Istanbul Technical University
Faculty of Computer and Informatics
Department of Artificial Intelligence and Data Engineering

Artificial Intelligence & Data Engineering Design Project Report

Title

Clustering Football Teams' Playing Styles
Using Multi-Step K-Means

Prepared By

150200319 Ege DEMİR

Supervisor

Associate Prof. Nazım Kemal ÜRE

June, 2024

İTÜ İSTANBUL TEKNİK ÜNİVERSİTESİ 1773
250 YIL 1773–2023

CONTENTS

SUMMARY

First of all, this project aims to divide these teams' playing styles into different clusters with unsupervised learning. The results of this research might be beneficial to teams' coaching staff, and tactical analysts. It can be utilized against a rival for a specific game by choosing the playing style that works best against the playing style of the opponents. This project is related to multiple disciplines such as football, soccer analytics, statistics, data science, and AI.

A publicly available data, which consists of more than 3 million data points from 1826 distinct matches has been utilized for this project. All matches played in England, Germany, Italy, Spain, and France's first-division football leagues in the 2016/2017 football season are in the scope of the data. Each data point in the data is a separate event in the game. The information of which event occurs where on the field by which player, is expected to provide valuable insight on the tactical characteristics of football teams. An event can be a pass, shot, duel, or run. Their frequencies and locations are the most crucial information for our model.

The methodology of this project is applying the proposed method of Multi-Step K-Means, after conducting comprehensive feature engineering steps. Recognizing the separate game phases of the game is crucial in this methodology. Creating a final dataset to apply the K-Means algorithm one more time by combining the probabilities of each team to be in each cluster, concerning the game phase it belongs, is the novel approach.

Project results consist of clustered teams, tactical patterns of respective clusters, predicted match outcomes for given teams, visualizations showing the distinct playing styles of football teams and average values for variables for each cluster, as well as an interactive user interface for coaches and football analysts to get tactical recommendations and analysis tools.

This project will be evaluated on several criteria. Clustering success, extensiveness of data preparation and visualization steps, interpretability of the clustering results, as well as the usability and performance of the user interface are among the evaluation test subjects.

As for the used technologies, Python programming language has been used throughout the project. NumPy and pandas libraries have been utilized for data manipulation and feature generation. For visualization, matplotlib and seaborn libraries have been put to use. The modeling part has been conducted using Scikit-learn. Finally, the Flask web framework has been taken advantage of for the User Interface, with the help of HTML.

The project's source code can be accessed via the link below:
https://github.com/egecjdemir/how_football_teams_play

1 INTRODUCTION

Tactical aspects of football, the most popular sport worldwide, have gained interest with the increased availability of novel data. Event data, that consist of location and time information about notable actions such as simple passes, shots, air duels, etc., have made the examination of football tactics with a statistical approach, and football analytics in general, a lot easier. Grouping the playing styles of football teams has become a popular problem in this process. Moreover, the evolution in learning from data techniques made unsupervised learning algorithms the preferred approach, rather than conventional statistical inference methods which were formerly the common solution.

As mentioned above, clustering the playing styles of football teams is the aim of this project. The significance of this project comes from its ability to help the tactical staff in football like coaches about which playing style to choose in a specific game or as a main plan for the whole football season. To achieve this goal, the evaluation of clustering results has been held by analyzing the match outcomes. This analysis revealed the effectiveness of each playing style against others. Coaches can choose the optimal playing style for their teams by exploring their rivals' playing styles and which playing style works best against them with the results of this project.

As a Machine Learning project about football, this project intersects with multiple disciplines. Sports analytics, AI, statistics, football, and data science are among the disciplines that are related to this project.

To briefly summarize the previous research on the area, it must be highlighted that while some of the earlier studies about grouping football teams' playing styles utilized statistical inference methods such as Principal Component Analysis, Chi-square, or Linear Regression; others utilized unsupervised learning techniques, mostly K-means algorithm. Event data, similar to the main source of this project, is one of the three most common data types, the others are basic match statistics, and tracking data. A literature survey about the area will be discussed in detail in the next section.

As previously mentioned, event data collected by Pappalardo and Massucco [1], which consists of more than 3 million data points from 1826 distinct matches has been utilized for this project. All matches played in England, Germany, Italy, Spain, and France's first-division football leagues in the 2016/2017 football season are in the scope of the data.

For this project, a customized multi-step K-Means clustering algorithm was used, building on previous research that primarily uses the K-Means algorithm [2]. This unique method has been developed to tackle the problem by first dividing the dataset into 4 because as Plakias et. al. stated [3], there are four distinct phases in a football game. After training each 4 subparts of the dataset with K-means where k equals 3, 4xm separate labels have been acquired for each team where m is equal to the number of matches that team has played (18 or 20 depending on the league). By using a majority voting technique, in other words by simply counting, probabilities of each cluster for a team have been obtained for each phase. In this way, a new dataset with n rows (number of teams) and 12 columns (4 datasets x 3 clusters) has been created. Each cell in the data shows the probability of a team being in that cluster for that phase in the game. Finally, training a K-means model 1 more time with the final dataset successfully divides the teams into 3 clusters.

Even though a method previously unapplied to this problem has been introduced in this project, it is not truly novel. It is simply a multi-step version of the most popular unsupervised learning algorithm. The novelty that this project aims to achieve, comes from the feature engineering steps rather than the modeling phase. The literature survey, which will be discussed in the next section, revealed that previous research lacks the extensiveness of the feature engineering that is constructed for this project. To generate new features, four distinct methods from four papers were combined. The idea of dividing the pitch into different zones from Diquigiovanni and Scarpa's paper [4], the concept of pass motifs from the work of Gyarmati et. al. [6], the approach of treating passes between football players as a graph and calculating the average connectivity score from Peña and Touchette's research [5], and the categorization of events into four different game phases—possession, out of possession, positive transition (events after stealing the ball), and negative transition (events after losing the ball) from the paper of Plakias et. al. [3]—were utilized. In addition to integrating these four ideas to create a more inclusive dataset, new features were developed by calculating the ratios of each event's different subtypes. In addition to integrating these four ideas to create a more inclusive dataset, new features were developed by calculating the ratios of each event's different subtypes. For example, passes were classified as side, backward, or forward based on pass angle information, and as short, middle distance, or long based on pass distance information. Ratios of each pass type were then calculated relative to the total number of passes after that. A similar process was applied to different event types like shots, runs, and duels. Consequently, 30 features for the in-possession dataset and 45 features for the out-of-possession dataset were acquired. This process of feature engineering, which broadens the dataset beyond previous projects, adds novelty to this project.

Moreover, this project proposes an original evaluation method for the problem by comparing the performance of each playing style against each other with respect to match results. It is expected to acquire reasonable evaluation results which can be commented on. The evaluation, which is considered a complication for all unsupervised learning projects due to the absence of labeled data, is aimed to be solved in this way. The comparison of the resulting clusters, as well as the comparison with chosen teams with the resulting clusters, are visualized by utilizing a dendrogram, scatter plot, heatmap, bar plot, stacked bar plot, and pie chart which will all be shown in the results and evaluation section of this report.

In order to provide an Interface for football coaches to use, the Flask web framework has been utilized. All the visuals mentioned above are available for coaches on the web page, as well as the win probability for each cluster against each other. Consequently, coaches can receive advice on their team's playing style by only choosing their rival team or the team they play in. They can also analyze their rivals courtesy of the visualizations showing the comparison between each cluster's average score for each event and the probability of their rivals being in each cluster as well as the closest teams to their rivals in terms of the playing style.

As the project results, the following are expected to be reached: different playing styles acquired with multi-step K-means method, characteristics of each of these playing styles analyzed with respect to features of the final datasets for different phases of the game, performance analysis of each playing style against each other using the match outcomes between the clusters, visual representation of clusters, playing characteristics, performance comparisons, and finally a User Interface to obtain recommend playing style for a specific competition or against a certain opponent, while observing the related visuals that supports the recommendation.

On the next pages of this report, the following will be discussed: the previous research on clustering the playing styles of football teams problem, descriptions of the utilized techniques such as K-means, Bisecting K-Means, and Principal Component Analysis, design constraints, and relative engineering standards concerning the problem, functional and non-functional requirements of the projects regarding the design constraints and relative engineering standards, evaluation methodology, system architecture showing the feature engineering and modeling processes, visual and numeric results of the project, and finally the conclusion text which also discusses what can be added to move this project forward.

2 BACKGROUND

Plakias et. al. [3], reviewed 40 papers about Identifying Soccer Teams' Styles of Play. Since that's exactly my problem, this paper is essential in my related work research. This survey reveals the fact that most of the research in the area uses conventional statistical methods such as Factor-PCA, Chi-square, and Linear Regression. There are also several pieces of research that use AI techniques like I will in my project. While most of them use older clustering methods like K-means, there is research where Deep Neural Networks based on Multi-Layer Perceptron and feature engineering are being used. Since I also used the K-Means algorithm and PCA, it was beneficial to explore initial papers for the methods they use, and it was beneficial to explore the latter paper for feature engineering which is an essential step in my project too. Apart from making my job easier in the literature survey, a concept from this paper made a huge impact on my project: game phases. The writers divide a single football match into 4, with respect to ball possession. When a team controls to ball, when their opponents control it, when they receive the ball recently, and when they lose it recently. Since it is visible to everyone watching the game that both teams have separate plans for these 4 phases, their playing style for each phase should be clustered separately. This idea that I inherited from this paper, resulted in the methodology of my project, multi-step K-means.

According to the same paper, all the research in the area focuses on at least 1 of these 3 targets: game style recognition, contextual variables, and game style effectiveness. First of all, recognition, which is the most common research target, only focuses on distinguishing different playing styles, and in some of these papers, identifying the characteristics of those playing styles. This target area will also be the main focus of my project. Secondly, contextual variables, focus on match details like the location of the game, rankings of the teams playing, and competition of the match, instead of event data like I use. These projects aim to reveal relations such as "Home teams are more likely to build up from the back.". Contextual variables are out of scope in my project. Lastly, the least researched area, effectiveness, compares how effective are these playing styles. Different papers use various metrics for comparison. To compare the effectiveness of the playing styles, I investigated the results between the games of each cluster of teams to figure out which playing style works better against which ones. According to both my research and the papers examined in this survey, this comparison has never been done before.

Diquigiovanni and Scarpa [4], form networks for teams using pass, dribble, tackle, and shot locations. Then it hierarchically clusters teams by comparing the similarity between the networks. It was beneficial to study this project because it inspired me to divide the pitch into various areas and represent the team as a graph. I used graph representation for only pass events while calculating the average connectivity score of the team. Even though I found their event dataset a lot more limited than mine overall, it influenced my project with two different ideas.

Similar to the paper mentioned above, Peña and Touchette [5] discuss the network representation of football teams. Unlike the other mentioned papers, it is not directly about clustering football teams' playing styles. They analyze the football matches from the 2010 World Cup by utilizing graph theory, primarily focusing on player performances. It was influential on my work for the same reason I mentioned above, average connectivity calculation of the teams' pass graphs.

In another research on the Spanish League, Gyarmati et. al. [6], cluster tactics with event data while focusing on only passes. Their algorithm examines the pass sequences, which they name "pass motifs", to search for teams' passing styles. After that, they use hierarchical clustering to analyze the similarities and differences of teams' passing styles. I directly used their pass motifs idea as a part of my feature generation process. It was quite effective in clustering results in terms of feature importance. For that reason, I can safely say that this work is one of the most important references in this report in a way that it is among the few papers that have a direct influence on my project. As I declared in the previous section, I have managed the take their work forward by combining it with a far more inclusive list of features as well as using various types of events like shots, duels, runs, etc. instead of using only pass data.

From now on, the papers that will be mentioned will not directly influence my project. However, they must be discussed in order to display the state-of-the-art solutions to clustering football teams' playing style problems. Lopez-Valenciano et. al. [7], explores the relationship between the variables they created which represent the attacking and defensive playing styles of teams, and the rankings of football teams in the Spanish League. They use Principal Component Analysis (PCA) for this task rather than ML methods, and they are not trying to cluster or classify teams. The similarity between the problem in this research and mine is representing playing style with feature engineering. Even though I generated many more features than them, it's still important to see their approach too. Their dataset is also limited to only 1 league, and conventional statistics of games such as the number of shots, number of recoveries, and ball possession percentage. This data is a lot more available and a lot less significant than the one I use, event data. I also utilized PCA, but it was a part of my data preprocessing steps, with the goal of avoiding the curse of dimensionality, since the number of columns is huge in my data. Besides that, PCA was also needed for visualizing the clustering results by squeezing the data into only 2 features.

In an older paper from 1988, Pollard et. al. [8], try to examine playing styles with only 6 variables. They use Principal Component Analysis (PCA) like most of the projects in this area. Their data handling and statistical inference methodologies are not significant, to be honest, however, this research shows the continuing interest of researchers in grouping playing styles.

Bialkowski et. al. [9], utilize spatiotemporal tracking data, which is more detailed and less available than event data. They examine the formations of football teams which shows the distributions of football players. They discover that formation is an essential attribute to reveal the playing style. Since I do not have access to tracking data, I had to find a way to represent the distribution of football players. I used the coordinates of passes for this. Even though it is not possible to represent the locations of the football players as comprehensively as tracking data, the information of which zones are more occupied, which is generated from the event data, still benefits the clustering process.

Ruan et. al. [10], use a similar approach to Pollard et. al. [8], by using variables from simple match statistics and applying Principal Component Analysis (PCA) to them, in order to analyze defensive playing styles. They take the previous research forward by applying also regression models with the aim of measuring the effectiveness of the defensive playing styles they identified with PCA. They measure the effectiveness with a popular statistic called Expected Goals (xG). With a different approach, by comparing the match outcomes in the games that each cluster played against each other, I measured the relative effectiveness of playing styles against each other.

