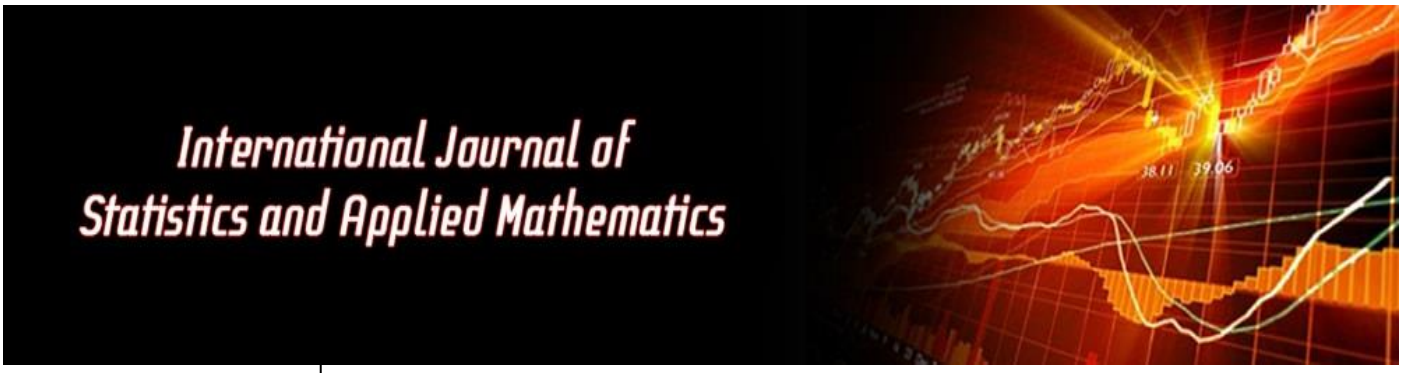
International Journal of Statistics and Applied Mathematics 2024; 9(2): 180-188



ISSN: 2456-1452
Maths 2024; 9(2): 180-188

www.mathsjournal.com
Received: 04-01-2024
Accepted: 06-02-2024

Rajarathinam A
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Corresponding Author:
Rajarathinam A
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Discriminant analysis for heart disease attributes

Rajarathinam A

DOI: https://doi.org/10.22271/maths.2024.v9.i2c.1703

Abstract
This study employs multivariate analysis to investigate the relationship between heart disease parameters and outcomes. Rigorous assessment of normality, multicollinearity, and covariance matrix equality ensures analysis validity. Data normalization via the Box-Cox method enhances normality, facilitating robust statistical analyses. Multivariate analysis of Variance (MANOVA) uncovers significant heart disease parameter variations across Outcome variables. Linear discriminant analysis (LDA) assesses heart disease parameters' capacity to classify individuals, emphasizing gender as a discriminating factor. Results highlight the importance of heart disease parameters in understanding population characteristics and their implications for medical research and clinical practice. The confusion matrix reflects the classification accuracy of a heart disease prediction model, achieving 72.9% overall accuracy in distinguishing between individuals with and without heart disease.

Keywords: Heart disease parameters, box-cox, manova, Wilks' lambda, and fisher discriminant analysis

Introduction
Heart disease remains a significant contributor to global morbidity and mortality rates, posing substantial challenges to public health systems and individual well-being. Understanding the multifaceted nature of heart disease is paramount, necessitating a comprehensive grasp of the intricate interplay between various physiological factors and lifestyle choices in its onset and progression. Efforts in research aimed at unravelling these complexities are crucial for advancing medical knowledge and refining clinical interventions to alleviate the burden of heart disease.

In assessing Heart disease health, researchers consider a range of factors, including Age, chest pain type, resting blood pressure, serum cholesterol levels, fasting blood sugar levels, resting electrocardiographic results, maximum heart rate achieved, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the number of significant vessels colored by fluoroscopy. Acknowledging the significance of these factors, researchers and healthcare practitioners strive to elucidate their correlations with demographic variables to enhance risk prediction models and tailor interventions effectively.

Discriminant analysis, a robust multivariate technique, offers a systematic approach to exploring the relationships between heart disease parameters and outcome characteristics. By identifying linear combinations of these parameters, discriminant analysis facilitates classifying individuals into distinct outcome groups based on observed characteristics. This analytical approach sheds light on the factors influencing heart disease biology and contributes to developing predictive models for disease risk stratification.

The main objectives of this work are as follows:
1. To identify linear combinations of heart disease parameters through discriminant analysis.
2. Determine the effectiveness of these combinations in discriminating between outcome groups.
3. Ascertain the relative importance of different heart disease parameters in distinguishing between disease and non-disease categories.
4. Assess the statistical significance of discriminant functions to evaluate differences among outcome groups based on heart disease parameters.

Review of literature

Durrant and Kaban (2010) [7] utilized a novel approach incorporating random projections with Fisher's Linear Discriminant (FLD) classifier for classification tasks. Unlike previous methods focusing solely on preserving pairwise distances under projection, the authors' approach emphasizes leveraging the inherent class structure within the data. This unique methodology does not impose sparsity or low-dimensional constraints on the dataset, allowing for more flexible and accurate classification. Through their research, the authors derive a tight upper bound on the estimated misclassification error over random projections, demonstrating improved generalization with larger training datasets. Moreover, they highlight that covariance misspecification errors are not exacerbated in the low-dimensional space, providing further evidence of the efficacy of their approach. Overall, the authors' development and utilization of this innovative technique offer significant advancements in classification analysis.

