Robert Lorenz, Thanippuli Appuhamilage Dimuth Indeewara, Elsa Maria Jose
Chair of Econometrics and Statistics, esp. in the Transport Sector

Predict Presence of a Heart Disease

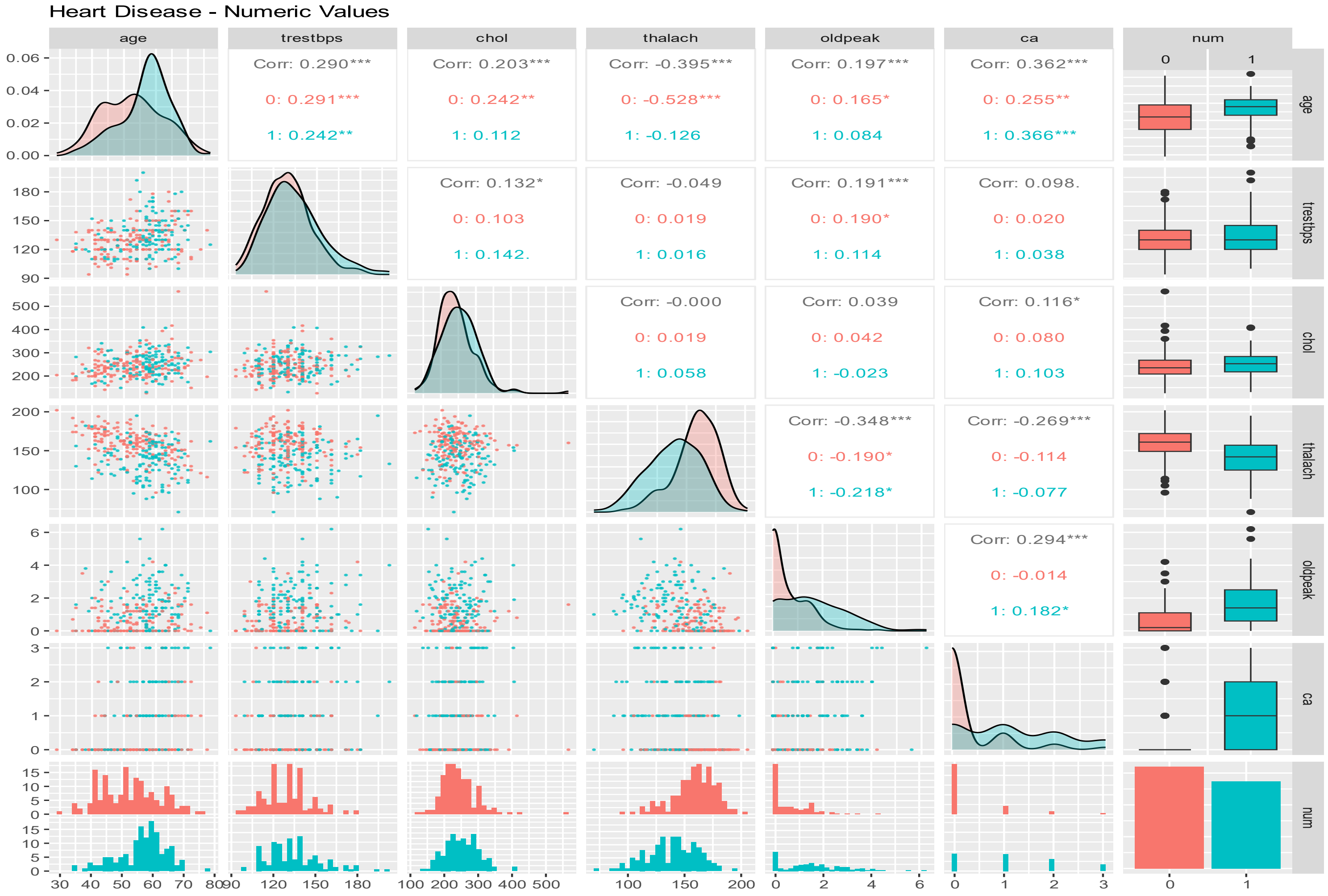
Dresden, Applied Multivariate Statistics / July 17th, 2024