Now that previous research on the problem has been discussed, the methods that have been used must also be addressed, starting with the main algorithm behind the project. K-Means is a popular machine learning technique for dividing a dataset into K unique, non-overlapping clusters. The approach was first introduced in 1967 by MacQueen [2]. It works by first randomly allocating a number of centroids equal to K, which stands for the initial centers of clusters. Each data point is allocated to the closest centroid, and the centroid of each cluster is recalculated using the average of the points assigned to it. Until the centroids stabilize, this assignment and recalculation process is repeated iteratively. Convergence suggests that the clusters are heterogeneous amongst themselves and rather homogeneous inside. K-means is very useful for cluster analysis in a variety of application domains, such as computer vision and market segmentation, due to its straightforwardness and effectiveness. Nevertheless, the approach has drawbacks as well, such as its sensitivity to the original centroid positions and its inability to handle non-spherical clusters or outliers. K-means is still an essential technique in the field of unsupervised learning, regardless of these challenges.

A variation of the classic K-means algorithm known as Bisecting K-means uses a hierarchical clustering method in place of the more common agglomerative one. As Di and Gou explained [11], Bisecting K-means starts with all points in a single cluster rather than a fixed number of clusters and splits the clusters iteratively until the desired number of clusters is obtained. The algorithm chooses a cluster to divide in each iteration according to a predetermined criterion, either the cluster's size or a measure of cluster variation. The fundamental K-means method is then used to partition the selected cluster into two, and this procedure is repeated. For several types of data distributions, the Bisecting K-means method is well-known for generating more balanced cluster sizes and frequently results in superior clustering quality. Even though the regular K-Means method has been used for simplicity in my project, the same datasets have also been trained with Bisecting K-Means, with the purpose of visualizing the clustering hierarchically with dendrograms.

A statistical method for dimensionality reduction that keeps as much of the data's variance as possible is principal component analysis, or PCA, first presented by Karl Pearson in 1901 [12], which divides a group of potentially related variables into a smaller number of uncorrelated variables known as principal components. These components, which aid in determining the directions of maximum variance in high-dimensional data, are obtained by computing the eigenvalues and eigenvectors of the covariance matrix of the data. With each subsequent component having the maximum variance allowed by the requirement that it be orthogonal to the preceding components, the first principal component provides the maximum amount of variance. This technique is especially helpful for processing and visualizing genetic data, as well as for improving the interpretability of predictive models, which is why it is being used in this project. Since there are many columns in my data, PCA is needed to overcome the curse of dimensionality. Moreover, it is essential in the visualization of the clustering results by squeezing the data to only 2 variables, making it possible to display the data in two-dimensional space.

When clustering datasets without prior knowledge of the group allocations, a common difficulty is figuring out the ideal number of clusters for the K-means clustering algorithm. This is where the Elbow Method comes in. Plotting the sum of squared distances from each point to its designated cluster center (SSE) versus the number of clusters is how the Elbow Method is implemented, as described by Humaira and Rasyidah [13]. When the number of clusters grows, this plot usually exhibits a rapid drop in SSE, which is followed by a decrease gradually or "elbow" where more clusters result in lesser SSE reductions. Then, at this "elbow" point—where incorporating further clusters does not yield appreciable increases in variance explained—the ideal number of clusters is chosen, thereby striking a balance between a simple model and effective clustering performance. This approach is used because of its empirical efficacy and visual simplicity, particularly in scenarios involving exploratory data analysis where computational efficiency is crucial.

Developed in 1987 by Peter J. Rousseeuw, the silhouette score is a powerful graphical tool for evaluating the integrity of clustering results. As explained by Rousseeuw [14], the silhouette score computes an object's similarity to its own cluster in relation to other clusters, hence aiding in the interpretation and validation of cluster analysis. For researchers and analysts engaged in cluster analysis, Rousseeuw's method highlights the significance of the silhouette score in assessing the overall clustering arrangement as well as in estimating the coherence inside a cluster.

A preprocessing method called standard scaling, also known as Z-score normalization, centers a dataset's characteristics such that its mean is zero and its standard deviation is one. This scaling technique is especially useful when a dataset's features have disparate units or vastly varying magnitudes. This method entails subtracting the mean for each data point, and then dividing the result by the standard deviation, as explained by Ali and Haraj [15]. As a result, the number of standard deviations that separate each feature's initial value from the mean is displayed. This transformation standardizes the range of independent variables, which accelerates the convergence of gradient descent-based algorithms. It is essential for algorithms that assume data is normally distributed. In situations involving principal component analysis, regression models, and support vector machines—where the input feature size has a substantial impact on the model's performance—standard scaling is frequently used.

Since logistic regression can convert the results of K-Means clustering into useful information about the tactical traits of football teams, it is selected as a key methodological technique in this study. Although helpful for clustering, K-Means does not reveal how each feature affects the cluster assignment. On the other hand, logistic regression provides a means of directly quantifying the impact of particular features by modeling the likelihood of cluster assignments using a logit function. The significance of these features is effectively indicated by the logistic regression coefficients, which represent the log odds of belonging to a cluster for each unit increase in a feature. According to Peng et al.[16], the logistic model is especially beneficial because it does not require the independent variables to be regularly distributed, linearly connected, or have equal variance within each group. Due to its ability to handle binary outcomes—like the cluster labels obtained from K-Means—it is highly ideal for categorical or binary variable analysis, such as the analysis of pass/fail or win/lose scenarios. Using logistic regression, one may measure the relative contributions of each tactical feature to the likelihood of a team's clustering outcome by considering the cluster labels as a dependent variable. As a result, the sign and magnitude of the logistic regression coefficients indicate which features have the greatest bearing on team strategy, which facilitates decision-making by emphasizing crucial tactical information. Logistic regression proves to be an invaluable tool in this attempt due to its capacity to produce straightforward measurements, or coefficients, for feature importance. It provides an in-depth understanding of football team tactics by guaranteeing not only the prediction of results based on specified qualities but also the evaluation of which variables are most important in affecting these outcomes.

In summary, there have been multiple researches with the purpose of grouping football teams' playing styles. While most of them only work on identifying playing styles, some researchers take the next step to discover the effectiveness of playing styles. Even though most of the papers used statistical inference methods like PCA, there is also a significant amount of research that uses unsupervised learning methods like my project. K-means is the most popular choice, followed by hierarchical clustering. While most of the projects utilize event data like my project or simple match statistics, some of them make use of tracking data which is the hardest one to get access to, and therefore, out of my project's scope.

In order to move the current state of the art forward, my project carries out a detailed feature engineering process, involving some of the previous papers' methods as well as original ones. Since I saw an improvable area with the used features in the previous research, I believe the original features I generated and merging them with previous work, made a difference. Moreover, using a multi-step version of the K-means algorithm instead of directly using it, as most of the previous research has done, was a step forward in acknowledging the different phases of the game. Clustering the playing styles for each game phase successfully captured the fact that teams have different tactical schemas for different phases. Finally, examining the effectiveness of each playing style against each other with the investigation of match outcomes, also contributed to the uniqueness of my project.

3 SYSTEM REQUIREMENTS

3.1 Design Constraints and Relevant Engineering Standards

Design constraints:

1. Extensibility and Reproducibility
   If incoming event data becomes accessible, the system shall be available to benefit from the new data. The system be able to accommodate the inclusion of additional league teams or match data in the future. Moreover, the same results will be shown in the upcoming sections of this report and must be able to be reached by simply running the provided code.

2. Cost Efficiency
   The system must minimize the usage of the computational resources. The algorithm should be carefully implemented to work well with limited hardware.

3. Interpretability and Usability
   As the project outcome, clearly, visualized and interpretable results must be reached. As well as a user interface can be used by football coaches to get tactical advice.

Relevant Engineering Standards:

1. ISO/IEC 25012:2008 Data Quality Assurance Criteria:
   Data quality in this project is critically dependent on compliance with ISO/IEC 25012:2008 [17]. Accuracy and broadness are crucial, so an in-depth analysis of the event dataset is required to look for any inconsistencies, errors, or missing data. Correlating position and timestamp information with recorded events such as passes, shots, and duels, as well as verifying the correctness of the data are all part of this process.

2. ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML):
   For the AI and machine learning components of the project, compliance with ISO/IEC 23053:2022 [18] is essential. The large-scale datasets must be handled by an efficient and effective unsupervised learning algorithm in order to cluster team playing styles. This involves developing algorithms that can handle enormous volumes of data and add new data sources as they become available. This flexibility is essential to the project's sustainability over time because it allows the inclusion of teams from leagues outside of its original scope and takes evolving team tactics into account. In order to provide coaching staff and tactical analysts with precise, useful insights, the AI model must be powerful as well as adaptable to these criteria.

3.2 Functional Requirements

1. The system shall create new features from existing features.
2. The system shall group the data by teams.
3. The system shall be extensible with the introduction of new data, by seamlessly building on the model with recently added data.
4. The system shall be reproducible by simply running the provided code on a different computer.
5. The system shall explore the feature importance and average values for the features with respect to found clusters, in order to discover the characteristics of playing styles.
6. The system shall provide visualized results.
7. The system shall provide an interface to choose the competition or team.
8. The system shall provide an interface to return the optimal playing style for the given team.

3.3 Non-Functional Requirements

1. Clustering results must be visualized in a clearly understandable way.
2. All classification scores for accuracy, precision, recall, and f1 score metrics, must be higher than 0.9 on the test set.
3. When the user interacts with the interface by providing a team or competition name, the recommended playing style a without a significant waiting time, 2 seconds at most.
4. At least 3 types of plots shall be created to visualize different aspects of clustering results.
5. When the user interacts with the interface by choosing the requested plot type and filtering teams or competitions, visualizations shall be shown without a significant waiting time, 5 seconds at most.
6. Silhouette score for all 5 clustering results, shall be greater than 0.25.
7. The interface shall be as simple as possible for accessibility purposes.

3.4 Evaluation Methodology

Evaluation is considered an issue for unsupervised learning, due to the absence of labeled data. Unlike supervised learning, the success of the algorithm can not be measured by the difference between the predicted labels and the actual labels, because the actual labels do not exist. Nevertheless, evaluating the success of an unsupervised learning task is still possible by examining the clustering results, as well as using techniques that compare Intra-cluster and Inter-cluster distances, such as silhouette scores. Furthermore, the project must ensure all functional and non-functional requirements listed above. These requirements contain criteria about the clustering success, extensiveness of data preparation and visualization steps, interpretability of the clustering results, along with usability and the computational performance of the user interface.

3.4.1 Evaluating Clustering Success

First of all, to validate the clustering success, clustering results must be checked to make sure that the multi-step K-Means model has successfully grouped the data. If there is only one cluster even though a predetermined number of clusters is greater than one, it means the data is not suitable for the K-Means algorithm, and possibly, it can not be clustered. Moreover, when the overall outcomes of the matches between different clusters are examined, if the win, draw, and lose percentages do not significantly differ from 1/3, it might reveal the failure in clustering. However, if they considerably differ from 1/3, which is the outcome of randomly chosen matches, it means that the multi-step K-Means method successfully grouped the data, and its results might reveal notable information about different playing styles of football teams. Furthermore, it must be checked if the silhouette scores of all five clustering tasks are greater than the chosen threshold, 0.25, indicating a satisfying clustering.

3.4.2 Evaluating the Extensiveness of Data Preparation and Visualization

Secondly, the extensiveness of data preparation and visualization steps should be evaluated. The new datasets grouped by teams, must include information about all the major event types available on the original data, such as pass, shot, acceleration, and duel. Moreover, there must be many newly generated features that are non-trivial, in other words, they must contain information which are not directly understandable from the original variables. Preferred pass motifs, more used areas in the football pitch, and average connectivity of the pass graph are some examples of newly generated non-trivial features. The created visuals should also be comprehensive enough. As the non-functional requirements dictated, there must be at least 3 separate types of plots to visualize the clustering results and analyze the distinct playing styles.

3.4.3 Evaluating Clustering Results' Interpretability

Third of all, the visuals created should provide interpretable information about the playing styles of football teams. Comparing the average value of a feature for each cluster should reveal patterns about the characteristics of found playing styles. If a pattern such as "The teams that prefer ABCD pass motif more also tend to prefer high passes rather than low passes.", it means that the aim of the project is reached. In the results and evaluation section of this report, the found patterns will be discussed and visualized. Furthermore, the success of the project will be evaluated based on the quality and quantity of the found patterns.

3.4.4 Evaluating the User Interface

Finally, the usability and the computational performance of the user interface will be evaluated. In the user interface, coaches will be asked to choose a team or a competition. When a team is chosen, the interface will offer the user 4 pages to view. In the first one, the system recommends the best playing style against the chosen team. This process must not exceed 2 seconds. On another page, the interface will generate a dendrogram plot to show the hierarchical clustering result, and a scatter plot to show the generated clusters in two-dimensional space, utilizing PCA. In both cases, the chosen team’s name will be highlighted. On the third page, the system will generate bar plots comparing the average values for each cluster alongside the average value for the chosen team, for chosen variables. In the fourth option, the interface generates a stacked bar plot, showing the probability for every team to be in each cluster. If a competition is chosen instead of a team on the start page, the dendrogram and scatter plot highlight all the teams playing in the chosen competition. In bar plots, the average value for that league is shown similarly to the team pick. Generating any one of the plots mentioned above, must not exceed the designated threshold, of 5 seconds. If the interface works flawlessly, it will pass the usability test. Additionally, an average loading time lower than 5 seconds for every plot will mark the interface as passed the performance test.

