Babeș-Bolyai University

LINEAR DISCRIMINANT ANALYSIS: APPLICATIONS AND CHALLENGES IN DATA ANALYSIS

Author: Turcu Ciprian-Stelian

OVERVIEW

- Introduction
- Theoretical Part
- Datasets
- Analysis Results
- Conclusions

INTRODUCTION

What is LDA ?

- A statistical method for dimensionality reduction and classification.
- Finds linear combinations of features that maximize class separability.

Why LDA ?

- Interpretable and computationally efficient.
- Works well for structured, labeled datasets.

Comparison with PCA:

- PCA: Maximizes variance without class labels.
- LDA: Maximizes separability using class labels.

THEORETICAL PART

1

Core Process:

- Calculates scatter matrices: within-class and between-class.
- Solves eigenvalue problems to find discriminant components.

2

Limitations and Solutions:

- Sensitivity to assumptions and outliers.
- Alternatives: QDA (class-specific covariance), RDA (regularization), Sparse LDA (feature selection).

3

Key Assumptions:

- Multivariate normality.
- Homogeneous covariance matrices.

DATASETS

Datasets :

Wine Quality, Mall Customers, Heart Disease, Diabetes, Online Retail.

Preprocessing:

- Standardization, through scaling of features
- Label encoding for categorical variables.
- K-means clustering for datasets without target labels. (Mall Customers, Online Retail)

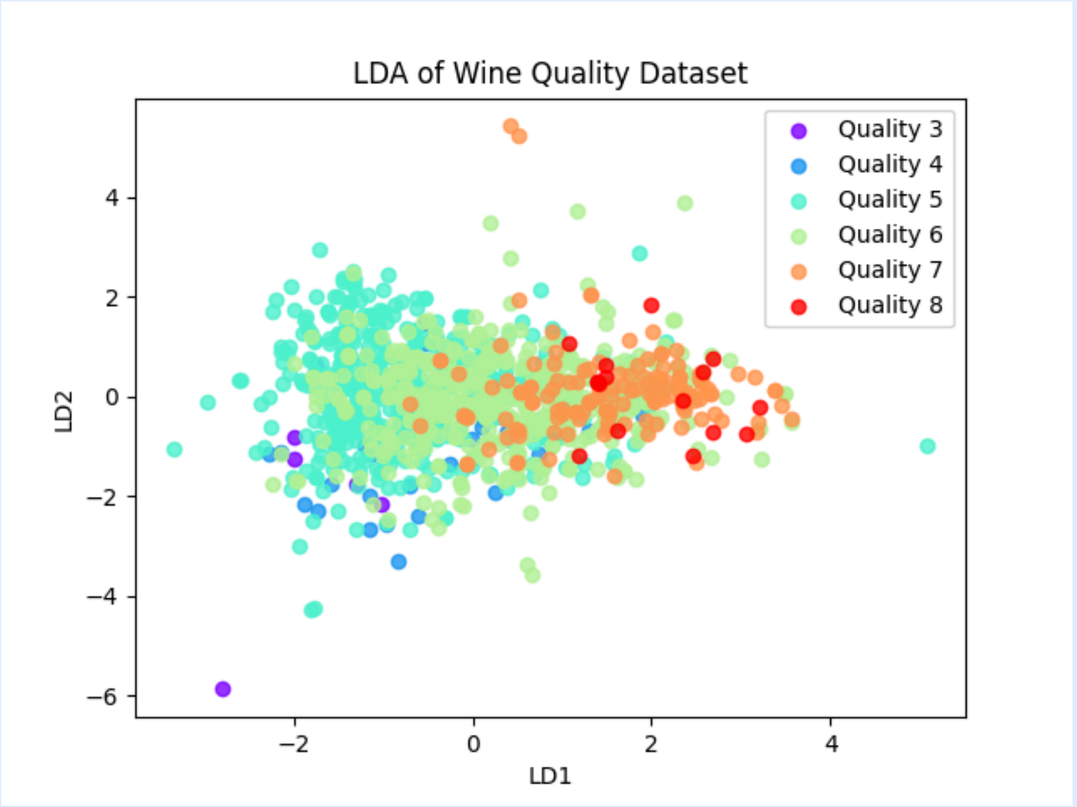
LDA Transformation:

- Reduce high-dimensional data to discriminant components while maintaining separability.