Ramayah *et al*. (2010) [20] present a step-by-step example, making it easier for readers to comprehend the intricacies of discriminant analysis. They carefully explain the necessary assumptions and procedures involved in discriminant analysis, including data preparation, model estimation, and interpretation of results.

Nainggolan *et al*. (2018) [16] used discriminant analysis and classified Hypertension women aged 27 to 54 years living in the village in the central district of Bogor. The results of the multivariate discriminant analysis showed that the level of Vo2 max is the only distinction maker in the incidence of hypertension.

ALKubaisi *et al*. (2019) [1] used discriminant analysis with three criteria to test the developed model, producing excellent projecting precision. The discriminant function has properly assessed and classified about 67% of the cases in the analysis. Also, the study included two discriminant tasks: The first was explained by 77%, and the second was presented by 23% of the Variance.

Dibal and Abraham (2020) [5] applied Fisher's linear Discriminant Analysis (FLDF) to health data on diabetic patients from the University of Port Harcourt Teaching Hospital, Rivers, Nigeria. He created a predictive discriminant model that classifies patients into one of two groups (Diabetic and Non-Diabetic). Fisher's linear discriminant function correctly classifies 65.4% of the total observation.

Garate-Escamila *et al*. (2020) [9] proposed a dimensionality reduction method and identified pertinent features related to heart disease by applying feature selection techniques. We evaluated six ML classifiers for validation using data from the UCI Machine Learning Repository's Heart Disease dataset, consisting of 74 features and a label. Among these, the Chi-square and principal component analysis (CHI-PCA) coupled with random forests (RF) yielded the highest accuracy rates, achieving 98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets. The ChiSqSelector technique derived features of anatomical and physiological significance, including cholesterol levels, maximum heart rate, chest pain indicators, features associated with ST depression, and heart vessel characteristics. Experimental findings underscored that combining chi-square with PCA demonstrates superior performance across most classifiers. Conversely, employing PCA directly from raw data yielded lower results, necessitating increased dimensionality for performance enhancement.

Ricciardi *et al*. (2020) [21] presented a comprehensive analysis of data mining techniques applied to a population of 10,265 individuals assessed for myocardial ischemia by the Department of Advanced Biomedical Sciences. With 22 features extracted, linear discriminant analysis (LDA) is utilized twice, employing the Knime analytics platform and R statistical programming language to classify patients into normal or pathological categories. The Knime analysis solely focuses on classification, while the R-based method incorporates principal component analysis (PCA) for feature creation before classification. Results indicate classification accuracies of 84.5% and 86.0%, respectively, with high specificity (>97%) and sensitivity (62-66%). This practical implementation demonstrates the utility of traditional data mining techniques in aiding clinical decision-making, leveraging PCA for feature reduction.

Ndako *et al*. (2020) [17] investigated if haematological measurements could differentiate between typhoid-positive and negative paediatric patients. Using Fisher's Linear Discriminant Method, 200 patients were analyzed. A discriminant score threshold of 0.0067 was established, with patients above classified as unfavourable and below as positive. Classification efficacy was assessed using retribution estimate and leaving-one-out approaches, indicating a 75.8% and 74.7% prevalence for typhoid-positive patients, respectively. These findings suggest a high prevalence of typhoid fever among paediatric patients, emphasizing the need for improved point-of-care diagnostics with robust positive predictive value.

Liberia *et al*. (2020) explored Discriminant Function Analysis (DFA) to evaluate the effectiveness of Indigenous health-and-wellness programs, particularly in the Eeyou Istchee territory, Canada. By analyzing various health parameters, DFA models were developed to discriminate between individuals with and without Type 2 Diabetes Mellitus (T2DM). The models exhibited high specificity (~97%) in classifying non-T2DM individuals. This research underscores the potential of DFA in point-of-contact evaluations for monitoring and assessing health interventions in rural and remote Indigenous communities, providing valuable insights for T2DM management and prevention strategies among the James Bay Cree population.

Ding *et al*. (2023) [6] introduced the Sparse Variables Selection Exponential Local Fisher Discriminant Analysis (SELFDA) model to address shortcomings in fault classification using Local Fisher Discriminant Analysis (LFDA). By automatically identifying key faulty variables through the minor absolute shrinkage and selection operator, SELFDA enhances fault diagnosis performance and model interpretability. It overcomes the Small Sample Size (SSS) problem by employing a matrix exponential strategy, ensuring full-rank within-class scatter matrices. This approach, tested on the Tennessee Eastman process and a real-world diesel working process, outperforms existing methods, demonstrating its effectiveness in practical industrial applications.

Rahamneh *et al*. (2023) [19] utilized discriminant analysis to distinguish between two types of Bowel and Esophageal cancer in Jordan, identifying significant variables such as sex, weight, and Platelets Count P.C. The correct classification rates for the first and second groups were 62.8% and 77% respectively, with misclassification rates of 37.2% and 23%. The proper classification ratio was 71.6%, with a false classification ratio of 28.4%. The method effectively identified vital independent variables for diagnosing both

cancer types, with correct classification probabilities of 66.4% and 77.6% for the first and second groups, respectively.

Henry *et al*. (2023) [11] investigated the spectral differences of tobacco leaves under macronutrient deficiencies. They employed information entropy and spectral derivatives methods to identify the most effective wavelengths for discrimination. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) algorithms were utilized to reduce data dimensionality and classify the symptoms. The study's findings revealed that the overall accuracy for classifying young, intermediate, and mature plants was 92%, 82%, and 75%, respectively. The results also indicated that nitrogen, sulfur, and magnesium deficiencies significantly impacted the classification accuracy. In contrast, deficiencies in phosphorus and potassium had minimal effect on the classification outcomes.

