Discriminating Heart Disease Presence Using Clinical Measurements: A Linear Discriminant Analysis Approach

Name: Rosu Liviu-Mihai

Group: 1107

1. Title and Abstract

Heart disease remains one of the leading causes of mortality worldwide, making early identification of at-risk individuals a critical public health objective. This study applies Linear Discriminant Analysis (LDA) to a clinical heart disease dataset in order to identify which medical variables best discriminate between patients with and without diagnosed heart disease. The dataset consists of individual-level observations including demographic information, physiological measurements, and diagnostic test results. Prior to analysis, variables were standardized and the outcome variable was encoded as a binary classification. Exploratory analysis was conducted to examine the relationship between selected risk factors and heart disease status. LDA was then applied to assess class separability and variable importance. The results indicate that LDA achieves meaningful separation between the two groups, with variables such as ST depression, chest pain type, and thallium test results contributing most strongly to discrimination. These findings suggest that multivariate linear techniques can effectively summarize and interpret complex clinical data, while also providing insights into the most relevant predictors of heart disease presence.

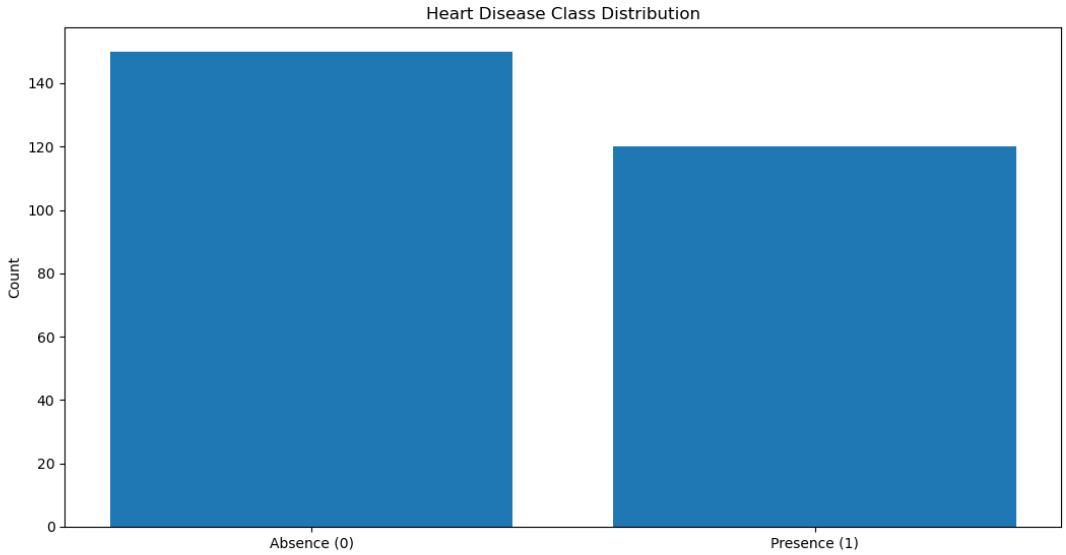
2. Introduction

Heart disease is a multifactorial condition influenced by demographic, physiological, and lifestyle-related factors. While individual clinical indicators such as blood pressure or cholesterol are commonly examined, these variables often interact in complex ways that are not easily captured through univariate analysis. Multivariate statistical methods provide a framework for analyzing such interactions simultaneously.

This study investigates whether patients with and without heart disease can be effectively distinguished using a set of clinical measurements. The research question guiding this analysis is:

Which clinical variables best discriminate between patients with and without heart disease, and how effectively can Linear Discriminant Analysis separate these two groups?

To address this question, Linear Discriminant Analysis (LDA) is employed due to its suitability for supervised classification problems and its interpretability.



