Data Mining II — D212

Task 2: Principal Component Analysis (PCA)

by

Prateep Kul

January 01, 2025

# Table of Contents

1. 1. Part I: Research Question
2. 2. Part II: Method Justification
3. 3. Part III: Data Preparation
4. 4. Part IV: Analysis
5. 5. References

# Part I: Research Question

The research question focuses on understanding the factors contributing to overweight patients, significantly impacting cost-effective treatment plans through PCA.

# Part II: Method Justification

Principal Component Analysis (PCA) reduces dimensionality by identifying components explaining variance. The technique uncovers relationships between variables and improves interpretability.

# Part III: Data Preparation

## 1. Identifying Continuous Data Variables

All numeric columns (except the target column) were dynamically selected for PCA.

# Select continuous columns for PCA  
feature\_columns = [col for col in data.columns if col != config.TARGET\_COLUMN]  
numeric\_columns = data[feature\_columns].select\_dtypes(include=['number']).columns  
  
X = data[numeric\_columns]  
y = data[config.TARGET\_COLUMN]

## 2. Standardizing Continuous Data Variables

Standardization ensures variables contribute equally to PCA by scaling them to mean 0 and standard deviation 1.

from sklearn.preprocessing import StandardScaler  
  
# Standardize the features  
scaler = StandardScaler()  
X\_scaled = scaler.fit\_transform(X)

## 3. Performing PCA

PCA was performed using the specified number of components, transforming the data to its principal components.

from sklearn.decomposition import PCA  
  
# Perform PCA  
pca = PCA(n\_components=config.PCA\_COMPONENTS\_RETAINED)  
X\_pca = pca.fit\_transform(X\_scaled)

## Visualizations

The following visualizations illustrate the distribution of features before and after standardization, as well as the explained variance by PCA components.

Visual not found: Figure 2: Distribution of Features (Before Standardization)

Visual not found: Figure 3: Distribution of Features (After Standardization)

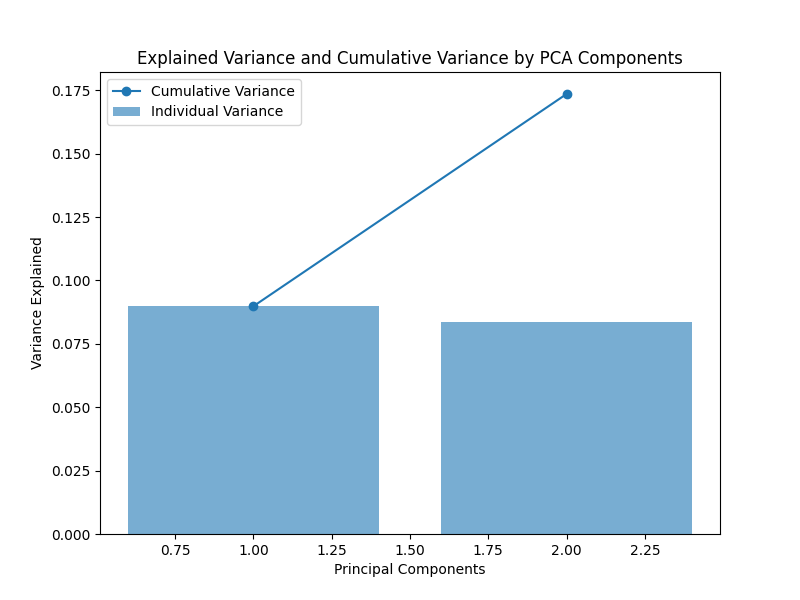


Figure 4: Explained Variance by PCA Components

# Part IV: Analysis

Visual not found: Figure 5: Correlation Heatmap

The analysis demonstrates how principal components capture maximum variance, revealing important patterns.

# References

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Retrieved from https://jmlr.org/papers/v12/pedregosa11a.html

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Retrieved from https://doi.org/10.1109/MCSE.2007.55

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Retrieved from https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

Waskom, M. L. (2017). Seaborn: Statistical Data Visualization. Retrieved from https://seaborn.pydata.org/

Oliphant, T. E. (2006). A Guide to NumPy. Retrieved from https://numpy.org/