Materials and Methods

Material

The dataset under scrutiny comprises 303 meticulously gathered patient data records, incorporating 14 distinct features alongside the target variable, shedding light on the dynamics of heart disease. These features encapsulate vital health indicators and clinical observations, comprehensively portraying factors influencing heart disease health. The dataset, meticulously assembled and accessible for analysis, offers valuable insights into the intricate dynamics of heart diseases, facilitating advancements in medical research and clinical practice. The dataset is available on Kaggle (https://www.kaggle.com/code/desalegngeb/heart-disease-predictions), allowing researchers, clinicians, and enthusiasts to explore and analyze the data. Each entry in the dataset reflects meticulous data collection processes, ensuring accuracy and reliability for subsequent analysis and interpretation. This dataset is a valuable resource for exploring the multifaceted aspects of heart disease and advancing our understanding of its underlying mechanisms.

Methods

Box-Cox method

The Box-Cox method is used in statistics and econometrics to transform non-normal data into approximately normal data. It is named after statisticians George Box and Sir David Cox and was introduced in 1964. Let $y=(y_1,y_2,...,y_n)'$ be the date on which the Box-Cox transformation is applied. Box and Cox (1964) defined their transformation as

$$y_i^{(\lambda)}=\begin{cases}\lambda^{-1}(y_i^{\lambda}-1) & if\ \lambda\neq 0\\ \log(y_i) & if\ \lambda=0\end{cases} \quad (1)$$

Such that for unknown λ

$$y^{(\lambda)}=X\beta+\varepsilon \quad (2)$$

Where $y^{(\lambda)}$ is the λ transformed data, X is the design matrix (possible covariates of interest), β is the set of parameters associated with the λ transformed data, and $\varepsilon=(\varepsilon_1,\varepsilon_2,...,\varepsilon_n)$ is the error term. Since the aim of Equation (1) is that

$$y^{(\lambda)} \square N(X\beta,\sigma^2 I_n) \quad (3)$$

Then, $\varepsilon \sim N(0,\sigma^2)$. The transformation in Equation (1) is only valid for $y_i>0, i=1,2,...,n$ modifications to be made when negative observations are present (Vélez *et al*., (2015) [24].

Multivariate analysis of variance

A manova technique (Johnson & Wichern, (1998) [13] is employed to test the significance of variation among all the five parameters considered simultaneously. The MANOVA model for comparing the population means vectors is as follows:

$$Y_{ij}=\mu+V_i+E_{ij} \quad (4)$$

Where, E_{ij} is a vector of random error distributed as $N_P(0,\Sigma)$. Here, the parameter vector μ is the overall mean and V_i represents the model's status in (4); each component of the observation vector Y_{ij} satisfies the univariate model, and the variance-covariance matrix Σ is the same for all populations.

Variance Inflation Factor

The variance inflation factor is used to measure how much the Variance of the estimated regression coefficient is inflated if the independent variables are correlated. VIF is calculated as

$$VIF=\frac{1}{1-R^2} \quad (5)$$

Where the tolerance is simply the inverse of the VIF; the lower the tolerance, the more likely the multicollinearity among the variables. The value of VIF=1 indicates that the independent variables are not correlated. If the value of VIF is 1 < VIF < 5, it specifies that the variables are moderately correlated. If the VIF value is above 5, there will be multicollinearity among the predictors in the regression model (Goldstein, (1993) [10] and Shrestha, (2020) [23]. Another one is the scatterplot graphical method that signifies the linear relationship between pairs of independent variables. It is essential to look for scatterplots that indicate a linear relationship between pairs of independent variables. The correlation coefficient is calculated using the formula:

$$r=\frac{n(\sum XY)-(\sum x)(\sum y)}{\sqrt{\left[n\sum X^2-(\sum)^2\right]\left[n\sum Y^2-(\sum Y)^2\right]}} \quad (6)$$

Where r is the correlation coefficient, n is the number of observations, X represents the first variable in the context, and Y is the second variable in the context. If the correlation coefficient value is higher with the pairwise variables, it indicates the possibility of collinearity (Young, (2018) [25].

Box's-M test

The Box's-M-test for homogeneity of covariance matrices, introduced in 1949, examines the covariance matrices derived from multivariate normal data considering one or more classification factors. This test assesses the similarity between the separate covariance matrices by comparing the product of their log determinants to the log determinant of the combined covariance matrix, similar to a likelihood ratio test. The test statistic employs a chi-square approximation.

Wilk's lambda

In discriminant analysis, Wilk's lambda is utilized to assess the contribution of each level of an independent variable to the model. This scale ranges from 0 to 1, where a value of 0 indicates complete discrimination, while a value of 1 signifies no discrimination. To test the impact of each independent variable, it is successively included and excluded from the model, generating a Λ statistic. The significance of the change in Λ is evaluated using an F-test; if the computed F-value exceeds the critical value, the variable is retained in the model (Onwukwe, (2014)) [18]. Thus, a non-significant Wilks' lambda value is always preferred.

$$Wilks\ lamda(\Lambda) = \frac{|W|}{|B+W|} \quad (7)$$

B is the between-groups matrix, and W is the within-group matrix. The Eigenvalue can be explained as the ratio of the between-groups sum of squares to the within-group sum of squares (McGarigal *et al*., (2000)) [15].