3. Data Description

- The variables included in the analysis consist of one binary target variable and multiple explanatory variables. The target variable, *Heart Disease*, indicates whether heart disease was detected (*Presence*) or not (*Absence*). The explanatory variables capture demographic characteristics (e.g., age and sex), clinical measurements (e.g., blood pressure, cholesterol, and maximum heart rate), and diagnostic test results (e.g., electrocardiogram findings and thallium stress test outcomes).
- **Age**: Age of the patient in years.
- **Sex**: Gender of the patient (1 = Male, 0 = Female).
- **Chest pain type**: Categorical indicator of chest pain experienced by the patient.
- **BP**: Resting blood pressure measured in mm Hg.
- **Cholesterol**: Serum cholesterol level measured in mg/dL.
- **FBS over 120**: Indicator of fasting blood sugar exceeding 120 mg/dL.
- **EKG results**: Resting electrocardiogram results categorized into clinically meaningful groups.
- **Max HR**: Maximum heart rate achieved during exercise.
- **Exercise angina**: Indicator of exercise-induced angina.
- **ST depression**: ST depression induced by exercise relative to rest.
- **Slope of ST**: Slope of the peak exercise ST segment.
- **Number of vessels fluro**: Number of major vessels colored by fluoroscopy.
- **Thallium**: Thallium stress test result.

These variables were selected because they are clinically relevant indicators commonly associated with cardiovascular risk and are expected

to jointly contribute to the discrimination between patients with and without heart disease.

Data Preprocessing

The dataset does not contain missing values; therefore, no imputation procedures were required. All variables were retained in the analysis.

Prior to applying Linear Discriminant Analysis, all explanatory variables were standardized to have zero mean and unit variance. Standardization was necessary because the variables are measured on different scales, and LDA is sensitive to differences in magnitude across predictors.

No additional variable selection or dimensionality reduction was performed prior to analysis. Instead, all available clinical variables were included to allow the LDA model to identify the linear combination of predictors that maximizes separation between the two outcome groups.

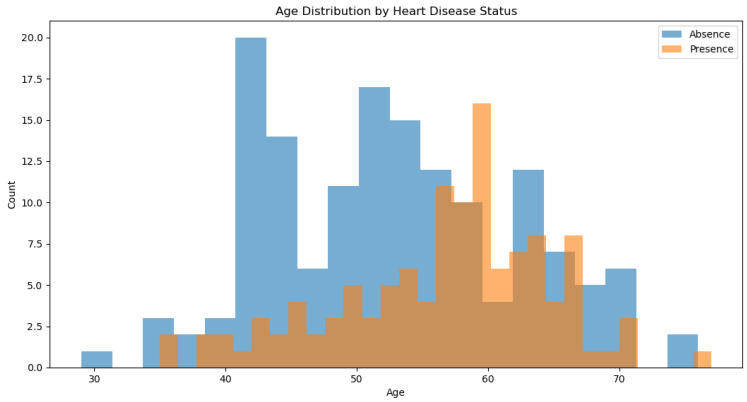
4. Methodology

Linear Discriminant Analysis is a supervised multivariate technique designed to find linear combinations of variables that best separate predefined groups. In the binary case, LDA produces a single discriminant axis that maximizes the ratio of between-group variance to within-group variance.

LDA assumes approximate normality within classes and similar covariance structures across groups. Although these assumptions may not be strictly met in clinical data, LDA remains effective and interpretable for exploratory classification tasks.

