June 1st, 2026

SUBSTITUTION PATTERNS AND THEIR SHORT-TERM EFFECTS IN THE DANISH SUPERLIGA

Yosef Tarek Hassan

Data Science Master Thesis

University of Southern Denmark

Supervisor: Panagiotis Tampakis

Abstract

Introduction: Football substitutions represent an important tactical component of football, yet the contextual factors associated with substitution impact remain relatively underexplored within football research. The purpose of the present study was therefore to explore how different substitution patterns were associated with short-term changes in match dynamics in the Danish Superliga, given the contextual factors in which the substitutions occurred.

Methods: The study was based on tactical second-half substitutions from the 2024/2025 Danish Superliga season. Seven contextual match variables were constructed: Match Minute, Opponent Strength, Score Difference, Number of Subs, Home/Away, Response, and Team Size (Red Cards). In addition, an xG-based impact score was developed to represent the short-term impact of substitutions using five-minute windows before and after each substitution. To examine contextual substitution patterns and variable influence, the study combined clustering analysis, adjusted rand index analysis, random forest for feature importance, and random forest for classification, across ten dataset variants containing different numerical and categorical representations of the seven variables.

Results: Overall, the clustering analyses identified several recurring contextual substitution patterns across the dataset variants. In particular, substitutions performed during the mid-second half, involving double substitutions and teams trailing in the match, were frequently associated with higher impact scores. Across the adjusted rand index and feature importance, the numerical representations of Match Minute and Opponent Strength consistently appeared most influential. The classification analyses further showed moderate but relatively consistent predictive performances, where the numerical and hybrid variants generally outperformed the fully categorical variants.

Conclusion: The findings suggest that the seven contextual match variables contained meaningful information related to short-term substitution impact, while also highlighting the complexity and unpredictability of football substitutions. The study further demonstrated that the results depended strongly on how the contextual variables were represented within the analytical framework. Consequently, the study should primarily be interpreted as an exploratory analysis of contextual substitution tendencies rather than a direct evaluation of substitution effectiveness.

Table of Contents

Introduction

Substitutions are a central tactical element in modern football. Coaches use substitutions to respond to developments in the match, manage fatigue, adjust playing style, and influence how the game unfolds (Pan et al., 2023). Unlike many other team sports where substitutions are frequent and unlimited, football has a limited number of substitutions, thereby increasing the strategic weight of each decision (Myers, 2012). Substitutions therefore represent one of the few direct interventions available to coaches during a game, and can play a decisive role in shaping match development and outcome (Wittkugel et al., 2022).

With the increased number of substitutions allowed in recent seasons, these decisions have become even more important. In 2020, the International Football Association Board introduced the option of five substitutions per team, a rule that was later permanently approved for top-level competitions (Fifa, 2023). This change has increased tactical flexibility and influenced how coaches manage their squads during matches. A recent study looking into international tournaments suggests that substitutions are increasingly used as deliberate attempts to change match dynamics rather than simply replacing tired players (Wei et al., 2024). In addition, the increased number of substitutions has expanded the possibility for coaches to perform broader tactical adjustments involving several players within shorter periods of the match, making double and triple substitutions a relevant aspect, when examining substitution patterns and their potential influence on match dynamics (Xiao & Zhang, 2024). Substitutions do not occur in isolation, and there are several different aspects that may be relevant to consider when examining how and why coaches choose to make them. The decision to substitute can vary depending on the timing within the match, the current scoreline and the broader flow or intensity of the game. Each of these factors may shape the development of the match.

An important factor could be the timing of the substitutions within the match, when examining their potential impact. Research analyzing large samples of professional matches has shown that substitutions are not evenly distributed across a match, but tend to cluster in specific phases, mostly in the second half (Sun et al., 2024). These patterns suggest that substitutions are often aligned with predictable shifts in match intensity, fatigue accumulation, and tactical adjustments. Substitutions made at different stages of the match may therefore serve distinct

strategic purposes, and their association with subsequent match dynamics may vary depending on when they occur.

Apart from the timing, another important factor could be the strength of the opposition. Teams may apply different tactical approaches depending on whether they are facing stronger or weaker opponents. Recent research suggests that substitution strategies vary depending on both match state and the relative strength of the opposition, particularly when teams are trailing during matches (Chen et al., 2025). Opponent strength may therefore represent an important contextual factor when examining substitution patterns and their association with subsequent match development.

Another factor could be the scoreline at the time of the substitution. A study found statistical evidence that making a substitution while losing is associated with an increased probability of subsequent goal scoring (Amez et al., 2021). Their results indicate that the relationship between substitutions and performance differs across match states, suggesting that substitutions may operate differently depending on whether a team is losing, drawing, or leading. Similarly, a study shows that substitution timing varies with score status, with coaches tending to substitute earlier when losing (Rey et al., 2015). A study suggests that the effectiveness of substitutions when trailing may depend on the timing of the intervention (Myers, 2012), while a study found no significant change in goal-scoring probability for trailing teams following a substitution (Silva & Swartz, 2016). In addition, behavioral evidence suggests that player effort varies with score difference, highlighting the relevance of score status in understanding substitution dynamics (Schneemann & Deutscher, 2017).

Match context may also include structural factors such as game location. Recent evidence indicates that physical performance differs between home and away matches, with significant variation observed in distance covered, high-speed running and sprinting distance (Chaize et al., 2024). This suggests that playing at home or away may influence performance responses and, consequently, the dynamics following substitutions.

Another possibly important contextual factor is numerical balance between teams. Red cards represent events that can change match dynamics. Studies have shown that receiving a red card is associated with a reduction in the teams goal-scoring rate and an increase in the opponents scoring probability (Badiella et al., 2023; Červený et al., 2018). Substitutions made under

conditions of numerical imbalance may therefore operate under fundamentally different constraints compared to substitutions made with 11 players, highlighting the importance of accounting for red cards when analyzing substitution patterns.

Taken together, existing research highlights that substitution decisions are based on a complex match context. Factors like the ones mentioned above may all influence how substitutions are made and how they relate to subsequent performance. Recent research has also begun to examine substitutions through more data-driven and contextual approaches. For example, a recent study investigated how different contextual match factors, including opponent strength and match state, influenced substitution strategies and tactical decision-making during matches (Chen et al., 2025). However, despite these more advanced contextual approaches, literature still rarely examines how broader substitution patterns consisting of multiple contextual factors are associated with subsequent changes in match dynamics. In addition, such analyses remain largely absent in the Danish Superliga, where detailed tracking and event data now provide new opportunities for investigation. These limitations form the basis for the present study. The aim of this project is therefore to explore how different substitution patterns, based on multiple contextual match factors, are associated with subsequent changes in match dynamics in the Danish Superliga. To examine this, the study applies an exploratory approach combining clustering techniques, expected goals (xG)-based performance measures, and random forest analysis to identify contextual substitution patterns and examine the relative influence of different match-related factors. Rather than focusing on isolated contextual variables or causal interpretations of substitutions, the project aims to examine how combinations of contextual factors relate to short-term changes in match dynamics. Accordingly, the study addresses the following research question:

How are different substitution patterns and contextual match factors, associated with subsequent short-term changes in match dynamics in the Danish Superliga?

Methods

The methodological workflow of this study was divided into seven main stages. The first stage focused on data collection, preprocessing, and construction of the primary substitution dataset, where substitutions, seven contextual match variables and xG-based impact score were prepared for further analysis. The Second stage consisted of the making of ten dataset variants, based on the seven variables, which would be used for the later analyses. The third stage consisted of a clustering analysis, where the ten dataset variants were used across different clustering methods, to identify different types of substitutions and examine the impact of the resulting clusters through the xG-based impact score. The fourth stage examined the influence of the abovementioned match variables on the clustering structure through Adjusted Rand Index (ARI) analysis. The fifth stage focused on random forest analysis, based on the ten variants, for feature importance of the match variables, while the sixth stage applied random forest classification to predict the impact score based on the match variables. Finally, the seventh stage consisted of a sensitivity analysis used to evaluate the robustness of the methodological choices and variant results throughout the study. All code used throughout the study has been uploaded to GitHub, and the link can be found in Appendix 1.

Preprocessing and Construction of the Primary Dataset

This study was made in collaboration with Divisionsforeningen, which provided access to nearly all data used throughout the project. The study was based on data from the 2024/2025 season in the Danish Superliga, which included 193 matches. Match data were obtained from Optas SRML files, which provided detailed information about substitutions, timestamps, goals, and other contextual match variables. In addition, xG data were extracted from Optas F70 files linked to each match. This data were used to construct an xG-based impact score measuring the short-term effect of substitutions, which will be described in more detail later. League standings from the season were collected from Superstats.dk (*Stillingen, 2024/2025 - SuperStats*).

As a preprocessing step, the dataset was restricted to tactical half time and second half substitutions. Information regarding substitution type and match period was directly obtained from Optas SRML files, where substitutions were already classified according to their reason and timing. First half substitutions and injury related substitutions were excluded to focus on substitutions assumed to primarily reflect tactical decision making, rather than forced changes

caused by either injuries or other rare early match situations. Out of 1739 total substitutions, 125 were injury related substitutions, while 44 occurred during the first half. However, these categories overlapped, as 40 substitutions were both first half and injury related. Therefore, a total of 129 substitutions were excluded, resulting in a final dataset consisting of 1610 tactical half time and second half substitutions used for later analysis (figure 1).

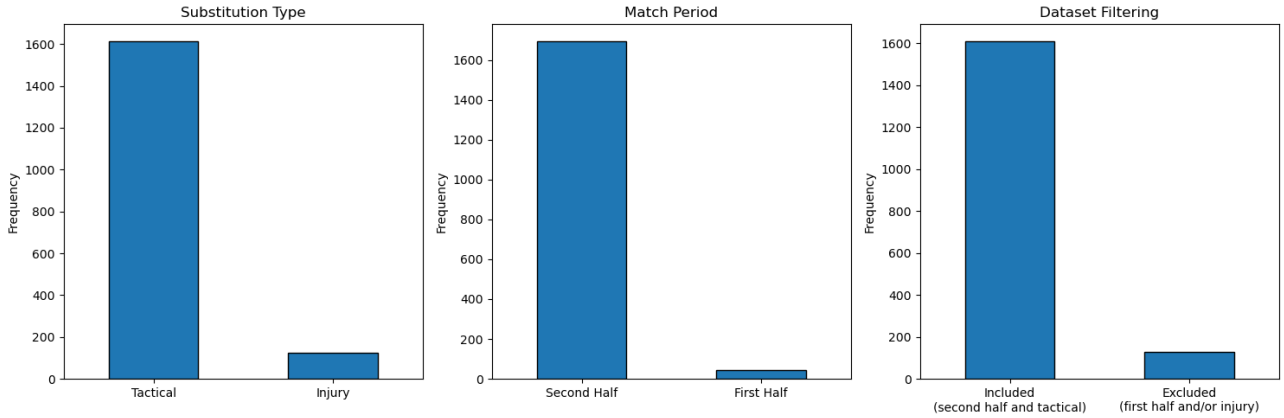


Figure 1: Distribution of substitutions by substitution type, match period, and dataset filtering criteria in the original dataset.

Following the preprocessing stage, additional variables were attached to each substitution, to construct the final dataset for analysis. These variables can broadly be divided into two categories. The first category consisted of identifier and match metadata variables, including Sub ID, Match ID, Match Day, Team Name, and Opponent Name, which were used to uniquely identify and contextualize each substitution within the dataset. The second category consisted of the seven contextual match variables used throughout the clustering and classification analyses. These variables were constructed to capture different tactical and situational aspects surrounding each substitution. Furthermore, the effectiveness of each substitution was evaluated through an xG-based impact score, which was used to measure the short-term impact of substitutions. The variables and the impact measure were selected and discussed in collaboration with Divisionsforeningen and will be described in more detail below.

Variable 1: Match Minute

The Match Minute variable represents the timing of each substitution within the match. It allows the analysis to distinguish between substitutions made in different phases of the match, rather than treating all substitutions as tactically equivalent regardless of timing. Alongside the main Match Minute variable, the exact minute and second of each substitution were also extracted. While the Minute and Second variables represented the exact timestamp of the substitution, Match Minute represented the corresponding playing minute of the match and was

the variable used throughout the analyses. For example, a substitution occurring at 73 minutes and 20 seconds was registered with Minute = 73, Second = 20, and Match Minute = 74. In addition, the Minute and Second variables were combined into a Total Seconds variable, representing the exact number of seconds played in the match at the time of the substitution. Total Seconds were primarily used in the construction of other contextual variables and outcome measures, including the Response variable, the Number of Subs variable, and the xG-based impact score, which will be described in more detail in their respective sections later. Match Minute was treated as a hybrid variable throughout the study, as both numerical and categorical representations were used across the different dataset variants later. More specifically, the variable was represented in one numerical form and three different categorical forms. This approach was chosen to examine how different representations of timing influenced the later clustering structure and the classification analyses.

Match Minute as a Numerical Variable

In its numerical form, Match Minute was used as a continuous variable representing the playing minute of the substitution event (the 74 example from above). Since differences in variable scales may influence the contribution of numerical variables within clustering and classification (Strobl et al., 2007; Wongoutong, 2024), Match Minute was standardized using z-score standardization through Scikit-learn's StandardScaler (SciKit Learn - SandardScaler), resulting in the numerical variable "Match Minute Num". The distribution of substitutions across the Match Minute variable is illustrated in figure 2 below. The detailed distribution with exact number of substitutions can be found in appendix 2.

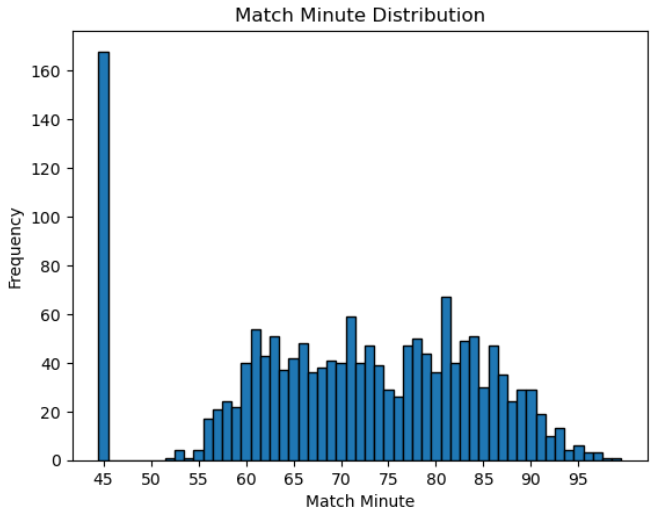
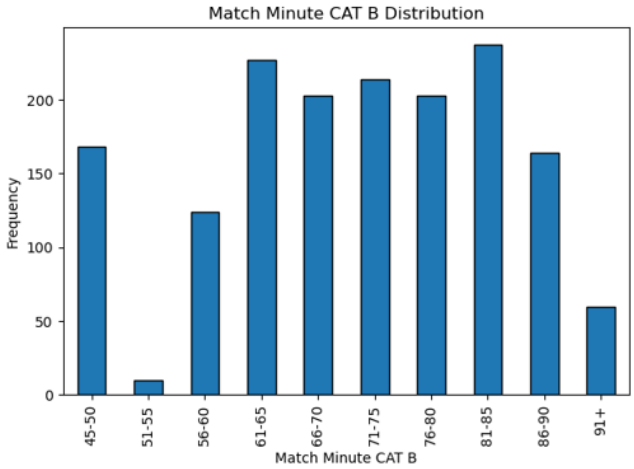


Figure 2: Distribution of match minutes for all included substitutions.