ANALYSIS RESULTS

Wine Quality

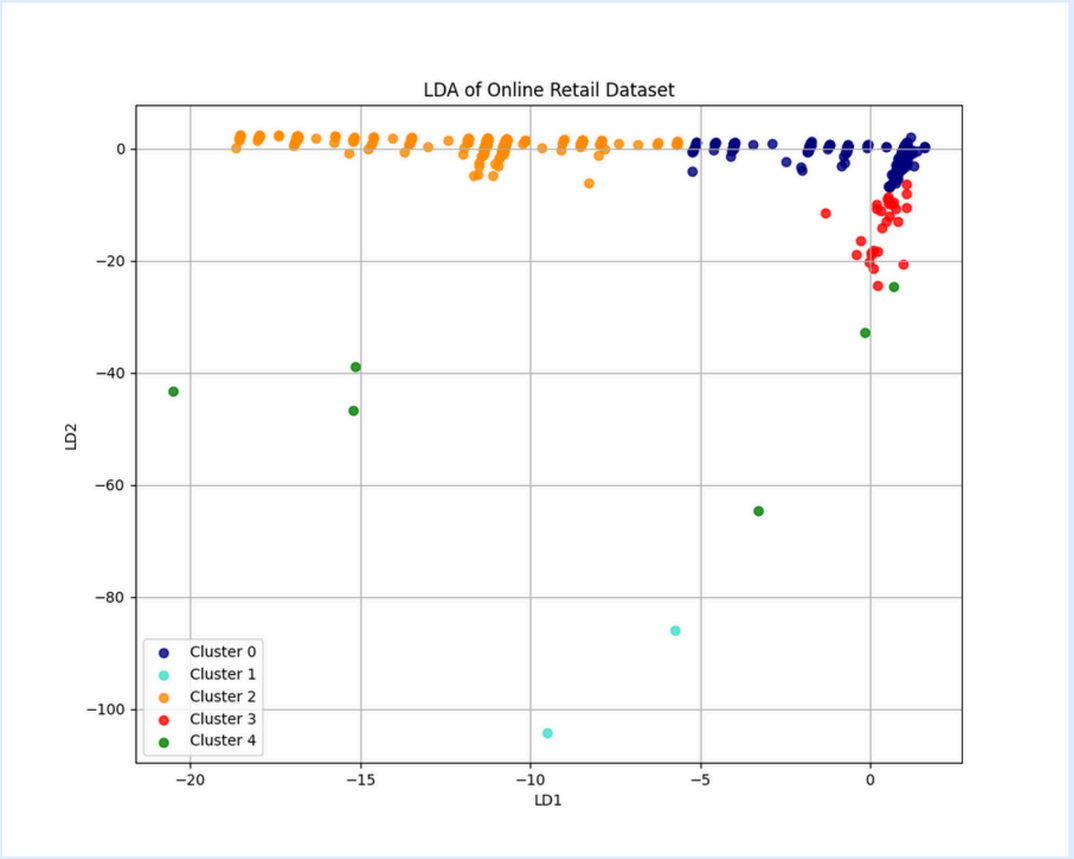
- LD1 separates high-quality (7, 8) from low-quality (3, 4).
- Alcohol and acidity are key discriminants.