Data

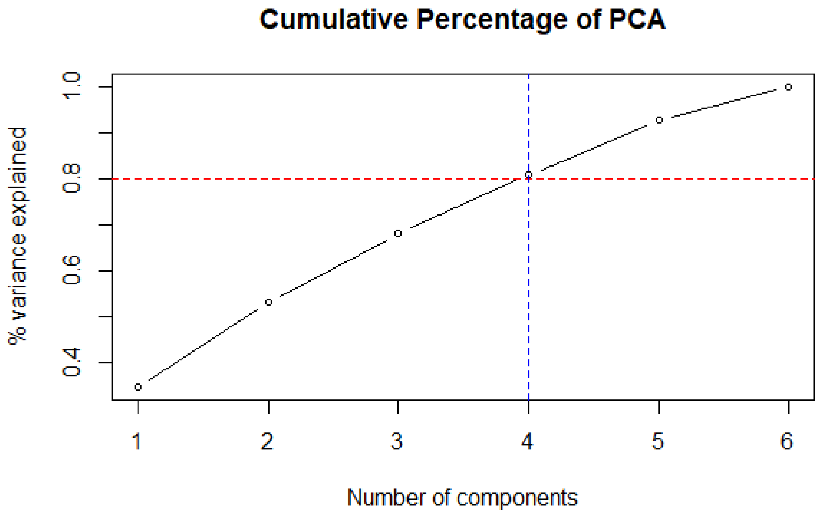
- **Heart Disease Data**
- **Region:**
  - Cleveland, Hungary, Switzerland, and the VA Long Beach
  - Only using Cleveland dataset – 303 observations
- **76 attributes – using a subset of 14**
  - 6 numeric
  - 8 categorical
- **Target Value**
  - presence of heart disease in the patient
  - Values: 0 – absence and 1, 2, 3, 4 – presence
- **Missing Values:**
  - 6 Values
  - Removed → 297 observations left

Variable Name	Description	Type
age	age in years	numeric
sex	sex (1 = male; 0 = female)	categorical
cp	chest pain type(1: typical angina - 2: atypical angina - 3: non-anginal pain - 4: asymptomatic)	categorical
trestbps	resting blood pressure (on admission to the hospital)	numeric
chol	serum cholestoral	numeric
fbs	fasting blood sugar > 120 mg/dl	categorical
restecg	resting electrocardiographic results	categorical
thalach	maximum heart rate achieved	numeric
exang	exercise induced angina	categorical
oldpeak	ST depression induced by exercise relative to rest	numeric
slope	the slope of the peak exercise ST segment	categorical
ca	number of major vessels (0-3) colored by flourosopy	numeric
thal	3 = normal; 6 = fixed defect; 7 = reversable defect	categorical
num	diagnosis of heart disease	target

Data


Heart Disease - Numeric Values
age
trestbps
chol
thalach
oldpeak
ca
num
Corr: 0.290***
0: 0.291***
1: 0.242**
Corr: 0.203***
0: 0.242**
1: 0.112
Corr: -0.395***
0: -0.528***
1: -0.126
Corr: 0.197***
0: 0.165*
1: 0.084
Corr: 0.362***
0: 0.255**
1: 0.366***
Corr: 0.132*
0: 0.103
1: 0.142.
Corr: -0.049
0: 0.019
1: 0.016
Corr: 0.191***
0: 0.190*
1: 0.114
Corr: 0.098.
0: 0.020
1: 0.038
Corr: -0.000
0: 0.019
1: 0.058
Corr: 0.039
0: 0.042
1: -0.023
Corr: 0.116*
0: 0.080
1: 0.103
Corr: -0.348***
0: -0.190*
1: -0.218*
Corr: -0.269***
0: -0.114
1: -0.077
Corr: 0.294***
0: -0.014
1: 0.182*
0
1
0.06
0.04
0.02
0.00
180
150
120
90
500
400
300
200
200
150
100
6
4
2
0
3
2
1
0
15
10
5
0
15
10
5
0
30 40 50 60 70 80
90 120 150 180
100 200 300 400 500
100 150 200
0 2 4 6
0 1 2 3
0 1

Dimension Reduction : Principal Componenet Analysis (PCA)

