Kmeans Clustering vs. Graph based clustering

Football dataset: [Data] Code for [Kmeans] Clustering Code for [Graph based] Clustering

Details about the chosen dataset:

The dataset consists of 177 rows and 22 columns, capturing comprehensive match statistics from a football season. It includes details such as the match date, participating teams (HomeTeam and AwayTeam), full-time (FTHG, FTAG, FTR) and half-time scores (HTHG, HTAG, HTR), referees, and various metrics like shots, fouls, corners, and yellow/red cards for both teams. The data is complete, with no missing values, making it ideal for analyzing team performance, referee influence, or patterns in match outcomes.

K-Means Clustering Report

Introduction

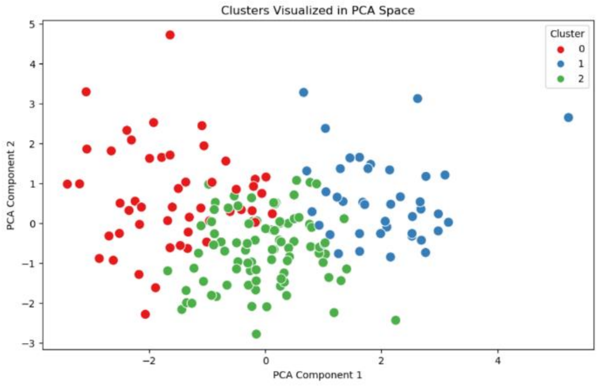
K-Means clustering is an unsupervised algorithm that groups data into distinct clusters. The goal is to minimize within-cluster variance. The challenge lies in selecting the optimal number of clusters, k, which is addressed using methods like the Elbow Method.

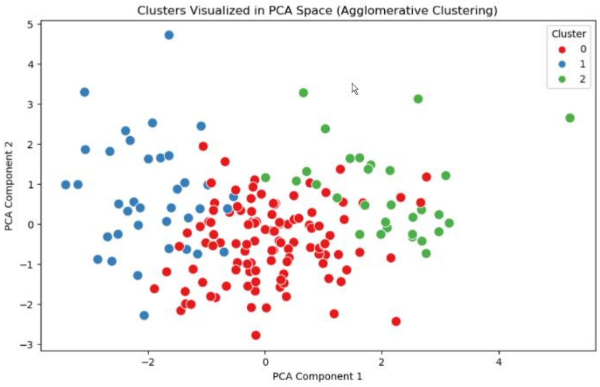
Process Overview

- The dataset (season-2425.csv) was loaded, and relevant numerical columns representing football match statistics were selected.
- Preprocessing: Missing values were removed, and the data was standardized using StandardScaler to ensure equal contribution from all features.
- PCA was applied to reduce the data to 2 dimensions for visualization.
- The Elbow Method was used to find the optimal number of clusters, with k = 3 determined as the best choice based on the inertia plot.
- K-Means clustering with k = 3 was applied, and cluster labels were added to the dataset.
- The clusters were visualized in 2D using PCA components, with distinct color-coding for each cluster.
- The Silhouette Score was calculated to evaluate clustering quality, with a higher score indicating good separation between clusters.
- The final clustered dataset was saved as kmeans_clustered_football_data.csv for further analysis.

Conclusion

K-Means clustering successfully segmented the dataset into three distinct clusters, providing insights into match statistics. The Elbow Method and Silhouette Score confirmed the effectiveness of the clustering. The results are saved for further exploration.





Graph-Based Clustering Report

Introduction

Graph-based clustering, such as **Agglomerative Clustering**, is a hierarchical clustering method that builds a tree of clusters based on data similarities. It is particularly useful when the number of clusters is unknown or the data structure is complex.

Process Overview

- The dataset (season-2425.csv) was loaded, and relevant columns representing football match statistics (goals, shots, fouls, etc.) were selected.
- Preprocessing :Missing values were removed, and the data was standardized using StandardScaler.
- PCA was applied to reduce the data to 2 components for easier visualization.
- Agglomerative Clustering was applied with n_clusters = 3 to group the data into three clusters.
- The fit_predict method was used to assign each data point to a cluster.
- The clusters were visualized in 2D using PCA components, with distinct colors for each cluster.
- The Silhouette Score was computed to evaluate the quality of the clustering. If multiple clusters were found, the score provided insight into cluster cohesion and separation. If all data points were assigned to one cluster, the score could not be calculated.
- The final clustered dataset was saved as graph_based_clustered_football_data.csv.

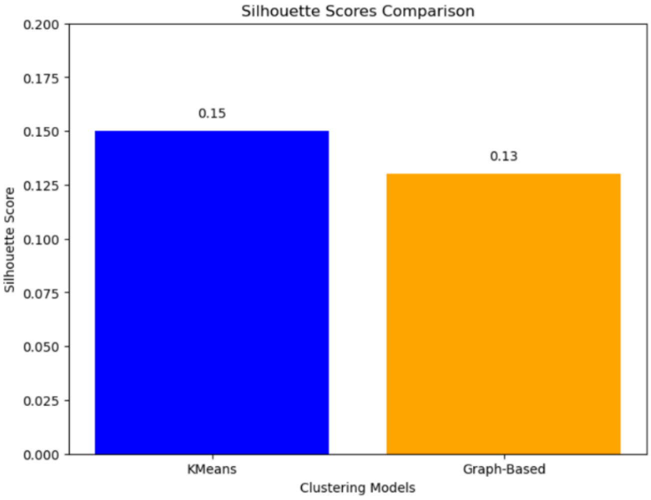
Conclusion

Agglomerative Clustering successfully grouped the football match statistics into three clusters. The clustering was visualized in 2D, and the Silhouette Score confirmed the clustering quality. The results have been saved for further analysis.

Comparison

The Silhouette Scores for K-Means and Agglomerative (graph-based) clustering models were 0.15 and 0.13, respectively. Both scores are relatively low, indicating that the clustering results are not highly well-separated, with some data points potentially being assigned to incorrect clusters.

However, the K-Means model shows a slightly better separation between clusters (higher score) compared to the Agglomerative Clustering model. This suggests that, for this dataset, K-Means might offer a marginally better grouping of data points, though both models may benefit from further tuning or exploration of alternative clustering methods.


Silhouette Scores Comparison
Silhouette Score
KMeans
Graph-Based
Clustering Models