ANALYSIS RESULTS

Online Retail

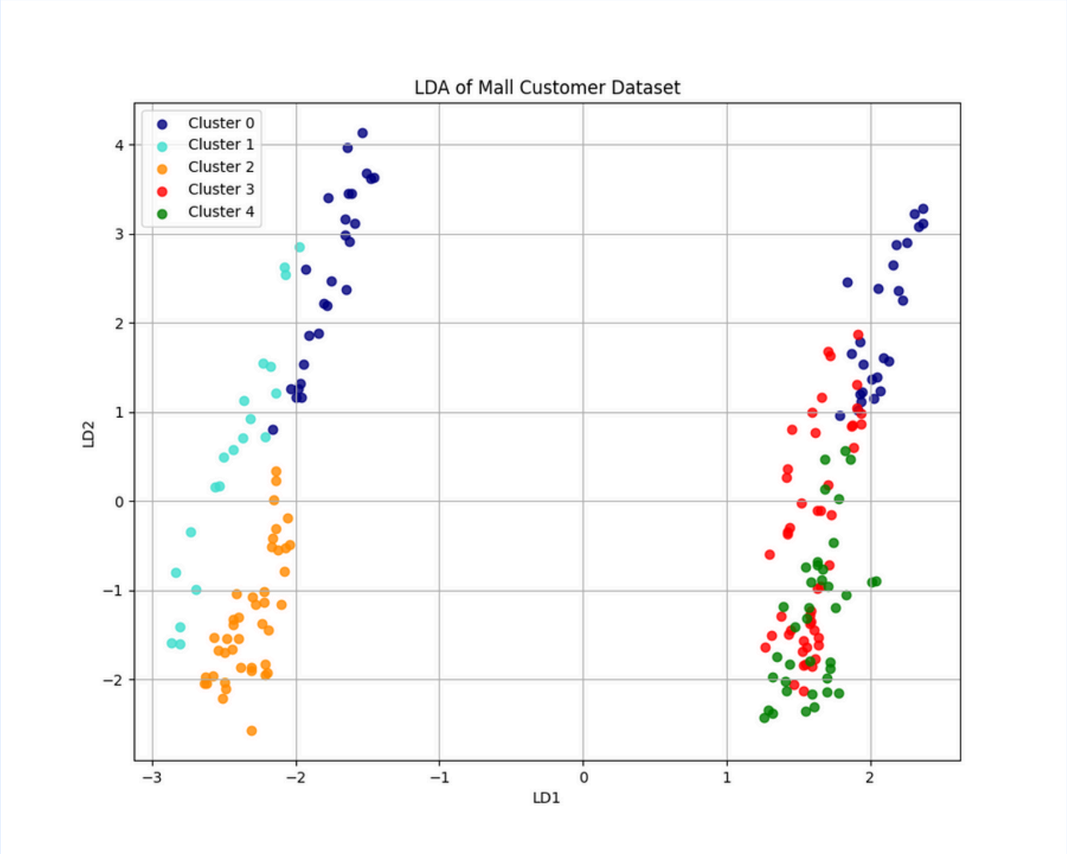
- LD1 captures spending behavior.
- LD2 captures regional variations.



ANALYSIS RESULTS

Mall Customer

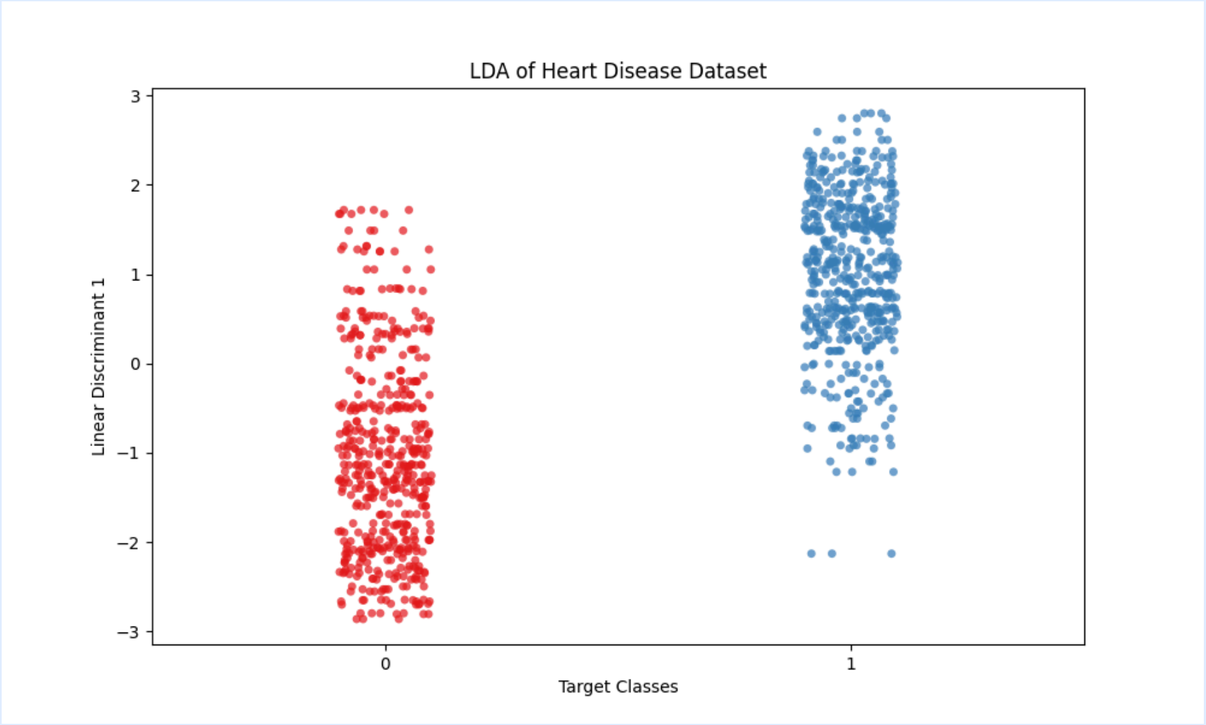
- LD1 distinguishes spending habits; LD2 captures demographics.
- High-spending vs. low-income clusters can be identified.