Match Minute as Category B

The first categorical representation of Match Minute, referred to as Match Minute CAT B, divided the second half into ten five-minute intervals. This representation was constructed to preserve a relatively detailed temporal structure, while reducing the continuous nature of the original numerical variable. The intervals were defined as: 45–50, 51–55, 56–60 … 86–90, and 91+. Although the first interval technically consisted of six minutes due to the inclusion of half-time substitutions registered as minute 45, no substitutions occurred between minutes 46 and 50 in the dataset. As a result, the 45–50 category exclusively consisted of half-time substitutions. The final category was included to account for substitutions occurring during stoppage time. The distribution of substitutions across the Match Minute CAT B variable is illustrated in figure 3.



Match Minute CAT B Intervals	Number of Substitutions
45-50	168
51-55	10
56-60	124
61-65	227
66-70	203
71-75	214
76-80	203
81-85	237
86-90	164
91+	60

Figure 3: Distribution substitutions across Match Minute CAT B intervals, where match minutes were grouped into 5-minute intervals. The table on the right shows the exact distribution of substitutions within each interval.

Match Minute as Category C

The second categorical representation of Match Minute, Match Minute CAT C, divided the second half into three broader tactical phases: Early (45–60), Mid (61–75), and Late (76–91+). Compared to Match Minute CAT B, this representation provided a more generalized description of Match Minute. This categorization was constructed to examine whether broader tactical phases of the second half could capture meaningful differences between substitutions. The distribution of substitutions across the Match Minute CAT C variable is illustrated in figure 4 below. The Early category consisted of 302 substitutions, while the Mid and Late categories consisted of 644 and 664 substitutions, respectively.

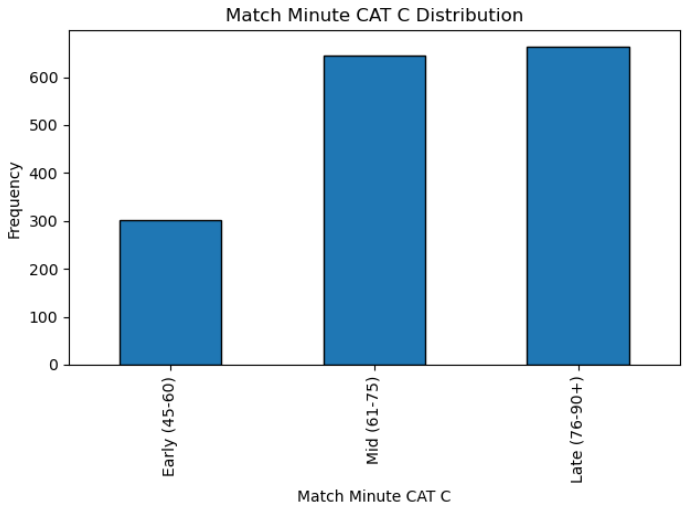
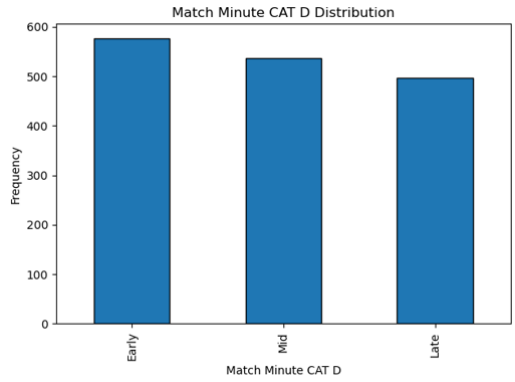


Figure 4: Distribution of substitutions across Match Minute CAT C intervals, where match minutes were grouped into the broader categories Early (45–60), Mid (61–75), and Late (76–90+).

Match Minute as Category D

The third categorical representation of Match Minute, Match Minute CAT D, divided substitutions into three equally sized categories labelled Early, Mid, and Late. Unlike Match Minute CAT C, where the intervals were manually defined based on tactical phases of the match, Match Minute CAT D was constructed using quantile-based categorization. This resulted in categories containing a more similar number of substitutions across the three groups. The distribution of substitutions across the Match Minute CAT D variable is illustrated in figure 5 below.



Match Phase	Number of Substitutions	Mean Match Minute	Min	Max
Early	577	57	45	66
Mid	536	73	67	79
Late	497	85	80	99

Figure 5: Distribution of substitutions across quantile-based Match Minute CAT D phases. The table on the right shows the exact distribution, mean match minute, and minimum and maximum match minute for each phase.

Variable 2: Opponent Strength

The Opponent Strength variable represented the strength difference between the team making the substitution and the opposing team. The variable was included to capture differences in

substitution context related to opponent quality, as teams may apply different tactical strategies depending on whether they are facing stronger or weaker opponents. Opponent Strength was treated as a hybrid variable throughout the study, similarly to Match Minute, as both numerical and categorical representations were used across the different dataset variants.

Opponent Strength as a Numerical Variable

In its numerical form, Opponent Strength was calculated as the difference between the pre-match league ranking of the team making the substitution and the pre-match league ranking of the opposing team. A positive value indicated that the team was facing a stronger opponent. For example, if the team making the substitution was ranked 8th and the opponent was ranked 3rd, Opponent Strength would be +5. In contrast, a negative value indicated that the team was facing a weaker opponent. For example, if the team making the substitution was ranked 3rd and the opponent was ranked 8th, Opponent Strength would be -5. For each match, the league table prior to the match was used to assign pre-match ranks to both teams. A value of 0 was only assigned to substitutions from the first matchday, as teams did not yet have a league position before the opening round. For the same reasons and in the same way as Match Minute, Opponent Strength was standardized, resulting in the numerical variable "Opponent Strength Num". The distribution of substitutions across the Opponent Strength variable is illustrated in figure 6 below. The detailed distribution with the exact number of substitutions across each Opponent Strength value can be found in appendix 3.

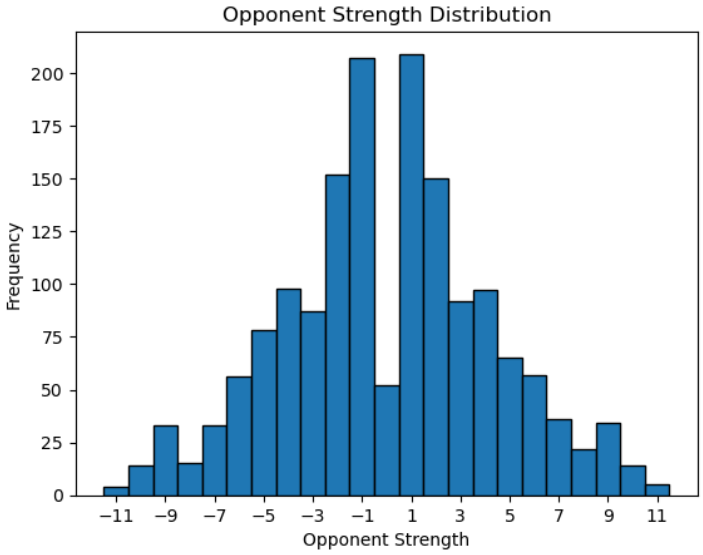


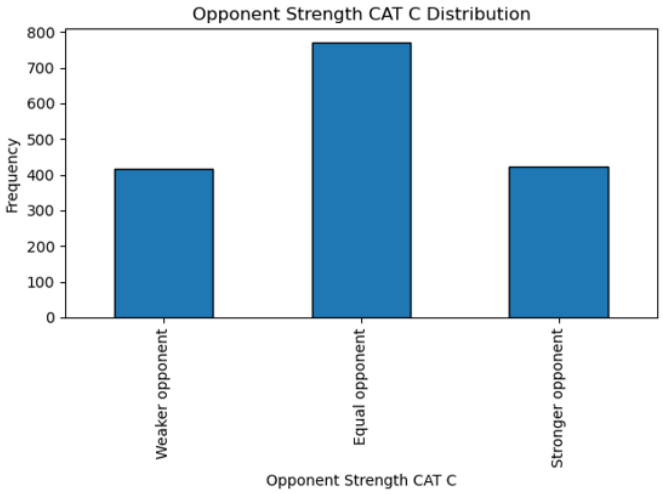
Figure 6: Distribution of substitutions across the numerical Opponent Strength variable, where negative values represent substitutions against weaker opponents and positive values represent substitutions against stronger opponents, based on pre-match league rankings.

Opponent Strength as Category B

The first categorical representation of Opponent Strength, Opponent Strength CAT B, used the same underlying values as the Opponent Strength Num variable, but treated each value as an individual category rather than a continuous numerical variable. As a result, each possible Opponent Strength value, ranging from -11 to +11, represented its own categorical group. This representation was constructed to preserve the detailed distinction between different levels of opponent quality, while removing the continuous numerical relationship between the values. Since Opponent Strength CAT B was based directly on the Opponent Strength Num variable, the distribution of substitutions across the categories was identical to the numerical representation illustrated in figure 6 and appendix 3.

Opponent Strength as Category C

The second categorical representation of Opponent Strength, Opponent Strength CAT C, was constructed in the same manner and for the same purpose as Match Minute CAT C, by grouping the variable into broader contextual categories. More specifically, Opponent Strength was divided into the three categories: "Weaker opponent", "Equal opponent", and "Stronger opponent". The distribution of substitutions across the Opponent Strength CAT C variable is illustrated in figure 7 below.

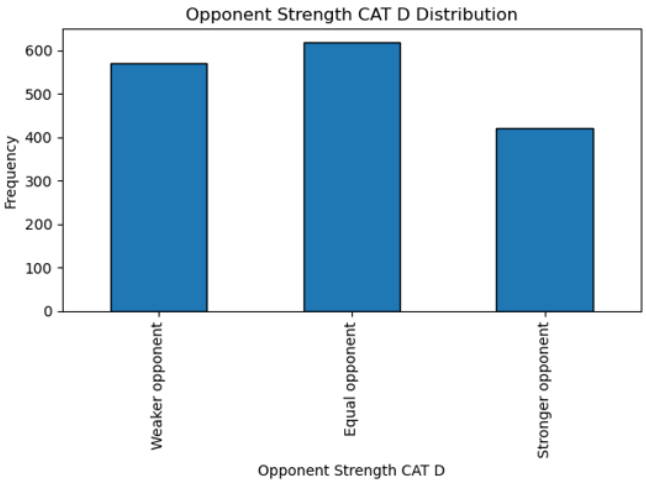


Opponent Strength	Number of Substitutions	Mean Opponent Strength	Min	Max
Weaker	418	-5	-11	-3
Equal	770	0	-2	2
Stronger	422	5	3	11

Figure 7: Distribution of substitutions across Opponent Strength CAT C categories based on pre-match league ranking differences. The table on the right shows the exact distribution, mean Opponent Strength, and minimum and maximum values for each category.

Opponent Strength as Category D

The third categorical representation of Opponent Strength, Opponent Strength CAT D, was constructed using quantile-based categorization, similarly to Match Minute CAT D. This resulted in three categories labelled "Weaker opponent", "Equal opponent", and "Stronger opponent", containing a more balanced number of substitutions across the groups. The distribution of substitutions across the Opponent Strength CAT D variable is illustrated in figure 8 below.

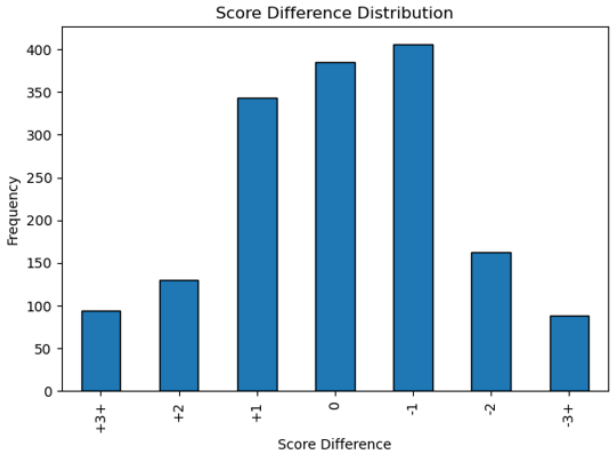


Opponent Strength	Number of Substitutions	Mean Opponent Strength	Min	Max
Weaker	570	-4	-11	-2
Equal	618	0	-1	2
Stronger	422	5	3	11

Figure 8: Distribution of substitutions across quantile-based Opponent Strength CAT D categories. The table on the right shows the exact distribution, mean Opponent Strength, and minimum and maximum values for each category.

Variable 3: Score Difference

The Score Difference variable represented the goal difference at the time of the substitution from the perspective of the team making the substitution. For each substitution, the cumulative number of goals scored by both teams prior to the substitution was calculated, after which the current goal difference for the team making the substitution was determined. To simplify the representation of Score Difference, while still preserving meaningful tactical distinctions, the variable was categorized into seven groups: +3+, +2, +1, 0, -1, -2, and -3+. Positive values represented situations where the team making the substitution was leading, negative values represented situations where the team was trailing, and 0 represented a draw. Goal differences of three or more goals were grouped into the +3+ and -3+ categories to reduce the influence of rare scorelines. The distribution of substitutions across the Score Difference variable is illustrated in figure 9 below.



Score Difference	Number of Substitutions
+3+	94
+2	130
+1	343
0	385
-1	406
-2	163
-3+	89

Figure 9: Distribution of substitutions across Score Difference categories at the time of substitution. The table on the right shows the exact distribution for each score difference category.

Variable 4: Number of Substitutions

The Number of Subs variable represented whether a substitution occurred as an isolated substitution event or as a part of a grouped substitution. The variable was constructed using the Total Seconds variable described previously. Substitutions made by the same team within 60 Total Seconds of each other were grouped into the same substitution sequence. Based on the number of substitutions within a sequence, observations were categorized as "Single", "Double", or "Triple or more". The distribution of substitutions across the Number of Subs variable is illustrated in figure 10 below. The exact distribution consisted of 614 substitutions categorized as "Single", 784 categorized as "Double", and 212 categorized as "Triple or more".