3.4.5 Conclusion

In conclusion, by concentrating on clustering approaches in football team data, this project successfully addressed the difficulties of evaluating unsupervised learning without direct label comparisons. The study showed that it could correctly cluster the data by using the silhouette score and a multi-step K-Means model, as seen by the notable differences in match results between clusters. The distinct football playing styles that emerged from the clusters were much easier to understand thanks to the intensive work that went into processing the data and creating detailed visuals. Moreover, a user interface was designed to be fast and functional, catering to users who need results quickly without sacrificing accuracy or detail. The project will be evaluated using these criteria, which are clustering success, extensiveness of data preparation and visualization steps, interpretability of the clustering results, usability, and the computational performance of the user interface, in the report's results and evaluation section. This project guarantees that the results are understandable and helpful for realistic football decision-making, in addition to demonstrating the application of advanced unsupervised learning methods in sports analytics. All things considered, it has effectively demonstrated how data-driven methodologies may reveal insights into team tactics, providing coaches and analysts with a useful tool.

4 SYSTEM ARCHITECTURE

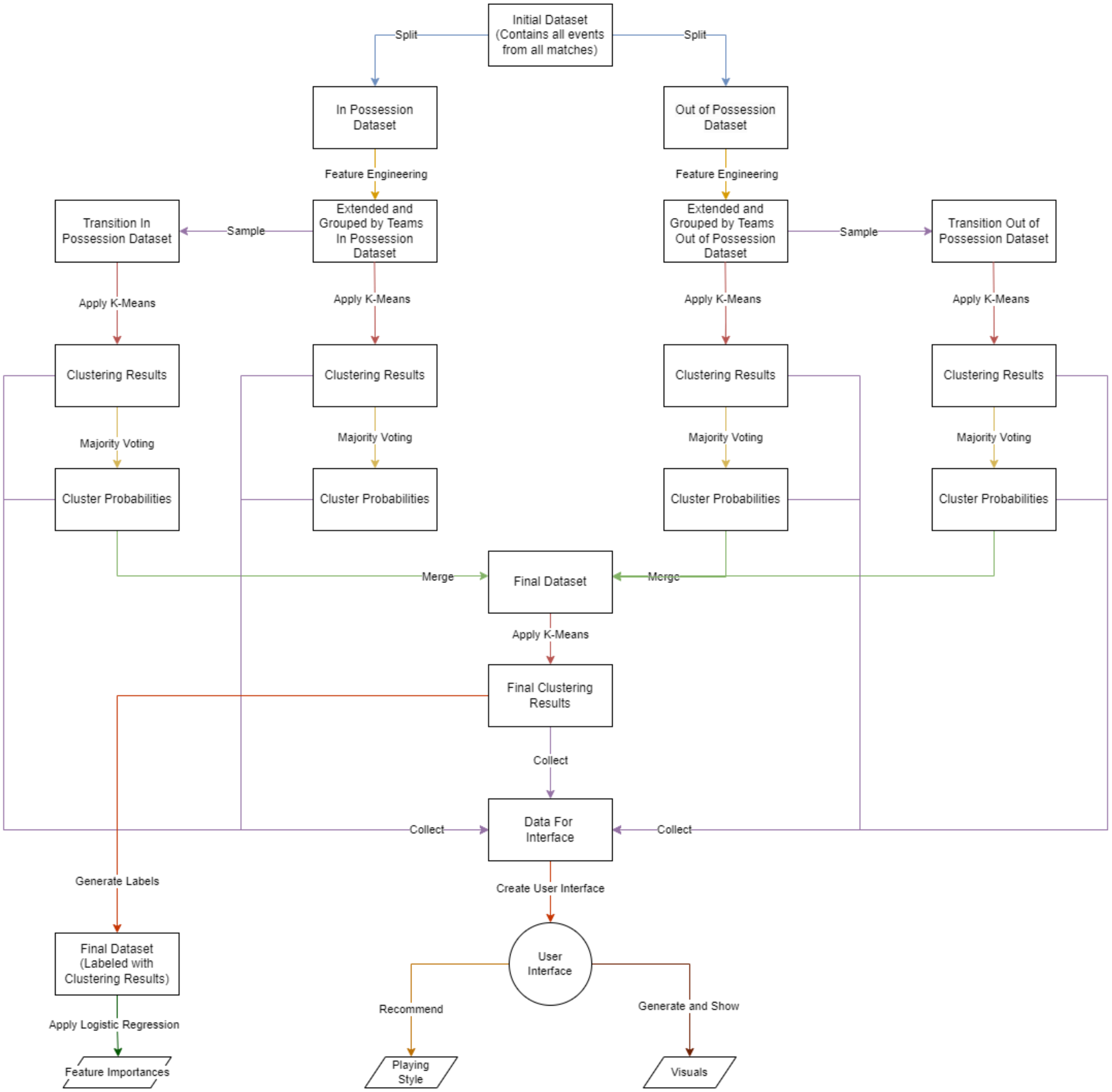

Figure 4-1: Diagram of the System Architecture

4.1 Splitting the Initial Dataset

The public event data that Massucco and Pappalardo [1] had first gathered was kept in a single table. It included every event from every game played in the first-division football leagues of France, England, Germany, Italy, Spain, and France in the 2016–17 football season. This dataset consists of over 3 million data points from 1826 different matches. The system architecture begins by dividing this initial dataset into two according to ball possession information, as seen in Figure 4-1. The first division of the data, named the in-possession dataset, contains events that require holding the ball, such as passes and shots, whereas the second data division contains events that require the absence of ball possession such as duels and runs. They are being used to capture the teams' attacking and defensive tactical schemas respectively.

4.2 Feature Engineering

As mentioned above, both In Possession and Out of Possession datasets are based on events, where each entry is a specific incident that took place in a certain football match. The data to be used in the K-Means model must be team-centric since clustering their playing styles is the ultimate goal of this project. Simply grouping by teams, results in a dataset only containing 98 rows, which is the number of teams in the data. It can not be expected for the model to find comprehensive patterns with that small amount of data entries. Treating each game of a team as separate teams is the preferred way to overcome this issue. For both the attacking and defensive phases of the game, a dataset with 3652 rows is acquired in this manner.

To generate extensive features in the team-grouped data, some ideas inherited from related studies are used, along with original ones. They will be discussed in this section.

4.2.1 Feature Generation Ideas Inherited From Related Studies

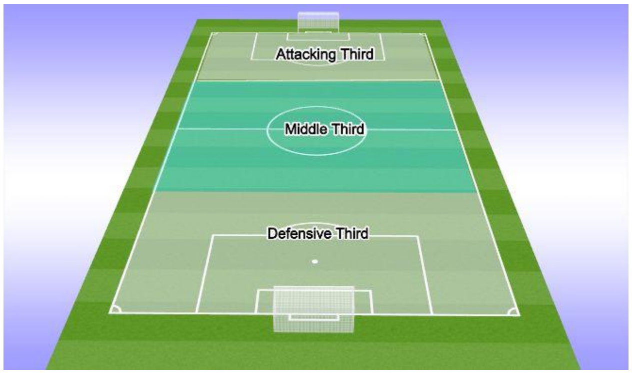

Figure 4-2: Three Zones in Football

First of all, the football pitch has been partitioned into 3 equal zones, as shown in Figure 4-2, a visual taken from the Sport Session Planner web page [19]. Since a pass between strikers in front of the opponent's goal does not indicate a similar action to pass from goalkeeper to left-back, they have to be treated differently. On another note, it must be declared that this idea is not original. Diquigiovanni and Scarpa [4], in their work on clustering football teams' playing style problem, divide the pitch with 3 horizontal and 3 vertical lines. They influenced me to do the same.
However, the number of zones in my project differs from theirs. I only use 3 zones instead of 9, to represent the most important information, distance to the goal, while also avoiding too many features, thus the curse of dimensionality.

Secondly, by investigating each sequence of 4 passes, and labeling them as “ABAB”, “ABCA”, “ABCB”, and “ABCD”, Gyarmati et. al. [6], examine the preferred “pass motifs” of football teams. “ABAB” means 4 passes between 2 unique players while “ABCD” means 4 passes between 4 unique players. The writers argue that by comparing the pass motif frequencies, distinct playing styles can be acquired, and comments like “The teams prefer ABAB pass motif aim to keep the ball as long as possible, while the teams prefer ABCD motif chooses a more direct approach.” Can be made. Every 4-pass sequence has been labeled and their total frequencies have been counted for each team in my project, benefitting this method from previous studies.

Third of all, Peña and Touchette's research [5], inspired me to make use of graph representation. By treating players as nodes, and passes as edges in the graph, it is possible to calculate the average connectivity score of the pass network. The lesser connectivity score indicates a dependence on few players in the build-up phase, while the higher connectivity score reveals the lack of pass centers and, therefore more flexible pass schemas. It can be said that the latter team can carry the ball to desired areas more easily.

Finally, the categorization of events into four different game phases—possession, out-of-possession, positive transition (events after stealing the ball), and negative transition (events after losing the ball)— idea comes from the research of Plakias et. al. [3]. As previously mentioned, the whole data manipulation flow starts with the data split. Since positive transition data is a subset of in-possession data, and negative transition data is a subset of out-of-possession data, right after the feature generation steps, transition datasets can be accessed by simply sampling the existing datasets. This process makes the representation of distinct game phases, along with searching for separate game plans for each game phase possible. This concept inherited from Plakias et. al. [3], occupies a central place for this project and it is the main reason behind developing the presented method, multi-step K-Means.

4.2.2 Original Feature Generation Steps

As mentioned before, the possession dataset contains information about pass and shot events, and the pass information is much more detailed, since pass is by far the most common event in football. Using the location info, the distance and the angle between the initial and final position of the ball are calculated for each pass. Based on pass angle information, passes were categorized as side, backward, or forward; depending on pass distance information, passes were categorized as short, middle distance, or long. After that, the ratios of each pass type were determined in relation to the overall number of passes. Moreover, in the original data, passes are labeled as simple pass, high pass, launch, cross, smart pass, etc. Since there are subevents with very low frequencies, and they are all subtypes of high pass, except for simple pass and smart pass, they should be treated as a high pass for generalizability purposes. A simple pass is essentially a low pass, and a smart pass can be both, so the ratios of high, low, and smart passes with respect to the total number of passes, are added as new features. Furthermore, the frequency of all these pass types is also calculated for three distinct zones. Shot data, on the other hand, lacks the detail level of pass data. Their locations are the only usable info. Therefore, they are classified as short, medium-distance, and long shots, according to their distance from the opponent’s goal. All these steps, along with the features inherited from previous papers, create a dataset with 30 features.

Moreover, defense data contains events such as duels, fouls, runs, clearances, touches, goalkeeper-related actions, etc. Similar to the passes, the angle and distance of the runs are also measured, and runs are classified as backward, forward, and side runs with respect to angle information. The average length of runs is also calculated for each team. Duels, which are the second most common event, involve details about the context of the duel, like whether it happened on air or ground, whether it was for defensive or attacking purposes whether it was a loose ball challenge. All of these subevents' total numbers are counted, along with their total number for each zone. Subtypes for fouls are counted similarly, resulting in 45 features in total.

4.3 Multi-Step K-Means Method

After grouping the data by teams while treating each match of every team as a distinct team, and generating new features as discussed in the previous chapter, datasets are ready for modeling. Starting with preprocessing steps like data scaling and PCA, the Multi-Step K-Means Method clusters the playing styles of football teams with the steps described below.

4.3.1 Data Scaling and Dimensionality Reduction

First, the Standard Scaler method is used to normalize every feature in the dataset. This procedure ensures that no feature has an overly significant influence on the clustering result by normalizing the features to have a zero mean and unit variance. Once the data has been normalized, Principal Component Analysis (PCA) is used to minimize the complexity of the data by transforming the original variables into a new set of variables. These variables capture the underlying variation in the data and allow for more accurate grouping because they are merely linear combinations of the originals.

4.3.2 Determining Optimal Clusters

To determine the ideal number of clusters for the K-means algorithm, the Elbow Method is utilized. Using this method, the total squared distances between each location and the designated cluster center are computed, and the values are plotted compared to the number of clusters. In order to balance cluster compactness and model complexity, the ideal number of clusters is indicated by the 'elbow' point, which is the point at which the drop in the sum of squared distances becomes less noticeable.

4.3.3 K-Means Clustering on Four Subsets

The K-means algorithm is applied separately to each of the four subsets that comprise the dataset, each of which represents a distinct stage of the game, as introduced before. Through segmentation, the teams' strategy during different phases of the game can be reflected in the clusters, offering comprehensive insights into tactical adjustments made throughout the encounter. By capturing the various tactics used at various stages, clustering within each subgroup seeks to provide detailed insights into gameplay.

4.3.4 Majority Voting

Following the clustering process, each team's most common cluster assignment for all of its matches is determined by a majority vote mechanism, which combines several cluster assignments into a single representative cluster for each team at the conclusion of each game phase. The possibility that a team will employ particular tactics in each phase is quantified by calculating the likelihood of each team being assigned to each cluster.

4.3.5 Creating the Final Dataset

The cluster probabilities for every team during the four-game phases are combined to create the final dataset. This gives each team a thorough profile that captures their main playing styles. This combined format streamlines the examination while maintaining crucial tactical actions demonstrated throughout several phases.

4.3.6 Final Application of K-Means

Scaling and PCA are applied once more to the final dataset before conducting another round of K-means clustering. This final clustering splits teams into groups based on the likelihood of being in each cluster in various game phases in an effort to blend diverse information into more comprehensive strategic insights and make the clusters more interpretable and actionable.

4.3.7 Extraction of Final Clustering Results

The final clustering results include the cluster allocations for each team as well as a detailed analysis of the statistical profiles and performance metrics for each cluster. These results show the win percentages of each cluster relative to the others, indicating the relative effectiveness of different techniques. The characteristics and tactics shared by the teams in each strategic group are revealed by computing the average values of each variable for the teams in each cluster. This comprehensive perspective of the clustering data provides useful information for sports analytics tactical planning and advances our understanding of team strategies and performance.

