Report: Exploring Clustering-based Anomaly Detection and K-means Clustering in Football Match Data

1. Abstract: This report explores the application of machine learning methodologies—specifically Clustering-based Anomaly Detection and K-means Clustering—using a detailed football match dataset. The objective is to analyse and contrast these techniques for uncovering anomalies and inherent groupings within match statistics. The dataset includes diverse metrics such as goals scored, shots taken, and fouls committed by both home and away teams.

2.Introduction

In this report, we investigate two machine learning techniques applied to football match data: Clustering-based Anomaly Detection and K-means Clustering. These methodologies are employed to identify anomalies and reveal natural groupings within the dataset, providing insights into team performances during matches.

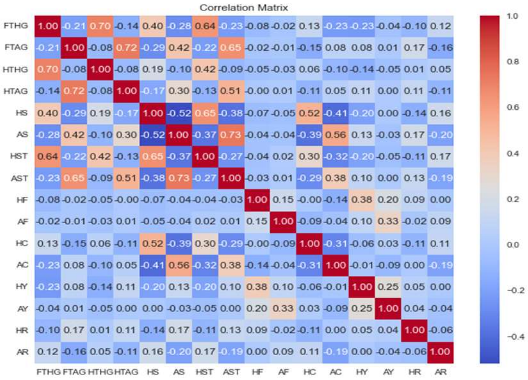
3.Dataset Overview

The football match dataset encompasses features like FTHG (Full Time Home Goals), FTAG (Full Time Away Goals), HS (Home Team Shots), AS (Away Team Shots), HST (Home Team Shots on Target), AST (Away Team Shots on Target), HF (Home Team Fouls), AF (Away Team Fouls), and more. These metrics offer valuable insights for clustering and anomaly detection analyses.

4.Exploratory Data Analysis (EDA)

Before applying machine learning techniques, exploratory data analysis (EDA) is conducted to understand the dataset's structure and relationships.

- **4.1. Summary Statistics**: Descriptive statistics provide an overview of each feature, including mean, standard deviation, minimum, maximum, and quartile values.
- **4.2. Pairplot Visualization**: Pairwise relationships between key features (FTHG, FTAG, HS, AS, HST, AST, HF, AF) are visualized using a pairplot.
- **4.3. Correlation Analysis:** A correlation matrix is computed to measure the linear relationship between features. The heatmap visualization aids in identifying strong correlations.



	FTHG	FTAG	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
FTHG	1.00	-0.21	0.70	-0.14	0.40	-0.28	0.64	-0.23	-0.08	-0.02	0.13	-0.23	-0.23	-0.04	-0.10	0.12
FTAG	-0.21	1.00	-0.08	0.72	-0.29	0.42	-0.22	0.65	-0.02	-0.01	-0.15	0.08	0.08	0.01	0.17	-0.16
HTHG	0.70	-0.08	1.00	-0.08	0.19	-0.10	0.42	-0.09	-0.05	-0.03	0.06	-0.10	-0.14	-0.05	0.01	0.05
HTAG	-0.14	0.72	-0.08	1.00	-0.17	0.30	-0.13	0.51	-0.00	0.01	-0.11	0.05	0.11	0.00	0.11	-0.11
HS	0.40	-0.29	0.19	-0.17	1.00	-0.52	0.65	-0.38	-0.07	-0.05	0.52	-0.41	-0.20	0.00	-0.14	0.16
AS	-0.28	0.42	-0.10	0.30	-0.52	1.00	-0.37	0.73	-0.04	-0.04	-0.39	0.56	0.13	-0.03	0.17	-0.20
HST	0.64	-0.22	0.42	-0.13	0.65	-0.37	1.00	-0.27	-0.04	0.02	0.30	-0.32	-0.20	-0.05	-0.11	0.17
AST	-0.23	0.65	-0.09	0.51	-0.38	0.73	-0.27	1.00	-0.03	0.01	-0.29	0.38	0.10	0.00	0.13	-0.19
HF	-0.08	-0.02	-0.05	-0.00	-0.07	-0.04	-0.04	-0.03	1.00	0.15	-0.00	-0.14	0.38	0.20	0.09	0.00
AF	-0.02	-0.01	-0.03	0.01	-0.05	-0.04	0.02	0.01	0.15	1.00	-0.09	-0.04	0.10	0.33	-0.02	0.09
HC	0.13	-0.15	0.06	-0.11	0.52	-0.39	0.30	-0.29	-0.00	-0.09	1.00	-0.31	-0.06	0.03	-0.11	0.11
AC	-0.23	0.08	-0.10	0.05	-0.41	0.56	-0.32	0.38	-0.14	-0.04	-0.31	1.00	-0.01	-0.09	0.00	-0.19
HY	-0.23	0.08	-0.14	0.11	-0.20	0.13	-0.20	0.10	0.38	0.10	-0.06	-0.01	1.00	0.25	0.05	0.00
AY	-0.04	0.01	-0.05	0.00	0.00	-0.03	-0.05	0.00	0.20	0.33	0.03	-0.09	0.25	1.00	0.04	-0.04
HR	-0.10	0.17	0.01	0.11	-0.14	0.17	-0.11	0.13	0.09	-0.02	-0.11	0.00	0.05	0.04	1.00	-0.06
AR	0.12	-0.16	0.05	-0.11	0.16	-0.20	0.17	-0.19	0.00	0.09	0.11	-0.19	0.00	-0.04	-0.06	1.00

Fig 4.3.1

5. Clustering-based Anomaly Detection

5.1. Purpose

Anomaly detection identifies unusual patterns or outliers within a dataset, potentially representing matches with statistically rare events or deviations.