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is an extension of discriminant analysis; it shares ideas and techniques with multiple analyses of Variance (MANOVA). LDA aims to classify cases into three or more categories using continuous or dummy categorical variables as predictors (Cramer, (2003) [4], Jang *et al*., (2015)) [12]. The term DA (Fisher, (1936)) [8] refers to numerous types of analyses. DA is the most popular statistical technique to classify individuals or observations into non-overlapping groups based on scores derived from a suitable "statistical decision function" constructed from one or more continuous predictor variables. While investigating the differences between the groups or categories, the necessary step is to identify the attributes with the most contributions to maximum reparability between known groups or categories to classify a given observation into one of the groups. For that purpose, DA successively identifies the linear combination of attributes known as canonical discriminant functions (equations) that contribute maximally to group separation. Predictive DA addresses the question of how to assign new cases to groups.

The form of the Equation or function is:

$$D_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + ... + \alpha_j X_{ik}$$

Where D is an independent variable and ^{D_i} is the value of discriminant score from the ith category (i=1, 2…n), $^{\alpha_j}$ is the discriminant coefficient of jth attributes (j=0,1,2,…,k), and X_{ik} is the kth independent variable of ith category. This function is similar to a regression equation or function. There $\alpha's$ are unstandardized discriminant coefficients analogous to the ones in the regression equation. These $\alpha's$ maximize the distance between the means of the dependent variable, and the standardized discriminant coefficients can also be used, like beta weight in regression.

Results and Discussion

The descriptive statistics in Table 1 provide insights into the heart disease parameters. On average, patients are approximately 54 years old, with a resting blood pressure of around 131 mmHg. Cholesterol levels are notably higher, averaging at 246 mg/dL. The heart rate averages 149 beats per minute, indicating potential variability. ST depression is relatively low, averaging at 1.04 mm. Generally, while the skewness and kurtosis values suggest slight deviations from a normal distribution, further analysis is needed to understand the relationship between these parameters and heart disease risk.

Table 1: Descriptive statistics for heart disease parameters

Repressors	Mean	Std. Devi.	Variance	Skewness	Kurtosis
Age	54.37	9.08	82.48	-0.20	-0.54
Resting Blood Pressure	131.62	17.54	307.59	0.71	0.93
Cholesterol	246.26	51.83	2686.43	1.14	4.51
Heart Rate	149.65	22.91	524.65	-0.54	-0.06
ST Depression	1.04	1.16	1.35	1.27	1.58

The Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted (Table 2) to assess the normality of the distributions for heart disease parameters before and after applying the Box-Cox transformation. Before the transformation, all variables exhibited statistically significant deviations from normality ($p< 0.05$), with varying degrees of skewness. However, following the Box-Cox transformation, there was an improvement in the normality of the distributions for most variables, as indicated by non-significant p-values ($p > 0.05$) in both tests. Specifically, Age, Resting Blood pressure, Cholesterol, Heart rate, and ST Depression. These results suggest that the Box-Cox transformation effectively normalized the distributions of heart disease parameters, rendering them more suitable for subsequent statistical analyses assuming normality, such as parametric tests.

Table 2: Normality test for Heart disease parameters

Before Box-Cox Method						
Repressors	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	DF	Sig.	Statistic	DF	Sig.
Age	0.076	303	0.000	0.986	303	0.006
Resting Blood Pressure	0.102	303	0.000	0.966	303	0.000
Cholesterol	0.055	303	0.025	0.947	303	0.000
Heart rate	0.071	303	0.001	0.976	303	0.000
ST Depression	0.185	303	0.000	0.844	303	0.000
After Box-Cox Method						
Repressors	**Kolmogorov-Smirnov**			**Shapiro-Wilk**		

	Statistic	DF	Sig.	Statistic	DF	Sig.
Age	0.108	204	0.071	0.957	204	0.125
Resting Blood Pressure	0.081	204	0.113	0.985	204	0.072
Cholesterol	0.040	204	0.200	0.988	204	0.080
Heart rate	0.104	204	0.841	0.959	204	0.075
ST Depression	0.121	204	0.533	0.940	204	0.093

In Figure 1, the histogram brown curve shows the Gaussian distribution, while the histogram shows the 303 Heart disease parameters distribution. The top bars in the histogram match nicely with the Gaussian distribution; therefore, after the Box-Cox method, the dataset was perfectly normally distributed. The points in the histogram plot form a bell-shaped line since the dataset's quintiles nearly match the dataset's quintiles, which would theoretically be the normally distributed dataset.