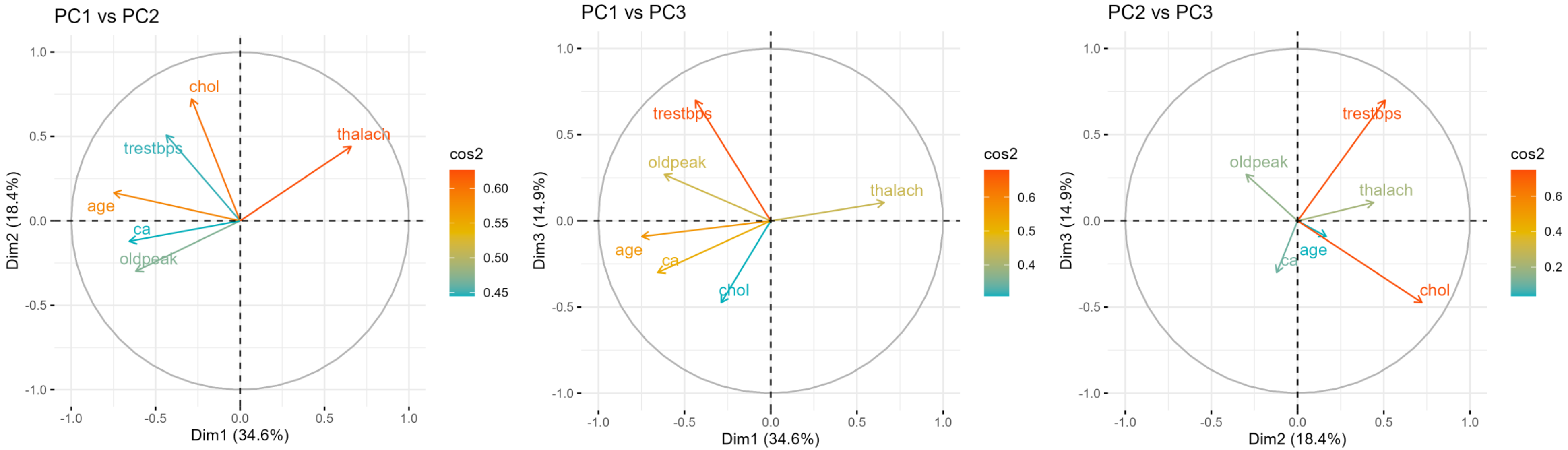


First four principal components are chosen, Since first four Principal components capture majority of variance (80+ %)

PC1 - Strong positive influence from ***thalach***, Strong negative from ***age***, ***oldpeak*** & ***ca***
PC2 - Strong positive influences from ***chol*** and ***thalach***
PC3 - Strong positive influences from ***trestbps*** and strong negative from ***chol***
PC4 - Strong positive influences from **oldpeak** and strong negative from **age**

trestbps, chol, thalach: Have substantial contributions across the dimensions, as indicated by their positions and cos2 values

age, ca, oldpeak: Contribute moderately, evident from their positioning and cos2 values.



Dimension Reduction : Factor Analysis Model

P = 6 , K = 2 , d = 4

Principal Component Method

Estimated loadings after varimax

	Factor 1	Factor 2
Age	- 0.585	-0.494
Trestbps	- 0.151	-0.652
Chol		-0.773
Thalach	0.787	
Oldpeak	- 0.685	
Ca	- 0.637	- 0.197

	Factor 01	Factor 02
SS loadings	1.866	1.313
Proportion Var	0.311	0.219
Cumulative Var	0.311	0.530

Maximum likelihood Method

	Communality	Uniqueness
Age	0.526	0.474
Trestbps	0.177	0.823
Chol	0.099	0.900
Thalach	0.995	0.005
Oldpeak	0.168	0.832
Ca	0.232	0.768