5.2. Methodology

- **5.2.1 Feature Selection**: Relevant features (FTHG, FTAG, HS, AS, HST, AST, HF, AF) are chosen to reveal anomalous patterns related to match statistics.
- **5.2.2 Data Preprocessing:** Features are scaled using RobustScaler to handle outliers and ensure fair comparison across different scales. Principal Component Analysis (PCA) is applied to reduce feature space dimensionality while retaining critical information.
- **5.2.3 Anomaly Detection Model**: The EllipticEnvelope model is employed to identify outliers based on selected features.

5.3 Result Visualization

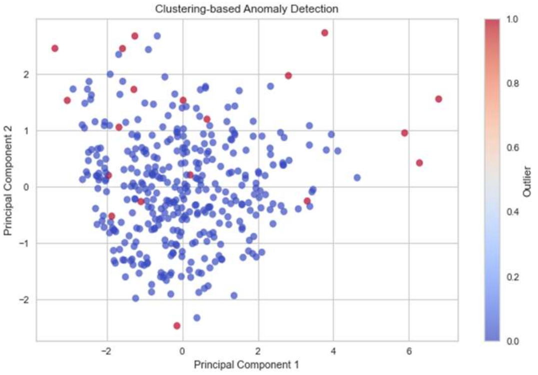


Fig 5.3.1

6. K-means Clustering

6.1 Purpose

K-means clustering partitions data into K clusters based on feature similarity, unveiling inherent groupings or clusters within the dataset.

6.2 Methodology

- **6.2.1 Feature Selection:** Relevant features (FTHG, FTAG, HS, AS, HST, AST, HF, AF) are chosen for clustering.
- **6.2.2 Data Preprocessing:** Features are scaled using RobustScaler for normalization.
- **6.2.3 K-means Clustering:** KMeans clustering is applied with the optimal K to partition data into clusters.
- **6.2.4 Determining Optimal K:** The Elbow method is used to determine the optimal number of clusters (K) for K-means clustering.

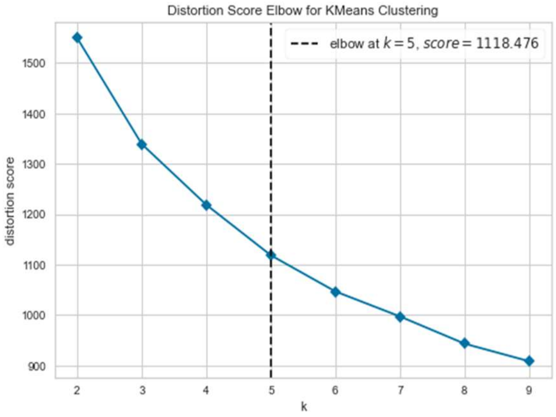


Fig 6.2.4.1

6.3 Result and Visualization

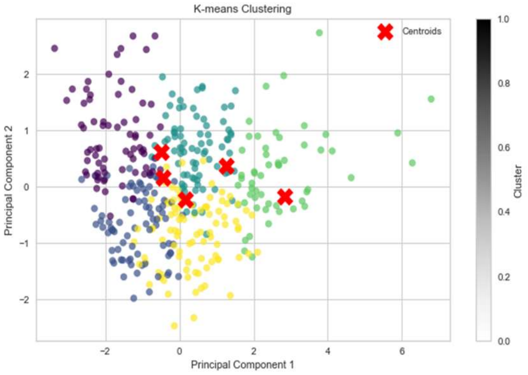


Fig 6.3.1

7. Comparison: Clustering-based Anomaly Detection vs. K-means Clustering

7.1 Feature Focus:

- Anomaly Detection**:** Identifies statistically unusual data points or outliers.
- K-means Clustering**:** Discovers inherent patterns and groupings based on feature similarity.

7.2 Output Interpretation:

- Anomaly Detection**:** Highlights rare events or anomalous match patterns.
- K-means Clustering**:** Reveals inherent groupings and patterns within the dataset.

7.3 Optimal K Determination:

- Anomaly Detection**:** Typically, doesn't require K determination as it's not based on cluster partitioning.
- K-means Clustering**:** Involves determining K using the Elbow method for optimal clustering**.**

8. Conclusion

In conclusion, Clustering-based Anomaly Detection and K-means Clustering offer valuable insights into football match data analysis. Anomaly detection identifies rare events, while K-means clustering unveils inherent groupings, aiding sports analytics, and decision-making. Each technique has unique advantages, contributing to a comprehensive understanding of team performances and match statistics.

Mohammed Nihad Kaipalli

22081746