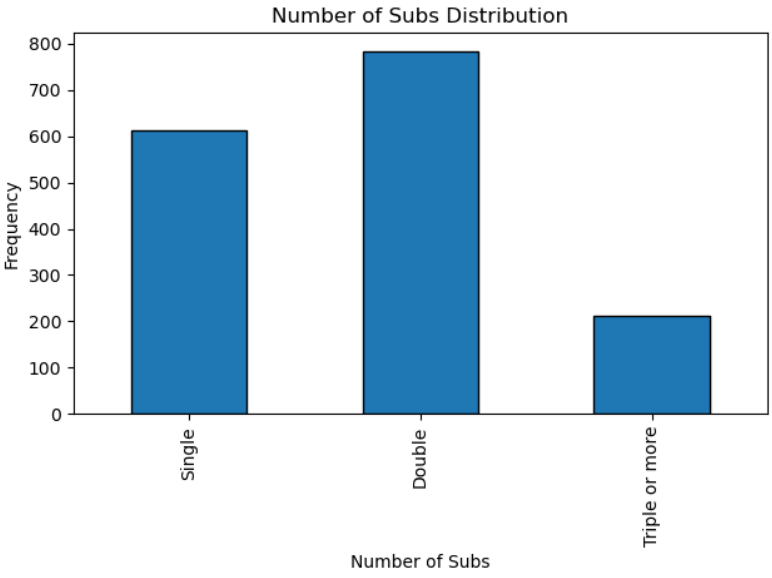


Figure 10: Distribution of substitutions across Number of Subs categories, including Single, Double, and Triple or more substitutions.

Variable 5: Home/Away

The Home/Away variable represented whether the substitution was made by the home team or the away team. Substitutions performed by the home team were categorized as "Home", while substitutions performed by the away team were categorized as "Away". The distribution of substitutions across the Home/Away variable is illustrated in figure 11 below. The exact distribution consisted of 798 home substitutions and 812 away substitutions.

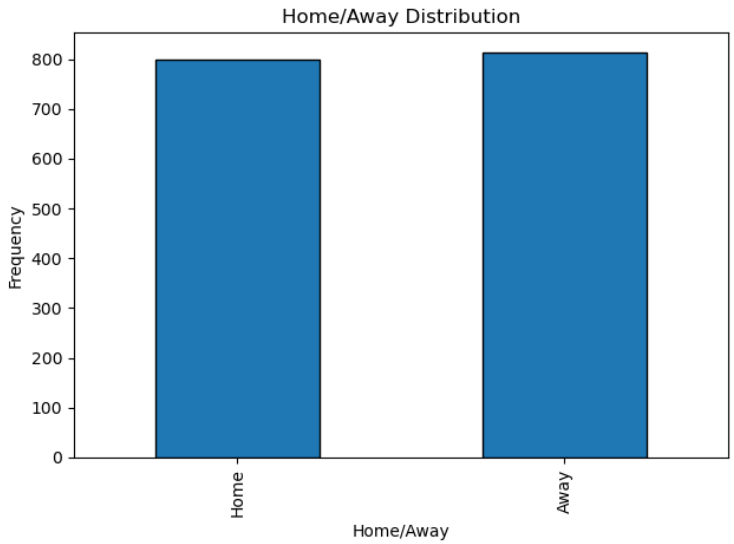


Figure 11: Distribution of substitutions across Home and Away matches.

Variable 6: Response

The Response variable represented whether the opposing team had made a substitution shortly before the current substitution event. The variable was included to capture potential reactive substitution behavior, where teams may adjust their tactical decisions in response to substitutions made by the opponent. The variable was constructed using the Total Seconds variable. For each substitution, the dataset was examined for opposition substitutions occurring within the previous five minutes (300 Total Seconds) before the current substitution event. If at least one opposition substitution had occurred within this time window, the substitution was categorized as "Yes", else, it was categorized as "No". The distribution of substitutions across the Response variable is illustrated in figure 12 below. The exact distribution consisted of 494 substitutions categorized as "Yes" and 1116 categorized as "No".

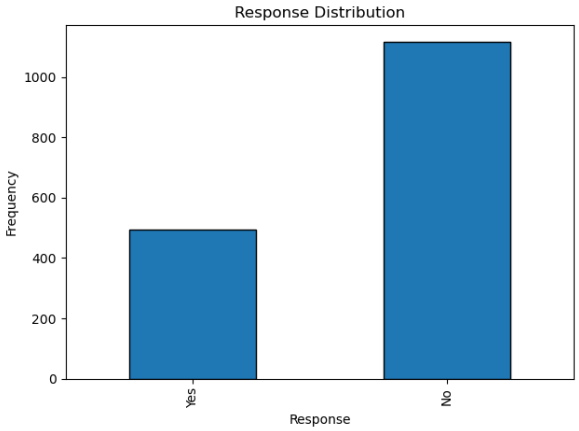


Figure 12: Distribution of substitutions across the Response categories, indicating whether the opposing team had made a substitution within the previous 5 minutes.

Variable 7: Team Size (Red Cards)

The Team Size (Red Cards) variable represented the numerical player balance between the two teams at the time of the substitution. For each substitution, the cumulative number of red cards received by both teams prior to the substitution was calculated. Based on the numerical player balance at the time of the substitution, each substitution was categorized into one of three groups: "Even", "Up a player", or "Down a player". Substitutions were categorized as "Even" when both teams had the same number of players on the field, "Up a player" when the team making the substitution had a numerical advantage, and "Down a player" when the team making the substitution had a numerical disadvantage. The distribution of substitutions across the Team Size (Red Cards) variable is illustrated in figure 13 below. The exact distribution consisted of 1540 substitutions categorized as "Even", 36 categorized as "Up a player", and 34 categorized as "Down a player".

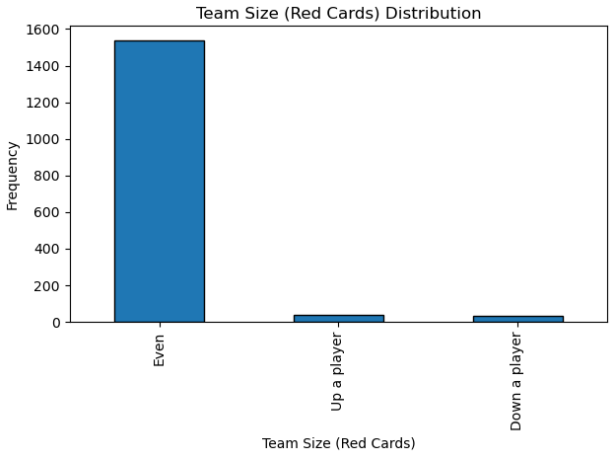


Figure 13: Distribution of substitutions across Team Size (Red Cards) categories, indicating whether teams were playing with an equal number of players, up a player, or down a player.

xG Score and Impact Score

After each substitution had been assigned values for the seven variables, the final step in the construction of the primary dataset consisted of assigning an impact measure representing the effectiveness of each substitution. This was done through the construction of an xG-based impact score. xG data were extracted from Optas F70 files, which contained xG-values for individual shot events throughout each match. These xG values were subsequently aggregated around each substitution event to construct an xG score representing the short-term offensive and defensive impact following a substitution. The xG score was calculated using the Total Seconds variable. For each substitution, xG values were examined within two separate five-minute windows: five minutes before the substitution and five minutes after the substitution. Within each time window, xG values were calculated both for the team making the substitution and for the opposing team. This resulted in four separate xG measures for each substitution: xG For Before, xG Against Before, xG For After, and xG Against After. The final xG Score was calculated as:

$$xG\ Score = (xG\ For\ After - xG\ For\ Before) - (xG\ Against\ After - xG\ Against\ Before)$$

As a result, positive xG score values indicated that the substitution was associated with an improved xG balance in the five minutes following the substitution, while negative values indicated a worsening xG balance. In this way, xG score attempted to capture the short-term impact of substitutions on match dynamics from both an offensive and defensive perspective. The distribution of substitutions across the xG score variable is illustrated in figure 14 below.

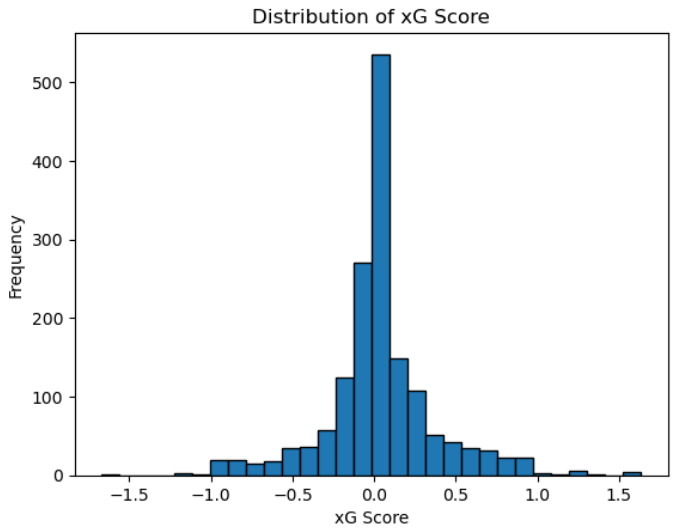


Figure 14: Distribution of xG Scores calculated from the change in xG before and after substitutions.

To construct the final outcome variable used throughout the clustering and classification analyses, the continuous xG score was converted into a categorical impact score ranging from 1 to 5. This categorization was performed using quantile-based binning, where the xG score distribution was divided into five approximately equally sized groups. As a result, substitutions with the lowest xG scores were assigned with an impact score of 1, while substitutions with the highest xG scores were assigned an impact score of 5. The relationship between xG score and impact score, including the distribution and range of xG values across the five impact score categories, is illustrated in table 1 below.

Impact Score	Number of Substitutions	Mean xG Score	Min xG Score	Max xG Score
1	322	-0.411	-1.668	-0.131
2	323	-0.054	-0.131	-0.004
3	322	0.014	-0.003	0.047
4	321	0.104	0.047	0.208
5	322	0.511	0.021	1.636

Table 1: Distribution of substitutions across the five Impact Score categories derived from quantile-based xG Score intervals. The table shows the number of substitutions, mean xG Score, and minimum and maximum xG Score for each Impact Score category.

Dataset Variants

Based on the seven contextual match variables, a total of ten different dataset variants were constructed and used throughout the clustering and classification analyses. The purpose of constructing multiple variants of the dataset, was to examine how different representations of the variables influenced the clustering structure, variable importance, and classification performance. The ten variants were created through different representations of the two hybrid variables, Match Minute and Opponent Strength. Depending on the variant, these variables were included either in their standardized numerical form or in one of their three categorical forms. The remaining five contextual variables were kept constant across all variants. To create a clear overview of the methodological structure, the ten variants were grouped into four overall variant categories based on the representations of Match Minute and Opponent Strength used in each group. The variant categories are presented in more detail below. A summary of the variants can be seen at table 2. The summary only includes Match Minute and Opponent Strength, as the other variables are used as categorical variables across all ten variants.

Dataset Variant A

In variant A, both Match Minute and Opponent Strength were included in their standardized numerical forms, referred to as Match Minute Num and Opponent Strength Num. The remaining five contextual variables were included in their original categorical forms. Because Match Minute and Opponent Strength were included as standardized numerical variables, the variant retained the exact relative distances between observations. For example, substitutions made in minutes 60 and 65 were treated as more similar than substitutions made in minute 60 and 85. Similarly, smaller differences in opponent quality were preserved numerically rather than grouped into broader categories. This allowed variant A to represent the most detailed interpretation of Match Minute and Opponent Strength among all ten variants.

Dataset Variants B, C and D

Variants B, C and D represented the fully categorical variants included in the study. In these variants, both Match Minute and Opponent Strength were included in their categorical forms. As a result, all seven variables were treated as categorical variables within these variants. The three variants differed according to which categorical representations of Match Minute and Opponent Strength were applied. Variant B used Match Minute CAT B and Opponent Strength CAT B, variant C used Match Minute CAT C and Opponent Strength CAT C, and Variant D used Match Minute CAT D and Opponent Strength CAT D.

Variants EB, EC and ED

Variants EB, EC and ED represented the hybrid variants, combining numerical and categorical representations of the two hybrid variables. In these variants, Match Minute was included in its standardized numerical form, Match Minute Num, while Opponent Strength was included in its three categorical representations. The three variants differed according to which categorical representation of Opponent Strength was applied. Variant EB used Opponent Strength CAT B, Variant EC used Opponent Strength CAT C, and Variant ED used Opponent Strength CAT D.

Variants FB, FC and FD

In Variants FB, FC and FD, Opponent Strength was included in its standardized numerical form, Opponent Strength Num, while Match Minute was included in its categorical representations. Variant FB used Match Minute CAT B, Variant FC used Match Minute CAT C, and Variant FD used Match Minute CAT D.

Variant	Description
Variant A	Using Match Minute Num and Opponent Strength Num
Variant B	Using Match Minute Cat B and Opponent Strength Cat B
Variant C	Using Match Minute Cat C and Opponent Strength Cat C
Variant D	Using Match Minute Cat D and Opponent Strength Cat D
Variant EB	Using Match Minute Num and Opponent Strength Cat B
Variant EC	Using Match Minute Num and Opponent Strength Cat C
Variant ED	Using Match Minute Num and Opponent Strength Cat D
Variant FB	Using Opponent Strength Num and Match Minute Cat B
Variant FC	Using Opponent Strength Num and Match Minute Cat C
Variant FD	Using Opponent Strength Num and Match Minute Cat D

Table 2: Overview of the 10 dataset variants used in the study, including the different numerical and categorical representations of the Match Minute and Opponent Strength variables.

Clustering

The clustering analysis was the first analytical stage of the study following the construction of the primary dataset and the different dataset variants. The purpose of the clustering analysis was to see whether distinct substitution patterns could be identified based on the seven match variables. Rather than defining substitution types beforehand, the clustering approach allowed substitutions with similar contextual characteristics to be grouped together in a more data-driven manner. Because the dataset contained both numerical and categorical representations of variables across the different variants, two different clustering approaches were applied. These consisted of K-prototypes clustering and K-modes clustering (Huang, 1998), which will be described in more detail in the following sections. An important methodological consideration in clustering analysis concerns the selection of the number of clusters, represented by the value K. To evaluate suitable cluster solutions across the different variants, the elbow method was applied (*Elbow Method*). The elbow method examines how the clustering cost changes as the number of clusters increases, where an optimal cluster solution is typically identified near the point where further increases in K, result in smaller reductions in clustering cost. As an example, the elbow method for Variant D is given in figure 15 below. The elbow plot showed that the clustering cost decreased up to around six clusters. After this point, the decreases became smaller as additional clusters were added. Six clusters were considered an appropriate solution for Variant D, as the curve began to level out around this point. The same procedure was subsequently applied across all ten variants to evaluate and determine the most appropriate number of clusters for each variant.

