	603.00
FASTINGBS	918.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00
MAXHR	918.00	136.81	25.46	60.00	120.00	138.00	156.00	202.00
OLDPEAK	918.00	0.89	1.07	-2.60	0.00	0.60	1.50	6.20
HEARTDISEASE	918.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00

We conduct an analysis of the means of various attributes between two groups: those with `HeartDisease` = 1 and those with `HeartDisease` = 0.

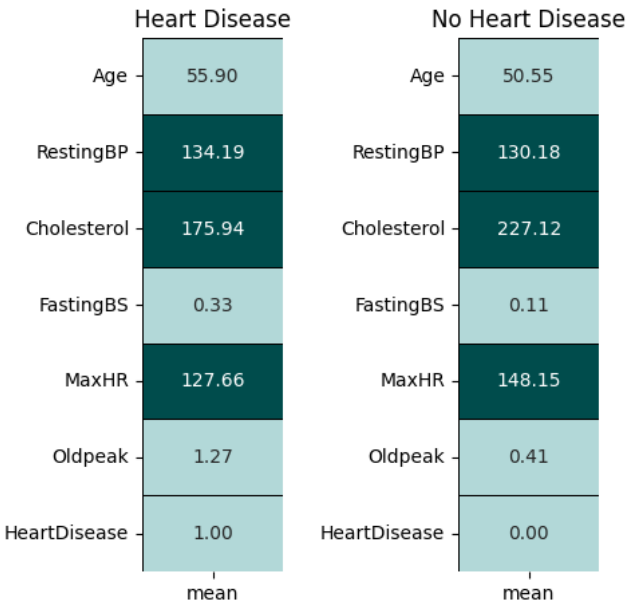

Fig. 1: Comparison of Feature Means by `HeartDisease` Status

III-B Testing Normality Assumptions

To visualize the distribution of the numerical features, we create density plots, as illustrated below.

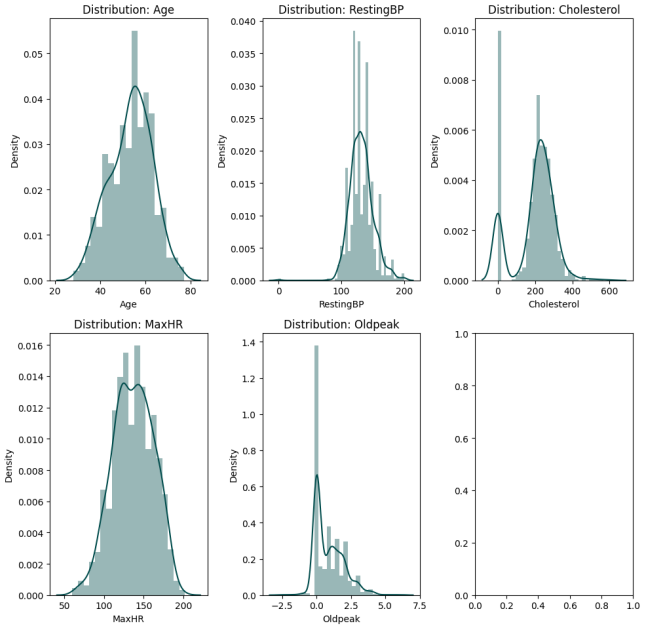

Fig. 2: Density Plots of Numerical Features

The resulting plots indicate that the variables exhibit an approximate normal distribution although a joint normal distribution in the data is elusive.

A Q-Q (Quantile-Quantile) plot [GW68] is a graphical tool used to assess if a dataset follows a particular theoretical distribution, like the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data is normally distributed, the points will align closely along a straight line; significant deviations from this line suggest departures from normality, such as skewness or heavy tails.

