Graph-Based and K-Means Clustering Analysis for Data Mining and Discovery

NAME: CYNTHIA CHINENYE UDOYE
STUDENT NO: 22029346

Introduction

Clustering is a widely used method in data mining to group items with similar characteristics (Tan et al., 2018). In this report, we analyse Premier League match data from the 2021-2022 season, which includes attributes such as goals scored, fouls, and match outcomes, to group teams based on their performance. Two clustering techniques are applied: Graph-Based Clustering and K-Means Clustering. These methods help identify patterns in team performance and uncover similarities between teams.

Graph-based clustering constructs graphs where nodes represent data points, and edges are formed based on proximity or similarity measures. This approach is particularly effective in detecting clusters with strong internal connections (Hastie et al., 2009). In contrast, K-Means clustering partitions data into distinct groups by minimizing intra-cluster variance, a method valued for its simplicity and efficiency in handling large datasets (Tan et al., 2018). To determine the optimal number of clusters, the Silhouette method was used, as it effectively measures the quality of clusters by evaluating their compactness and separation (Shahapure and Nicholas, 2020).

The report includes a detailed data preparation process, the application of both clustering techniques, and a comparison of their results. Metrics such as silhouette scores and visualizations are used to evaluate the quality of the clusters. This analysis demonstrates how data mining techniques can extract valuable insights from sports datasets.

Methodology

To analyse team performance, we applied two clustering methods: **Graph-Based Clustering (Spectral Clustering)** and **K-Means Clustering**. The following steps were carried out:

1. **Data Preparation**:
   The dataset for this analysis consists of two CSV files containing Premier League match data from the 2021-2022 season, retrieved from Kaggle. It includes statistics for 380 matches across 20 teams. The first file contains match-level data (e.g., goals, fouls, cards), while the second provides weekly team rankings, goal differences, and points. Key data preparation steps included:
   - **Aggregating Stats**: Match statistics were separated into home and away team performances. Metrics like goals scored, goals conceded, fouls, and shots were aggregated for each team.
   - **Merging Data**: Combined aggregated stats with weekly rankings to create a team-level dataset summarizing performance, including contextual features like goal differences, points, and ranks.
   - **Feature Selection and Scaling**: Key features selected for clustering were normalized to ensure comparability and prevent domination by larger values and included:
     - **Performance**: Goals scored, goals conceded, goal differences, points.
     - **Discipline**: Fouls, yellow cards, red cards.
     - **Activity**: Shots, shots on target, corners.
     - **Context**: Team rankings.
2. **Clustering Process**:
   - Determined the optimal number of clusters using the silhouette method.
   - Evaluated cluster quality with silhouette scores.
   - Applied Principal Component Analysis (PCA) to reduce dimensions for 2D cluster visualization.
3. **Cluster Analysis**:
   - Grouped teams within each cluster and analysed metrics such as performance, discipline, and activity.
   - Compared clustering results to highlight differences in team groupings and approaches.

Results

The following table summarizes the clustering results comparing Spectral Clustering and K-means Clustering based on key metrics, cluster characteristics, and key observations for both methods.

Table 1: Comparative Summary of Clustering Results for Spectral Clustering and K-means Clustering

Metric	Spectral Clustering	K-means Clustering
Optimal Clusters (k)	9	2
Silhouette Score	0.372 (moderate cluster separation)	0.483 (better-defined clusters)
Key Teams (Cluster)	- **Cluster 0**: Diverse mid-performing teams (e.g., Man United, Brighton, Wolves).	- **Cluster 0**: High performers (Chelsea, Liverpool, Man City).
	- **Cluster 1:** Top performers (Liverpool, Man City).	- **Cluster 1**: All other teams (e.g., Man United, West Ham, Everton).
	- **Clusters 2-8**: Individual teams with distinct characteristics (e.g., Chelsea, Arsenal, Norwich).	
Cluster Characteristics	Granular clustering capturing subtle differences between teams.	Simplified segmentation with clear performance distinction.
Teams per Cluster	- **Cluster 0**: 11 teams	- **Cluster 0**: 3 teams.
	- **Cluster 1**: 2 teams	- **Cluster 1**: 17 teams.
	- **Clusters 2-8**: Single teams in each cluster.	
Key Observations	More granular clustering with detailed separation, ideal for nuanced analyses.	Simplified clustering, better suited for high-level segmentation.

