Name: Đỗ Bá Hoàng Minh
ID: 22BI13278
PCA report

1. Datasets

The two chosen datasets for this labwork are Iris Extended and Heart Disease:

Dataset	Samples	Features	Type
Iris Extended	1200	21	Multiclass classification
Heart Disease	1025	14	Binary classification

1.1. Iris Extended

1.1.1. Feature types

Feature	Type	Description
Species	Categorical - Qualitative	Class/Label (Setosa, Versicolor, Virginica)
soil_type	Categorical - Qualitative	Soil category
elevation	Numerical - Continuos	Elevation value
Sepal-length, sepal_width, petal_length, petal_width	Numerical - Continuos	Measurements
sepal_area, petal_area	Numerical - Continuos	Derived features
ratios	Numerical - Continuos	Engineered features

1.1.2. Data quality

- No missing value found
- All continuous features are numeric and consistent
- Categorical features require encoding

1.1.3. Preparation

- Encode categorical attributes
- Normalize numerical features using StandardScaler

1.2. Heart Disease

1.2.1. Feature types

Feature	Type
target	Binary label (0 = no disease, 1 = disease)
age, cholesterol, resting_bp, max heart_rate, oldpeak	Numerical - Continuous
Sex, fasting_blood_sugar, chét_pain_type, slope, thalassemia, vessels_colored	Categorical - Qualitative

1.2.2. Data Quality

- No missing value found
- Categorical attributes were encoded
- Numerical features were standardized

1.2.3. Preparation

- Encode categorical attributes
- Normalize numerical features using StandardScaler

1.3 Statistical Measures

For both datasets, the mean, variance, covariance, and correlation were computed using the formulas.

Encoding is needed before calculating statistics.

Most correlated feature pairs:

Dataset	Most correlated features	Observation
Iris	petal_length - petal_area	Strong positive correlation
Heart Disease	Max_heart_rate - target	Strong negative correlation

This shows that engineered features and cardiovascular indicators strongly influence class separation.

2. Principal Component Analysis

2.1. PCA methods

Steps:

- Standardize data
- Compute covariance matrix
- Extract eigenvalues and eigenvectors
- Sort by descending eigenvalues
- Select top-k components

2.2. Influence of the Number of Principle Components (k)

The influence of the number of selected principal components (k) was analyzed for both datasets by observing the cumulative explained variance and the resulting data distribution.

Iris Extended:

k	Cumulative Variance
1	57.23%
2	74.79%
3	81.99%
4	87.28%
5	91.46%
6	93.99%

According to the table, we can assure that k = 3 is optimal.

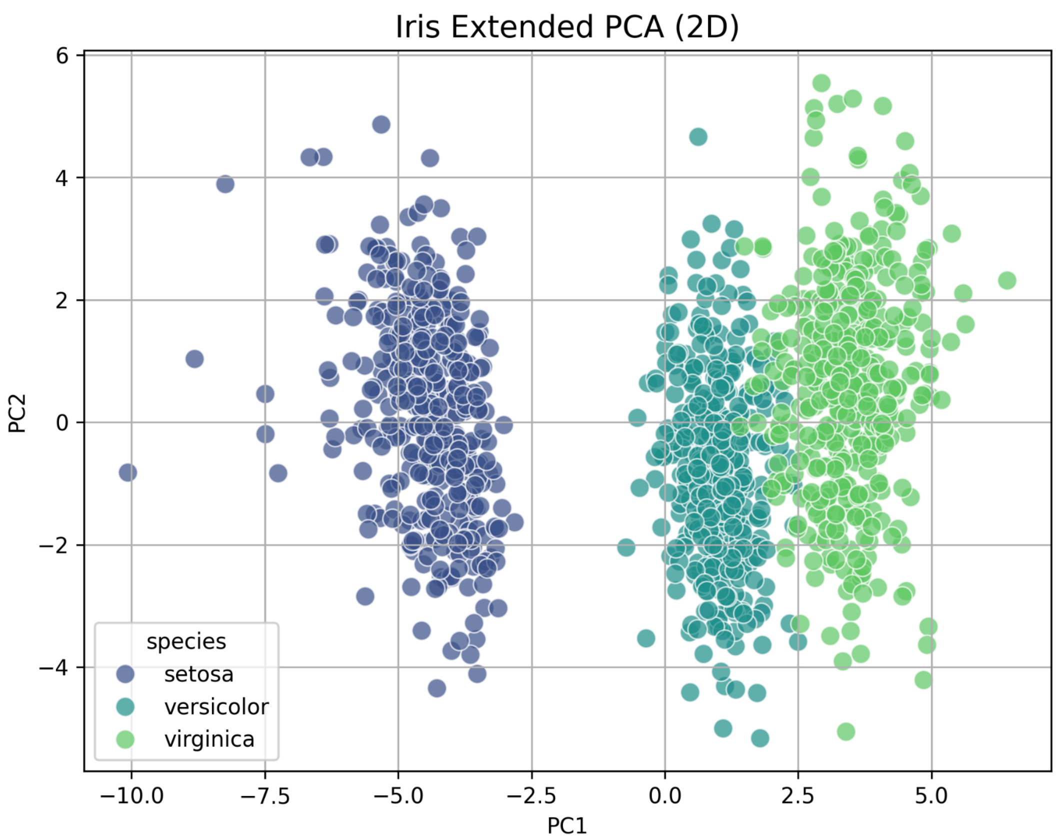
Doing the same with Heart Disease, we can conclude that it needs a much larger K than Iris Extended (k = 8)