4.4 Supervised Learning

Analyzing the importance of features is crucial for this project’s result since the primary goal is to understand the tactical characteristics of teams. Football teams’ play tendencies can be revealed by examining which feature affects the clustering model more. Since the K-Means model does not expose the weights of the variables, another way to check the feature importance is needed. This project proposes utilizing a supervised learning algorithm, after labeling the data with the clustering outcomes. If accuracy, precision, recall, and f1 score metrics are all satisfying with the designated test set, variable weights of the classification algorithm can represent the importance scores of the K-Means method’s features. After experimenting with several classification algorithms, Logistic Regression is chosen as the final model, because for all metrics mentioned above, the Logistic Regression model provided extremely high scores, and it is the model that can supply feature importance scores in the simplest way, via coefficient weights.

4.5 User Interface

The user interface created for this project takes advantage of the thorough clustering that was obtained at the end of the project, as well as the detailed clustering data received from the K-means analysis conducted during each of the four-game phases. In addition to the cluster to which each team is assigned, the clustering findings also include the win percentages of each cluster relative to the other clusters and the average values of each variable for both individual teams and clusters.

When using the interface, users can choose a league or an opponent team to examine. This choice is crucial since it enhances the system's strategic planning skills by adjusting the data and visualizations that follow the user's specific needs.

4.5.1 Team Analysis

4.5.1.1 Playing Style Recommendation

The first page offers suggestions for the best way to play against the team that has been chosen. Using the accumulated clustering data, a comparative examination of how various clusters have performed against it forms the basis of this recommendation.

4.5.1.2 Cluster Visualization

The results of the hierarchical clustering are displayed as a dendrogram. This graphic aids in illustrating how similar the teams' playing styles are. Furthermore, the clusters are shown in two dimensions using a scatter plot that reduces dimensionality using PCA. The selected team is highlighted in both plots, making it simple to find and examine in relation to the cluster distribution as a whole.

4.5.1.3 Comparison of Cluster-Based and Team-Based Statistics

Bar charts comparing the average values for each cluster and the average values for the chosen team across the selected variables are displayed on the third page. This study gives a clear strategic picture by highlighting both the advantages and shortcomings of the chosen team as well as all the clusters in several areas.

4.5.1.4 Cluster Probabilities

A stacked bar plot showing the probability distribution of each team among the various clusters is displayed on the last page. When evaluating the league's or tournament's overall strategic environment, this map is especially helpful.

4.5.2 League Analysis

If a league is selected rather than a single team, all of the teams taking part in that competition will be highlighted in the dendrogram and scatter plot. A comparative examination at the level of competition is made possible by this wider perspective. Likewise, the league average values will be shown in the bar plots, along with cluster average values. Making it possible to analyze the overall playing characteristics of the chosen league.

In order to enable deeper insights into how different playing styles interact and compete against one another in various circumstances, the interface design makes sure that users have a comprehensive tool to investigate football tactics. The interface helps coaches and analysts make well-informed decisions for individual opponents or leagues through interactive and graphically rich examinations.

5 RESULTS AND EVALUATION

5.1 Results

5.1.1 Elbow Method

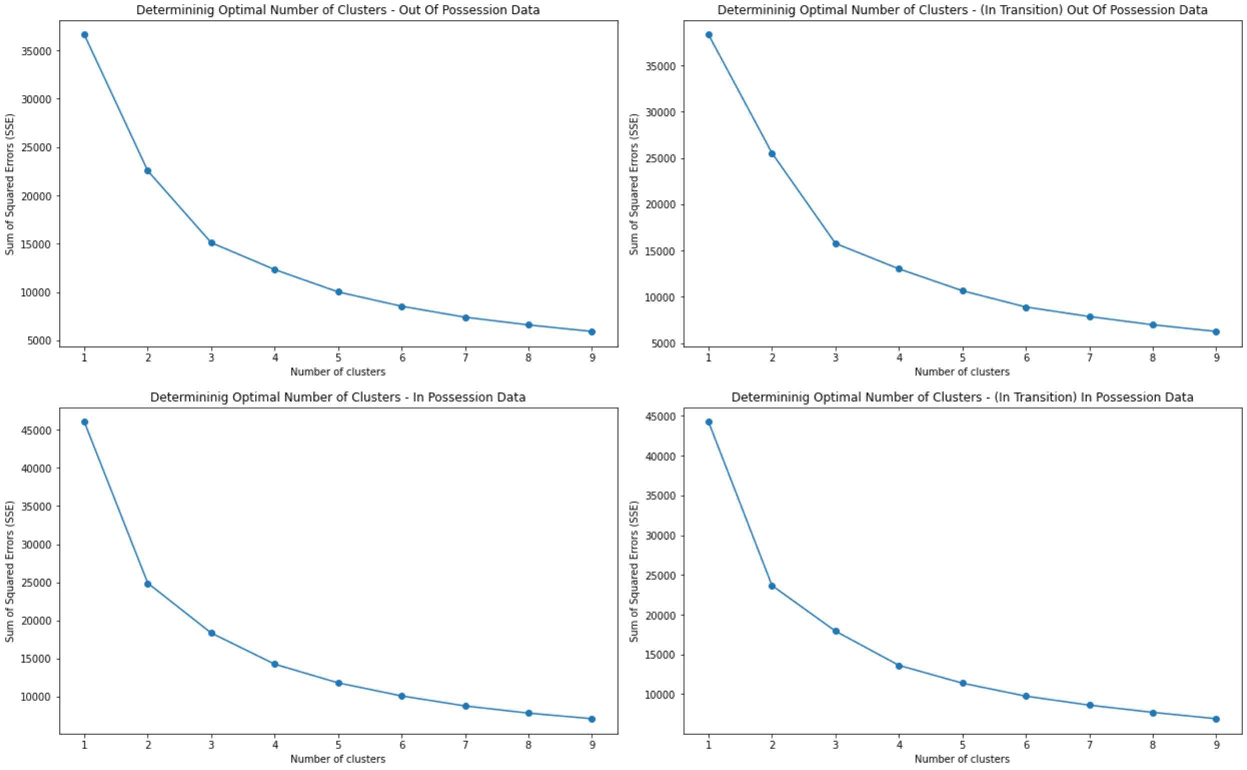

Figure 5-1: Deciding the Number of Clusters with Elbow Method

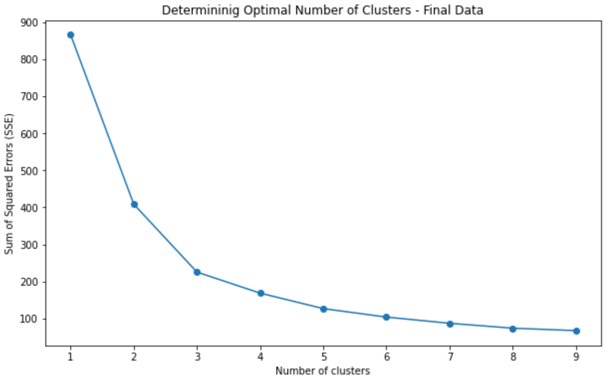

Figure 5-2: Elbow Method on the Final Dataset

As stated several times in previous sections, when the sum of squared error seems flattened, after a huge decrease, that point is named as the "elbow" point, and the number of clusters that cause it must be chosen. According to this rule, the optimal number of clusters for both out-of-possession datasets should be 3, as shown at the top of Figure 5-1. However, deciding the elbow point is not that simple for the in-possession datasets. 2 or 3 can both be chosen as k value and even though 2 looks more like an elbow point, 3 is chosen to decide a uniform number of clusters as well as avoid an unnecessarily simple model.

Similar to out-of-possession datasets, the elbow method conducted on the final data, which is constructed by the probability of each team being in every cluster in the previous models, reveals the optimal number of clusters as 3. Thus, the teams will be separated into 3 groups, as seen in Figure 5-2.

5.1.2 K-Means Clustering Results

5.1.2.1 Dendrogram

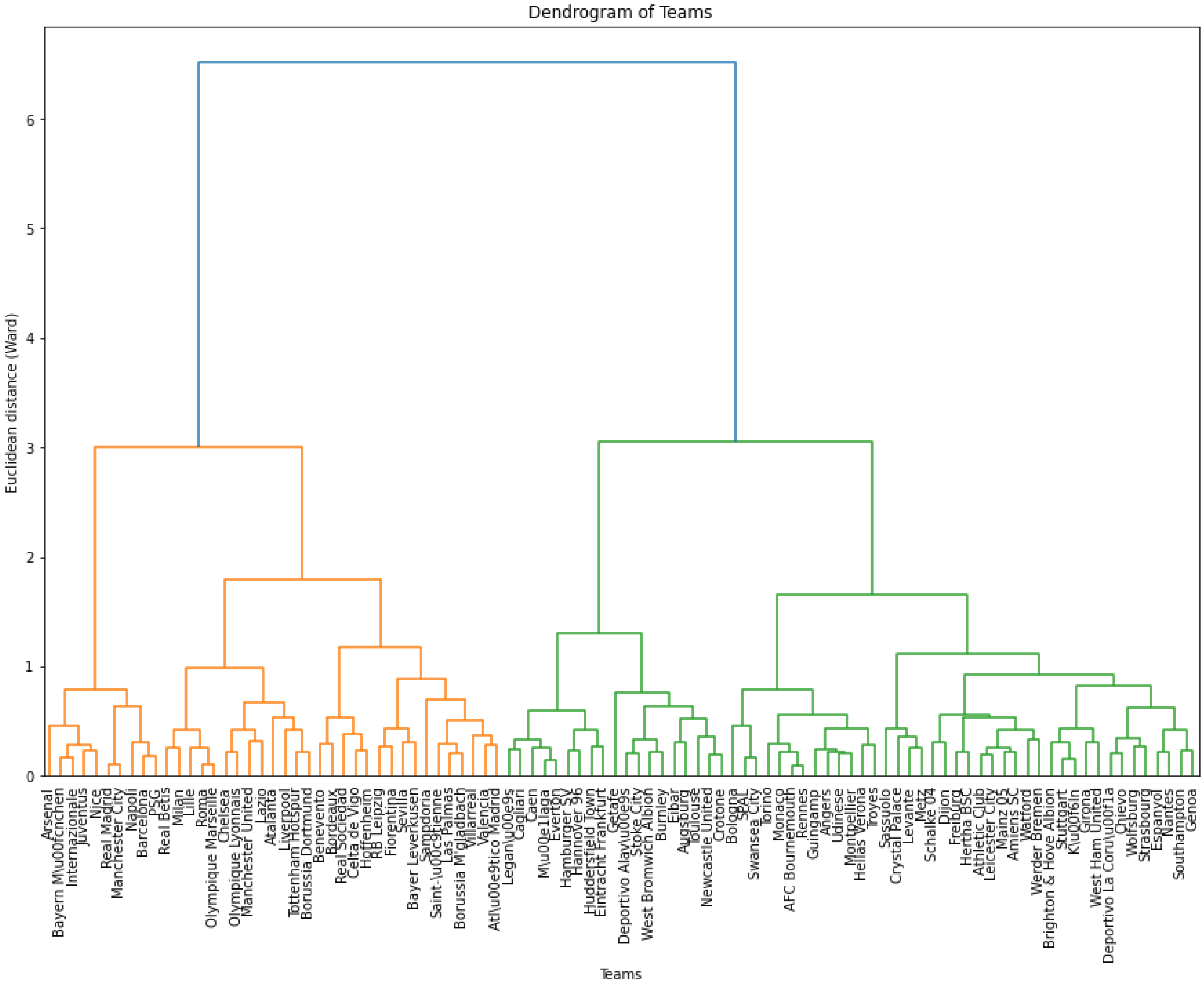

Figure 5-3: Dendrogram showing the result of Hierarchical Clustering

Using the Bisecting K-Means method, football teams have been clustered hierarchically, as seen in Figure 5-3. This dendrogram, visualizing the sub-groups within the clusters, can be beneficial to coaches wondering about the teams with similar tactical patterns with their opponents.
Consider this scenario: You are the coach of Monaco, and you have an upcoming match against PSG, an undefeated team for a long time. Since it is impossible to find an example reference game to get inspiration on how to beat them, you can analyze the games that Barcelona lost, the team resembles the playing style of PSG the most.

Moreover, for football analysts, it can be beneficial to examine a more detailed clustering of teams. The original grouping creates 3 clusters, as the elbow method dictated. But if some analysts require to classify with more detail, they can do so by simply viewing this dendrogram. Furthermore, this visual also shows the teams with the most unique styles. Teams such as Arsenal, Bordeaux, and Getafe have no siblings at the lower level of the tree, therefore, there is not a team that plays greatly similar with them.