We make separate plots to test normality. In the first plot, we include categorical features and in the second plot we omit them. The resulting plots look as follows:

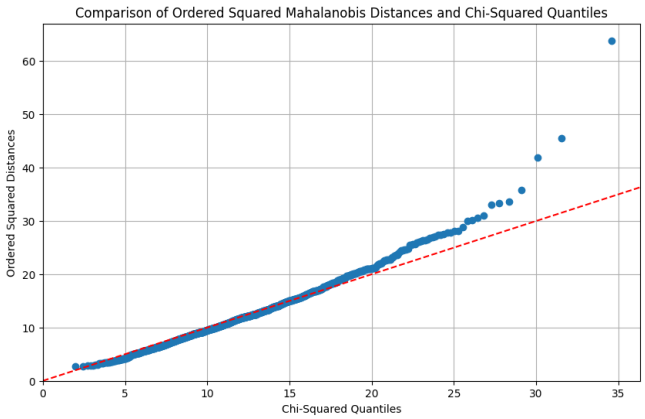

Fig. 3: Q-Q Plot with Categorical Features

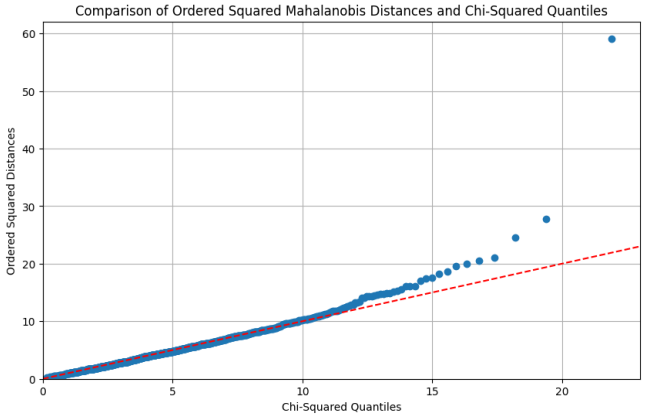

Fig. 4: Q-Q Plot without Categorical Features

In both the cases, we note a significant deviation from a normal distribution. Hence, we conclude that the data does not follow a joint normal distribution.

III-C Target Variable Visualisation

We wish to analyse the dependence of our target variable (`HeartDiseases`) with other numerical and categorical data.

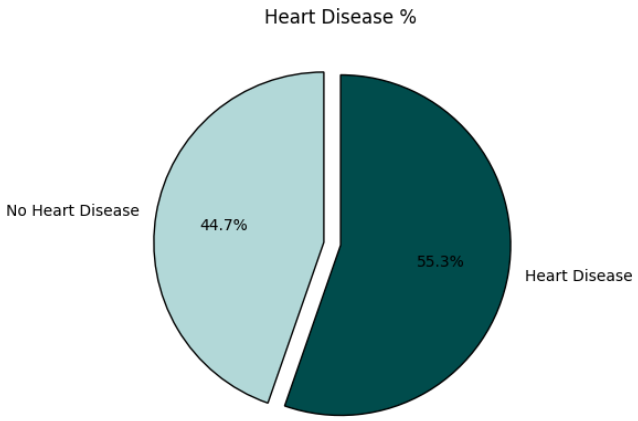

Fig. 5: `HeartDisease`

Among the 918 patients in our dataset, 508 are diagnosed with heart disease, while 410 are not. This distribution indicates that our dataset is fairly well-balanced and thus sampling bias will not be an issue while training classification models.

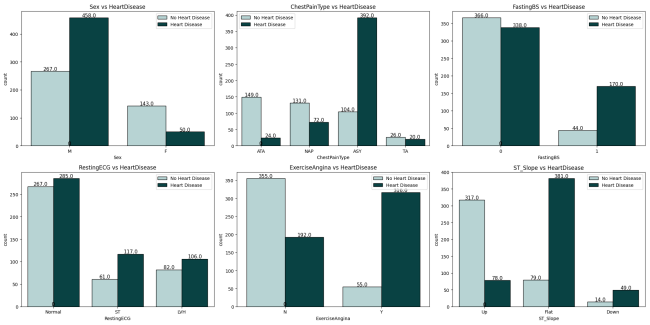

Fig. 6: Categorical Features vs `HeartDisease`

To identify potential correlations between categorical variables and `HeartDisease`, we created the plots shown in Figure 6. Key observations include:

1) `Sex` appears to influence heart disease prevalence, with men showing a higher incidence, while women have a lower incidence compared to those without heart disease.
2) The presence of the ASY `ChestPainType` is strongly associated with a higher likelihood of heart disease.
3) `FastingBS` status is complex; both patients with elevated and normal fasting blood sugar levels show substantial heart disease incidence.
4) `RestingECG` does not indicate a specific category that is predominantly associated with heart disease;

all three categories exhibit a high number of cases.