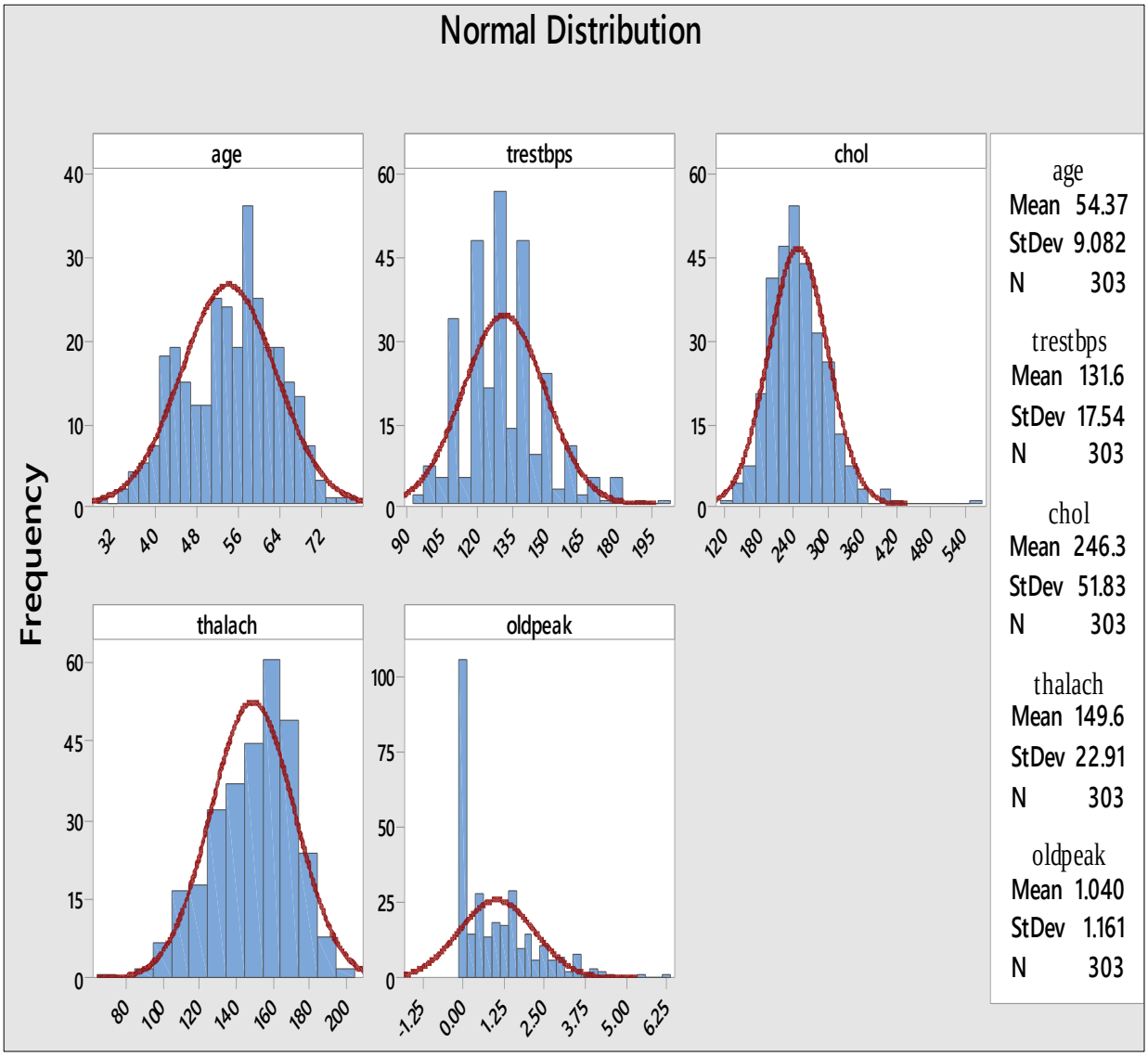


Fig 1: Normality plot for Heart disease parameters

Table 3: Multivariate analysis of variance for heart disease parameters

Effect	Value	Sig.
Pillai's Trace	0.994	0.000
Wilks' Lambda	0.006	0.000
Hotelling's Trace	168.257	0.000
Roy's Largest Root	178.282	0.000

Various multivariate tests, including Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's most significant root tests, were utilized to assess the collective variation of all five heart disease parameters across outcome groups. The outcomes of these tests are presented in Table 3. These MANOVA statistics offer insights into the multivariate effects of the analysis. Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root are all measures of the significance of the overall model. In this case, the extremely low p-value (0.000) indicates that the model has a significant overall effect. The values of these statistics (Ranging from 0.006 to 0.994) suggest the proportion of Variance in the dependent variables explained by the independent variables in the model.

