Project Report

"Heart Disease Dataset"

Course: Data Visualization

Instructor: Ilyes Jenhani

Prepared by:

Rania Manel MOUHOUBI

Lina MEHDI

Date:

November 21, 2024

Our Project :

We selected a **heart disease** dataset containing a variety of variable types, including continuous, categorical, and binary variables. This dataset was validated by the instructor. Throughout the project, we conducted multiple analyses, including PCA, CA, MCA, and FAMD, complemented by various graphical visualizations and in-depth interpretations.

Link to the html file: Heart-Disease-Analysis.html

GitHub Link : https://github.com/LynaMahdi/Comparative-analysis-of-PCA-CA-and-MCA/blob/main/projet.ipynb

Note: the merges have been done manually what explain the few commits in the repo.

Comparative Analysis of Methods

From our analyses, we can draw the following observations and insights:

1. **Continuous Variable Distribution**
    a. From the distribution of our continuous variables, we observe that most variables follow an approximate normal distribution. This is beneficial for statistical and machine learning models that assume normality, such as linear regression or PCA.
    b. However, the distribution across different categories is not homogeneous. For example, certain categories have significantly fewer observations than others, which indicates class imbalance. This could potentially bias our analysis and reduce the robustness of the results.
2. **Correlation Analysis**
    a. Visualizing the heatmap of correlations between continuous variables reveals relatively low correlation values, with a maximum of only around 0.58
    b. Additionally, the target variable shows weak correlations with other features, indicating that the existing variables may not sufficiently explain the target outcome. This suggests the need to augment the dataset with additional features or perform feature engineering to uncover hidden relationships.
3. **PCA (Principal Component Analysis)**
    a. PCA on continuous variables allowed us to capture approximately **75% of the variance** with only **3 components**, which is a good dimensionality reduction outcome. This reduces the complexity of the dataset while retaining most of the information.
    b. Although we did not identify clear patterns or clusters in the reduced PCA space, the transformation could still be advantageous for training machine learning models by simplifying the data structure and reducing noise.
4. **CA (Correspondence Analysis)**
    a. For categorical data, performing CA on a single variable was not very informative since we captured **100% of the variance** due to the limited number of categories.
    b. By combining two variables (e.g., `ca` and `cp`), we captured **83% of the variance** with the first component and the remaining **16% with the second component**, effectively reducing the dimensionality of the categorical dataset.
    c. This method allowed us to visualize relationships between categorical variables, such as the association between chest pain levels (`cp`) and age. For instance, **higher levels of chest pain (cp) appear to correlate with older age groups**, suggesting that aging might be linked to more severe health conditions.
5. **MCA (Multiple Correspondence Analysis)**

    a. Captures less variance in the first components (around 25% with two components).

b. Focuses on relationships between categorical variables and identifies groupings or outliers.
c. Highlights contributions of rare or unique categories, though more dimensions may be needed for complete representation.
d. Suitable for visualizing categorical data and identifying shared profiles or distinctions.

6. **FAMD (Factorial Analysis of Mixed Data)**
    a. FAMD captured a relatively low proportion of variance in the first components, reflecting the challenge of balancing contributions from continuous and categorical variables in a mixed dataset.
    b. Continuous variables dominated the contributions to the first components, while categorical variables added less structured information and potentially introduced noise.
    c. This result emphasizes the need for careful preprocessing and feature engineering when working with mixed data. For example, creating new variables based on existing ones or transforming categorical variables to reduce redundancy and enhance interpretability.

Proposed Solutions

1. **Feature Engineering**: Create new variables that better capture relationships within the data. For instance, derive interaction terms or combine existing variables into composite features.
2. **Additional Data**: Enrich the dataset with external variables or measures that could improve the predictive power and correlations with the target variable.
3. **Class Balancing**: Address class imbalance through resampling techniques (e.g., oversampling minority classes or undersampling majority classes) to ensure fair representation across categories.
4. **Advanced Preprocessing**: Experiment with transformations for both continuous (e.g., normalization) and categorical variables (e.g., one-hot encoding or hierarchical grouping) to better align with the analysis methods.

Conclusion

The comparative analysis highlights the strengths and limitations of each method:

- **PCA** is highly effective for continuous data, capturing significant variance with fewer components.
- **CA** is best suited for categorical variables, offering insights into relationships between categories.
- **FAMD** provides a holistic view for mixed datasets but requires careful tuning to overcome the challenges of integrating different variable types.

.