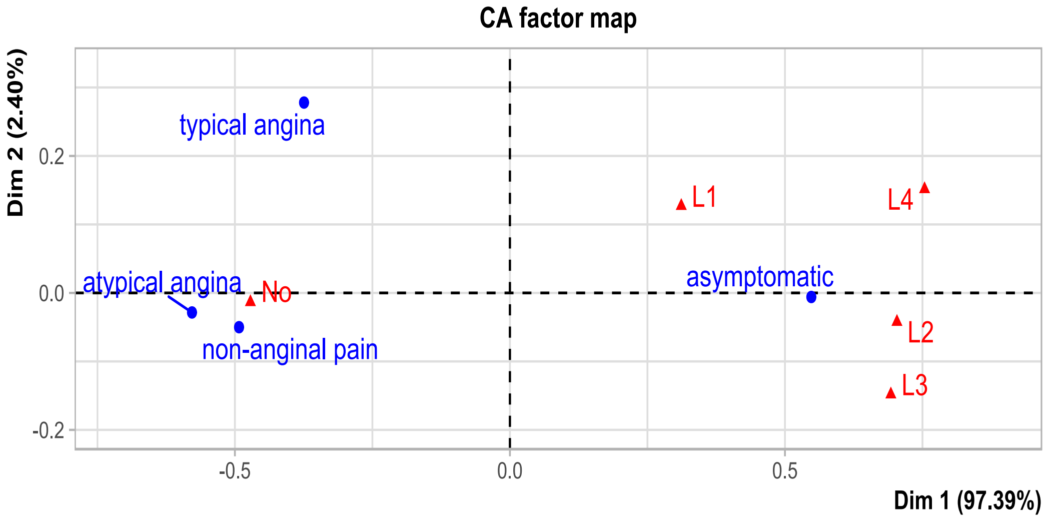
	Factor 01	Factor 02
SS loadings	0.526	0.474
Proportion Var	0.177	0.823
Cumulative Var	0.099	0.900

In the ML method, the loadings for before and after varimax rotation indicated the similar results

The p-value (0.000977) suggests that, not enough factors to capture the full dimensionality of the data set

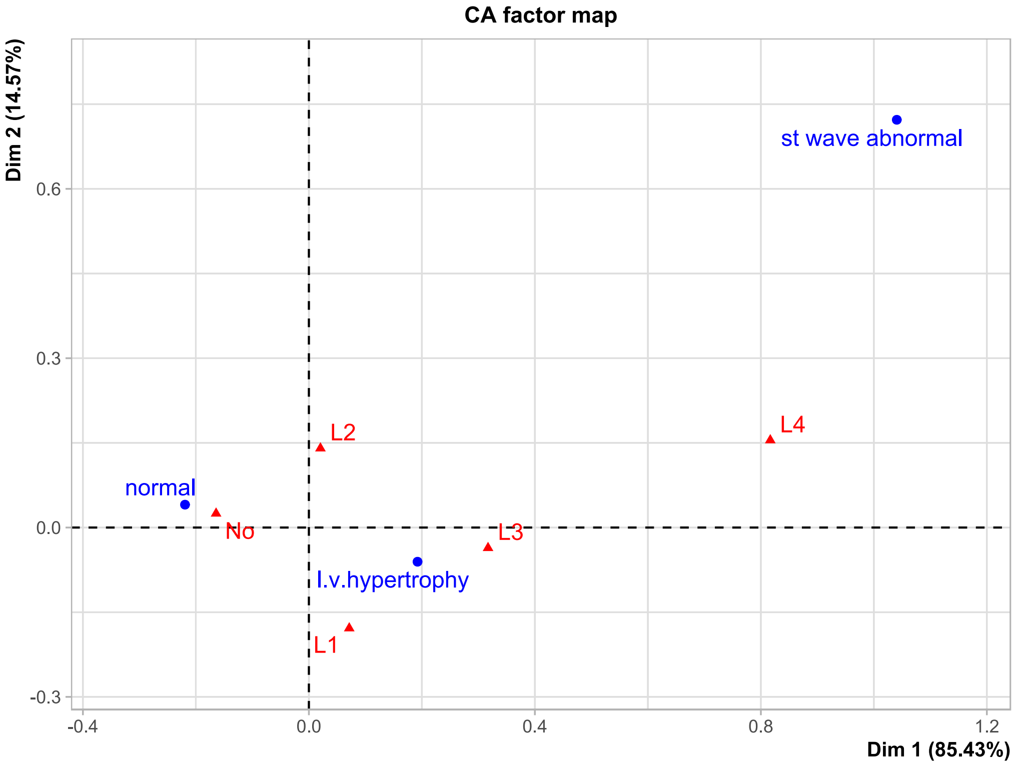
Correspondence Analysis

CP vs NUM



Typical angina is most closely associated with the highest level of heart disease

