UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Educational Data Analysis

ACADEMIC WORK REPORT

Dimensionality Reduction

Francesco Pio Capoccello
Daniele Borghesi

ACADEMIC YEAR 2023-2024

Contents

1 Introduction

This report illustrates the results that can be obtained by applying dimensionality reduction techniques, such as *Exploratory Factor Analysis* (EFA), and *clustering* to educational data. In particular:

- **Exploratory Factor Analysis (EFA)** is a powerful tool for exploring and understanding the hidden structure in data. It seeks to identify the underlying patterns or factors that influence the correlations between observed variables. The main objective of EFA is to simplify a large number of variables into a smaller number of unobserved factors. These factors represent common or latent characteristics among the variables and help explain correlations among them.
- **Clustering** is a data analysis technique that groups similar data into clusters, so that items within one cluster are more similar to each other than those in other clusters. It is used to discover hidden patterns and structures within data by clustering similar data points together. Clustering helps simplify complex data, identify relationships between observations, and make decisions based on similarities between data.

Regarding the data, we used the dataset related to the *Reflex project* (Competences and job values of university graduate students). In particular we worked on a portion of the dataset considering only the features related to the job values, all of which are of integer type: *Work_Autonomy*, *Job_Stability*, *Opportunity_Things*, *High_Income*, *Meeting_Challenges*, *Good_Prospects*, *Have_Yourself*, *Acknowledgment*, *Opportunity_Yourself*, *Work_Balance*.

2 Exploratory Factor Analysis

This section describes the steps to perform a Exploratory Factor Analysis (EFA) of our set of selected features. Specifically, in section 2.1 we proceeded to calculate the Kaiser-Meyer-Olkin index for the available feature. Next, the eigenvalues of the extracted factors were analyzed in subsection 2.2. The factors were then analyzed and interpreted in subsection 2.3.

2.1 Kaiser-Meyer-Olkin Index analysis

The Kaiser-Meyer-Olkin Adaptability Index (KMO) is a measure used in Factor Analysis to assess the feasibility of data for EFA. KMO indicates whether the observed variables in our dataset are suitable for factor analysis. A KMO value close to 1 suggests that the data are appropriate for EFA, while a value close to 0 suggests that the data may not be suitable. A value between 0.5 and 0.6 may be sufficient to apply EFA.

Table 1: KMO Values

Feature name	***KMO Value***
Work_Autonomy	0.86
Job_Stability	0.78
Opportunity_Things	0.79
High_Income	0.78
Meeting_Challenges	0.75
Good_Prospects	0.82
Have_Yourself	0.69
Acknowledgement	0.76
Opportunity_Yourself	0.76
Work_Balance	0.69
Global	**0.76**

Table 1 shows the KMO scores computed for each feature and globally. As can be seen, the values obtained are all greater than or equal to 0.69. These values suggest that our data are suitable for performing the EFA.

2.2 Eigenvalues analysis

After verifying that the data are suitable for EFA through the KMO Index calculation, we decided to apply dimensionality reduction in factors.

According to Kaiser's criterion, to decide how many factors to maintain, it is necessary to control the value of eigenvalues. When eigenvalues are greater than 1, they indicate that the factor explains more variance than a single variable. Usually, only factors with eigenvalues greater than 1 are retained. If an eigenvalue is very close to 1, it might indicate that that factor explains approximately the same amount of variance as a single variable.

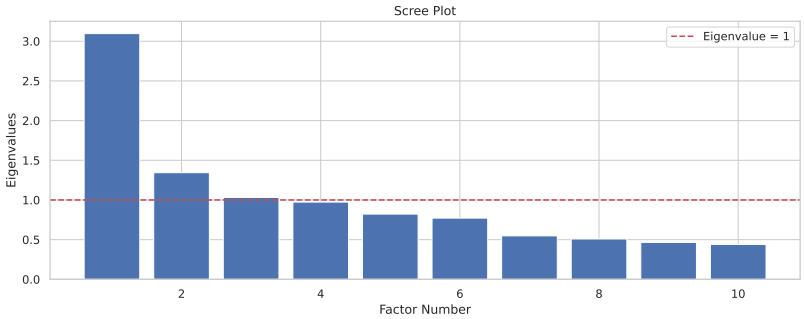

Figure 1: Eigenvalues obtained for factors

Each bar in figure 1 represents the eigenvalue of a factor, which indicates how much a factor explains the variance in the data: a visual guidance on how many dimensions (factors) to retain. In our case, the graph clearly shows that the first three factors' eigenvalues exceed the threshold of 1. We decided to maintain these trhee factors, capable to explain substantial amount of variance in the original data.