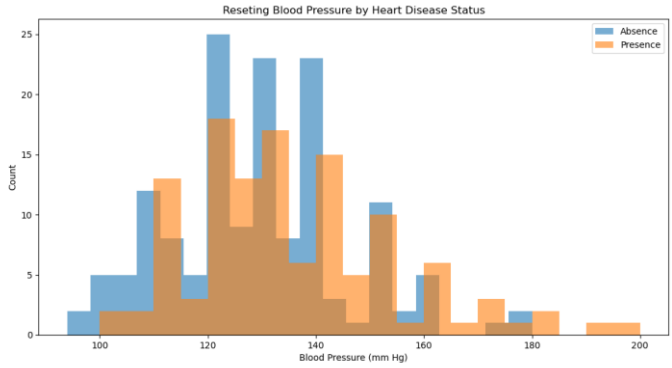
5. Results

Age as a potential risk to heart disease



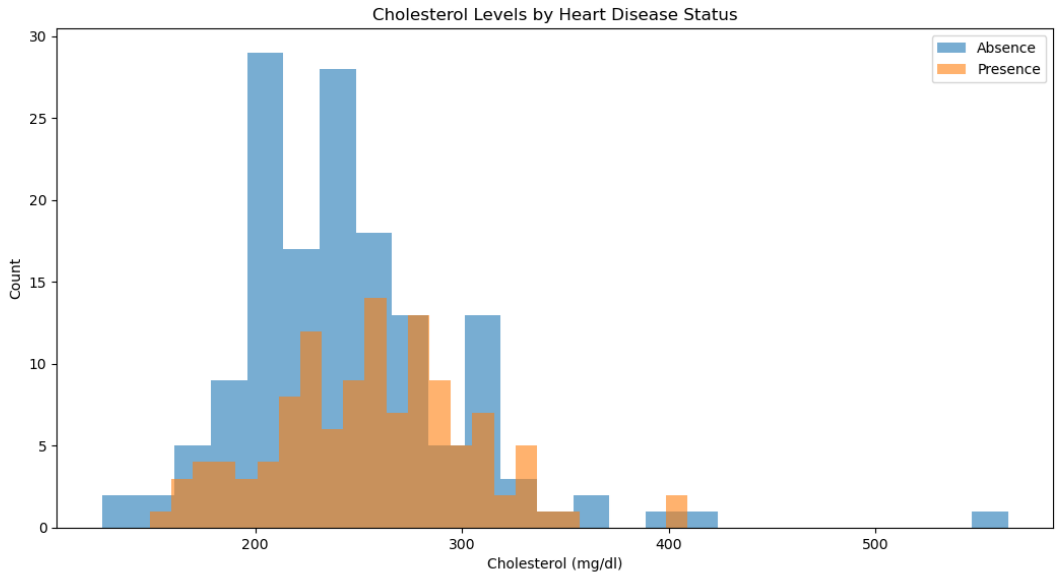
Age emerges as an important demographic factor associated with heart disease presence. The exploratory analysis indicates that older patients are more frequently diagnosed with heart disease, which is consistent with medical literature linking aging to cumulative cardiovascular risk. While age alone does not perfectly separate the two groups, its contribution becomes more meaningful when combined with other clinical variables in a multivariate framework such as Linear Discriminant Analysis.

Heart disease related to blood pressure



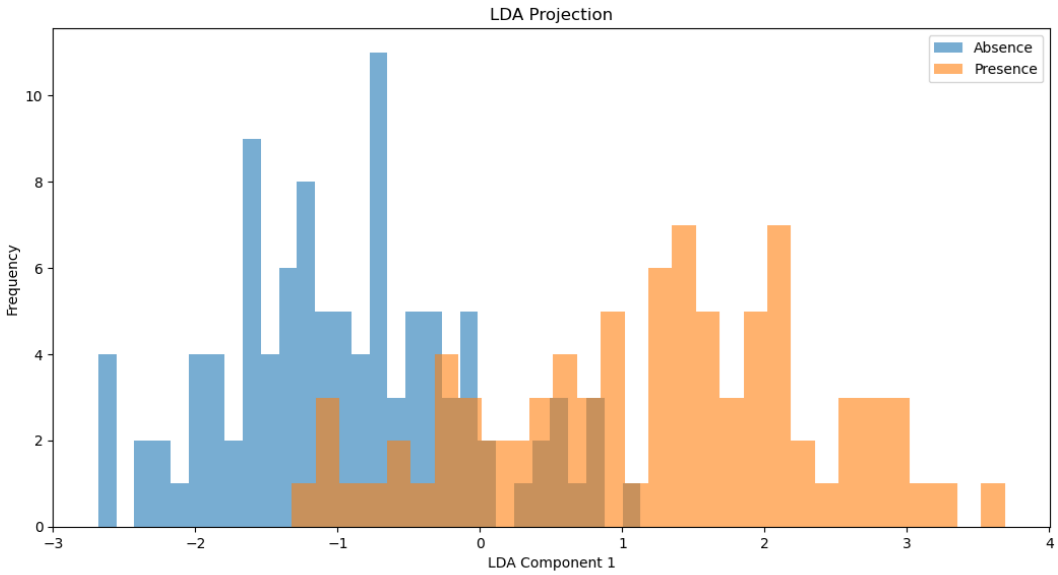
This illustrates the distribution of resting blood pressure for patients with and without heart disease. The distribution for patients with heart disease is shifted toward higher blood pressure values compared to those without the condition. Although there is considerable overlap between the two groups, higher resting blood pressure appears more frequently among patients diagnosed with heart disease, suggesting a positive association between blood pressure and heart disease presence.