5) `ExerciseAngina` definitely bumps the probability of being diagnosed with heart diseases.
6) For `ST_Slope` values, a flat slope is highly indicative of heart disease, while a downward slope suggests elevated risk, though with fewer occurrences.

Between patients with diagnosed heart disease, we'd also like to see the distribution of various categorical features.

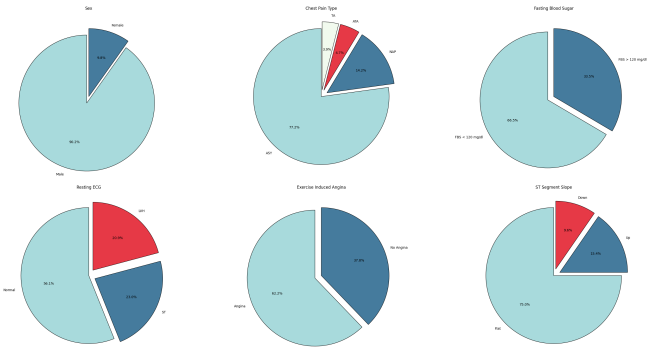

Fig. 7: Categorical Features vs Positive `HeartDisease` cases

Figure 7 tells us that 90.2% of patients with heart disease are men, 77.2% have asymptomatic chest pain, 66.5% have fasting blood sugar less than 120mg/dl, 56.1% have normal resting ECG, 62.2% have exercise induced angina, and 75% have flat ST-Segment slope.

We additionally seek to examine the relationships among various features to identify any underlying correlation patterns. To that end, we make scatter plots (Figure 8) for each pair of numerical variables. Additionally, each point in the scatter plot is also coloured • if the patient with that data has heart disease and • otherwise.

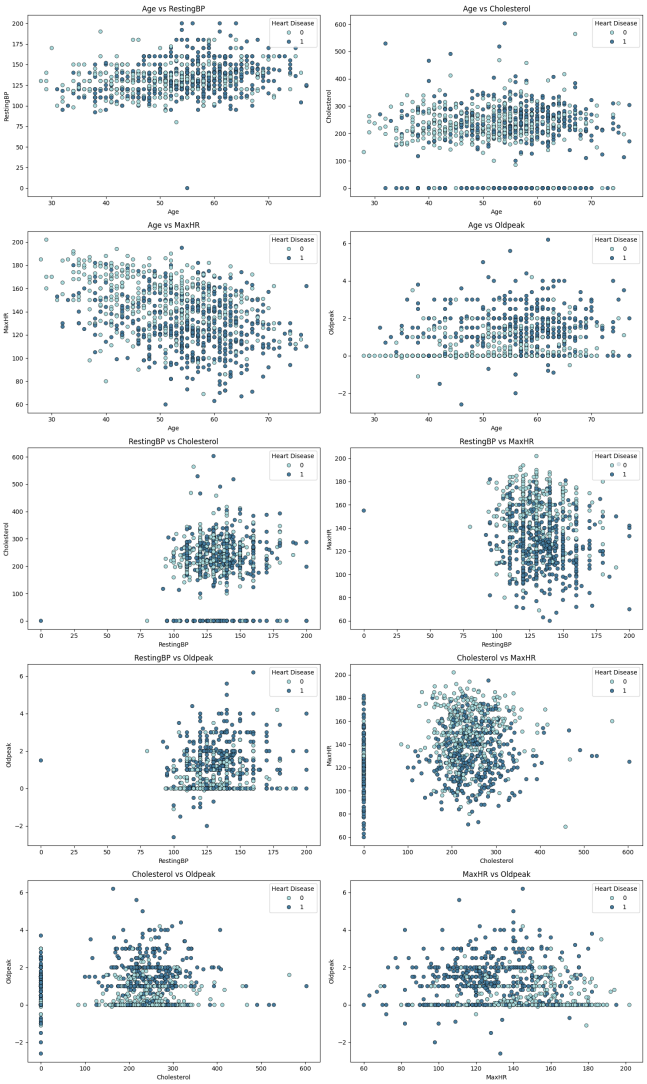

Fig. 8: Scatter Plots between Numerical Features

We visually observe the following trends from the scatter plots:

1) **Age vs RestingBP:**
   - Resting blood pressure values are scattered across the age range, with no clear trend solely based on these two variables.
   - For individuals aged 50 and above, RestingBP values between 100 and 175 are common among heart disease cases.
2) **Age vs Cholesterol:**
   - Cholesterol levels vary widely across different ages.
   - For individuals with cholesterol levels between 200 and 300, heart disease cases are notably more frequent.