2.3 Factors analysis

Following the selection of then number of factors, we decided to analyze them to check the contribution of each feature.

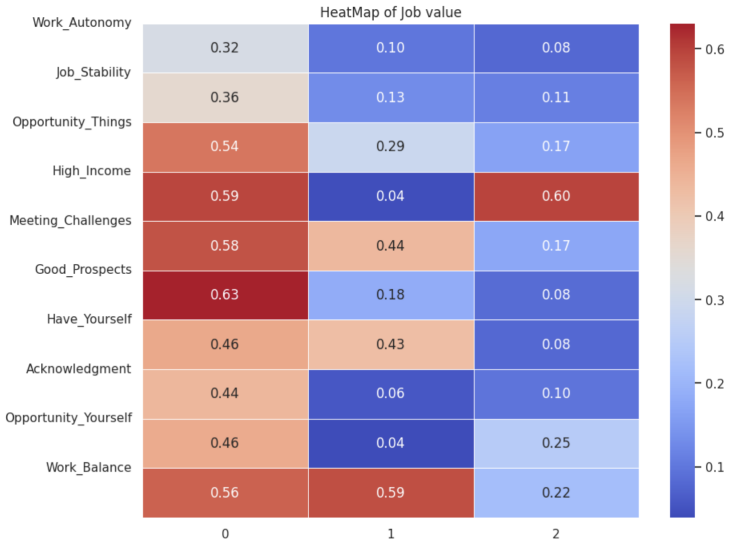

Figure 2: HeatMap obtained for 3 factors

Figure 2 shows a *HeatMap* about the contribution of each original feature to each of the three chosen factors. By looking at the values in the HeatMap, it is possible to identify which variables have more or less influence on the extracted factors.

As can be seen, the first factor (which has a higher eigenvalue) includes the contribution of several features, particularly *Opportunity_Things*, *Meeting_Challenge* and *Good_Prospects*. Other features that are included to a greater extent in the first factor are *Work_Autonomy*, *Job_Stability*, *Have_Yourself*,

Acknowledgment and *Opportunity_Yourself*. At the same time, the *Work_Balance* feature contributes mainly to the second factor, and the *High_Income* feature to the third.

3 Cluster analysis

The following section shows the results of applying two clustering techniques: hierarchical (subsection 3.1), and K-Means (subsection 3.2). In particular, a comparison of the two obtained clustering was performed, highlighting the differences between them.

3.1 Hierarchical clustering

As for hierarchical clustering, we performed the algorithm using Euclidean distance. The four main linkage methodologies were compared: single, complete, average and Ward.

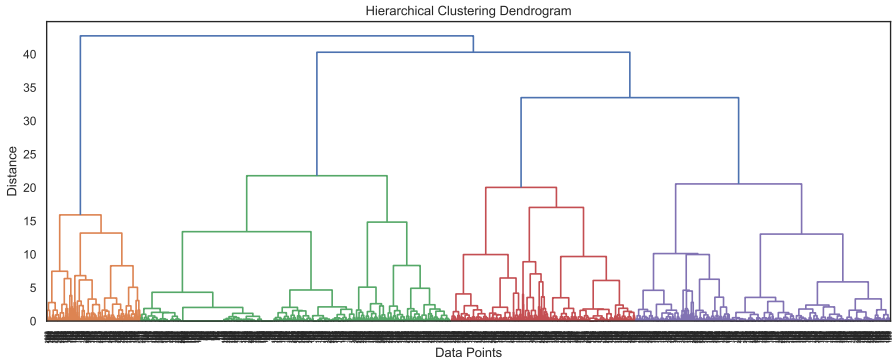

Figure 3: Hierarchical clustering obtained on three factors (Ward method)

A visual analysis of the dendrograms of the 4 linkage types shows significant differences among the clustering obtained.[1] In particular, Ward's method is the only linkage type that shows 4 distinct and separate clusters, as shown in Figure 3.

The dendrogram shows the 3 factors identified in section 2 lead to a division of instances into four main categories, which could represent four possible classes.