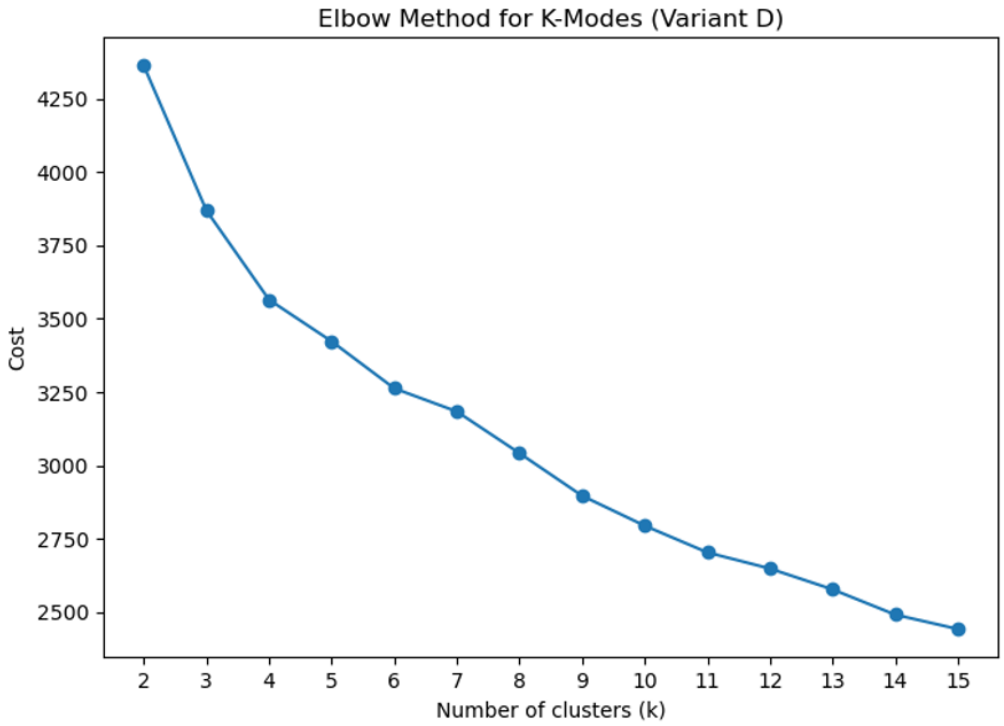


Figure 15: Example of elbow method results for the K-Modes clustering analysis in Variant D, showing the relationship between the number of clusters (k) and clustering cost.

K-prototypes and K-modes Clustering

K-prototypes clustering was applied for the variants containing both numerical and categorical variables. More specifically, this clustering approach was used for Variants A, EB, EC, ED, FB, FC and FD. For the numerical variables included in the different variants, similarity between substitutions was calculated using Euclidean distance, as done in K-means clustering. Euclidean distance measures the numerical distance between observations across the included numerical variables. For example, substitutions made in minutes 60 and 62 would have a Euclidean distance of 2, whereas substitutions made in minutes 60 and 85 would have a Euclidean distance of 25. As a result, substitutions closer in timing were considered more similar within the clustering structure. For the categorical variables, similarity was calculated using matching dissimilarity, as done in K-modes clustering, where observations were compared based on whether categorical values matched or differed between substitutions. For example, substitutions sharing the same categorical values, such as both being categorized as “Home” and “Single Substitution”, would be considered more similar than substitutions with differing categorical values. In this way, substitutions with more similar characteristics were grouped closer together within the clustering structure. K-modes clustering was applied for the fully categorical variants, consisting of Variants B, C and D, where all seven variables were

represented as categorical variables. Unlike K-prototypes clustering, K-modes clustering is designed specifically for categorical data and therefore does not include numerical distance calculations.

Following the clustering, each cluster within the different variants was assigned a mean impact score based on the substitutions belonging to the given cluster. This allowed the clusters to be evaluated according to their average short-term impact on match dynamics. In addition, the clusters within each variant were compared and described based on their dominant contextual characteristics. As a result, clusters could be interpreted as distinct substitution patterns, such as “Late Non-Reactive Double Substitutions in Winning Away Games vs Weaker Opponents” or “Mid-Second Half Reactive Single Substitutions in Losing Home Games vs Equal Opponents”. This will be described more in the results section.

Adjusted Rand Index

Once the clustering analyses had been completed, the next step was to examine how strongly each of the seven variables influenced the clustering structure across the different variants. More specifically, a leave-one-out approach was used, where clustering was first performed using all included variables within a given variant. Afterwards, one variable at a time was removed, after which the clustering procedure was repeated and compared to the original clustering solution. The similarity between the original and modified cluster solutions was evaluated using ARI (SciKit Learn - Adjusted Rand Index). ARI measures the similarity between two clustering structures while adjusting for similarity occurring by chance. The ARI scale ranges from 1 to -0.5, where a value of 1 indicates identical clustering solutions, values close to 0 indicate similarity expected from random labeling, and negative values indicate especially discordant clustering solutions. Within the present study, lower ARI values following the removal of a variable indicated that the clustering structure changed more, suggesting that the removed variable had a greater influence on the clustering solution. In contrast, higher ARI values indicated that the clustering structure remained more similar after removal of the variable, suggesting a smaller influence on the clustering structure. As a result, each of the seven variables received an ARI score within each of the ten variants. In addition, a mean ARI score across all ten variants was calculated for each variable to provide an overall indication of its influence on the clustering structure. These results will also be described further in the results section.

Random Forest Analysis

Following the clustering and ARI analyses, random forest analysis was applied to further examine the relationship between the seven variables and the impact score. While the ARI analysis focused on how strongly the variables influenced the clustering structure, the random forest analysis focused on how strongly the variables contributed to predicting the impact score across the different variants. For each of the ten variants, the seven variables were used as input variables, while the impact score served as the target variable. Since several of the variables were categorical, categorical representations were transformed using one-hot encoding (Samuels, 2024) prior to the random forest analyses, as the Scikit-learn random forest framework cannot directly process categorical variables (*1.10. Decision Trees*). To evaluate variant performance, the dataset was divided into training, validation, and test datasets using stratified sampling to preserve the distribution of the impact score categories across the different subsets. Approximately 70% of the observations were assigned to the training dataset, while the remaining observations were divided equally between the validation and test datasets. The random forest analyses subsequently formed the basis for the feature importance and classification analyses described in the following sections.

Feature Importance

As part of the Random Forest analysis, feature importance (*Feature Importance with Random Forests*) values were calculated to examine how strongly each of the seven variables contributed to predicting the impact score across the ten variants. Variables contributing more to the prediction of the impact score receive higher feature importance values. Since several variables were transformed using one-hot encoding prior to the analyses, feature importance values were initially calculated for the individual encoded variables. Afterwards, the encoded feature importance values belonging to the same original variable were combined to obtain an overall feature importance score for each of the seven variables. As a result, each variable received a feature importance score within each of the ten variants. In addition, a mean feature importance score across all ten variants was calculated for each variable to provide an overall indication of its contribution to predicting the impact score. These results will be described further in the results section.

Classification

As the second part of the random forest analysis, classification analysis was performed to examine how accurately the seven variables could predict the impact score across the different variants. More specifically, the random forest variants attempted to classify each substitution into one of the five impact score categories based on the variables. Variant performance was evaluated using accuracy, precision, recall, and F1-score. Accuracy measured the overall proportion of correctly classified substitutions across the five impact score categories. Precision measured the proportion of correctly classified substitutions among the substitutions predicted to belong to a given impact score category. Recall measured the proportion of substitutions within a given impact score category that were correctly identified by the variant. F1-score combined precision and recall into a single performance measure. In addition, classification reports were generated for each variant to examine performance across the individual impact score categories. These results will be described further in the results section.

Sensitivity Analysis

As the final stage of the methodological workflow, sensitivity analyses were made to examine the robustness of the methodological choices and results throughout the study. More specifically, four separate sensitivity analyses were made focusing on the clustering structure, ARI analysis, feature importance analysis, and classification performance. Firstly, sensitivity analyses were performed on the clustering by examining alternative cluster solutions around the selected K values identified through the elbow method. This was done to evaluate whether small changes in the number of clusters influenced the overall clustering structure and the resulting cluster mean impact scores. Secondly, sensitivity analyses were performed on the ARI results using the alternative cluster solutions described above. This allowed the study to examine whether the observed influence of the seven variables on the clustering structure remained relatively stable across different cluster solutions. Thirdly, sensitivity analyses were made for the feature importance results by reducing the number of impact score categories from five to three and two categories, respectively. This was done to see whether the feature importance results changed when the number of impact score categories was reduced. Finally, sensitivity analyses were also made for the classification analyses using the same three-category and two-category versions of the impact score. Test accuracy was calculated to examine how the classification results changed when fewer impact score categories were used.

Results

Overview of The Results Section

Due to the large number of variants and analytical outputs generated throughout the study, the results section primarily focuses on the overall patterns and key findings observed across the variants. All detailed results for the ten dataset variants are presented in appendices 4–13, respectively. This includes elbow method plots, cluster structures and descriptions, variable distributions across clusters, ARI and feature importance results, as well as complete random forest classification outputs, including validation/test performance and detailed precision, recall, and F1-scores. The results section is therefore structured around the main analytical stages of the study. Firstly, the clustering results are presented, including the selected number of clusters and the differences in mean impact score across the identified cluster structures. Secondly, the ARI results are presented to examine how strongly the seven variables influenced the clustering structures across the variants. Thirdly, the feature importance results are presented to evaluate which variables contributed most to predicting the impact score. Fourthly, the classification results are presented through overall variant performance and impact score specific classification metrics. Finally, the results section concludes with the sensitivity analyses, which evaluate the robustness of the clustering structures, variable influence, feature importance results, and classification performance across alternative methodological configurations.

Clustering

The clustering results across the ten variants are presented in figure 16 below. The detailed cluster descriptions can be found in appendices 4-13. The figure shows the minimum and maximum mean impact scores identified among the clusters within each variant, while the number displayed above each line represents the difference between the minimum and maximum mean impact scores. As described previously, the number of clusters for each variant was selected using the elbow method. Across all ten variants, the selected cluster solutions ranged from 6 to 8 clusters.

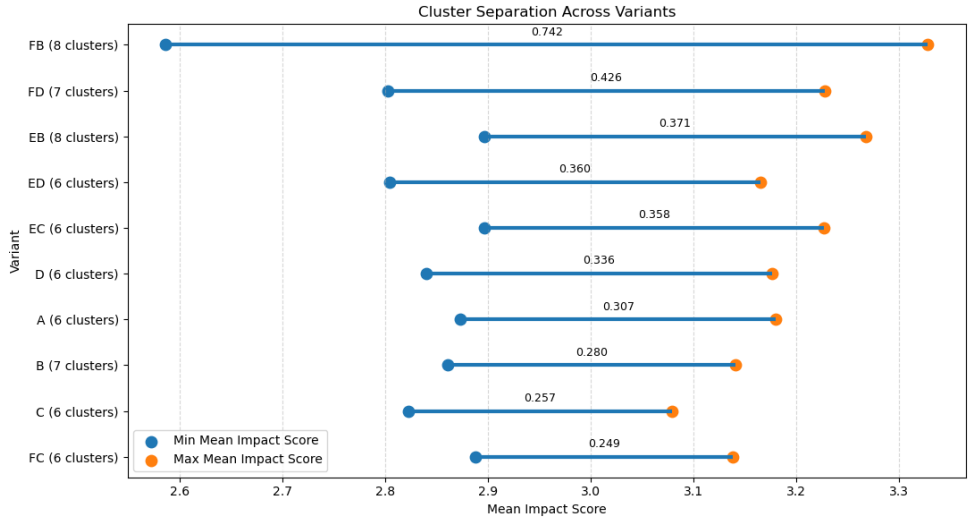


Figure 16: Cluster separation across all dataset variants, showing the difference between the minimum and maximum mean Impact Score observed across clusters within each variant.

Overall, the clusters showed several similar substitution patterns and descriptions across the different variants. Table 3 summarizes how often each substitution characteristic was included in the clusters with the highest and lowest mean impact scores across the ten variants. Team Size (Red Cards) was excluded from the summarized profiles, as all identified clusters were dominated by substitutions performed under even team-size conditions.

Substitution Type	Included in Best Impact Score	Included in Worst Impact Score
Match Minute		
Early	2	0
Mid	8	4
Late	0	6
Response		
Non-Reactive	10	6
Reactive	0	4
Number of Subs		
Single	1	5
Double	8	4
Triple or more	1	1
Score Difference		
Drawing	3	3
Loosing	4	2
Winning	3	5
Home/Away		
Home Game	4	7
Away Game	6	3
Opponent Strength		
Weaker Opponent	6	3
Equal Opponent	2	4
Stronger Opponent	2	3

Table 3: Overview of the most frequently represented substitution characteristics within the clusters with the highest and lowest mean Impact Scores across all dataset variants.

Across the six variables included in the summarized cluster profiles, a total of 324 substitution combinations were theoretically possible to make. However, due to the selected number of clusters across the ten variants, as presented in Figure 16 above, a maximum of only 66 clusters could therefore be made. Despite this large theoretical combination of substitutions, several similar substitution patterns and descriptions still appeared across the different variants. Based on the summarized frequencies in Table 3, it was possible to generate hypothetical highest and lowest impact substitution profiles across the variants. The hypothetical highest-impact substitution could be described as a "Mid-Second Half, Non-Reactive, Double Substitution, in Losing Away Games vs Weaker Opponents", while the hypothetical lowest-impact substitution could be described as a "Late, Non-Reactive, Single Substitution, in Winning Home Games vs Equal Opponents".

Interestingly, the real cluster with the highest mean impact score identified across all variants was found in Variant FB (Cluster 6) and was described as "Mid-Second Half, Non-Reactive, Double Substitutions, in Winning Away Games vs Weaker Opponents", differing only in Losing/Winning from the hypothetical highest-impact profile. In addition, cluster 6 in Variant FD had the third-highest mean impact score across all identified clusters and was completely identical with the hypothetical highest-impact substitution. Likewise, cluster 2 in Variant A had one of the lowest mean impact scores overall and were identical with the hypothetical lowest-impact substitution. Despite the large number of possible substitution combinations, the highest and lowest mean impact score clusters identified across the variants therefore remained highly similar to the hypothetical highest and lowest impact substitution profiles.