Comparison:

Dataset	K for approx 80% variance	Structure
Iris Extended	3	Low dimensional, well structured
Heart Disease	8	High dimensional, complex

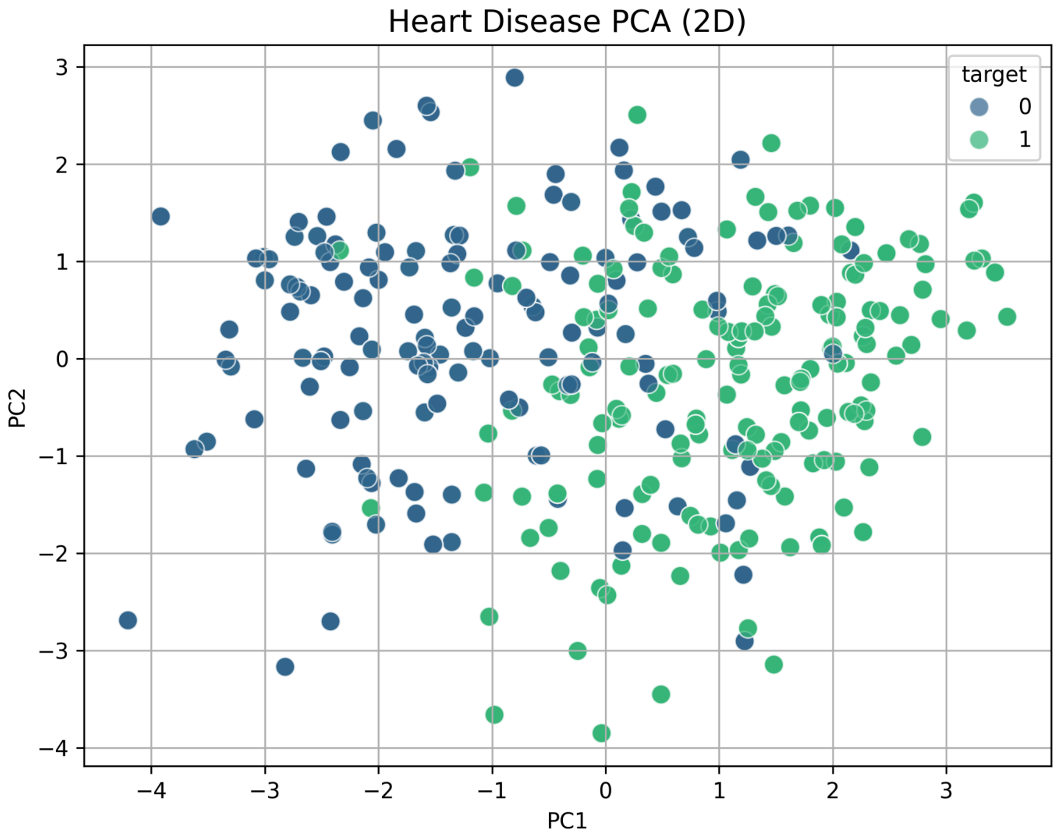
2.3. PCA Visualization and Analysis

2.3.1. Iris Extended



The Iris dataset shows very clear cluster separation using only PC1 and PC2. The three species form compact and well-separated clusters, indicating strong low-dimensional structure.

2.3.2. Heart Disease



The two-dimensional projection shows partial class separation along PC1, but significant overlap remains, indicating high data complexity. PC1 is strongly related to cardiovascular risk factors, while PC2 represents secondary variability.

3. Discussion

The two datasets exhibit very different structures. Iris Extended has highly correlated features and is easily separable in low-dimensional space. In contrast, Heart Disease is high-dimensional, with overlapping classes, making it more challenging for machine learning models.

4. Conclusion

PCA effectively reduced dimensionality, enhanced visualization, and revealed intrinsic data complexity. It demonstrates that dimensionality reduction is crucial for understanding dataset structure and improving learning efficiency.