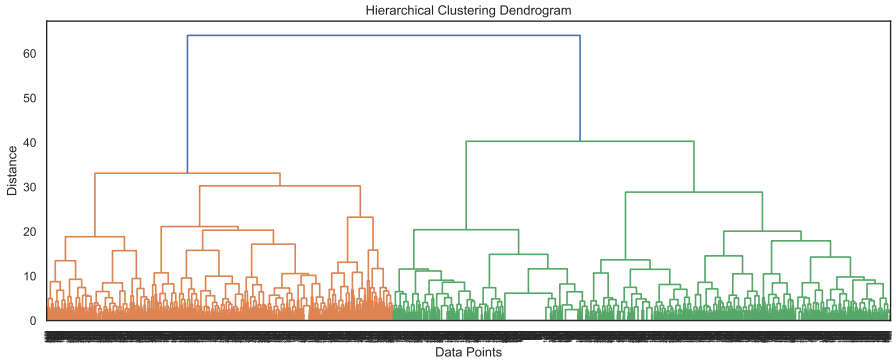

Figure 4: Hierarchical clustering obtained on the original features (Ward method)

Figure 4 shows that even keeping the original features, 4 clusters can be identified. However, these are less distinguishable and separated than those identified by clustering performed on the factors. In particular, keeping the original features makes a division into only two clusters more feasible.

In conclusion, hierarchical clustering shows the existence of at least four distinct classes among the instances. It also shows how reducing the dimensionality of the features into three factors allows these potential classes to be delineated in a more pronounced manner.

[1] Dendrograms related to clustering performed with linkages of type single, complete, and average can be seen within the notebook `clustering.ipynb` attached to this report

3.2 K-Means clustering

The results obtained with hierarchical clustering showed that the reduction of features into three factors leads to the identification of 4 distinct main clusters. In this regard, we decided to perform K-Means clustering with a number of centroids equal to 4, to test whether the 4 clusters identified with the hierarchical technique are equally identifiable with a centroid-based strategy such as K-Means. Also in this case, the euclidean distance was maintained.

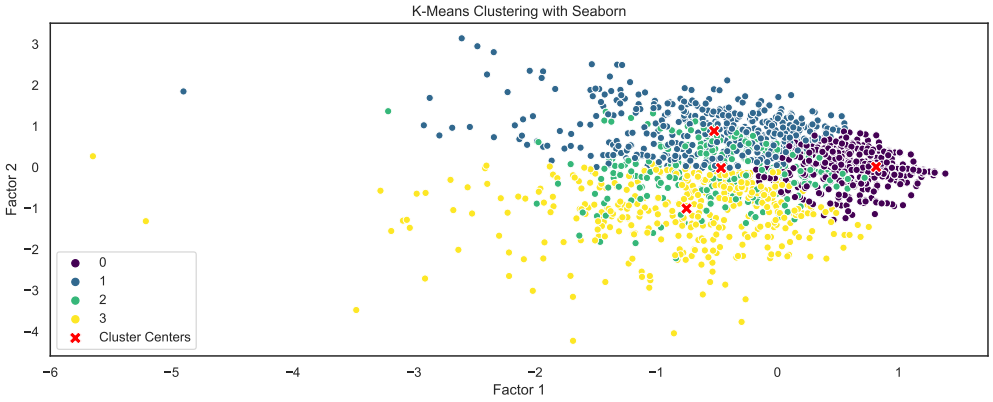

Figure 5: K-Means clustering obtained on three factors with $k = 4$

Figure 5 shows how the 4 clusters hypothesized with hierarchical clustering are also easily discernible with a centroid-based clustering strategy such as K-Means. Two-dimensional visualization of the first and second factors (the most important ones) shows how the 4 clusters are clearly visible and distinguishable. Any overlaps between clusters are hypothetically due to the 2-dimensional visualization: a 3-dimensional visualization would allow the clusters to be visualized more prominently.

4 Conclusions

In this report, an Exploratory Factor Analysis (EFA) (section 2) and a cluster analysis (section 3) were performed on a group of features. The objective was to see if reducing the number of features into a smaller amount of factors might be possible. In addition, we analyzed whether this reduction in dimensionality could lead to improvements in clustering.

The results showed that EFA allows reducing dimensionality from 10 features to 3 factors while maintaining a high level of information. In addition, the clustering results showed that the reduced dimensionality dataset provided higher quality hierarchical clustering than the original dataset. The centroid-based clustering confirmed the visibility of the four clusters hypothesized with hierarchical clustering.