5.1.2.2 Scatter Plot

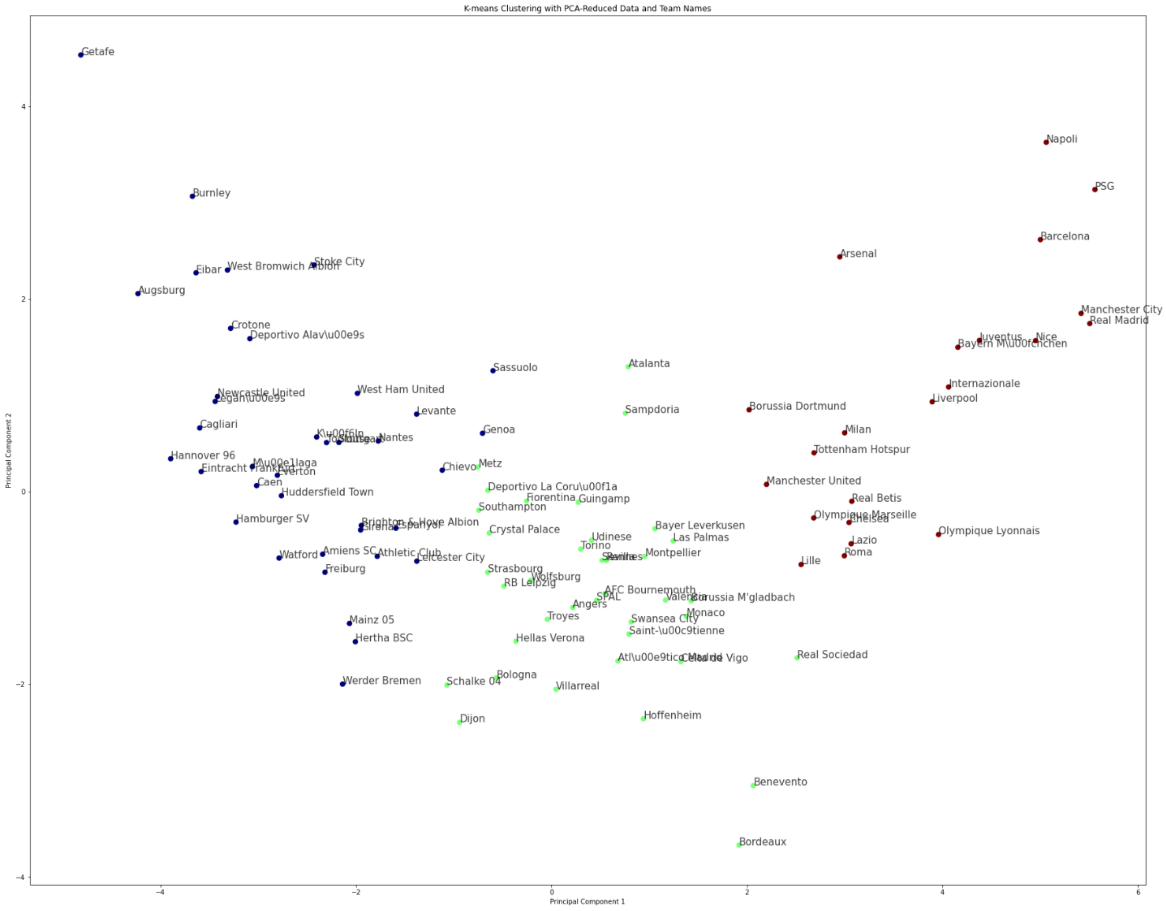

Figure 5-4: Scatter Plot showing the Clustering of Teams

The scatter plot above is also created via the Bisecting K-Means approach, even though the final model still uses K-Means, for uniformity reasons. This plot validates the findings of the dendrogram. Barcelona and PSG are located near each other, implying a similar tactical pattern, additionally, Arsenal, Bordeaux, and Getafe are located distant from all other teams, indicating a unique playing style.

Even though it is possible, predicting the subgroups is harder in the scatter plot, compared to the dendrogram. But it is much easier to compare a team's playing style with any other team since they are all placed in the same plane.

5.1.2.3 Cluster Probabilities

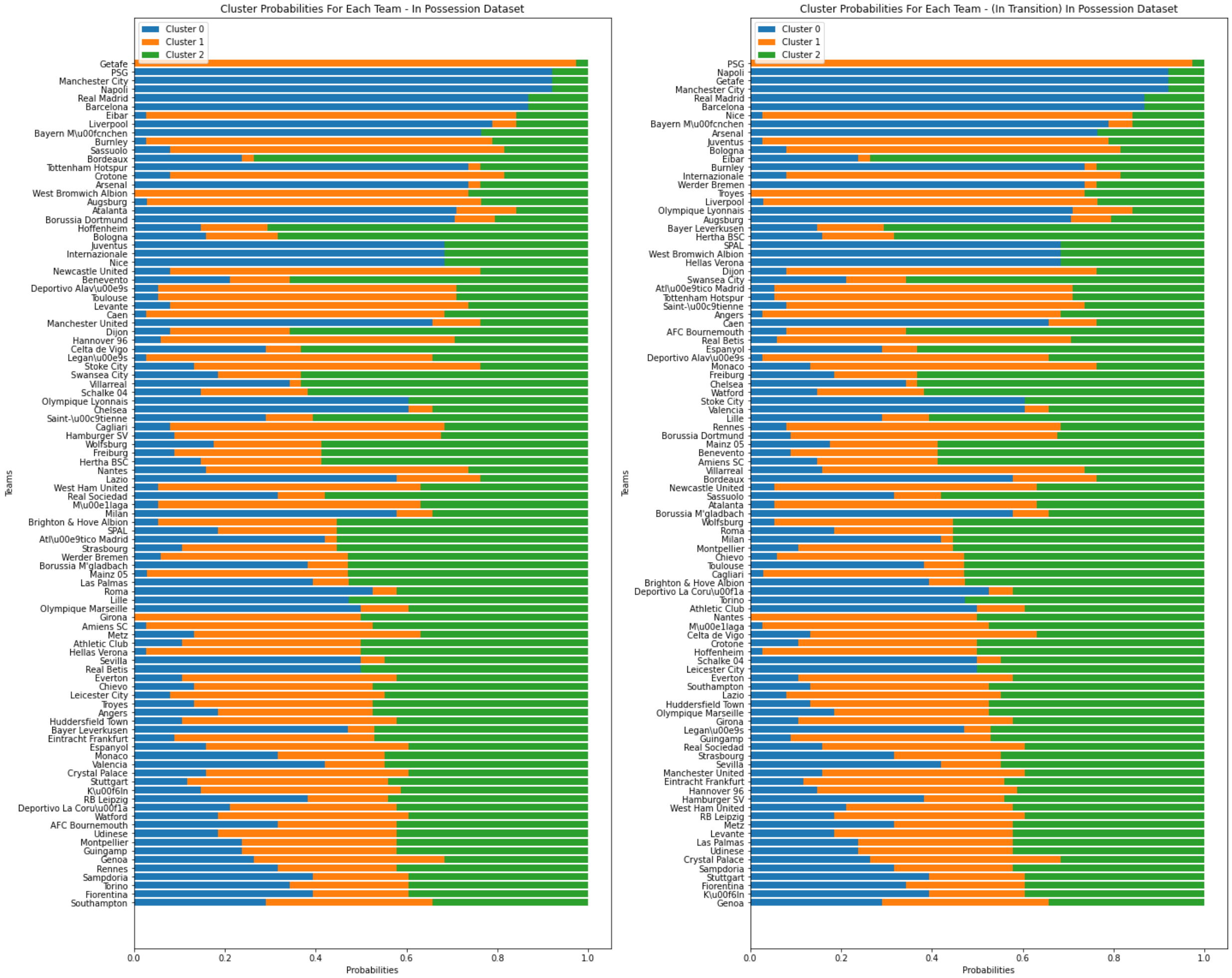

Figure 5-5: Stacked Bar Plot Showing the Probabilities for Each Team to be In Each Cluster For In Possession Datasets

The stacked bar plot shown on the left side of Figure 5-5, shows the frequency of the teams playing with each style, with respect to ball possession. It can be assumed that cluster 0, visualized with the color blue, contains the teams that control the ball throughout the game, whereas cluster 1, visualized with the color orange, contains the teams that prefer a more defensive approach. Barcelona, Napoli, and Manchester City are examples of the first one, while Getafe and Burnley are examples of the latter. It is interesting to see that more popular and successful teams are grouped together. The assumption regarding the patterns of the clusters will be tested in upcoming chapters.

Seeing this graph with both general in-possession and after-the-transition in-possession datasets' probabilities provides the distinct approaches of the teams in the transition phase. For example, PSG is undoubtedly a blue cluster team in the overall game whereas it is an orange cluster game in the transition phase, showing how some teams are quite similar with each other overall but they approach different phases of the game in distinct ways.

The cluster probabilities visuals are also good for visualizing which teams maintain their tactical patterns throughout the season, and which teams change their strategies on a game basis, with respect to who their opponent is. As seen in the left side of Figure 5-6, Napoli is the most loyal team to their out-of-possession tactics. They almost never change their game plan. Bayer Leverkusen on the other hand, treating each game separately, is impossible to categorize.

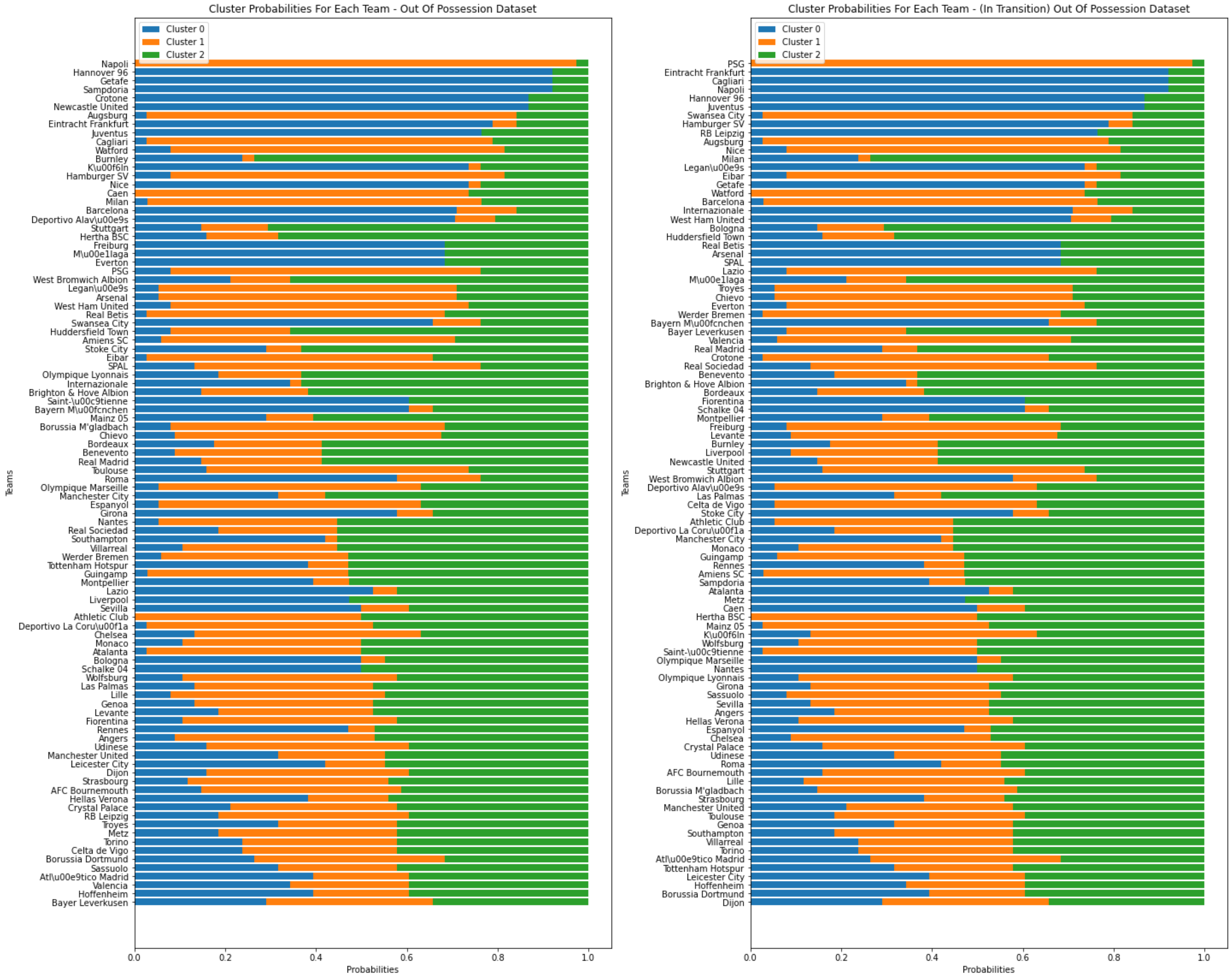

Figure 5-6: Stacked Bar Plot Showing the Probabilities for Each Team to be In Each Cluster

For Out of Possession Datasets

5.1.3 Win Percentages Results

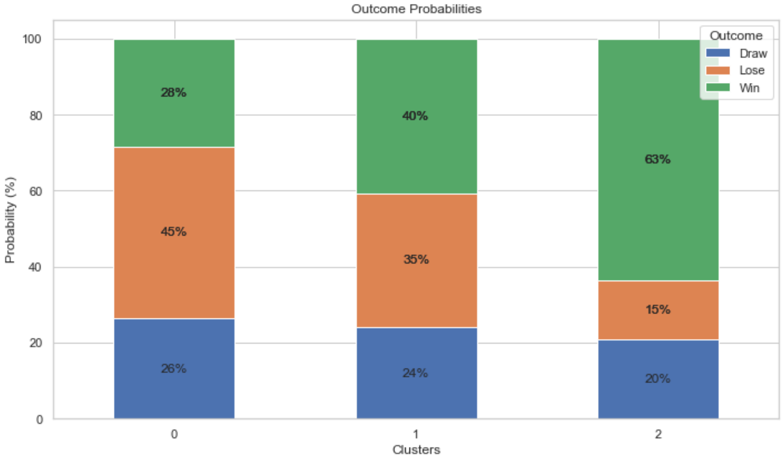

Figure 5-7: Probability of Each Match Outcome for Each Cluster

As shown in Figure 5-7, cluster 2 is the most successful, among three, and cluster 0 is the worst performing one. Moreover, cluster 2 dominates the other 2, in terms of winning probabilities. It is the best-performing cluster by far, no matter its opponent's cluster, as seen in Figure 5-8.