Cholesterol levels by heart disease status



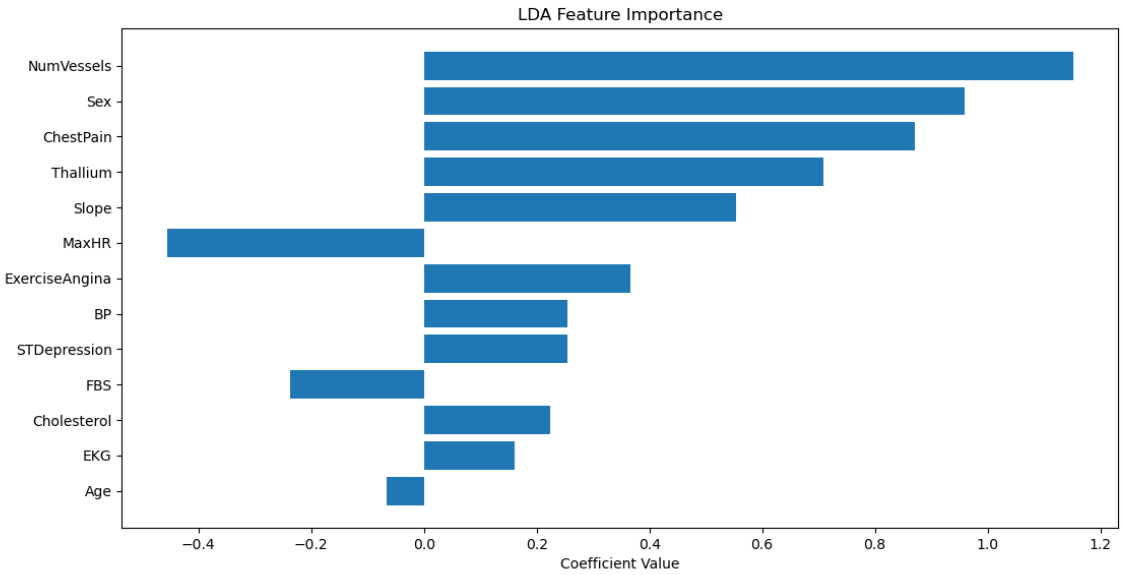
This figure displays the distribution of serum cholesterol levels for patients with and without heart disease. Patients diagnosed with heart disease tend to exhibit higher cholesterol values on average, with a heavier concentration in the upper range of the distribution. While both groups overlap substantially, we can see that patients with cholesterol level above 200 (200 or below being a desirable level) values are more frequently observed among patients with heart disease, indicating a positive association between elevated cholesterol and disease presence.