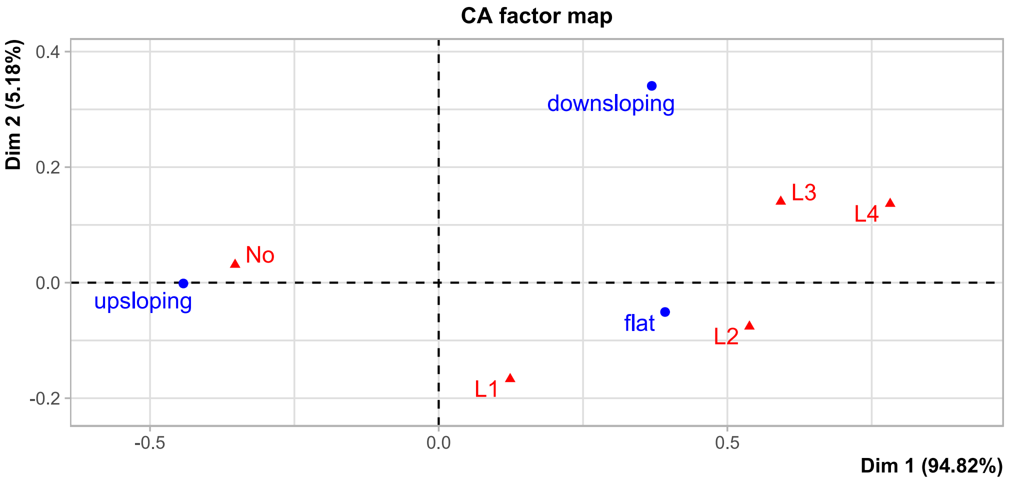
REST ECG vs NUM



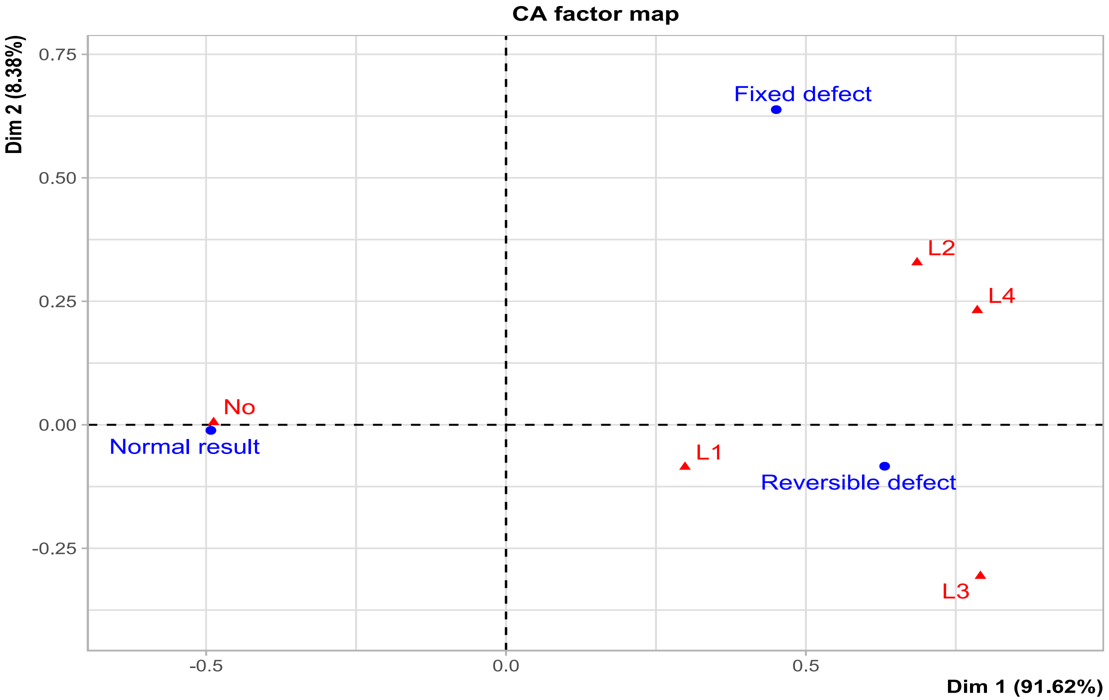
individuals with st wave abnormal findings are likely to show the highest risk of heart disease

Correspondence Analysis

SLOPE vs NUM



THALLIUM STRESS vs NUM



Down sloping ECG results are associated with higher levels of heart disease, indicating more severe conditions.

Fixed defects in test results are associated with higher levels of heart disease, indicating more severe conditions.