3) **Age vs MaxHR:**
   - Maximum heart rate (MaxHR) decreases with age, which aligns with expected physiological trends.
   - Heart disease cases are more common at MaxHR values below 140, especially in older age groups.
4) **Age vs Oldpeak:**
   - Oldpeak values do not show a strong relationship with age.
   - Higher Oldpeak values are slightly more common among individuals with heart disease, particularly in the older age groups.
5) **RestingBP vs Cholesterol:**
   - Resting blood pressure values between 100 and 175 span both heart disease and non-heart disease cases.
   - Cholesterol values in the range of 200 to 300 are associated with a higher likelihood of heart disease.
6) **RestingBP vs MaxHR:**
   - A modest negative correlation is observed, with individuals who have higher RestingBP tending to have slightly lower MaxHR.
   - Heart disease cases are concentrated at lower MaxHR values, regardless of resting blood pressure, especially when RestingBP is between 100 and 175.
7) **RestingBP vs Oldpeak:**
   - There is no distinct relationship between RestingBP and Oldpeak values.
   - Higher Oldpeak values appear more frequently in heart disease cases, particularly when RestingBP is within the range of 100 to 175.
8) **Cholesterol vs MaxHR:**
   - Individuals with cholesterol levels between 200 and 300 often have lower MaxHR values, with a concentration of heart disease cases in this range.
9) **Cholesterol vs Oldpeak:**
   - Higher cholesterol levels, particularly between 200 and 300, along with higher Oldpeak values, show an increased incidence of heart disease.
10) **MaxHR vs Oldpeak:**
   - An inverse relationship is observed: as MaxHR decreases, Oldpeak values tend to increase.
   - Heart disease cases tend to have lower MaxHR values (especially below 140) and higher Oldpeak values.

Summary

1) MaxHR and Oldpeak appear to be significant indicators for heart disease: lower MaxHR and higher Oldpeak values correlate with heart disease cases.
2) Cholesterol levels, especially between 200 and 300, show an association with heart disease, though the relationship is weak.
3) RestingBP values between 100 and 175 consistently appear among heart disease cases across various feature pairings.
4) Age on its own does not distinctly separate heart disease cases, though age combined with other factors such as MaxHR provides more insight.
5) Overall, these patterns suggest that the risk of heart disease in this dataset is influenced by a combination of factors, with MaxHR, Oldpeak, Cholesterol, and RestingBP being noteworthy indicators.

One way to numerically see the relationship between features is to make a sample correlation matrix as in 9 The sample correlation between two random variables samples $X = (X_i)_{1 \le i \le n}$ and $Y = (Y_i)_{1 \le i \le n}$ is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

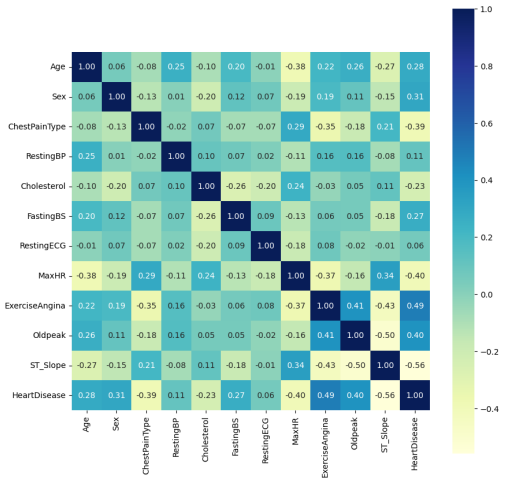

Fig. 9: Correlation matrix

Heart Disease Correlations

`ST_Slope` has a strong negative correlation with heart disease (-0.56), meaning lower values increase heart disease risk. `ExerciseAngina` is positively correlated (0.49), linking exercise-induced angina to higher heart disease likelihood. `Oldpeak` has a positive correlation (0.40), with higher values indicating greater risk. `Age` has a moderate

positive correlation (0.28), suggesting slightly higher risk with age. Finally, MaxHR is negatively correlated (-0.40), with lower maximum heart rates linked to increased heart disease risk.