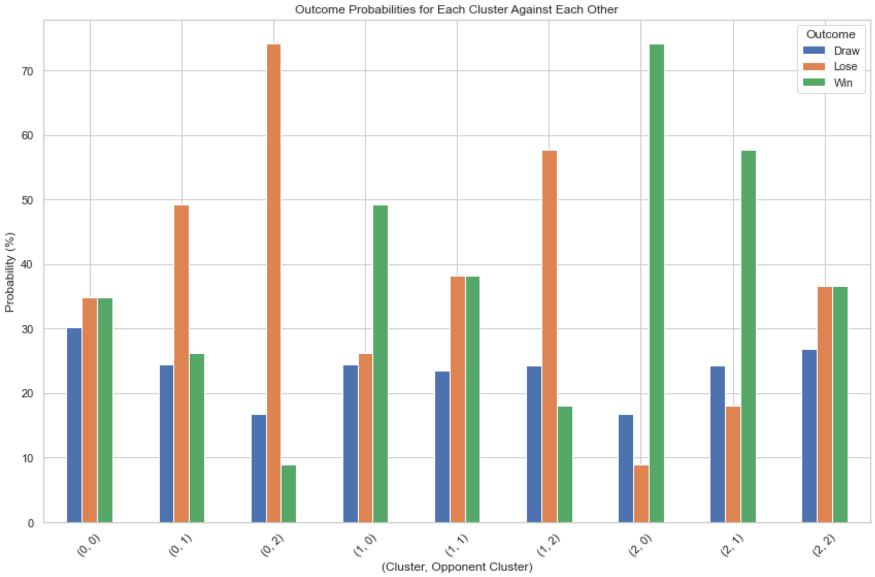

Figure 5-8: Probability of Each Match Outcome for Each Cluster Against Each Other

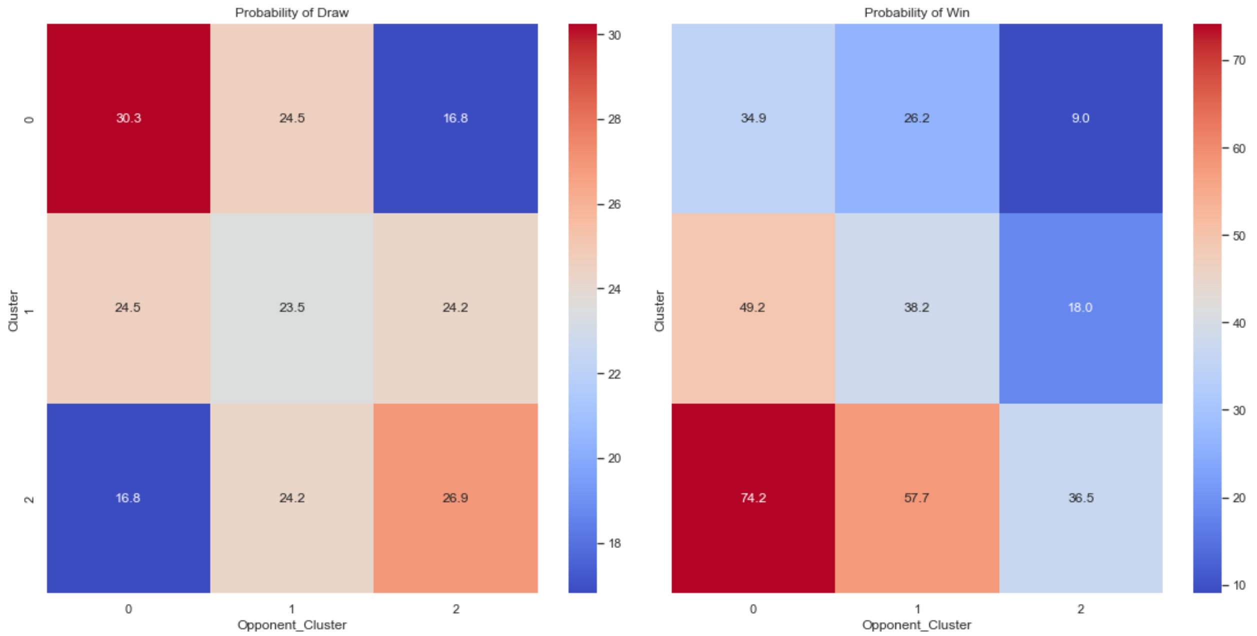

Figure 5-9: Deciding Optimal Strategy Against Each Cluster

Table 5-1: Deciding Optimal Strategy Against Each Cluster – Goal is Winning

Opponent’s Cluster	Optimal Cluster
0	2
1	2
2	2

Table 5-2: Deciding Optimal Strategy Against Each Cluster – Goal is Drawing

Opponent’s Cluster	Optimal Cluster
0	0
1	0
2	2

The heatmap in Figure 5-9, as well as Table 5-1, emphasize the dominance of cluster 2. If the goal is winning the game, the optimal cluster is always cluster 2, independent of the rival’s cluster. If the draw is aimed on the other hand, the cluster must be determined based on the rival’s cluster. If the rival’s cluster is 2, the optimal cluster is likewise 2. Otherwise, the best option is cluster 0. For both goals, cluster 1 must never be chosen.

5.1.4 Comparing Variable Statistics of Clusters

Since in total of 75 bar plots are generated in this section and demonstrating all of them would significantly harm the report’s readability, some of the chosen graphs will be represented. Their capability to reveal the tactical patterns is aimed to be shown in this way.

It must be foretold that any score demonstrated in this section, is the average value of the variable in question.

5.1.4.1 Variable Statistics for In-Possession Datasets

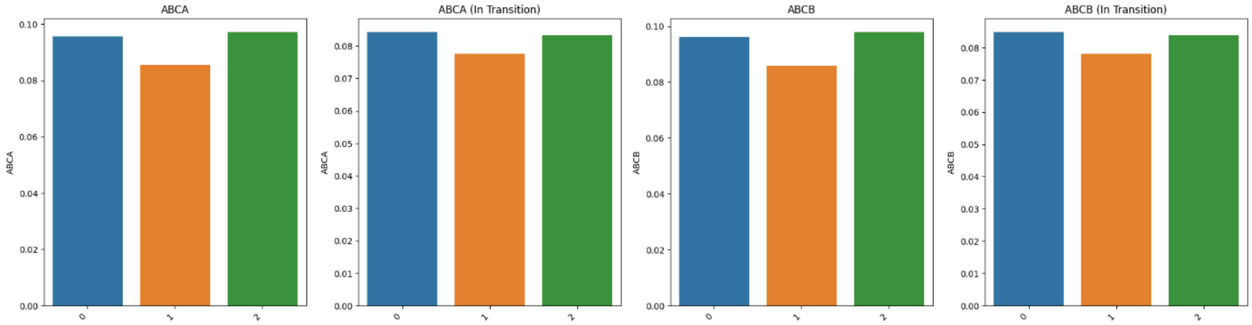

Figure 5-10: Comparing Pass Motifs

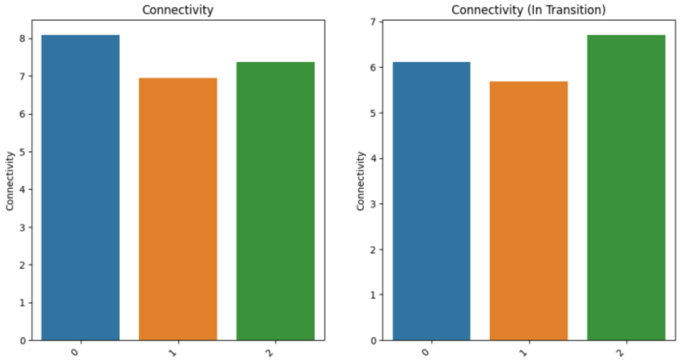

Figure 5-11: Comparing Average Connectivity Score

The bar plots shown in Figures 5-10 and 5-11 reveal the pattern between pass motifs and pass connectedness. ABCA and ABCB pass motifs represent the triangular-shaped pass sequences in the game. Preferring them more correlates with high connectivity.

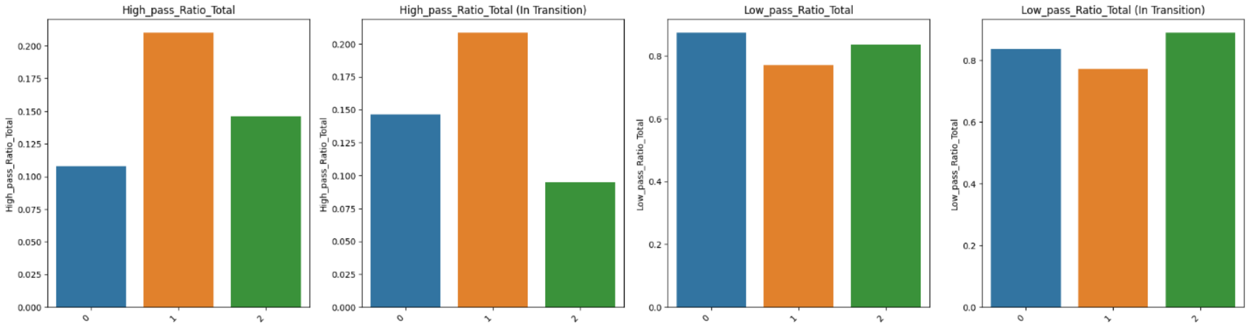

Figure 5-12: Comparing Pass Height Preferences

A similar pattern can be seen in pass height preferences, as seen in Figure 5-12. The plots for low pass ratio, look a lot like the bar plots above. The less connected teams who also don't usually form triangles in their pass sequences, tend to prefer long passes more frequently. The reason behind this is simple, if a team can not carry the ball forward with simple low passes, they can try the other option, risky, high passes. It can be said there is a trade-off regarding the choice between low versus high passes. One is more reliable and the other is more rewarding.

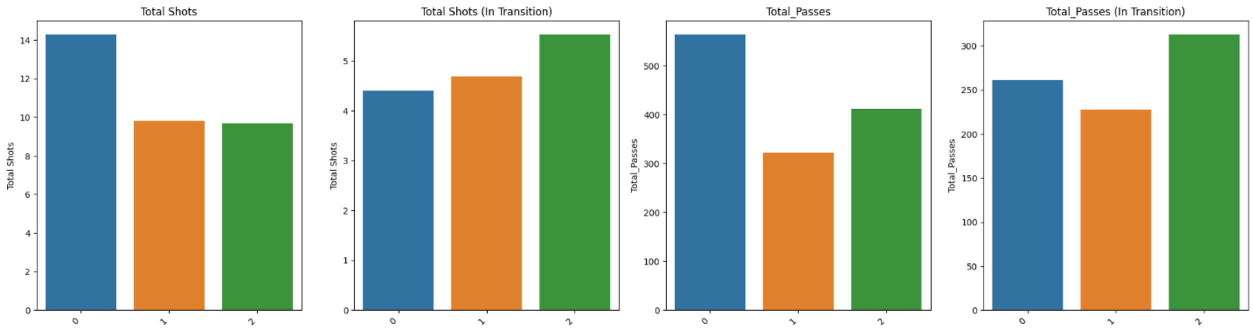

Figure 5-13: Comparing Average Number of Passes and Shots

As expected, the orange cluster passes the ball less often, as seen in Figure 5-13. Their high pass preference decreased their total pass amount by lowering the pass success rate. Moreover, the effect of separate plans for different game phases is also visible. Teams in the blue cluster tend to pass and shoot less in the transition phase, indicating a more direct passing approach.

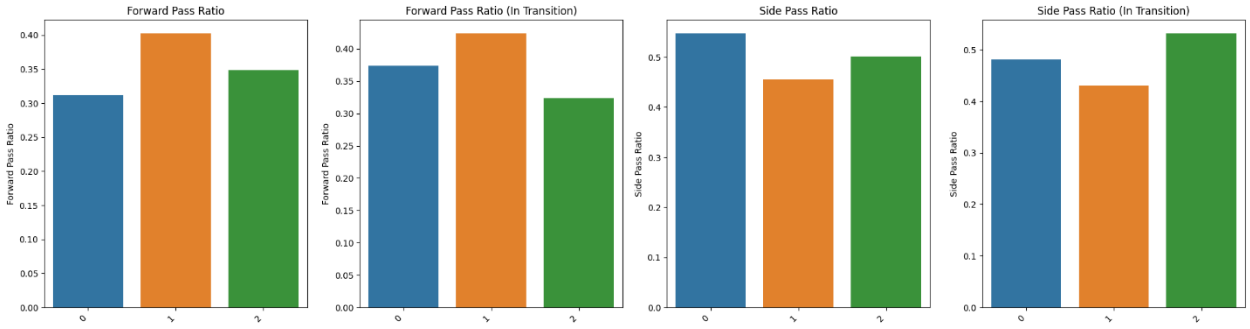

Figure 5-14: Comparing the Pass Direction Preferences

Trivially, the forward pass ratio and side pass ratio seem mutually exclusive, as shown in Figure 5-14. Additionally, the findings of these plots validate the previous ones. The orange cluster's preference for the direct pass option overlaps with its lack of total passes and connectedness as well as the great frequency of high passes.

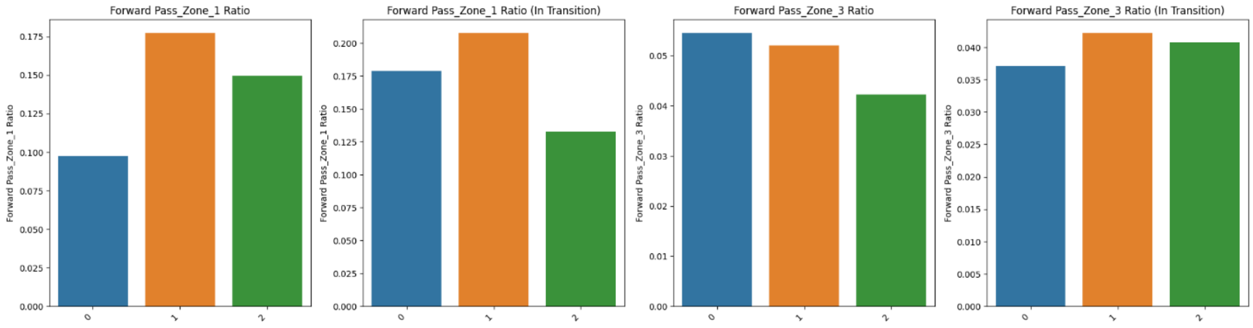

Figure 5-15: Comparing the Pass Direction Preferences Across Different Regions

As visualized in Figure 5-15, the blue class is less likely to prefer forward passes while the ball is near to its goal, whereas it passes forward more often further on the pitch. The pattern for more passing, and possession aiming teams to prefer side passes in the build-up phase while switching to a more direct approach further in the pitch has been revealed via the comparison of forward pass ratios

Between the first and third zones.

