Unlocking Patterns: Clustering Analysis with KDD

Nick Kornienko

September 21, 2023

Abstract

This research paper encompasses a clustering analysis using the KDD methodology on a multifaceted dataset, aiming to reveal inherent patterns and groupings within the data. The structured approach of KDD allowed for comprehensive exploration, transformation, and interpretation of the various variables in the dataset, ultimately leading to the discovery of insightful patterns.

1 Introduction

The exploration of patterns within diverse datasets is a fundamental aspect of data science. This research focused on a meticulous application of the KDD methodology to perform a clustering analysis and interpret the inherent patterns within a dataset containing demographic and socioeconomic variables.

2 Selection and Preprocessing

The dataset, consisting of variables like Sex, Marital status, Age, Education, Income, Occupation, and Settlement size, was carefully selected and preprocessed. The preprocessing phase revealed a dataset with no missing values, enabling a seamless transition to the subsequent phases of the KDD process.

3 Transformation and Data Mining

Normalization was pivotal in the transformation phase due to the diversity of the dataset's variables, preparing the data for the k-Means clustering algorithm. The optimal number of clusters was determined using the Silhouette score, and meticulous data mining unveiled distinct patterns and groupings within the dataset.

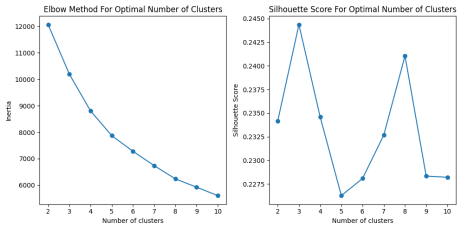

Figure 1: Silhouette Score depicting the optimal number of clusters

Figure 2: 2 Component PCA depicting the Clusters

4 Interpretation and Evaluation

The interpretation phase focused on understanding the inherent groupings and patterns within the dataset, and the evaluation was aided by quantitative and visual representations provided by Silhouette scores and PCA plots, showcasing the cohesion and separation of the discovered clusters.

5 Conclusion

The KDD methodology facilitated a structured and comprehensive exploration, transformation, and interpretation of the dataset. The insights and patterns revealed through this methodology provide a profound understanding of the dataset's inherent groupings and a foundation for subsequent detailed analysis and research.