Feature Correlations

MaxHR and Age have a moderate negative correlation (-0.38), aligning with the typical decline in maximum heart rate with age. ST_Slope and Oldpeak show a strong negative correlation (-0.50), suggesting that as ST_Slope decreases, Oldpeak tends to increase, indicating a potential link between ST depression and ST slope during exercise.

Modeling Insights

ST_Slope, ExerciseAngina, Oldpeak, MaxHR, and Age may be key predictors for heart disease models, while low-correlation variables like Cholesterol and RestingBP could be weighted less or excluded.

IV Exploratory Data Analysis Summary

Our analysis of the dataset yields several key insights related to HeartDisease, focusing on how various features correlate with the presence or absence of heart disease.

IV-A Categorical Features Observations

The analysis of categorical features reveals clear patterns:

- Sex: Men exhibit a higher incidence of heart disease compared to women.
- ChestPainType: The presence of asymptomatic chest pain (ASY) is strongly associated with heart disease, while other types show less correlation.
- FastingBS: Both elevated and normal fasting blood sugar levels are common in heart disease cases, indicating a complex relationship.
- RestingECG: No specific ECG type is predominantly associated with heart disease, though all categories show some prevalence.
- ExerciseAngina: This is a strong indicator of heart disease; patients with exercise-induced angina have a higher incidence of heart disease.
- ST_Slope: A flat ST segment slope is highly indicative of heart disease, and a downward slope suggests a significant risk.

IV-B Numerical Feature Observations

Analysis of the numerical features through mean comparisons, density plots, and scatter plots highlights several key points:

- Age: Patients over 50 tend to have higher rates of heart disease, especially those with specific ranges of RestingBP and Cholesterol.
- Cholesterol: Levels between 200 and 300 mg/dl are common in heart disease cases, though overall correlation with HeartDisease is weak.
- MaxHR: Heart disease is more prevalent among patients with a MaxHR below 140 beats per minute.
- Oldpeak: Higher Oldpeak values are positively correlated with heart disease, often indicating ST depression related to abnormal ST slope patterns.

IV-C Correlation Analysis

The correlation matrix further refines our understanding:

- ST_Slope shows a strong negative correlation with HeartDisease (-0.56), indicating lower values are associated with higher heart disease risk.
- ExerciseAngina is positively correlated with HeartDisease (0.49), reinforcing its role as a risk factor.
- MaxHR and Age have a moderate negative correlation (-0.38), in line with the typical decrease in maximum heart rate with age.
- Oldpeak and ST_Slope show a strong negative correlation (-0.50), suggesting that higher ST depression correlates with a flatter ST slope.

IV-D Implications for Modeling

The insights above suggest several features as important predictors for heart disease:

- Key predictors: ST_Slope, ExerciseAngina, Oldpeak, MaxHR, and Age emerge as critical features for predictive models due to their strong or moderate correlations with heart disease.
- Lesser impact: Features with lower correlations, such as Cholesterol and RestingBP, may have a reduced role in model development or may even be excluded.

These findings set a clear direction for developing machine learning models, with a focus on optimizing the contribution of highly correlated variables to improve predictive accuracy.

V Principal Component Analysis

V-A Principal Component Analysis

Principal Component Analysis (PCA) [WEG87] is a dimensionality reduction technique that transforms a dataset into a new coordinate system defined by its principal components, which are the directions of maximum variance. It starts by centering the data and computing the covariance matrix, from which eigenvalues and eigenvectors are derived. The eigenvectors corresponding to the largest eigenvalues form the new axes, allowing us to reduce the dataset's dimensions by selecting the top k components that capture the most variance, thereby simplifying the data while retaining essential information.