5.1.4.2 Variable Statistics for Out-of-Possession Datasets

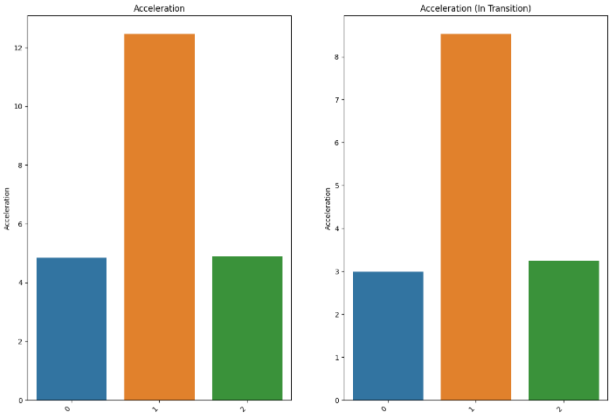

Figure 5-16: Comparing the Number of Accelerations per Game

Analyzing the number of accelerations per game provides information about the tactical and physical intensity of various clusters, with the orange cluster being of special interest. Teams in this cluster have a lot of accelerations, which suggests that they play an aggressive style of possessionless ball play that is dynamic and fast in order to retake possession. This can be a useful tactic to throw off the other team's rhythm and induce turnovers, which are essential for starting counterattacks.

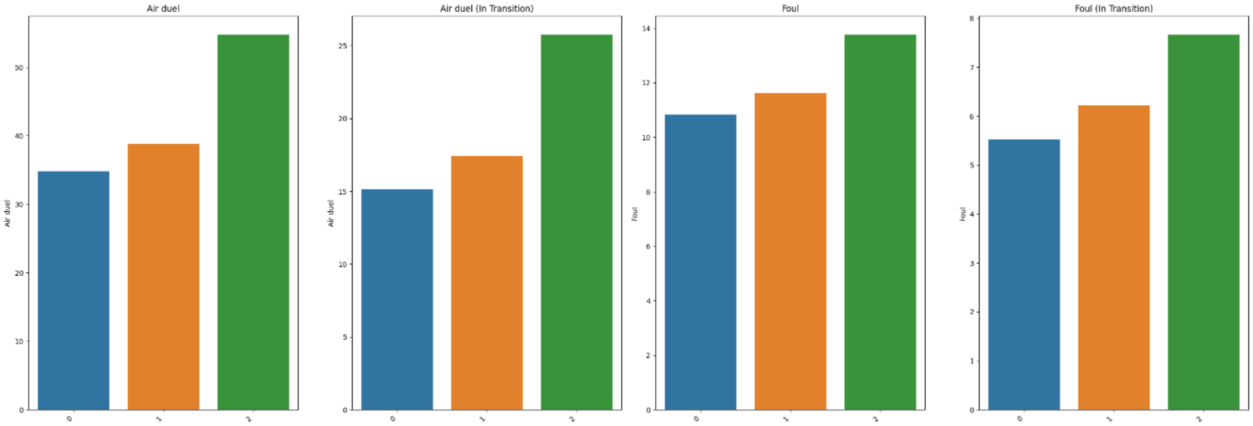

Figure 5-17: Comparing Air Duels and Total Fouls per Game

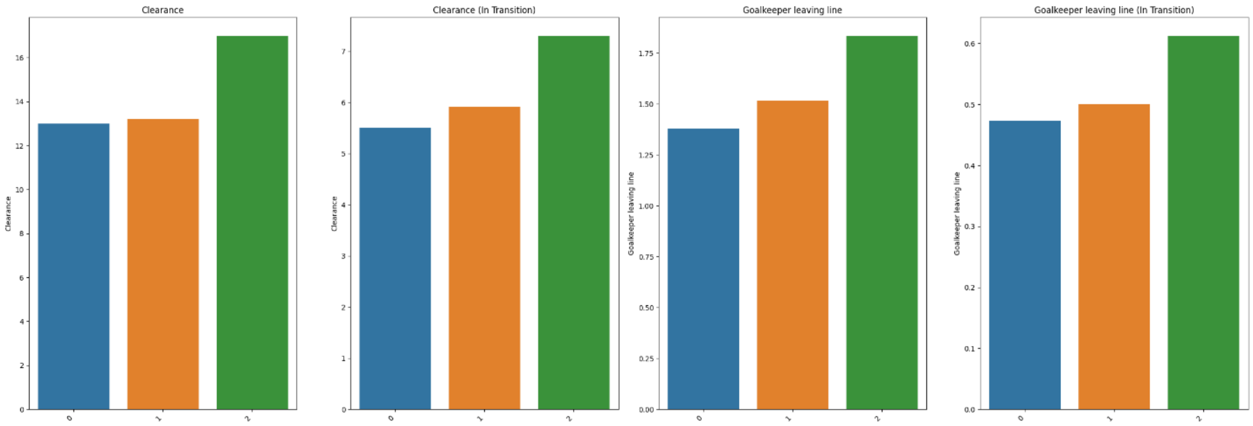

Figure 5-18: Comparing Clearances and Number of Times the Goalkeeper Left His Line per Game

The numbers from Figure 5-18 give a thorough picture of the defensive tactics used by various groups and show how frequently air duels, fouls, clearances, and goalkeeper actions occur. Teams with defensive-focused clusters, like the green cluster, typically participate in more of these exercises, indicating a forceful, possibly more aggressive style of defense. These tactics might show that a team places a high priority on resilience and defensive organization, which frequently enables them to better withstand offensive pressure and limit scoring opportunities for their opponents.

5.1.5 Supervised Learning Results

After labeling the data with the clusters found by the K-Means algorithm, several supervised learning methods have been tested in order to acquire the variables' feature importance scores. A comparison of the models' performances and the resulting feature importance scores are shown below.

5.1.5.1 Comparing the Performance of Several Supervised Learning Algorithms

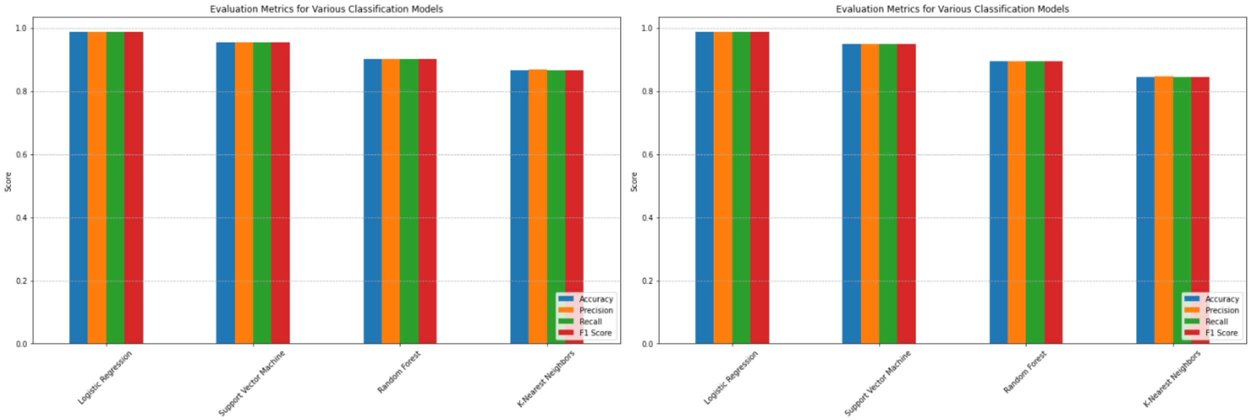

Figure 5-19: Classification Models' Performances for In Possession (Left) and In Transition In Possession (Right) Data

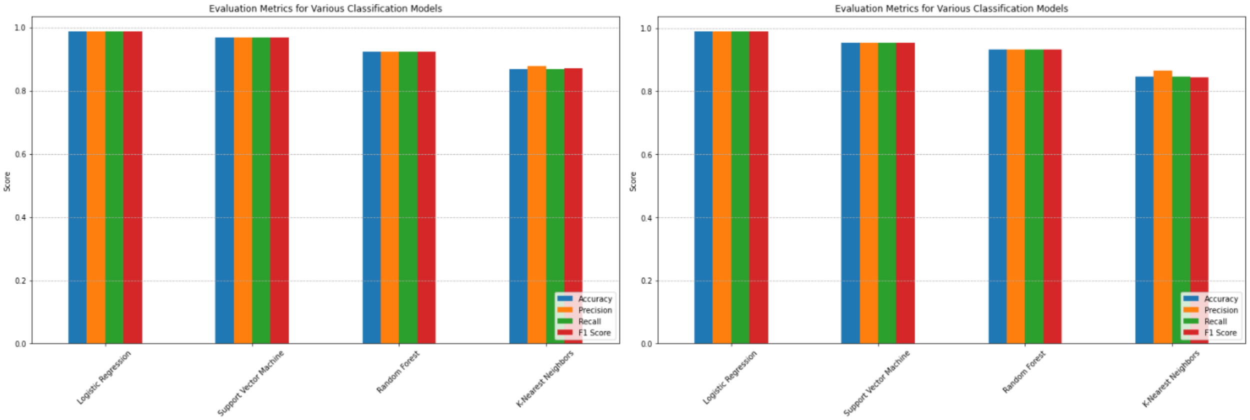

Figure 5-20: Classification Models’ Performances for Out of Possession (Left) and In Transition Out of Possession (Right) Data

As shown in Figures 5-19 and 5-20, for all datasets, Logistic Regression is the best-performing method. Nearly perfect scores for all metrics, on the unseen test set. This result justifies the chosen model used in the feature importance part.

Additionally, one of the non-functional requirements is easily satisfied by Logistic Regression. Even the lowest score for any dataset and metric is greater than 0.98 and therefore greatly exceeds the chosen threshold of 0.9.

5.1.5.2 Feature Importances

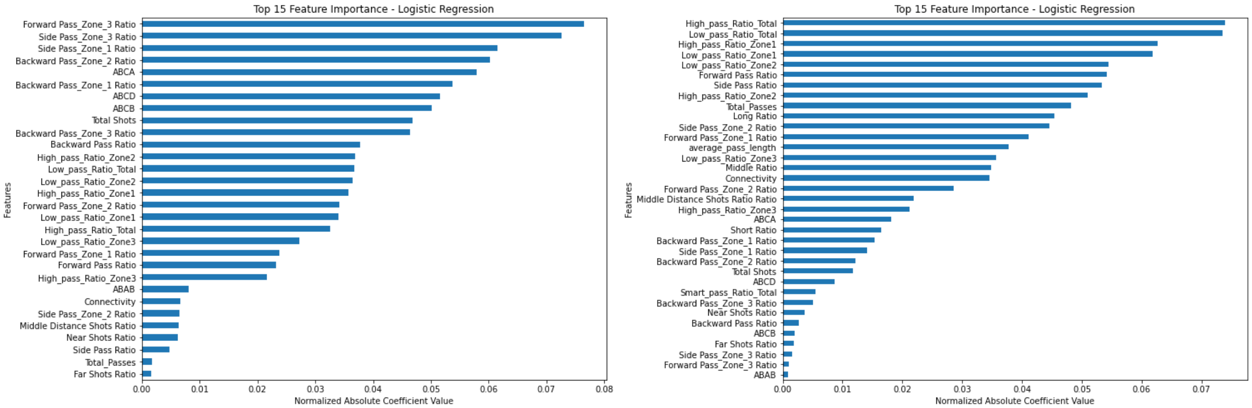

Figure 5-21: Importance Score of Top 15 Features for In Possession (Left) and In Transition In Possession (Right) Data

The K-Means model learned from different aspects of the data, for different phases in the game. Pass directness is the most important aspect, followed by pass motifs in possession games, whereas pass height is more significant for the positive transition phase, according to graphs in Figure 5-21. In both graphs, the influence of pitch division is visible. These results emphasize the relevance of pitch and phase partition methods, as well as the concept of pass motifs.

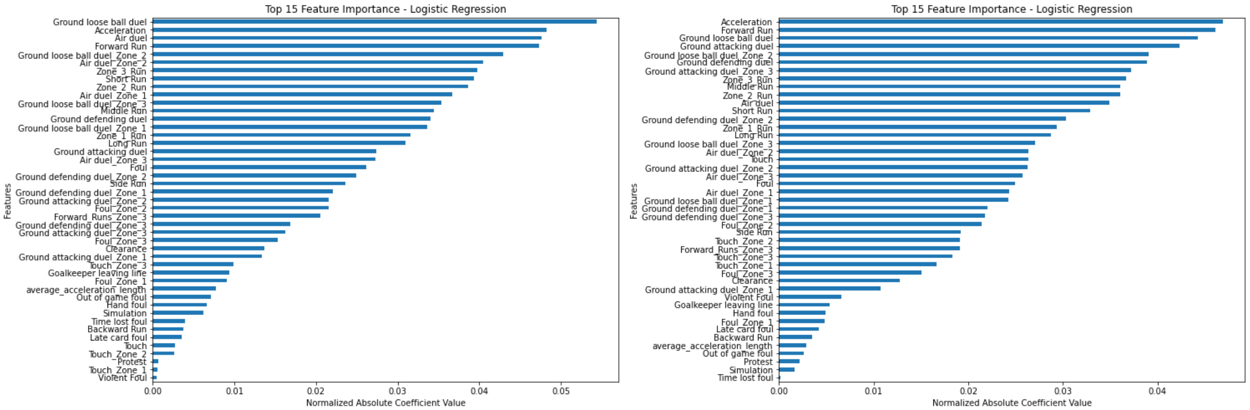

Figure 5-22: Importance Score of Top 15 Features for Out of Possession (Left) and In Transition Out of Possession (Right) Data

For out-of-possession datasets, the partition of the football pitch method does not affect as much as in-possession data, which is not surprising at all. The developers of the method were working on passing data after all. Game phase separation, on the other hand, is still relevant even though it is not as powerful. Acceleration and forward run variables are the most influential for the transition phase, whereas loose ball duels are more important overall.