Adjusted Rand Index

The variables ARI results across the ten variants are presented in figure 17 below. Lower ARI values indicate that removing a variable resulted in larger changes to the clustering structure, suggesting a stronger influence on the final cluster solution. In contrast, higher ARI values indicate that removing a variable resulted in smaller changes to the clustering structure, suggesting a more limited influence on the final clusters. To compare the influence of different representations of Match Minute and Opponent Strength, the variables were separated into their original representations, numerical representations (Num) and categorical representations (Cat).

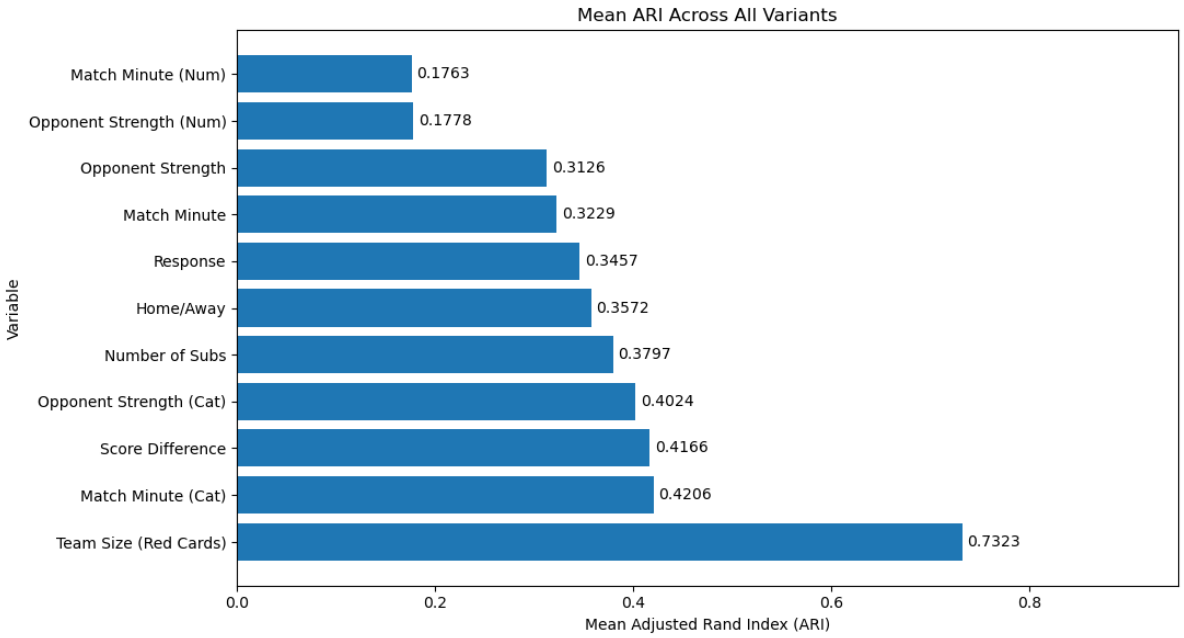


Figure 17: Mean Adjusted Rand Index (ARI) for each variable across all dataset variants. Match Minute and Opponent Strength were included as their overall, numerical and categorical representations.

Overall, the numerical representations of Opponent Strength and Match Minute produced the lowest mean ARI values across the variants, with mean ARI values of 0.1778 and 0.1763, respectively. This indicates that the numerical versions of these variables had the strongest overall influence on the clustering structures. The categorical representations of Opponent Strength and Match Minute generally produced higher ARI values compared to their numerical counterparts. Opponent Strength (Cat) and Match Minute (Cat) produced mean ARI values of 0.4024 and 0.4206, respectively, while the combined overall representations produced mean ARI values of 0.3126 and 0.3229. This indicates that the numerical representations contributed more strongly to the clustering structures than the categorical representations. The remaining contextual variables generally produced moderate ARI values across the variants. Score Difference produced a mean ARI value of 0.4166, while Number of Subs, Home/Away, and Response produced mean ARI values between 0.3457 and 0.3572. Team Size (Red Cards) produced the highest overall mean ARI value of 0.7323. In addition, five variants produced ARI values of 1.000 when this variable was removed. This indicates that removing Team Size (Red Cards) resulted in either very limited or no observable changes to the clustering structures in these variants.

Feature Importance

The variables mean feature importance results across the ten variants are presented in figure 18 below. Higher feature importance values indicate that a variable contributed more strongly to predicting the impact score in the random forest analyses. As in the ARI results, Match Minute and Opponent Strength were separated into their original form, numerical representations (Num) and categorical representations (Cat) to compare the influence of the different variable representations across the variants.

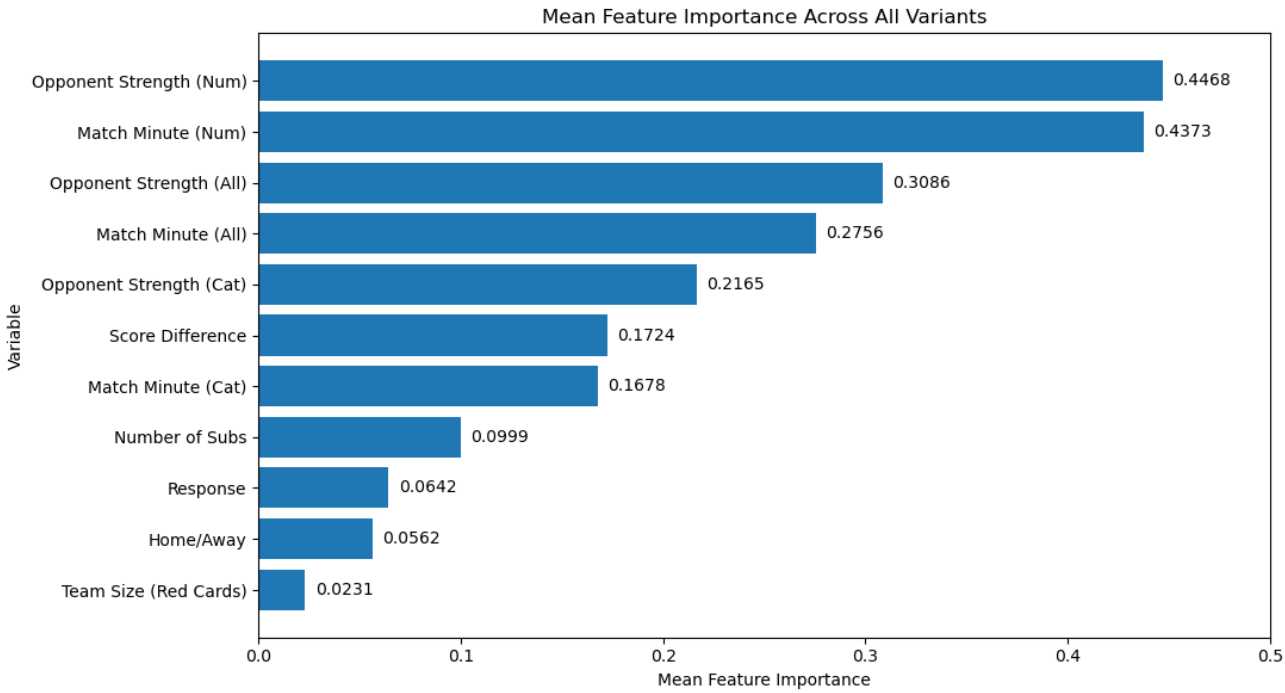


Figure 18: Mean feature importance for each variable across all dataset variants based on the Random Forest analysis. Match Minute and Opponent Strength were included as their overall, numerical and categorical representations.

Overall, the numerical representations of Opponent Strength and Match Minute produced the highest mean feature importance values across the variants, with mean values of 0.4468 and 0.4373, respectively. Several of the hybrid variants produced particularly high feature importance values for the numerical variables, including 0.5385 for Match Minute (Num) in Variant EC and 0.5336 for Opponent Strength (Num) in Variant FC. The combined overall representations of Opponent Strength and Match Minute produced lower mean feature importance values of 0.3086 and 0.2756, respectively, while the categorical representations produced lower values of 0.2165 for Opponent Strength (Cat) and 0.1678 for Match Minute (Cat). This indicates that the numerical representations generally contributed more strongly to classification performance than the categorical representations. Among the remaining

contextual variables, Score Difference produced the highest overall mean feature importance value of 0.1724. In the fully categorical variants, Score Difference produced particularly high feature importance values of 0.2546 in Variant C and 0.2684 in Variant D. Number of Subs produced relatively consistent feature importance values, with a mean value of 0.0999. Response and Home/Away generally produced lower feature importance values, with mean values of 0.0642 and 0.0562, respectively. Team Size (Red Cards) consistently produced the lowest feature importance values across all variants, with a mean value of 0.0231.

Classification

The random forest classification results across the ten variants are presented in figure 19 below. Classification performance was evaluated using random forest test accuracy and macro F1-score. As mentioned earlier, detailed classification reports for all variants are presented in appendices 4-13.

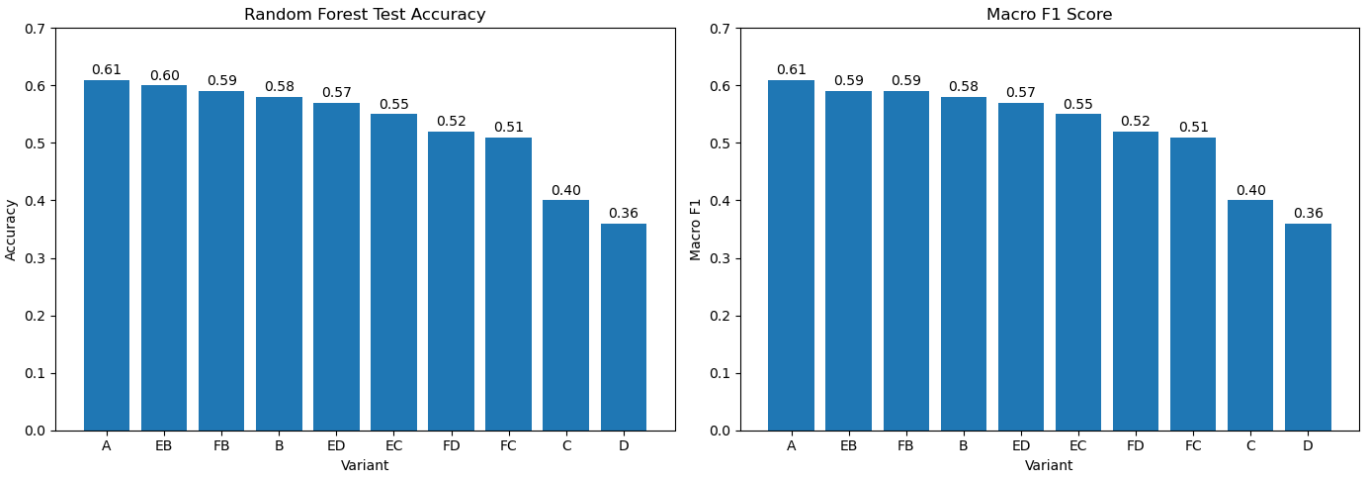


Figure 19: Random Forest test accuracy and macro F1-score across all dataset variants.

Overall, the highest classification performances were observed in the variants containing numerical representations of Match Minute and/or Opponent Strength. Variant A produced the highest overall classification performance, with both a random forest test accuracy and macro F1-score of 0.61. Variants EB and FB also produced relatively high performances, with test accuracies of 0.60 and 0.59, respectively. In contrast, the fully categorical Variants C and D produced the lowest classification performances across the variants. Variant C produced a test accuracy and macro F1-score of 0.40, while Variant D produced values of 0.36 for both metrics. The remaining variants generally produced intermediate performances ranging between 0.51

and 0.58. The mean classification results for the individual impact score categories are presented in figure 20 below. Overall, impact score 5 produced the highest mean F1-score of 0.58 across all variants, while impact score 1 produced the lowest mean F1-score of 0.47. Impact score 2 produced the second highest mean F1-score of 0.55, while impact scores 3 and 4 both produced mean F1-scores of 0.52.

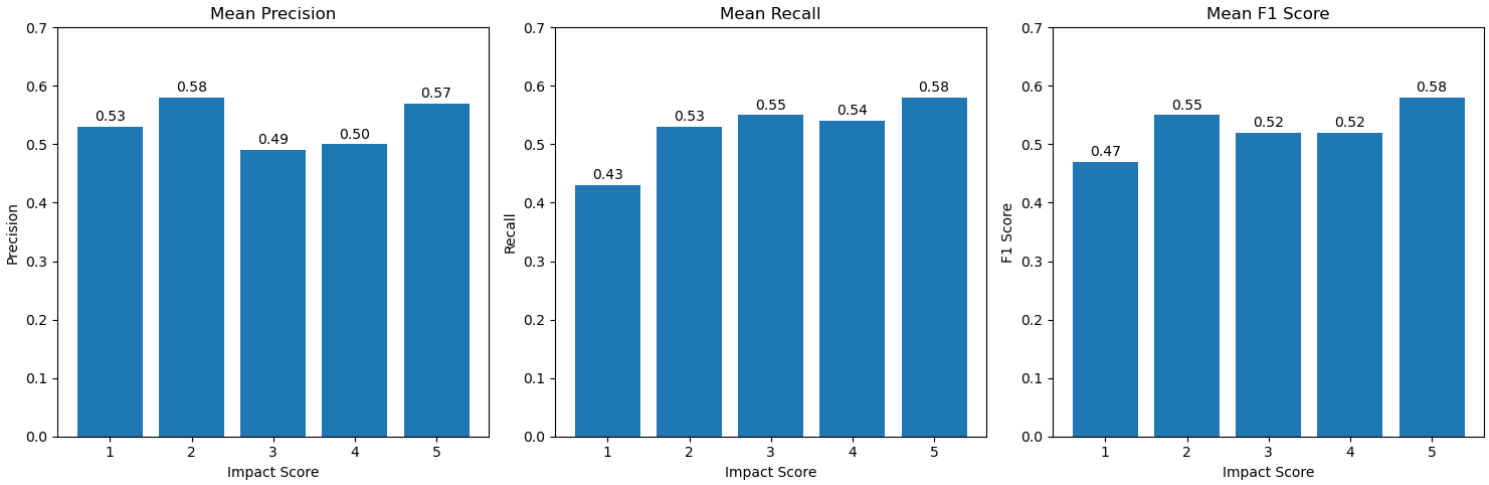


Figure 20: Mean precision, recall, and F1-score for each Impact Score category across all dataset variants.