We get the following values of eigenvalues and eigenvectors:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Age	-0.33	0.08	0.51	-0.00	-0.15	-0.27	-0.30	0.30	-0.43	0.26	-0.30
Sex	-0.22	-0.27	-0.21	-0.22	0.83	-0.13	-0.12	0.23	-0.10	-0.06	-0.05
ChestPainType	0.29	0.02	0.42	-0.24	0.17	0.40	-0.53	0.05	0.41	0.16	0.10
RestingBP	-0.17	0.27	0.52	0.39	0.43	-0.00	0.25	-0.43	0.08	-0.19	-0.03
Cholesterol	0.16	0.60	-0.06	0.18	0.10	-0.00	0.16	0.70	0.16	-0.15	0.00
FastingBS	-0.20	-0.34	0.39	-0.35	-0.09	0.12	0.63	0.30	0.19	0.05	0.16
RestingECG	-0.11	-0.40	-0.04	0.61	-0.01	0.59	-0.04	0.25	-0.14	-0.01	-0.12
MaxHR	0.40	0.17	-0.03	-0.25	0.17	0.34	0.33	-0.13	-0.36	0.31	-0.50
ExerciseAngina	-0.42	0.17	-0.26	0.12	0.04	0.03	0.06	-0.07	0.44	0.70	-0.12
Oldpeak	-0.36	0.37	-0.08	-0.21	0.01	0.43	-0.04	-0.07	-0.42	0.03	0.56
ST_Slope	0.43	-0.15	0.10	0.30	0.17	-0.28	0.10	0.06	-0.22	0.50	0.54
Eigenvalue	2.768	1.466	1.158	1.001	0.872	0.844	0.808	0.625	0.558	0.502	0.408
Explained Variance	0.25	0.38	0.49	0.58	0.66	0.74	0.81	0.87	0.92	0.96	1.00

TABLE III: Principal Component Analysis (PCA) Loadings

Component Contributions

- **PC1**: This component seems to be heavily influenced by MaxHR (-0.61) and Age (-0.52). These features might indicate that older individuals with lower maximum heart rates tend to vary together in this dataset.
- **PC2**: Influenced significantly by RestingBP (0.69) and Cholesterol (0.61), suggesting these features correlate positively in the dataset. High values in these features may indicate higher cardiovascular risk.
- **PC3**: The feature ChestPainType has a strong positive contribution (0.76), which could indicate that different types of chest pain are significant in differentiating this component from others.

Visualization of Principal Components To assess the relationship between heart disease presence and absence, we visualized the first two principal components (PC1 and PC2) on a 2D and 3D scatter plot, with each data point color-coded to indicate the presence or absence of heart disease.

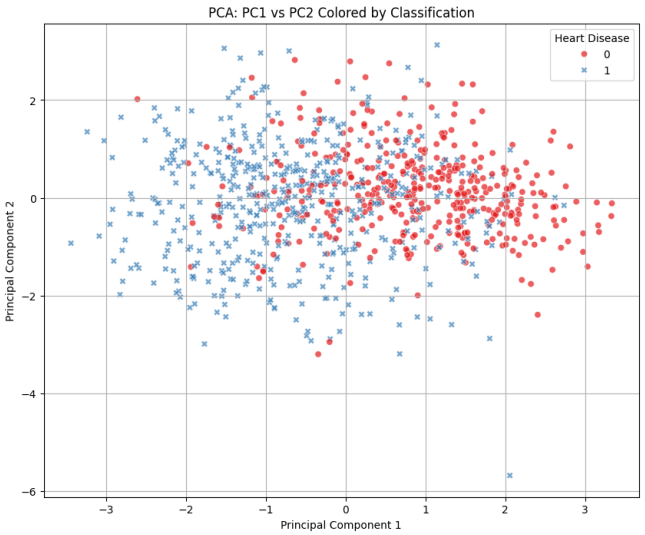

Fig. 10: First 2 principal components Scatter plot

The scatter plot reveals some clustering of cases with and without heart disease. However, there is no clear, linear boundary that fully separates the two classes using only PC1 and PC2.