Discriminant Analysis (DA)

Comparison of DA Algorithms Training Accuracy: no Cross-Validation(CV) vs. CV

	Original Data	PCA	Original Data - CV	PCA - CV
LDA with prior	0.852	0.848	0.838	0.838
QDA with prior	0.862	0.862	0.835	0.805
LDA without prior	0.845	0.845	0.842	0.839
QDA without prior	0.862	0.855	0.835	0.812

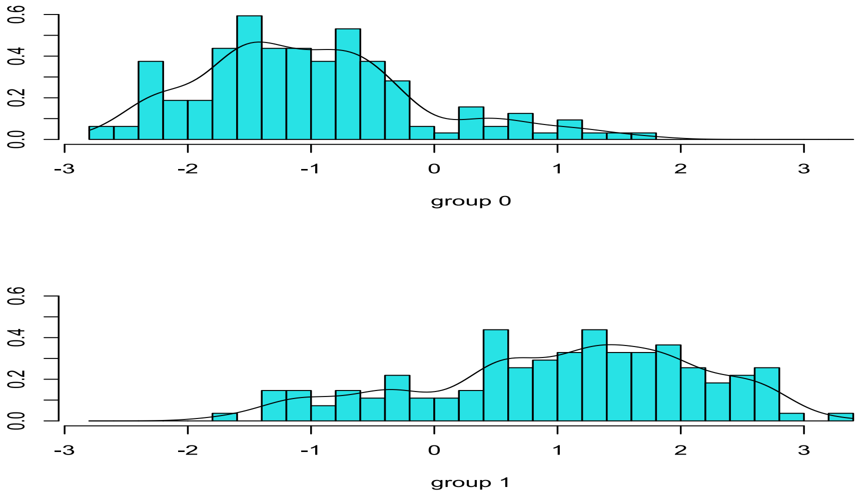


Figure: Distribution of variable num of data with PCA LAD with prior

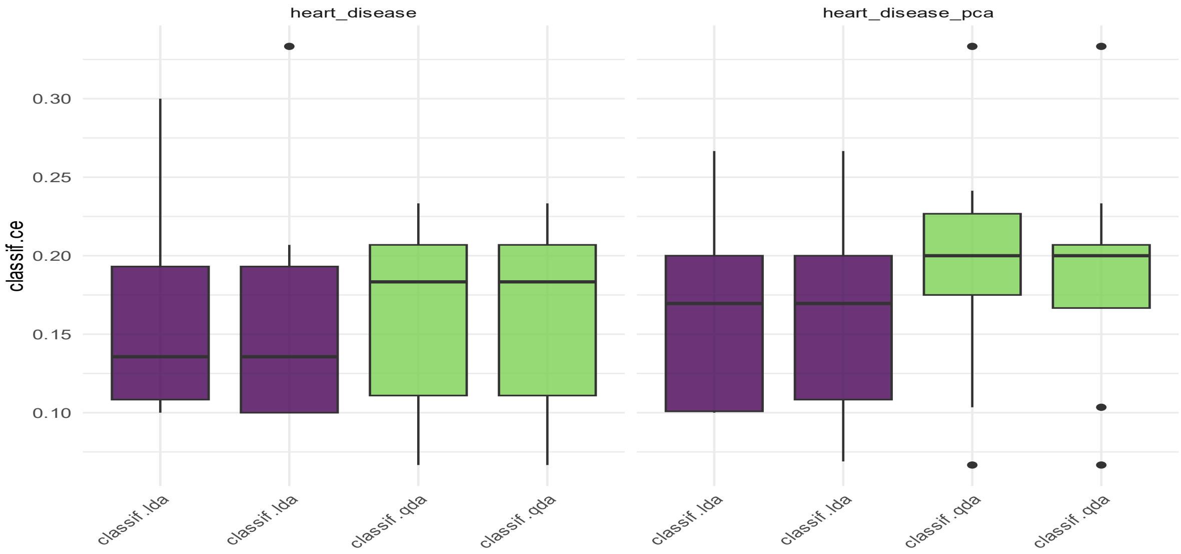


Figure: Pairwise with (left) and without (right) prior for original data and data with PCA for LDA and QDA

Sources

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. Making sense of factor analysis. Sage. 2003

Watkins, Marley W. A step-by-step guide to exploratory factor analysis with R and RStudio. Routledge, 2020.