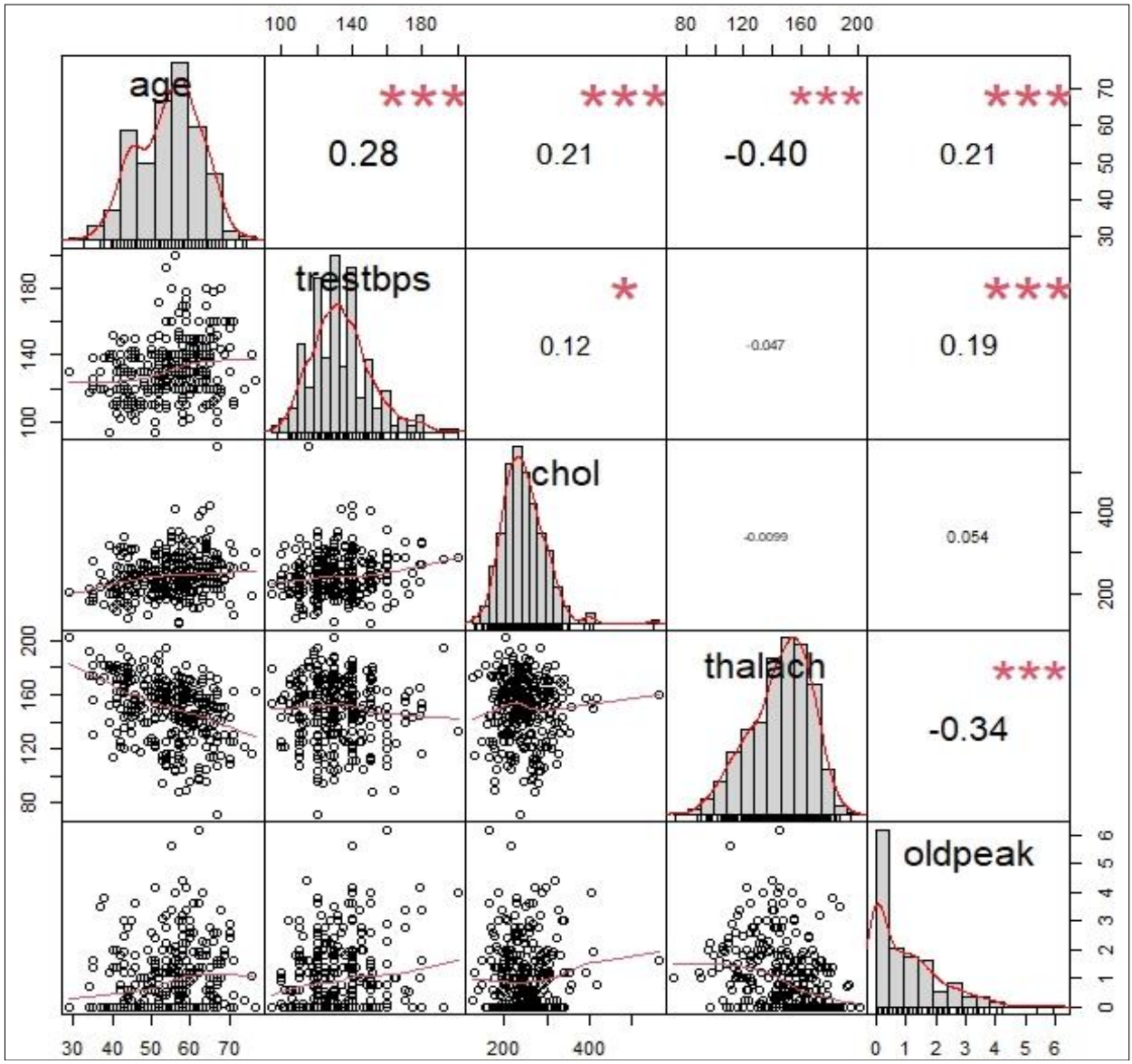


Fig 2: Correlation matrix for Heart disease parameters

The Pearson correlation between the study variables' Heart disease parameters has been calculated. It is depicted in Figure – 2. The upper triangular matrix shows the Pearson correlation and its significance level (as stars). Each significance level is associated with a symbol p-values 0.001 (***), 0.01 (**), and 0.05 (*). The results reveal significant correlations between Age and resting blood pressure (0.28), cholesterol, and ST depression (0.21). Additionally, Age exhibits a strong negative correlation with Maximum Heart Rate (0.40). ST Depression shows positive correlations with Age (0.21) and Cholesterol (0.91) and a negative correlation with Maximum Heart Rate (-0.34). These findings indicate interdependencies among the independent variables across multiple measures (p measures).

Table 4: Multicollinearity for Heart disease parameters

Repressors	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	0.06	0.32		0.20	0.84		
Age	0.00	0.00	-0.01	-0.15	0.88	0.74	1.35
Resting BP	0.00	0.00	-0.06	-1.18	0.24	0.89	1.13
Cholesterol	0.00	0.00	-0.06	-1.11	0.27	0.94	1.06
Heart rate	0.01	0.00	0.31	5.44	0.00	0.75	1.33
ST depression	-0.13	0.02	-0.31	-5.75	0.00	0.85	1.18

The coefficients in Table 4 depict the relationship between the model's independent variables (heart disease parameters) and the dependent variable. The standardized coefficients (Beta) indicate the strength and direction of the relationship, while the t-values and significance levels (Sig.) indicate the statistical significance of each coefficient. Collinearity statistics such as Tolerance and VIF assess multicollinearity among the independent variables. In this model, heart rate and ST depression exhibit statistically significant relationships with the dependent variable, with heart rate demonstrating a positive relationship and ST depression showing a negative relationship. However, Age, resting blood pressure, and cholesterol do not show statistically significant relationships. Furthermore, all variables exhibit acceptable levels of

multicollinearity, as indicated by Variance Inflation Factor (VIF) values in Table 5.

Table 5: Results of Box's M method

Box's M	79.376
Approx.	5.197
df1	15
df2	340980.412
Sig.	0.071

Box's M test results in Table 5 are a diagnostic tool used to assess the equality of covariance matrices across groups. It evaluates whether the assumption of homogeneity of covariance matrices (Homoscedasticity) is violated. The p-value (Sig.) of 0.071 suggests no significant violation of this assumption, indicating that the covariance matrices are approximately equal across groups. Therefore, the assumption of homogeneity of covariance matrices is met, ensuring the reliability of interpreting the results.

Table 6: Wilks Lambda Test Statistics

Test of Function (s)	Wilks' Lambda	Chi-square	DF	Sig.
1	0.722	97.423	5	0.000

In Table 6, "Function 1" refers to the specific function being tested. The Wilks' Lambda value of 0.722 indicates the proportion of Variance in the dependent variables not accounted for by the independent variables. The associated Chi-square statistic of 97.423, with 5 degrees of freedom (DF), results in a highly significant p-value (Sig.) of 0.000. This suggests that the overall model or the specific function being tested significantly affects the dependent variables.