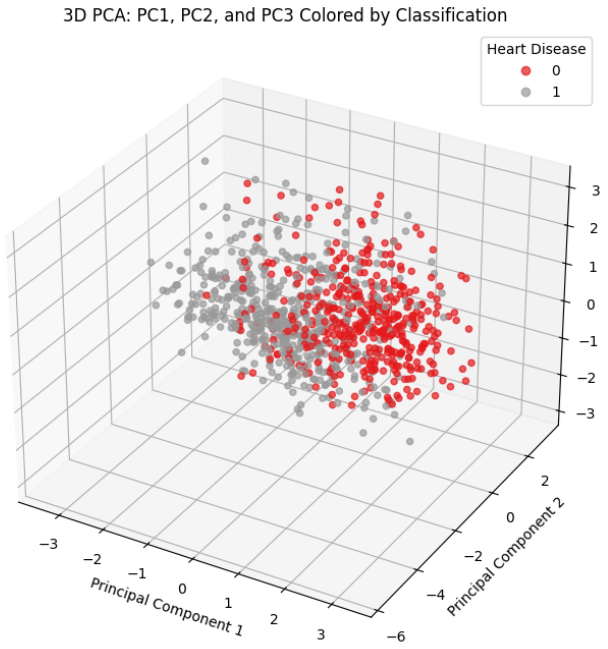

Fig. 11: First 3 principal components Scatter plot

While the 3D scatter plot of the first three principal components provides a richer view of the data, there is still noticeable overlap between the classes. This indicates that distinguishing between the presence and absence of heart disease may require more complex modeling techniques,

such as non-linear classifiers or additional features, to capture underlying patterns. Higher-dimensional analysis or advanced models could potentially enhance separation between the classes, improving classification accuracy.

VI Classification Models

VI-A Data Scaling

We scale the data to ensure uniform treatment of features with differing units and ranges—such as `Age` (in years) and `FastingBS` (in mg/dL)—to prevent any single feature from dominating model performance.

Normalization is applied to right-skewed features like `Oldpeak`, transforming values to a 0 to 1 range, which prevents wider-ranging features from overshadowing others. For normally distributed features, including `Age`, `RestingBP`, `Cholesterol`, and `MaxHR`, we standardize by adjusting the mean to 0 and standard deviation to 1, ensuring equitable contributions. These scaling techniques create a balanced feature space, enhancing model performance and promoting fair contributions in heart disease prediction.

VI-B k-Nearest Neighbours Algorithm

The K-Nearest Neighbors (KNN) algorithm [CH67] is a simple, non-parametric machine learning method used for classification and regression tasks. It operates on the principle of proximity, classifying a new data point based on the majority label of its k nearest neighbors in the feature space. For instance, in a classification task like heart disease prediction, KNN identifies the k most similar historical patient records and assigns a label based on the majority class among these neighbors. The choice of k significantly influences the model's performance, where a smaller k might capture finer details and potentially overfit, while a larger k smooths out predictions but risks underfitting. As KNN relies on distance calculations, the data is typically scaled, and Euclidean distance is a common metric used to find the closest points. KNN is valued for its interpretability and simplicity, though it can be computationally intensive with large datasets.

Results

Solid Performance but Sensitive to Dimensionality: KNN achieved a high cross-validation ROC AUC score of 92.68%, with a slightly lower accuracy of 83.15% on the test set. This difference suggests that KNN's effectiveness may be impacted by the complexity and dimensionality of the dataset. However, its strong performance in detecting heart disease cases (Class 1) indicates that KNN is capable of accurately identifying patients with heart disease. KNN's sensitivity to data structure may lead to reduced performance in high-dimensional spaces, a limitation when working with complex datasets.