5.2 Evaluation

5.2.1.1 Clustering Success

First of all, for all 5 datasets, the K-Means model successfully grouped the data into a chosen number of clusters. In this way, the suitability of the data for clustering is validated, which is the prerequisite of the upcoming evaluations. Moreover, win percentage results have revealed the clustering results' capacity to predict the match outcomes, which were hidden in the K-Means model. For example, as seen in Figure 5-7, cluster 2 have won 63% of their games. A random grouping would predict a 33% percent probability for each match outcome. This great difference proved that the K-Means model in fact learned some patterns from the data, instead of just grouping randomly. The capability of match outcome prediction of the model has also been proven, which indicates the usability of the system.

Table 5-3: Silhouette Scores

Dataset Name	**Silhouette Score**
In Possession	0.32
Positive Transition	0.34
Out of Possession	0.36
Negative Transition	0.37
Final Dataset	0.45

Furthermore, the silhouette scores for each dataset are higher than the designated threshold 0.25, as seen in Table 5-3. These scores prove that demonstrating a respectable degree of clustering process success. The groups within the datasets show a slight but positive separation as indicated by these scores. Although the clustering can be considered successful in part, the scores' close proximity to one another—all of them are below 0.5—indicates that the data points inside each cluster are not clearly distinguishable from one another. This emphasizes how difficult it is to separate the data into individual categories because of the considerable overlap between the clusters. In spite of this, the clustering indicates the complex nature and proximity of the data points in the supplied datasets and offers a valuable framework for analysis.

5.2.1.2 Extensiveness of Data Preparation and Visualization

It is evident from the assessment process for the degree of data preparation and visualization that the project has both successfully met and surpassed its predetermined criteria. First, the feature importance part of the report provides clear confirmation of the value of the newly generated non-trivial features. Preferred pass motifs, regions of the football field that are used more frequently, and the average connectivity of the pass graph are examples of features that were found to be effective, as discussed in the feature importance part. The clustering model was able to derive significant insights from these variables.

In addition, the project's visualization component exceeded objectives by utilizing a wide variety of plot types. Beyond the mandatory triple of distinct graphical representations, this project employed scatter plots, dendrograms, bar plots, stacked bar plots, and heatmaps to proficiently present the results. This visual diversity improves the interpretability of the clustering results in addition to facilitating a thorough analysis of the various playing styles.

Finally, it is impossible to overstate the significance of the massive feature engineering efforts, which produced 75 new features. These efforts had a crucial role in improving the dataset's richness and, as a result, the clustering process' efficiency. Additionally, a solid analysis and recommendation system, which is the ultimate goal, is possible thanks to the comprehensiveness of the visualization techniques used, which guarantee usability and enable a deeper understanding of the data. This thorough approach to data manipulation and presentation highlights how well the project met its overall goals and non-functional requirements.

5.2.1.3 Clustering Results' Interpretability

As previously demonstrated in the results chapter, the patterns such as "The teams that prefer ABCB pass motif less, also tend to prefer high passes rather than low passes.", and "The less connected teams tend to prefer forward passes more frequently, indicating more direct playing style." can be easily reached. The K-means clustering model's capability to produce highly interpretable results that reveal insight into football teams' playing styles, has been proven this way.

The findings acquired from the clustering model, are valuable resources for in-depth tactical study since they clearly highlight different patterns and tactical tendencies. Through the examination of the clustered data, coaches and analysts are able to identify the strategic tendencies of various teams. These results are made more useful by the illustrations, which include dendrograms, scatter plots, bar plots, stacked bar plots, and heat maps. These plots make these patterns intuitive to comprehend and visually appealing. With the help of each visualization, teams can better understand how various tactical frameworks affect their operations, which makes comparisons and strategic planning easier.

Because of its interpretability, the clustering model's findings are both practically and statistically meaningful, allowing stakeholders in the sports industry to make well-informed decisions based on the detected playing styles. The thorough analysis included in the results section demonstrates the K-means model's usefulness as a reliable analytical tool for studying football teams' behavior and conducting strategic analyses.

5.2.1.4 User Interface

Firstly, the interface is functional in a way that it generates similar plots to the ones shown in the results section, in a user-specific manner. It shows the chosen team or league’s average values in cluster statistics, and it highlights the chosen team’s name in the scatter plot for example. It also successfully recommends a playing time against a chosen team or league, which was the ultimate goal of the project. It can be stated that the User Interface successfully achieved its usability requirements.

Table 5-4: Average Response Time for Each Plot Type in User Interface

Plot Type	**Average Response Time (seconds)**
Generating Scatter Plot	3.42
Generating Dendrogram	9.39
Generating Bar plot	0.01
Generating Stacked Bar Plot	3.94
Recommending a Playing Style	2.12

Similarly, its performance criteria are also matched. As seen in Table 5-4, every task successfully stayed under the designated time threshold, which was 5 seconds for plot generation and 2 seconds for tactic recommendation, with the only exception of dendrogram generation tasks. Since all plots except the dendrogram, can be generated by simply processing previously saved clustering labels, they are expected to work fast. Dendrogram, on the other hand, requires running another hierarchical clustering model, which takes more time than simply drawing it.

Overall, the User Interface can be accepted as successful, in terms of usability, and performance.

6 CONCLUSIONS AND FUTURE WORKS

6.1.1 Conclusions

The implementation of a multi-step K-means clustering technique to analyze and comprehend the playing styles of football teams in multiple leagues has been thoroughly investigated in this study. Through the application of modern data analytics and machine learning techniques, specifically unsupervised learning, the study revealed profound insights into how teams modify their tactics at various phases of the match. This nuanced approach demonstrated how different tactical implementations affect game outcomes in addition to highlighting their range of characteristics.

This research is novel because of its multi-dimensional feature engineering method, which systematically combines event data to uncover underlying tactical patterns. Through the comprehensive collection of actions, ranging from passes and shoots to duels and runs, the study provided a detailed perspective of the game that was beyond conventional analysis. This large dataset made it possible to do a more thorough clustering procedure, which in turn made it possible to identify discrete strategic groups whose effectiveness could be measured in terms of match outcomes.

For football analysts and coaches, the predictive framework created for this research has proven to be a valuable tool. It gave planners and decision-makers of football, a strategic advantage by enabling the prediction of match outcomes based on recognized playing styles. The model gained credibility from the experimental validation of the clustering approach using game outcomes, which confirmed its ability to have a significant effect on real-world football strategies.

Furthermore, the project's extensive use of sophisticated visualizations improved the interpretability of complex datasets and opened up opportunities for public access to the findings. These visual aids gave users clear and useful information, bridging the knowledge gap between technical statistical analysis and practical tactical planning. Utilizing state-of-the-art web technologies, the user interface provided an interactive data exploration, enabling users to obtain personalized tactical recommendations derived from the in-depth investigation.

With this project's implementation of a multi-step K-means algorithm, The teams were categorized according to comparable playing styles, and the clustering technique additionally supplied a clear distinction in playing styles that are directly related to game performance by identifying the variables that separate each cluster.
The consequences of this study go beyond academic research; coaches and analysts pursuing a more profound understanding of game mechanics will find practical uses for it. The techniques developed for this project can be modified to keep up with football's ongoing evolution, providing that tactical evaluations will always be beneficial.

In conclusion, this project makes a contribution to the field of sports analytics by creating a system that combines conventional football analysis with machine learning. It provides a novel viewpoint on how data-driven methods may be used to improve our comprehension of sports and opens the door for more advancements in the field of tactical analysis. This work bridges the gap between sports science theory and practice by advancing academic understanding while also offering useful tools that may be utilized in professional sports environments.

6.1.1.1 Future Work

The creation of a model that takes into consideration the tactical changes that are dependent on the score and the temporal dynamics that occur during football games is an essential path for future research. Football teams routinely modify their strategy according to the score at any given point in the game and on different game periods. Recognizing these tactical shifts—which might differ dramatically from the beginning to the end of a game—offers vital insights into a team's tactical adaptability and decision-making mechanisms. Such a model would improve our knowledge of tactical adaptations in football by examining how plans change during a game and in reaction to scoring situations. This work can be taken even further with the presentation of real-time data, creating a dynamic model that integrates live game data in addition to static match data analysis. Coaches would be able to make tactical modifications on the spot based on the actual game conditions. This would require a model that is exceptionally responsive and capable of processing data rapidly, in order to give actionable insights during the match

There is a great chance to improve the depth and precision of our clustering approach by incorporating other data types, especially tracking data. Tracking data provides temporal and spatial information on player movements that may provide previously unknown details about tactical choices and playing styles. This data would provide a broader collection of variables for clustering and might help uncover playing styles that were previously unknown, enabling a more thorough examination of team behavior and strategy.

Future studies should also focus on expanding the scope of the investigation to include more football leagues from various nations. Due to differences in facilities climatic conditions, and culture, different leagues may display distinct tactical preferences and tendencies. The generalizability and efficacy of the approach in many footballing environments could be examined by examining a wider range of leagues. This would improve the model's resilience while also offering insights that are applicable globally, expanding the strategic toolkit that coaches and analysts have access to on a global scale.

Examining sophisticated machine learning models—particularly deep learning methods—is also important in order to manage the complexity and high dimensionality of data from football games. Intricate processes for classification and clustering may be made available by deep learning, improving model accuracy and enabling the discovery of more subtle patterns in the data. These methods may be more proficient at handling the complexities and nuances of large amounts of data, which could result in a more complex and useful grouping of playing styles.

Finally, creating a way for the model to incorporate user feedback is a crucial step in making sure the model is accurate and relevant. The model's end users, coaches, and analysts, may provide feedback that would be utilized to continuously improve and modify the model's parameters. Through this iterative process, the model's practical applicability in real-world contexts would increase as it becomes more closely aligned with growing tactical theories and user needs.

Future research can significantly improve the model's robustness, broaden its application, and guarantee that it remains at the forefront of football tactical analysis by addressing these shortcomings. This ongoing development is essential to keeping the model relevant and useful in a sport that is changing constantly.

7 REFERENCES

[1] L. Pappalardo and E. Massucco, “Soccer match event dataset”, 2019, doi: 10.6084/M9.FIGSHARE.C.4415000.

[2] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, 1967. [Online]. Available: https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed

[3] S. Plakias et. al, “Identifying Soccer Teams’ Styles of Play: A Scoping and Critical Review”, JFMK, vol. 8, no. 2, p. 39, Mar. 2023. [Online]. Available: doi: 10.3390/jfmk8020039

[4] J. Diquigiovanni and B. Scarpa, “Analysis of association football playing styles: An innovative method to cluster networks”, Statistical Modelling, vol. 19, no 1, pp. 28-54, Feb. 2019. [Online]. Available: doi: 10.1177/1471082X18808628.

[5] J. L. Peña and H. Touchette, “A network theory analysis of football strategies”, 2012, doi: 10.48550/ARXIV.1206.6904.

[6] L. Gyarmati, H. Kwak, and P. Rodriguez, “Searching for a Unique Style in Soccer”, 2014. [Online]. Available: doi: 10.48550/ARXIV.1409.0308.

[7] A. Lopez-Valenciano et. al., “Association between offensive and defensive playing style variables and ranking position in a national football league”, Journal of Sports Sciences, vol. 40, no 1, pp. 50-58, Jan. 2022. [Online]. Available: doi: 10.1080/02640414.2021.1976488.

[8] R. Pollard, C. Reep, S. Hartley, “The Quantitative Comparison of Playing Styles in Soccer”, in Science and Football, pp. 309-315, 1988.

[9] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, “Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data”, in 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China: IEEE, Dec. 2014, pp. 9-14. [Online]. Available: doi: 10.1109/ICDMW.2014.167.

[10] L. Ruan, H. Ge, Y. Shen, Z. Pu, S. Zong, and Y. Cui, “Quantifying the Effectiveness of Defensive Playing Styles in the Chinese Football Super League”, Front. Psychol., vol. 13, no. 899199, Jun. 2022. [Online]. Available: doi: 10.3389/fpsyg.2022.899199.

[11] J. Di and X. Gou, “Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization”, JCP, pp. 588-595, 2018, doi: 10.17706/jcp.13.6.588-595.

[12] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space”, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no 11, pp. 559-572, Oct. 1901, doi: 10.1080/14786440109462720.

[13] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm”, in Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia: EAI, 2020. doi: 10.4108/eai.24-1-2018.2292388.

[14] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, Oct. 1987, doi: 10.1016/0377-0427(87)90125-7.

[15] P. J. M. Ali and R. H. Faraj, “Data Normalization and Standardization: A Technical Report”, 2014, doi: 10.13140/RG.2.2.28948.04489.

[16] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, “An Introduction to Logistic Regression Analysis and Reporting”, The Journal of Educational Research, vol. 96, no 1, pp. 3-14, Sep. 2002, doi: 10.1080/00220670209598786.

[17] ISO/IEC 25012:2008, "Data Quality Assurance Criteria," International Organization for Standardization/International Electrotechnical Commission, Standard, 2008.

[18] ISO/IEC 23053:2022, "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)," International Organization for Standardization/International Electrotechnical Commission, Standard, 2022.

[19] sportsessionplanner.com, “Vertical Zones”, 2024. [Online]. Available: https://www.sportsessionplanner.com/s/suVGh/Vertical-Zones.html. [Accessed: 04- Jun- 2024].