Sensitivity Analysis

Clustering

The sensitivity analysis of the clustering results is presented in Table 4. The analysis compared the original and alternative cluster solutions across the ten variants. The "Original Difference" and "Alternative Difference" columns represent the difference between the lowest and highest mean impact score across the clusters within each variant, while the "Change" column represents the difference between the original and alternative impact score differences across the clusters. For example, if the cluster with the lowest mean impact score had a value of 2.85 and the cluster with the highest mean impact score had a value of 3.15, the resulting impact score difference across the clusters would be 0.30. Detailed results, including minimum and maximum mean impact scores for each variant, are presented in Appendix 14.

Variant	Original k	Alternative k	Original Difference	Alternative Difference	Change
A	6	7	0.307	0.233	- 0.074
B	7	5	0.208	0.247	- 0.033
C	6	9	0.257	0.516	+ 0.259
D	6	4	0.336	0.354	+ 0.018
EB	8	10	0.371	0.592	+ 0.221
EC	6	9	0.358	0.520	+ 0.162
ED	6	8	0.360	0.430	+ 0.070
FB	8	10	0.742	0.431	- 0.311
FC	6	8	0.249	0.318	+ 0.069
FD	7	8	0.426	0.209	- 0.217

Table 4: Sensitivity analysis of cluster separation across the original and alternative numbers of clusters (k) for all dataset variants. Original Difference and Alternative Difference represent the difference between the minimum and maximum mean Impact Score across clusters, while Change represents the difference between the two separation values.

Overall, the sensitivity analysis showed both positive and negative changes in the impact score differences across the clusters when alternative numbers of clusters were applied. Several variants produced relatively limited changes between the original and alternative cluster solutions, indicating relatively stable impact score differences across different numbers of clusters. Variants D, B, FC, and ED showed the smallest overall changes, with changes of +0.018, -0.033, +0.069, and +0.070, respectively. Other variants produced larger changes in the impact score differences across the clusters when alternative cluster solutions were applied. The largest positive increases were observed in Variants C and EB, which increased by +0.259 and +0.221, respectively. In contrast, the largest decreases were observed in Variants FB and FD, which decreased by -0.311 and -0.217, respectively. Overall, most of the variants retained relatively similar impact score differences across the original and alternative cluster solutions, although some variants showed moderate sensitivity to changes in the number of clusters.

Adjusted Rand Index

The sensitivity analysis of the ARI results is presented in figure 21. The analysis compares the mean ARI values across the original cluster solutions, the alternative cluster solutions that were mentioned in the section above, and the combined mean values across both cluster solutions.

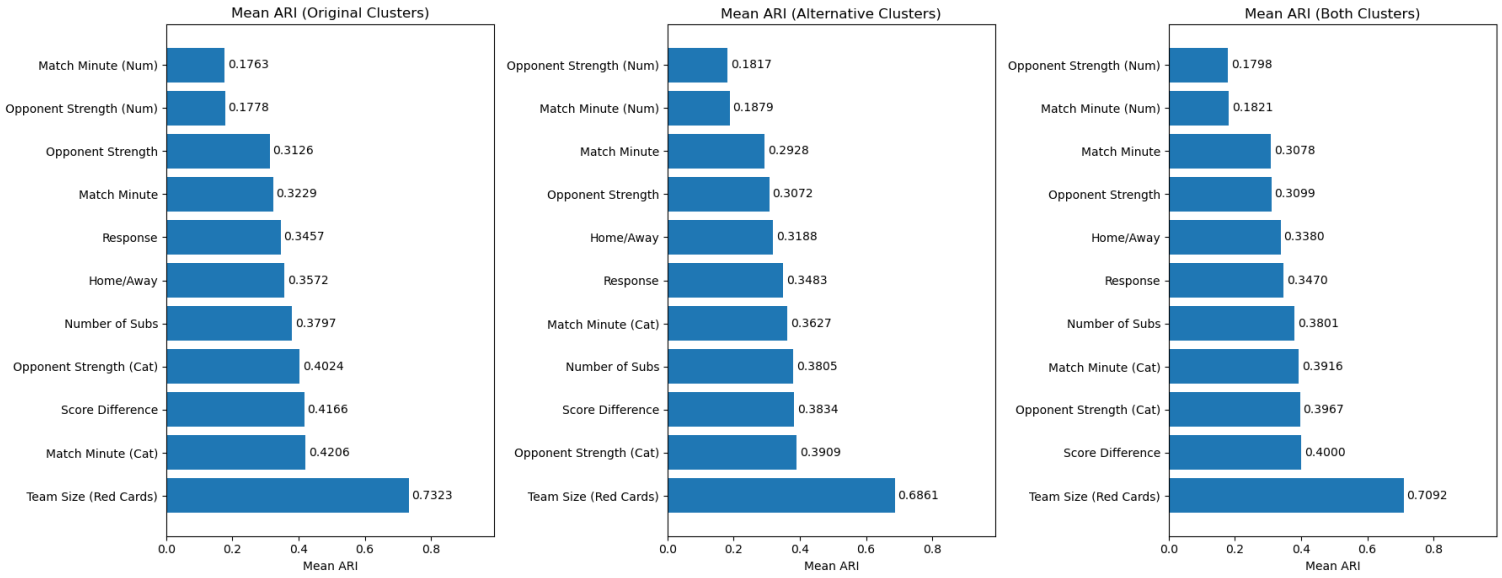


Figure 21: Mean Adjusted Rand Index (ARI) across all variables for the original clusters, alternative clusters, and the combined mean of both clustering solutions.

Overall, the ARI results remained relatively consistent across the original, alternative, and the mean cluster solutions, with the same overall variable patterns being observed across all three analyses. The numerical representations of Opponent Strength and Match Minute consistently produced the lowest ARI values, with combined mean ARI values of 0.1798 and 0.1821, respectively, indicating that these variables remained the strongest contributors to the clustering structures. In contrast, the categorical representations of Match Minute and Opponent Strength consistently produced higher ARI values, with combined mean values of 0.3099 and 0.3078, respectively. Some variables showed more noticeable differences between the original and alternative clusters. The largest decrease was observed for Match Minute (Cat), which decreased from 0.4206 in the original clusters to 0.3627 in the alternative clusters. Smaller decreases were also observed for Team Size (Red Cards), Score Difference and Home/Away. Despite these changes, the overall ranking and distribution of the variables remained relatively stable across the three analyses.

Feature Importance

The sensitivity analysis of the feature importance results is presented in figure 22. The analysis compared mean feature importance values across the original five impact score categories, the reduced three-category impact score, the reduced two-category impact score, and the combined mean across all three categorizations.

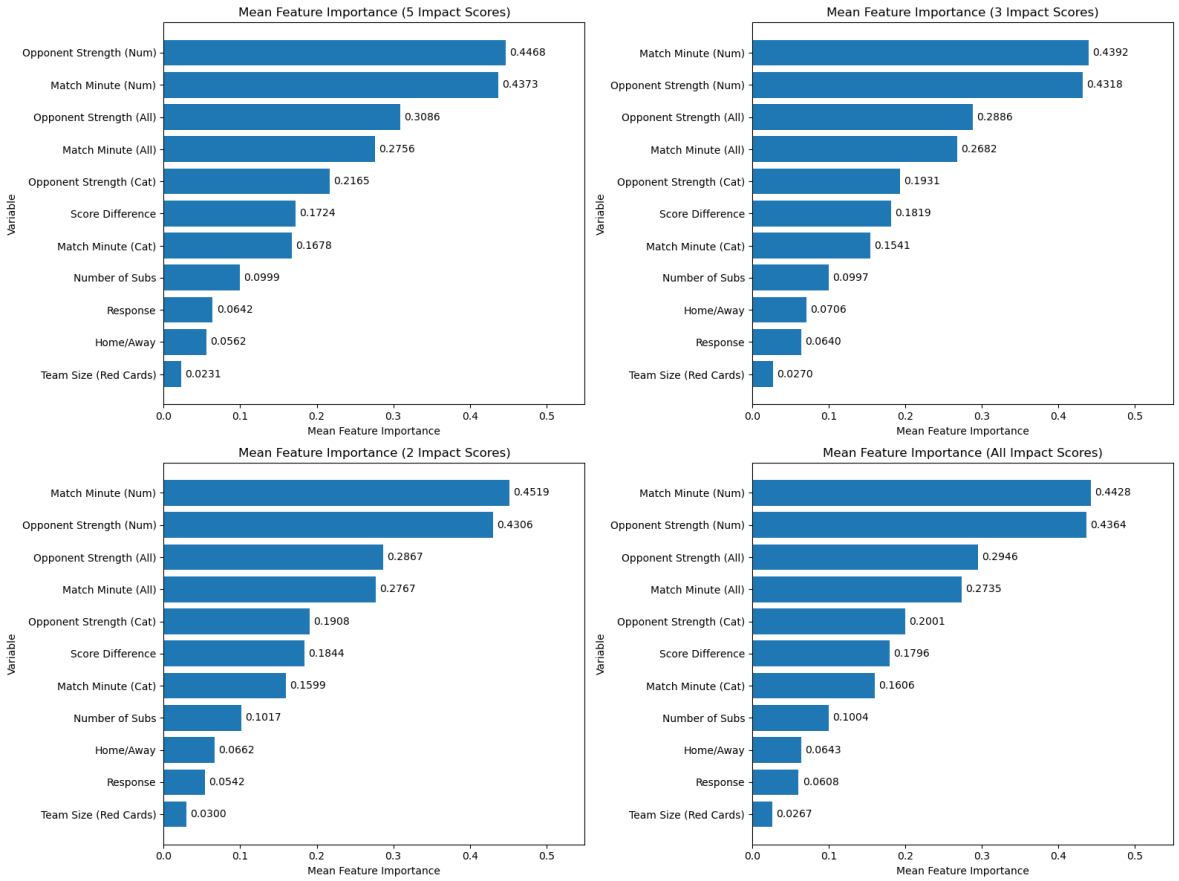


Figure 22: Mean feature importance across all dataset variants for the Random Forest analyses using 5, 3, and 2 Impact Score categories, as well as the combined mean across all analyses.

Overall, the feature importance results remained highly stable across the different impact score categorizations. Match Minute Num and Opponent Strength Num consistently produced the highest mean feature importance values across all four analyses. In the combined results, Match Minute Num produced a mean feature importance value of 0.4428, while Opponent Strength Num produced a value of 0.4364. The overall representations of Opponent Strength and Match Minute also remained among the most important variables across the sensitivity analyses, with combined mean feature importance values of 0.2946 and 0.2735, respectively. The categorical representations of Opponent Strength and Match Minute produced lower combined mean feature importance values of 0.2001 and 0.1606. The remaining contextual variables showed lower and relatively stable feature importance values across the different impact score

categorizations. Score Difference produced a combined mean feature importance value of 0.1796, while Number of Subs produced a value of 0.1004. Home/Away and Response produced lower combined mean values of 0.0643 and 0.0608, respectively, while Team Size (Red Cards) consistently produced the lowest feature importance values across all analyses, with a combined mean feature importance value of 0.0267. Overall, the sensitivity analysis showed that reducing the number of impact score categories did not substantially change the ranking of the most and least important variables.

Classification

Since the macro F1 scores and random forest test accuracies produced highly similar results across all variants, the sensitivity analysis of the classification performance was made using random forest test accuracy only. The sensitivity analysis results are presented in figure 23 and compare the test accuracies across the original five-category impact score, the reduced three-category impact score, the reduced two-category impact score, and the combined mean test accuracy across all categorizations.

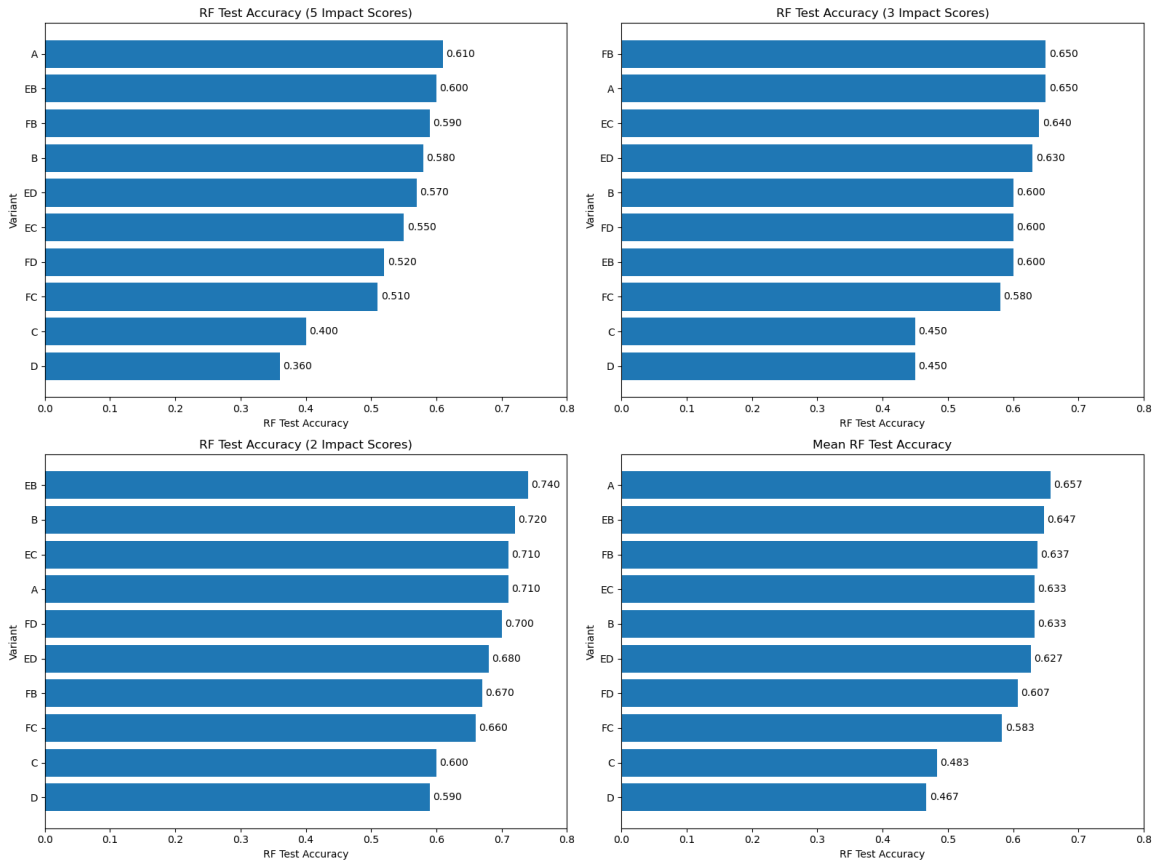


Figure 213: Random Forest test accuracy across all dataset variants for the classification analyses using 5, 3, and 2 Impact Score categories, as well as the combined mean test accuracy across all classification analyses.