Table 7: Eigenvalues for the first function

Function	Eigenvalue	% of Variance	Cumulative%	Canonical Correlation
1	.386	100.0	100.0	.528

In Table 7, the Eigenvalue of 0.386 indicates the amount of Variance explained by Function 1. With a percentage of Variance of 100.0%, Function 1 accounts for the entire Variance in the data, as reflected by the Cumulative%. The Canonical Correlation of 0.528 represents the correlation between the observed and canonical variables derived from the function.

Table 8: Canonical Discriminant Function Coefficients

	Constant	Age	Resting BP	Cholesterol	Heart rate	ST Depression
Function 1	0.002	0.008	0.002	-0.03	0.59	2.145

The coefficients in Table 8 indicate the weights assigned to each variable in the canonical discriminant function. These coefficients signify the magnitude and direction of the relationship between each predictor variable (heart disease parameters) and the discriminant function. Positive coefficients suggest a positive association with the function (heart rate and ST depression), while negative coefficients (Age, resting blood pressure, and cholesterol) imply a negative association. The values reflect the relative importance of each variable in discriminating between groups or explaining the variability in the data.

Table 9: Fisher linear Discriminant Function

Repressors	Non-Disease	Disease
(Constant)	0.862	0.859
Age	0.294	0.284
Resting BP	0.047	0.044
Cholesterol	0.456	0.493
Heart rate	1.540	0.807
ST depression	-83.691	-86.289

The Fisher linear discriminant coefficients results in Table 9 represent the relationship between each heart disease parameter and disease status (non-disease or disease) in a linear discriminant analysis. Positive coefficients indicate an increase in the value of the heart disease parameter associated with the specified disease status, while negative coefficients indicate a decrease. The constant term represents the intercept of the linear discriminant function for each disease status group. The equations representing the relationship between each heart disease parameter and disease status in the linear discriminant analysis are as follows.

Non-Disease (Y_{ND}) = 0.862 + 0.29*4$_{Age}$ + *0.047**$_{\text{Resting BP}}$ + 0.456*$_{Cholesterol}$ + *1.540**$_{\text{Heart Rate}}$-83.691*$_{\text{ST Depression}}$

Disease (Y_D) = 0.859 + 0.284*$_{Age}$ + *0.044**$_{\text{Resting BP}}$ + 0.493*$_{Cholesterol}$ + *0.807**$_{\text{Heart Rate}}$-86.289*$_{\text{ST Depression}}$

These equations, Y_{ND} and $Y_{D,}$ represent the discriminant scores for non-disease and disease groups based on the given heart disease parameters.

The coefficients of each heart disease parameter in the linear discriminant weights are depicted in Figure 3. This discriminant function exhibits a notably high positive correlation with stress depression followed by age, resting blood pressure, and cholesterol. Additionally, Heart rate heart disease parameters demonstrate negative correlations with the outcome of the heart disease. Consequently, another five heart disease parameters are significantly influenced by gender.

The confusion matrix in Table 10 outlines the classification performance of a model for heart disease prediction. It accurately distinguishes between individuals with and without heart disease, achieving an overall accuracy of 72.9%.

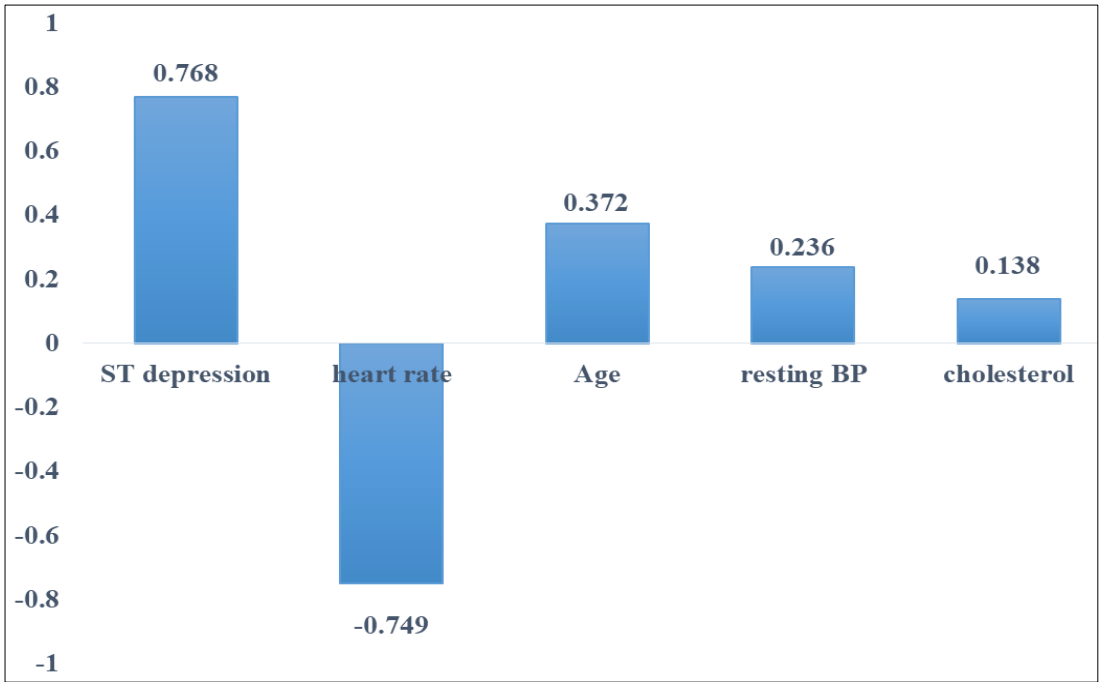


Fig 3: Association of different heart disease parameters with the discriminant function