LDA Projection



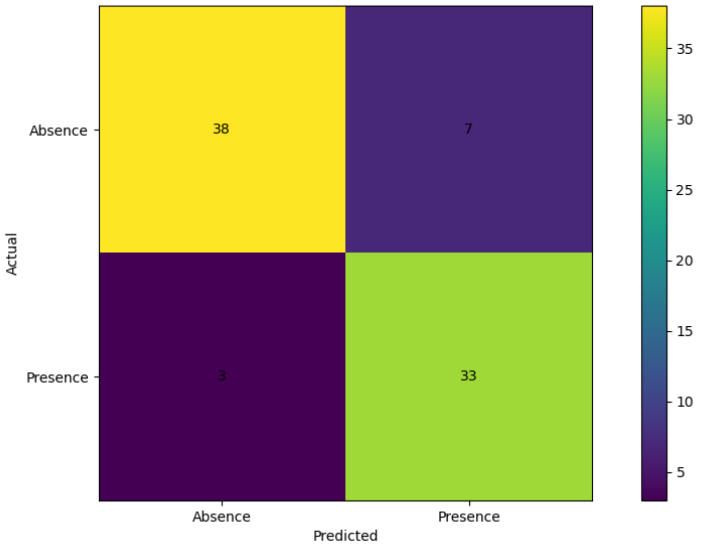
This figure presents the projection of observations onto the first linear discriminant axis. Patients with heart disease cluster toward one end of the axis, while those without heart disease are concentrated toward the opposite end. Despite some overlap, the separation indicates that the clinical variables collectively provide useful discriminatory information for distinguishing heart disease status.

LDA Coefficients indicating variable importance



This figure displays the coefficients of the Linear Discriminant Analysis, which indicate the relative contribution of each variable to the discriminant function. Variables with larger absolute coefficient values have a greater influence on the separation between patients with and without heart disease. The results suggest that measures related to exercise-induced cardiac stress, age, and resting physiological indicators contribute most strongly to class discrimination, while other variables play a more limited role.

Confusion matrix



This confusion matrix summarizes the model's ability to correctly classify patients with and without heart disease. A substantial proportion of observations are correctly classified along the main diagonal, indicating that the LDA model achieves reasonable predictive performance. Misclassifications occur in both directions, suggesting that while the model captures general patterns in the data, overlap between classes limits perfect separation.

6. Discussion

This study examined whether clinical and demographic variables can discriminate between patients with and without heart disease using Linear Discriminant Analysis. The results indicate that heart disease status can be reasonably distinguished using a linear combination of predictors, supporting the relevance of LDA for the research question. Exploratory analyses showed differences in age, blood pressure, and cholesterol distributions between groups, motivating the multivariate approach.

The projection onto the first linear discriminant axis revealed clear directional separation between patients with and without heart disease, although some overlap remained. LDA coefficient analysis indicated that multiple variables jointly contribute to class discrimination, rather than a single dominant factor. The confusion matrix further showed that the model correctly classified a majority of observations, while misclassifications reflected overlapping clinical profiles.

A key strength of this analysis is the integration of exploratory visualization and multivariate classification using standardized variables. However, the analysis is limited by LDA's linearity assumptions and the presence of overlapping class distributions and relatively small dataset (271 observations), which may reduce the robustness of the results and limit the generalizability of the conclusions to broader populations. Future work could explore nonlinear classification methods or incorporate additional predictors to improve classification performance.

7. Conclusion

This study applied Linear Discriminant Analysis to examine whether demographic and clinical variables can distinguish between patients with and without heart disease. The results demonstrate that a linear combination of predictors provides meaningful discriminatory power, as evidenced by the separation observed in the LDA projection and the classification performance reflected in the confusion matrix. Exploratory analyses further revealed that factors such as age, blood pressure, and cholesterol are associated with heart disease presence, although none are sufficient as standalone predictors.

Overall, the analysis highlights the importance of considering multiple clinical indicators simultaneously when assessing heart disease risk. While the LDA model achieved reasonable classification performance, overlap between groups and data limitations underscore the complexity of cardiovascular diagnosis. The findings suggest that multivariate statistical techniques are valuable tools for understanding heart disease patterns and can support more informed clinical and analytical decision-making.

8. References

Dataset used: https://www.kaggle.com/datasets/neurocipher/heartdisease

Fisher, R. A. (1936). *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, 7(2), 179–188.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.