ANALYSIS RESULTS

Hearth Disease

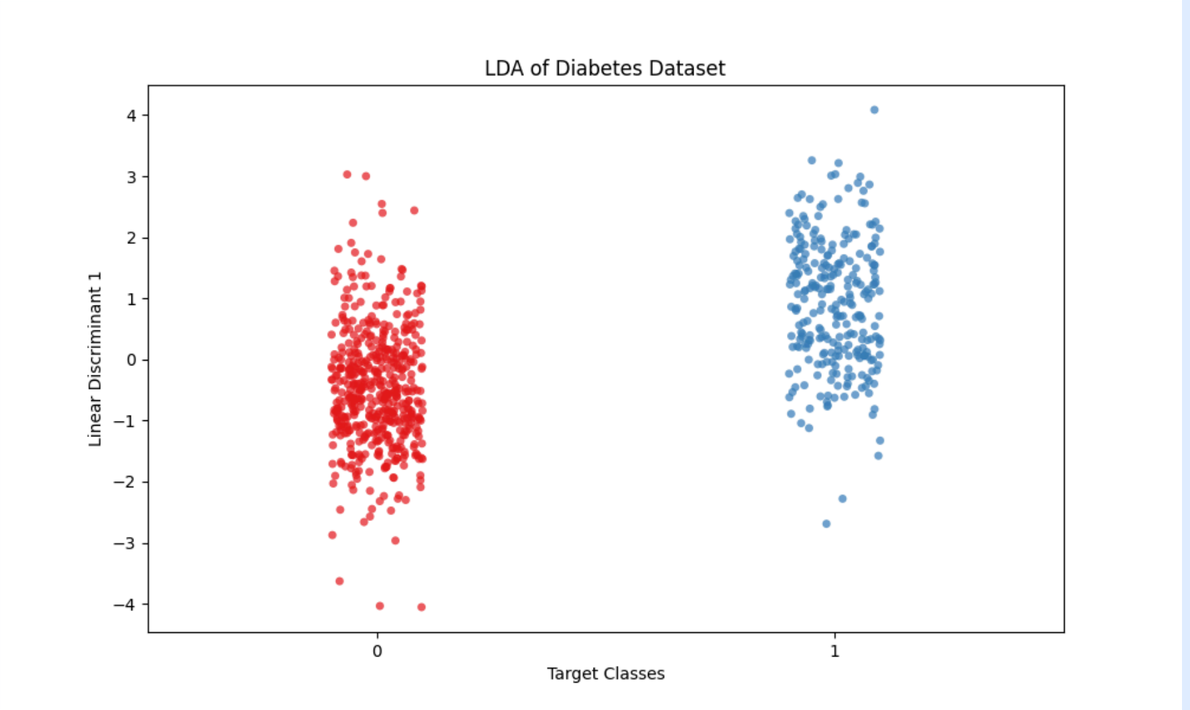
- Cholesterol levels and blood pressure are the most important indicators.
- Clear boundary using the clinical features to separate the two groups.



ANALYSIS RESULTS

Diabetes

- Glucose levels and BMI (Body Mass Index) are the strongest predictors.
- Clear divison between diabetic and non-diabetic individuals.



CONCLUSIONS & FUTURE WORK

Strengths:

- Clear interpretability and computational efficiency.
- Effective in dimensionality reduction and class separability.

Limitations:

- Requires assumptions like normality and covariance homogeneity.
- Struggles with non-linear separability and outliers.

Future Directions:

- Explore advanced methods: Sparse LDA, QDA, and RDA.
- Integrate with neural networks for hybrid, real-time applications.

Babeș-Bolyai University

THANK YOU

Turcu Ciprian-Stelian | 2024