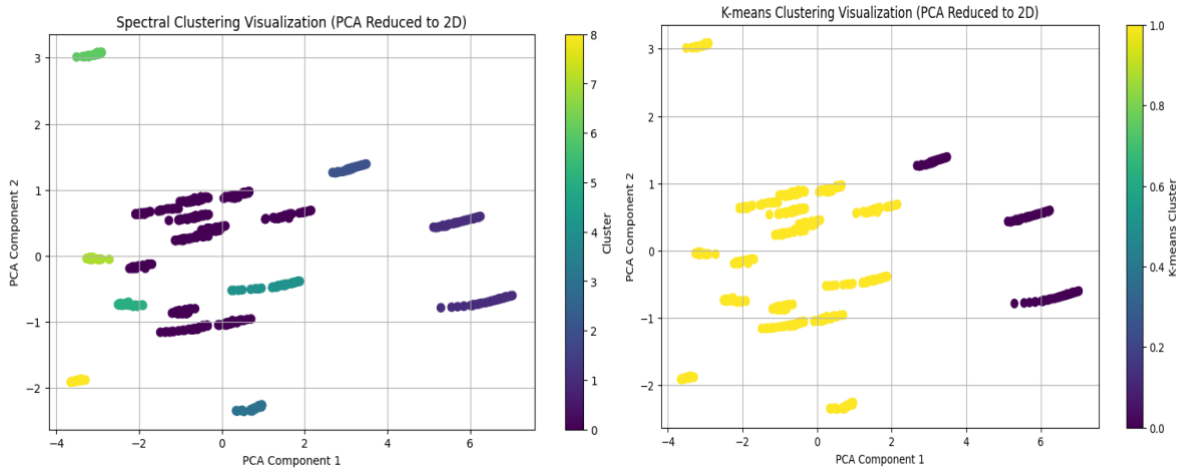

Figure 1: PCA Visualization of Spectral Clustering and K-means Clustering.

Discussion and Conclusion

The results reveal distinct clustering behaviours between Spectral Clustering and K-means, as shown in the PCA visualizations, where K-means clusters appear more compact and separated, as reflected in the results table. Spectral Clustering identified nine clusters, capturing finer distinctions between teams like Liverpool and Man City in Cluster 1 and mid-performing teams such as Man United and Wolves in Cluster 0. The presence of single-team clusters (e.g., Chelsea and Arsenal) suggests Spectral Clustering can highlight unique team behaviours but risks over-segmentation.

In contrast, K-means produced a clearer, simplified segmentation with only two clusters. Cluster 0 grouped high-performing teams (e.g., Chelsea, Liverpool, Man City), while Cluster 1 combined all other teams despite performance variations. The PCA visualization shows more distinct cluster boundaries for K-means, supported by its higher silhouette score (0.483 compared to 0.372), indicating better-defined clusters overall.

Both methods effectively highlighted patterns in team performance, with Spectral Clustering excelling in detailed segmentation and K-means in broader grouping. The choice between them depends on the analysis goal: Spectral Clustering for in-depth insights and K-means for a clearer, high-level view. Ultimately, both approaches offer valuable tools for exploring performance patterns in sports analytics.

References

Hastie, T., Tibshirani, R. and Friedman, J.H., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer. Available at: https://r3.vlereader.com/Reader?ean=9780387848587 [Accessed 28 December 2024].

Kaggle Dataset: Premier League Match Data. Available at: https://www.kaggle.com/datasets/evangower/premier-league-match-data [Accessed 18 December 2024].

Shahapure, K.R. and Nicholas, C., 2020. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, pp. 1-6. IEEE. Available at: https://ieeexplore-ieee-org.ezproxy.herts.ac.uk/document/9260048

Tan, P.-N., Steinbach, M., Kumar, V., and Karpatne, A. (2019) *Introduction to Data Mining EBook: Global Edition.* Pearson Education, Limited. Available at: https://ebookcentral.proquest.com/lib/herts/detail.action?docID=5720020 [Accessed 21 December 2024].

Colab Notebook: Code for *Graph-Based and K-Means Clustering Analysis Report* (IPYNB file). Available at: https://colab.research.google.com/drive/1vcVjpqnXSIaqVQBAAoJO5wWE6ubw9JQ9?usp=sharing