Overall, all variants showed increased random forest test accuracies when the number of impact score categories was reduced. The highest classification performances across the variants were generally observed in the two-category impact score analyses, where all variants produced random forest test accuracies between 0.59 and 0.74. In comparison, the original five-category impact score analyses produced lower test accuracies ranging from 0.36 to 0.61. Variant EB produced the highest test accuracy overall, with a value of 0.74 in the two-category impact score analysis. Variants A and EC also produced relatively high performances of 0.71, while Variant B produced a value of 0.72. In contrast, the fully categorical Variants C and D consistently produced the lowest performances across all impact score categorizations, although both variants also showed clear increases in test accuracy when the number of impact score categories was reduced. The mean test accuracy results across all categorizations followed similar overall patterns as the original classification results. Variant A produced the highest overall mean test accuracy of 0.657, followed by Variants EB and FB with values of 0.647 and 0.637, respectively. Variants C and D produced the lowest overall mean test accuracies of 0.483 and 0.467, respectively.

Discussion

Discussion of Variables, Impact Measures and Dataset Variants

Although the seven variables and the ten dataset variants provided a structured framework for analyzing substitution patterns, the construction of these variables involved several simplifications and methodological assumptions. The selection and construction of the variables were developed in collaboration with Divisionsforeningen, to ensure practical and tactical relevance within the context of professional football. Nevertheless, different representations and categorizations of the variables could potentially have produced different clustering structures and classification results. The following section therefore discusses the construction and rationale behind the contextual variables, the impact measures and the dataset variants, and the potential biases and limitations associated with these choices.

Match Minute

Match Minute represented one of the central contextual variables throughout the study. As previous literature has shown, substitution behavior and tactical decision-making are closely related to substitution timing, where substitutions also tend to cluster during specific phases of

the second half (Rey et al., 2015; Sun et al., 2024). However, several methodological considerations were associated with the construction and representation of the variable. One important consideration concerned the choice between numerical and categorical representations of Match Minute across the different dataset variants. In its numerical representation, Match Minute preserved the exact temporal distance between substitutions, allowing substitutions made in nearby minutes to be treated as more similar than substitutions occurring further apart in time. From a football perspective, this may provide a realistic representation of substitution timing, as tactical urgency and match dynamics often develop gradually throughout the match. In contrast, the categorical representations grouped substitutions into broader temporal phases such as early, mid or late substitutions. These categorizations may better reflect broader tactical phases of the match but simultaneously reduce the temporal precision of the variable. For example, substitutions made in minutes 60 and 61 could be placed into different categories despite occurring only one minute apart. Another methodological consideration concerned the difference between the tactically defined categories used in Match Minute CAT C and the quantile-based categories used in Match Minute CAT D. While the tactical categories were constructed to reflect broader football-specific phases of the second half, the quantile-based categories prioritized a more balanced statistical distribution of observations across the categories. As a result, the tactical categories may better reflect how coaches interpret different phases of a match, whereas the quantile-based categories may create more statistically balanced analytical conditions.
The treatment of half-time substitutions also represented a potential limitation. Although half-time substitutions were included within the broader Match Minute structure, these substitutions differ from substitutions made during active gameplay, as they occur during a scheduled break where coaches have additional time for tactical adjustments. An alternative approach could therefore have been to treat half-time substitutions as a separate contextual category rather than integrating them into the broader temporal structure of the second half.

Opponent Strength

Opponent Strength represented another central contextual variable throughout the study, as previous literature has shown that teams may apply different tactical approaches and substitution strategies depending on the strength of the opposition (Chen et al., 2025). However, several methodological considerations and assumptions were associated with the construction

of the variable. One important consideration concerned the decision to calculate Opponent Strength using the pre-match league standings at the time of each match, rather than the final league standings used in studies, such as in Chen et al. (2025). Using pre-match standings allowed the variable to reflect the contextual information available to coaches at the time of the substitution and thereby represented a more dynamic and real-time interpretation of opponent quality. However, this approach also introduced potential instability, particularly during the early stages of the season where league standings may not accurately reflect the actual strength of the teams. For example, after round 8, FCK and AaB were separated by only four league positions despite ultimately finishing first and last in the league, respectively. As a result, substitutions performed against these teams early in the season may have been assigned opponent strength values that did not fully reflect the long-term competitive difference between the teams. In contrast, using final league standings may provide a more stable estimate of overall team quality across the season, as it is less affected by early-season randomness. However, such an approach also introduces a form of bias, as the final league standings were not known at the time when the substitutions occurred. Consequently, the choice between dynamic and final league standings represents a methodological trade-off between contextual realism and long-term stability of the variable. Another methodological consideration concerned the use of league position difference in general to represent Opponent Strength. While league position provided a simple way of comparing teams, teams separated by the same number of league positions could still have very different point differences between them. For example, two teams separated by four league positions could either be very close or far apart in points depending on the stage of the season. An alternative approach could therefore have been to use point differences between teams, such as the Cann Table (*Cann Tabel, Superliga, 2024/2025 - SuperStats*.dk) instead of league positions alone.

Score Difference

Score Difference represented the current match state at the time of the substitution and was included because previous literature has shown that substitution behavior and tactical decision-making may differ depending on whether teams are leading, drawing, or trailing during matches (Amez et al., 2021; Rey et al., 2015). However, several methodological considerations were associated with the construction of the variable. One consideration concerned the decision to group score differences into predefined categories rather than using

the exact goal difference numerically. This simplified the match state into broader tactical situations while also reducing the influence of rare extreme scorelines. In particular, goal differences of three or more goals were grouped into the categories +3+ and -3+, as such situations occurred relatively infrequently within the dataset. While this reduced some of the detail in the variable, it likely created more stable and interpretable categories. Another consideration concerned the assumption that substitutions made within the same score category reflected relatively similar tactical contexts. In practice, however, the tactical meaning of a one-goal lead may differ depending on other contextual factors. For example, leading 1-0 in the 50th minute may still encourage relatively balanced or attacking tactical behavior, whereas leading 1-0 in the 90th minute may instead lead to more defensive game management and time-saving strategies. Consequently, Score Difference should be viewed as a simplified representation of match state rather than a complete reflection of the tactical situation surrounding each substitution.

Number of Substitutions

Number of Subs represented the number of substitutions made by the same team within a short time interval and was included because previous literature has suggested that substitution strategies and tactical flexibility have become increasingly important following the introduction of the five-substitution rule (Xiao & Zhang, 2024). In particular, the increase from three to five substitutions has created more opportunities for coordinated tactical substitution windows. To capture these coordinated substitution sequences, substitutions occurring within 60 seconds of each other were grouped together rather than treated as entirely isolated events. However, the choice of a 60-second interval was ultimately a methodological assumption, as alternative time intervals could also have been applied. A shorter interval may have separated substitutions that were still part of the same tactical adjustment, while a longer interval may have grouped substitutions that were not directly related.

Home/Away

Home/Away was included because previous literature has shown that match location may influence player performance and match dynamics (Chaize et al., 2024). However, the variable was represented as a simple binary distinction between home and away matches, which simplified several potentially important contextual differences. For example, the variable did not account for factors such as crowd size, stadium atmosphere, or specific tactical

considerations related to certain away environments. In practice, coaches may approach substitutions differently depending on the atmosphere of a specific stadium, such as being less likely to introduce an inexperienced player during an intense away match environment. Nevertheless, the binary representation provided a simple and interpretable way of including match location within the contextual framework of the study.

Response

Response was included to capture whether a substitution may have been made as a reaction to a recent substitution by the opposing team. However, this variable was also one of the most uncertain variables in the dataset. A substitution was categorized as a response if the opposing team had made a substitution within the previous five minutes, but this does not necessarily mean that the substitution was actually caused by, or tactically connected to, the opponents substitution. The five-minute window therefore represented a methodological assumption rather than a direct measure of reactive coaching behavior, as both shorter and longer time intervals could potentially have been applied. An alternative approach could have been to use a bell-curve inspired model, where substitutions occurring around an estimated most likely response window were considered more likely to represent an actual reaction, while substitutions occurring either too soon or too late after the opponents substitution were considered less likely to be connected. In this context, substitutions occurring very shortly after the opponents substitution may in many cases already have been planned beforehand, as players generally require time to warm up before entering the match. However, such an approach would still have relied on subjective assumptions regarding when a substitution should realistically be considered a tactical response. Consequently, Response should be interpreted as a rough indicator of possible reactive substitution behavior rather than a direct measure of actual tactical response.

Team Size (Red Cards)

Team Size (Red Cards) was included to capture whether a team was playing with a numerical advantage or disadvantage at the time of the substitution, as red cards may influence tactical behavior and match dynamics (Badiella et al., 2023; Červený et al., 2018). However, the variable was simplified into the categories “Even”, “Up a player”, and “Down a player”, which did not account for differences in the timing or context of the red card. For example, playing with ten players from the 20th minute may create a very different tactical situation compared to

receiving a red card shortly before a late substitution. In addition, red card situations occurred relatively infrequently within the dataset compared to the other contextual variables.

xG- and Impact Score

xG Score and the derived impact score represented the central outcome variables of the study and were constructed to capture the short-term impact of substitutions. xG were chosen because previous literature had shown that xG provides a more informative measure of performance than goals alone, as it reflects the quality of created and conceded chances rather than only the outcome of shots (Mead et al., 2023; Rathke, 2017). This may be particularly relevant when examining short-term periods surrounding substitutions, since actual goals occur relatively infrequently. One key methodological consideration was the length of the time windows before and after each substitution. A longer pre-substitution window could have offered a more stable picture of match dynamics, but longer post-substitution windows would increase the risk that the measured impact was affected by other events such as additional substitutions, goals, red cards, or tactical adjustments. Longer windows would also reduce the number of substitutions with enough remaining match time to complete the full window. In the dataset, about 2% of substitutions had fewer than five minutes left, rising to around 11% with a hypothetical 10-minute window. Therefore, the five-minute windows represented a trade-off between capturing enough short-term match development and limiting the influence of unrelated later events. Another methodological consideration concerned the assumption that short-term changes in xG balance reflected substitution impact. While changes in xG may indicate shifts in attacking or defensive performance, substitutions do not occur in isolation, and match dynamics may simultaneously be influenced by factors such as momentum shifts, tactical adjustments, individual player actions, or random match events. However, in some situations, the resulting xG changes may still be directly related to the substituted player. For example, the substitution associated with the lowest xG Score in the dataset, involved the substituted player conceding a penalty shortly after entering the pitch. However, such situations likely represent exceptions rather than the general pattern.

The construction of the impact score also introduced several considerations. By transforming the continuous xG Score into five quantile-based categories, the study created a more balanced distribution across the outcome variable. Consequently, the impact score should be viewed as a

simplified representation of short-term substitution impact rather than a complete measure of tactical substitution success.

Dataset Variants

The use of multiple dataset variants represented an important methodological component of the study, as several contextual variables could reasonably be represented in different ways. Based on the numerical and categorical representations of Match Minute and Opponent Strength, a total of 16 theoretical dataset combinations could have been constructed. However, only ten dataset variants were ultimately included, as several combinations were considered too similar or unlikely to provide meaningful additional insight. This introduced an important methodological consideration, as the selection of dataset variants partly relied on researcher judgment regarding which combinations were considered relevant to investigate. Different selections of dataset variants may therefore potentially have highlighted different structural patterns within the data. Another consideration concerned the trade-off between numerical precision and categorical simplification. Numerical representations preserved more detailed information, while categorical representations simplified the variables into broader groups. However, categorical transformations also introduced subjective decisions regarding category boundaries and grouping structures, which may influence the resulting analytical patterns. The use of multiple dataset variants therefore also functioned as a robustness assessment, allowing the study to examine whether broader analytical tendencies remained relatively stable across different contextual representations rather than depending entirely on a single dataset structure.

Discussion of Clustering

The clustering analysis showed that several relatively similar substitution patterns appeared across the different dataset variants, despite differences in variable representations and clustering structures. This suggests that the clustering analyses were able to identify broader contextual substitution tendencies rather than purely random groupings of substitutions. Many of the clusters associated with higher mean impact scores involved substitutions performed during the mid-second half, often through double substitutions and while trailing or facing weaker opponents. In contrast, lower-impact clusters more often involved late, single substitutions, performed while leading. Interestingly, the hypothetical highest-impact substitution profile identified across the variants (table 3) generally involved substitutions

performed while trailing. This aligns relatively well with the findings of Amez et al. (2021), who found statistical evidence that substitutions made while losing were associated with increased probability of subsequent goal scoring. Although the present study examined short-term xG-based impact scores rather than actual goals, the overall tendency still points in the same direction, suggesting that substitutions performed while trailing may be associated with more aggressive or tactically impactful match behavior.

An important consideration concerned the use of K-prototypes and K-modes clustering. Since the dataset contained both numerical and categorical contextual variables across the variants, these clustering methods allowed mixed contextual data to be analyzed together. However, the clustering structures were likely also influenced by the numerical representations of Match Minute and Opponent Strength observed throughout the study (Wongoutong, 2024). Within K-prototypes clustering, numerical variables are evaluated using Euclidean distance, meaning that substitutions occurring close together numerically are treated as more similar than substitutions further apart. In contrast, categorical variables are only evaluated according to whether categories match or differ. Consequently, the numerical variables likely preserved more detailed contextual information than the categorical representations, which may partly explain why the variants containing numerical representations generally produced stronger cluster separation than the fully categorical variants. Another important factor was the selection of the number of clusters. Although the elbow method provided guidance regarding suitable cluster solutions, no perfectly clear elbow points were observed, meaning that the final cluster selection still involved some degree of researcher judgment. Nevertheless, the selected cluster solutions remained relatively consistent across the variants, ranging between six and eight clusters. Interestingly, this aligns relatively well with the findings of a study, who also applied clustering-related methods and identified five dominant substitution clusters within elite football (Wu et al., 2024). Despite methodological differences between the studies, the relatively similar number of identified clusters may suggest that substitution behavior in professional football tends to organize around a limited number of broader tactical patterns.

The sensitivity analysis further supported the overall stability of the clustering results. Although some variants showed moderate changes in cluster separation when alternative numbers of clusters were applied, the overall substitution patterns and contextual tendencies remained relatively similar across the original and alternative cluster solutions. In addition, the

selected cluster solutions generally remained within a relatively narrow range across the variants. This suggests that the clustering results were not entirely dependent on one specific cluster configuration, but instead reflected broader tendencies within the data.

Although meaningful differences between clusters were identified, the differences in mean impact score generally remained moderate across most variants. However, this does not necessarily weaken the clustering results. Football is a highly complex and unpredictable sport, where short-term match dynamics are likely influenced by many additional factors beyond the seven variables included in the present study. Consequently, the clustering results may be better interpreted as identifying broader contextual substitution tendencies rather than fixed tactical outcomes.