Table 10: Classification table

Target	Predicted Group Membership		Total
	0	1	
Non-Disease	93	45	138
Disease	37	128	165
Overall Accuracy: 72.9%			

Conclusion

In this study, the Heart disease parameter dataset analysis reveals valuable insights into population characteristics. The dataset represents a diverse population and includes comprehensive information on heart disease parameters. Descriptive statistics highlight variability and distribution characteristics, while normality tests indicate the effectiveness of transformation methods. Also, this study highlights significant correlations among key heart disease parameters such as Age, blood pressure, cholesterol levels, and ST depression. Discriminant analysis effectively identifies combinations of these parameters that discriminate between outcome groups, enhancing our understanding of heart disease biology. Multivariate analysis reveals variations in heart disease parameters across demographics, emphasizing the importance of individual characteristics in risk assessment and intervention strategies. Normalizing data using the Box-Cox method ensures the validity of our analyses.

The confusion matrix reflects the classification accuracy of a heart disease prediction model, achieving 72.9% overall accuracy in distinguishing between individuals with and without heart disease. These findings contribute to developing more accurate predictive models and targeted interventions to alleviate the burden of heart disease on public health and individual well-being.

References

1. AlKubaisi M, Aziz WA, George S, Al-Tarawneh K. Multivariate discriminant analysis managing staff appraisal case study. Acad Strategic Manag J. 2019;18(5):1-12.
2. Box GE. A general distribution theory for a class of likelihood criteria. Biometrika. 1949;36(3/4):317-346.
3. Box GE, Cox DR. An analysis of transformations. J R Stat Soc Ser B Stat Methodol. 1964;26(2):211-243.
4. Cramer D. Advanced quantitative data analysis. McGraw-Hill Education (UK); c2003.
5. Dibal PN, Abraham AC. On applying linear discriminant function to evaluate data on diabetic patients at the University of Port Harcourt Teaching Hospital, Rivers, Nigeria. Am J Theor Appl Stat. 2020;9(3):53-56.
6. Ding Z, Xu Y, Zhong K. Exponential Local Fisher Discriminant Analysis with Sparse Variables Selection: A Novel Fault Diagnosis Scheme for Industry Application. Machines. 2023;11(12):1066.
7. Durrant RJ, Kabán A. Compressed Fisher linear discriminant analysis: Classification of randomly projected data. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining; c2010, 1119-1128.
8. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen. 1936;7(2):179-188.
9. Escamila GAK, Hassani ELAH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked. 2020;19:100330.
10. Goldstein R. Conditioning Diagnostics: Collinearity and Weak data in regression. J Oper Res Soc. 1993;35(1):85-86.
11. Henry J, Veazie P, Furman M, Vann M, Whipker B. Spectral discrimination of micronutrient deficiencies in greenhouse-grown flue-cured tobacco. Plants. 2023;12(2):280.
12. Jang DH, Cook ACM, Kim Y. Graphical methods for the sensitivity analysis in discriminant analysis. Commun Stat Appl Methods. 2015;22(5):475-485.
13. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. Biometrics. 1998;54(3):1203.
14. Liberda EN, Zuk AM, Martin ID, Tsuji LJ. Fisher's linear discriminant function analysis and its potential utility as a tool for the assessment of health-and-wellness programs in indigenous communities. Int J Environ Res Public Health. 2020;17(21):7894.

15. McGarigal K, Stafford S, Cushman S. Discriminant analysis. In: Multivariate statistics for wildlife and ecology research; c2000. p. 129-187.
16. Nainggolan O, Tjandrarini DH, Indrawati L. Discriminant analysis to predict hypertension in women aged 25-54 years. Glob J Health Sci. 2018;10(10):93.
17. Ndako JA, Olisa JA, Ifeanyichukwu IC, Okolie CE, Ojo SK, Jegede SL. Predictive evaluation of pediatric patients based on their typhoid fever status using linear discriminant model. Med Hypotheses. 2020;144:110264.
18. Onwukwe CE, Ogbonna EN, Ayeni A. Application of Wilks' lambda and hotelling's T_2 with manova on drug addiction and drug abuse data. Adv Life Sci Technol. 2014;24:53-59.
19. Rahamneh AAA, Jresat SS, Zubaidi F, Al-Hawary SIS. Using the linear discriminant analysis method to classify types of bowels and esophageal cancer in Jordan. Inf Sci Lett. 2023;12(3):1299-1305.
20. Ramayah T, Ahmad NH, Halim HA, Chiun MSRMZ. Discriminant analysis: An illustrated example. Afr J Bus Manag. 2010;4(9).
21. Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A, *et al*. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics Journal. 2020;26(3):2181-2192.
22. Selvajothi P, Packiaraj R. Dataset of platelet parameters and red cell distribution width. Mendeley Data; c2021. DOI: 10.17632/5t8dr6d73f.1
23. Shrestha N. Detecting multicollinearity in regression analysis. Am J Appl Math Stat. 2020;8(2):39-42. DOI: 10.12691/jams-8-2-1
24. Vélez JI, Correa JC, Ramos MF. A new approach to the Box-Cox transformation. Front Appl Math Stat. 2015;1:12. DOI: 10.3389/fams.2015.00012
25. Young DS. Handbook of regression methods. Chapman and Hall/CRC; c2018.