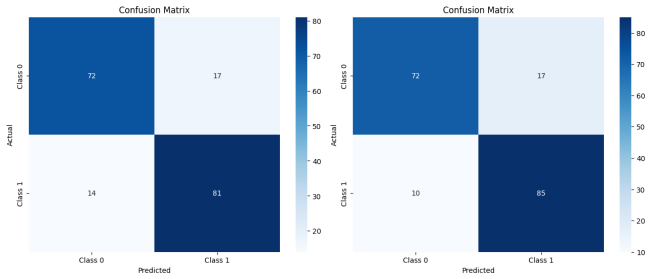

Fig. 12: k-NN

Fig. 13: Random Forest

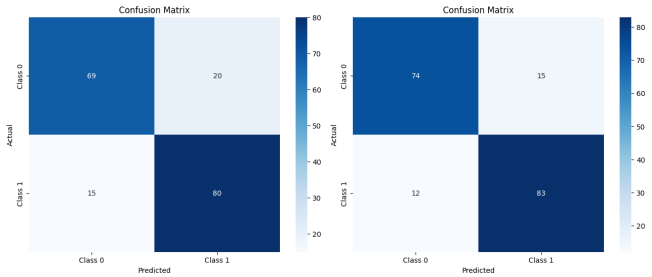

Fig. 14: SVM

Fig. 15: Logistic Regression

Fig. 16: Confusion Matrices

VI-C Random Forest

Random Forest [Bre01] is an ensemble learning technique primarily used for classification and regression tasks in machine learning. It operates by constructing a multitude of decision trees during training and outputting the mode of their predictions (for classification) or the mean prediction (for regression). This method enhances predictive accuracy and controls overfitting by averaging the results from various trees, which reduces variance without significantly increasing bias. Each tree is built using a random subset of the data and a random subset of features, promoting diversity among the trees. Random Forest is particularly valued for its robustness, ability to handle large datasets with high dimensionality, and its capacity to provide insights into feature importance, making it a popular choice in various domains, from finance to healthcare.

Results

Best Performing Model: With an accuracy of 84.78% and an ROC AUC score of 93.15%, the Random Forest model

demonstrates strong predictive performance on the heart disease dataset. It achieves a balanced precision and recall across both classes, effectively identifying both heart disease and non-heart disease cases. This balance makes it well-suited for medical applications where minimizing both false positives and false negatives is critical. Additionally, the ensemble approach of Random Forest reduces the risk of overfitting, making it a robust choice for heart disease prediction tasks.

VI-D Support Vector Machine (SVM)

Support Vector Machine (SVM) [Cor95] is a supervised learning algorithm used for classification and regression tasks. It identifies the optimal hyperplane that separates data points of different classes, maximizing the margin between them, which are known as support vectors. SVM can handle both linear and non-linear classifications by using kernel functions to transform data into higher dimensions. Its effectiveness in high-dimensional spaces and robustness against overfitting make SVM a popular choice for applications such as image recognition, bioinformatics, and text classification.

Results

Lower Accuracy, but Competitive ROC AUC: With an accuracy of 80.98% and a cross-validation ROC AUC of 92.42%, the SVM model shows some limitations in overall accuracy while maintaining good predictive power for heart disease cases. The lower accuracy suggests that SVM might have slightly higher false-positive rates for non-heart disease cases (Class 0), potentially flagging some individuals without heart disease as at-risk. This could lead to unnecessary follow-ups in a clinical setting, so SVM may be best used alongside models with higher generalization for both classes.

VI-E Logistic Regression

Logistic regression is a statistical method used for binary classification that predicts the probability of an event based on one or more predictor variables. It employs the logistic function to model the relationship between the dependent binary variable and independent variables, outputting values between 0 and 1. This technique is valued for its simplicity and interpretability, making it suitable for various applications in fields like healthcare, finance, and social sciences, where it is used for tasks such as risk assessment and decision-making. For multiclass scenarios, logistic regression can be extended to multinomial logistic regression.

Results

Strong Baseline with High Interpretability: Logistic Regression achieved an accuracy of 85.33% and a balanced ROC AUC of 91.07%, making it a strong baseline model for heart disease prediction. The model's balanced performance in precision and recall, combined with its interpretability, makes it particularly valuable for clinical use, where understanding the relationship between predictors and outcomes is key. Logistic Regression's interpretable coefficients allow identification of features most associated with the risk of heart disease, offering insights that clinicians can use for diagnostic purposes.

VII Discussion

Currently, we have not done data engineering as i using the results of the correlation analysis to increase the performance of the classification models but as we can see, the models perform pretty well even without that engineering. We can incorporate those findings to build stronger models and that can be future point of action.

References

[CH67] T. Cover and P. Hart. "Nearest Neighbor Pattern Classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[GW68] Ramanathan Gnanadesikan and Martin B Wilk. "Probability plotting methods for the analysis of data". In: *Biometrika* 55.1 (1968), pp. 1–17.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[Cor95] Corinna Cortes. "Support-Vector Networks". In: *Machine Learning* (1995).

[Bre01] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[Sor21] Federico Soriano. *Heart Failure Prediction Dataset*. Accessed: 2024-11-03. 2021. URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data.