Discussion of Adjusted Rand Index

The ARI analysis showed relatively consistent patterns across both the original and alternative cluster solutions. Overall, the numerical representations of Match Minute and Opponent Strength generally produced the lowest ARI values and therefore appeared to have the strongest influence on the clustering structures. As discussed in the clustering section, this was likely partly related to the structure of K-prototypes clustering itself, where numerical variables preserve more detailed distance information than categorical variables.

Interestingly, the categorical representations of Match Minute and Opponent Strength generally produced some of the highest ARI values among the remaining contextual variables, only exceeded by Team Size (Red Cards) and Score Difference. This suggests that when Match Minute and Opponent Strength were transformed into broader categorical variables, they no longer appeared substantially more influential than the other variables. In this way, the results may indicate that Match Minute and Opponent Strength were not necessarily the strongest variables overall, but rather that their influence depended heavily on the numerical detail preserved within their continuous representations. The remaining variables generally produced relatively similar mid-range ARI values across the variants. This may reflect the multidimensional nature of substitution behavior, where several contextual factors likely contribute simultaneously to the clustering structures rather than one single variable fully determining the substitution context. In contrast, Team Size (Red Cards) consistently produced very high ARI values across nearly all variants, suggesting that the variable had very limited influence on the clustering

structures. A likely explanation is the highly uneven distribution of the variable within the dataset, where most of substitutions occurred under even team-size conditions. Overall, the sensitivity analysis supported the stability of the ARI results, as the same overall variable patterns remained visible across the alternative cluster solutions. This strengthens the interpretation, that the observed variable influences were not entirely dependent on one specific clustering configuration.

Discussion of Feature Importance

The feature importance analysis showed highly stable patterns across both the original variants and the alternative impact score categorizations used in the sensitivity analysis. Overall, the numerical representations of Match Minute and Opponent Strength consistently produced the highest feature importance values, while Team Size (Red Cards) consistently produced the lowest values. The stability of these rankings across the sensitivity analyses strengthens the interpretation that the overall feature importance patterns were relatively robust.

Interestingly, the overall feature importance patterns showed several similarities to the ARI results, where the numerical representations of Match Minute and Opponent Strength also appeared highly influential. However, unlike the clustering analyses, random forest models evaluate variables through repeated tree-based splits rather than distance between observations. As a result, numerical variables and variables containing many categories may gain an advantage within the model structure due to the larger number of potential splits available. This is also supported by a study, who showed that random forest importance measures may favor continuous variables and categorical variables with many categories, which in some cases may lead to misleading interpretations of variable importance (Strobl et al., 2007). This is particularly relevant within the present study, where the analyses combined continuous variables with categorical variables containing different numbers of categories, such as Score Difference with seven categories compared to Response and Home/Away with only two categories. Although the numerical variables were standardized prior to the analyses, Strobl et al., (2007) found that scaling may reduce this type of bias, but not remove it entirely. Consequently, the high feature importance values observed for Match Minute Num, Opponent Strength Num, and to some extent Score Difference, should not necessarily be interpreted as objective measures of the true tactical importance of the variables. Nevertheless, the categorical representations of Match Minute and Opponent Strength still produced relatively high feature

importance values compared to most of the remaining contextual variables. This suggests that the two variables still contained meaningful contextual information even when transformed into broader categorical representations. However, the large drop in importance compared to the numerical representations also indicates that a larger part of their influence was linked to the detailed information preserved within their continuous structures.

Finally, Score Difference remained relatively influential across most variants, particularly within the fully categorical variants. This may suggest that Score Difference represented one of the more consistently important tactical variables across the different variants, although its relatively large number of categories may also have contributed to the high feature importance values. In contrast, Team Size (Red Cards) consistently produced the lowest feature importance values, likely due to the highly uneven distribution of red card situations within the dataset. Overall, the feature importance analyses suggest that the observed influence of the variables depended strongly on how the variables were represented within the analytical framework. This overall pattern was visible across both the original analyses, the sensitivity analyses, and the earlier ARI results.

Discussion of Classification

The classification analysis showed moderate but relatively consistent predictive performances across the different variants. Overall, the variants containing numerical representations of Match Minute and/or Opponent Strength generally produced the highest classification performances, while the fully categorical variants consistently produced the lowest performances. This overall pattern remained visible across the sensitivity analyses, although the exact ranking of the variants changed between the different impact score categorizations. The lower performances of the fully categorical variants may suggest that important contextual information was lost when Match Minute and Opponent Strength were transformed into broader categorical representations. In this way, the results indicate that substitution contexts may exist more naturally along continuous gradients rather than within a limited number of broader tactical categories. For example, small differences in substitution timing or opponent strength may still contain meaningful contextual information that becomes lost when grouped into broader categories such as “Early/Mid/Late” or “Stronger/Equal/Weaker”. At the same time, the strong performances of the numerical and hybrid variants should also be interpreted in relation to the methodological considerations discussed in the feature importance section,

where continuous variable representations may gain advantages within the random forest framework. The sensitivity analysis further showed that all variants improved their classification performances when the number of impact score categories was reduced from five to three and two categories. This was expected, since fewer target categories simplify the classification task. The relatively high test-accuracies observed in the two-category analyses may therefore indicate that the contextual variables were better suited for distinguishing between broader positive and negative substitution outcomes, than predicting highly specific impact score categories.

Overall, the classification analyses suggest that the seven variables contained meaningful predictive information related to substitution impact, but also that football substitution outcomes remain highly complex and difficult to predict precisely. While the variants were able to identify broader contextual tendencies related to substitution impact, the moderate classification performances across the original five-category analyses indicate that a large part of the variation in short-term substitution impact was likely influenced by additional tactical, technical, and situational factors not included within the present study.

Discussion of Methods and Limitations of the Study

The methodological framework used in the present study made it possible to examine football substitutions from several different analytical perspectives. However, the study should primarily be viewed as an exploratory analysis of contextual substitution tendencies rather than a direct evaluation of whether certain substitutions were objectively successful or unsuccessful. Football substitutions take place in highly dynamic match situations influenced by many tactical and situational factors at the same time. Consequently, the identified clusters, variable influences, and classification performances should mainly be interpreted as broader contextual patterns and associations related to substitution impact within the dataset rather than direct causal relationships.

One of the main limitations of the study concerns the relatively limited amount of contextual and player-specific information included within the analyses. Although the seven variables captured several important tactical aspects surrounding substitutions, many potentially important factors were not included in the dataset. For example, the analyses did not account for player quality, player positions, tactical systems, fatigue, individual player form, weather, or

coach-specific tactical tendencies. Since substitutions are highly player-dependent actions, these omitted factors may have influenced both the clustering structures and the predictive performances observed throughout the study. This may partly explain why several of the analyses produced moderate rather than highly distinct results, despite relatively stable overall tendencies across the variants.

Another limitation concerns the use of data from only a single Danish Superliga season. As a result, the identified substitution patterns may partly reflect league-specific tactical tendencies and substitution strategies unique to the 2024/2025 season. Substitution behavior may differ across leagues due to differences in tactical styles, coaching philosophies, squad depths, and match intensity. At the same time, limiting the dataset to a single league and season also created a more homogeneous analytical environment, which may have reduced some external variation unrelated to the contextual variables themselves.

The study also involved many researcher-driven methodological choices. Several analytical decisions relied on assumptions, including the construction of the contextual variables, the category boundaries, the five-minute xG windows, the 60-second substitution grouping interval, the selected dataset variants, and the final cluster selections. Although many of these choices were supported through previous literature and discussions with Divisionsforeningen, different methodological decisions could potentially have produced different analytical patterns and results. Consequently, the findings of the study should also be interpreted in relation to the methodological framework in which they were created.

At the same time, the use of multiple dataset variants and sensitivity analyses represented an important strength of the study. Rather than relying on one single variable representation or analytical structure, the study repeatedly examined whether broader analytical tendencies remained visible across alternative contextual representations, cluster structures, and impact score categorizations. Although several individual results changed across the analyses, many of the broader tendencies remained relatively consistent throughout the study, which strengthens the overall robustness of the findings.

Conclusion

The purpose of the study was to explore how different substitution patterns were associated with short-term changes in match dynamics in the Danish Superliga, given the contextual factors in which the substitutions occurred. To investigate this, the study combined clustering analysis, Adjusted Rand Index analysis, random forest for feature importance, and random forest classification across multiple dataset variants containing different representations of seven contextual match variables.

The clustering analyses identified several recurring contextual substitution patterns across the dataset variants, particularly substitutions performed during the mid-second half, involving double substitutions and teams trailing in the match, which were frequently associated with higher impact scores. The ARI analyses further showed that the numerical representations of Match Minute and Opponent Strength generally had the strongest influence on the clustering structures. Similar tendencies were observed in the feature importance analyses, where the numerical representations consistently produced the highest importance values across the variants. Finally, the classification analyses demonstrated moderate but relatively consistent predictive performances, where the numerical and hybrid variants generally outperformed the fully categorical variants. At the same time, the study also demonstrated that the analytical results depended strongly on how the variables were represented within the different analytical frameworks. When Match Minute and Opponent Strength were transformed into broader categorical representations, their influence and predictive performances generally decreased. This suggests that important contextual information may have been lost through broader categorical simplifications.

Overall, the findings suggest that the seven variables contained meaningful information related to short-term substitution impact, while also highlighting the complexity and unpredictability of football substitutions. Consequently, the study should primarily be interpreted as an exploratory analysis of contextual substitution tendencies rather than a direct evaluation of substitution effectiveness. Future research could potentially expand the framework by including multiple seasons, additional leagues, and more detailed player- and team-specific variables, in order to capture a larger part of the contextual complexity surrounding football substitutions.

References

1.10. Decision Trees. Scikit-Learn. Retrieved 15 May 2026, from https://scikit-learn/stable/modules/tree.html

Amez, S., Neyt, B., Van Nuffel, F., & Baert, S. (2021). The right man in the right place? Substitutions and goal-scoring in soccer. *Psychology of Sport and Exercise*, *54*, 101898. https://doi.org/10.1016/j.psychsport.2021.101898

Badiella, L., Puig, P., Lago-Peñas, C., & Casals, M. (2023). Influence of Red and Yellow cards on team performance in elite soccer. *Annals of Operations Research*, *325*(1), 149–165. https://doi.org/10.1007/s10479-022-04733-0

Cann tabel, superliga, 2024/2025—SuperStats. Retrieved 26 May 2026, from https://superstats.dk/stilling/cann?runde=32&aar=2024%2F2025

Červený, J., van Ours, J. C., & van Tuijl, M. A. (2018). Effects of a red card on goal-scoring in World Cup football matches. *Empirical Economics*, *55*(2), 883–903. https://doi.org/10.1007/s00181-017-1287-5

Chaize, C., Allen, M., & Beato, M. (2024). Physical Performance Is Affected by Players' Position, Game Location, and Substitutions During Official Competitions in Professional Championship English Football. *Journal of Strength and Conditioning Research*, *38*(12), e744–e753. https://doi.org/10.1519/JSC.0000000000004926

Chen, T., Chen, L., Li, R., & Lv, K. (2025). The Dominant factors and optimization plan of soccer substitutions in Chinese Football Association Super League under the Five-substitution rule. *PLOS One*, *20*(6), e0326656. https://doi.org/10.1371/journal.pone.0326656

Elbow Method. GeeksforGeeks. Retrieved 28 May 2026, from https://www.geeksforgeeks.org/machine-learning/elbow-method-for-optimal-value-of-k-in-kmeans/

Feature Importance with Random Forests. GeeksforGeeks. Retrieved 28 May 2026, from https://www.geeksforgeeks.org/machine-learning/feature-importance-with-random-forests/

Fifa. (2023). *The IFAB permanently approves five-substitute option in top-level competitions*. https://inside.fifa.com/innovation/media-releases/origin1904-p.cxm.fifa.com/the-ifab-permanently-approves-five-substitute-option-in-top-level

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, *18*(4), e0282295. https://doi.org/10.1371/journal.pone.0282295

Myers, B. R. (2012). A Proposed Decision Rule for the Timing of Soccer Substitutions. *Journal of Quantitative Analysis in Sports*, *8*(1). https://doi.org/10.1515/1559-0410.1349

Pan, P., Li, F., Han, B., Yuan, B., & Liu, T. (2023). Exploring the impact of professional soccer substitute players on physical and technical performance. *BMC Sports Science, Medicine and Rehabilitation*, *15*, 143. https://doi.org/10.1186/s13102-023-00752-x

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(Proc2). https://doi.org/10.14198/jhse.2017.12.Proc2.05

Rey, E., Lago-Ballesteros, J., & Padrón-Cabo, A. (2015). Timing and tactical analysis of player substitutions in the UEFA Champions League. *International Journal of Performance Analysis in Sport*, *15*(3), 840–850. https://doi.org/10.1080/24748668.2015.11868835

Samuels, J. I. (2024). *One-Hot Encoding and Two-Hot Encoding: An Introduction*. https://doi.org/10.13140/RG.2.2.21459.76327

Schneemann, S., & Deutscher, C. (2017). Intermediate Information, Loss Aversion, and Effort: Empirical Evidence. *Economic Inquiry*, *55*(4), 1759–1770. https://doi.org/10.1111/ecin.12420

SciKit Learn—Adjusted Rand Index. Scikit-Learn. Retrieved 28 May 2026, from https://scikit-learn/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

SciKit Learn—SandardScaler. Scikit-Learn. Retrieved 13 May 2026, from https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Silva, R. M., & Swartz, T. B. (2016). Analysis of substitution times in soccer. *Journal of Quantitative Analysis in Sports*, *12*(3), 113–122. https://doi.org/10.1515/jqas-2015-0114

Stillingen, 2024/2025—SuperStats. Retrieved 28 May 2026, from https://superstats.dk/stilling/runde?season=2025&round=32

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

Sun, R., Wang, C., Qin, Z., & Han, C. (2024). Temporal features of goals, substitutions, and fouls in football games in the five major European league from 2018 to 2021. *Heliyon*, *10*(5), e27014. https://doi.org/10.1016/j.heliyon.2024.e27014

Wei, X., Shu, Y., Liu, J., Chmura, P., Randers, M. B., & Krustrup, P. (2024). Analysing substitutions in recent World Cups and European Championships in male and female elite football – influence of new substitution rules. *Biology of Sport*, *41*(3), 267–274. https://doi.org/10.5114/biolsport.2024.134755

Wittkugel, J., Memmert, D., & Wunderlich, F. (2022). Substitutions in football—What coaches think and what coaches do. *Journal of Sports Sciences*, *40*(15), 1668–1677. https://doi.org/10.1080/02640414.2022.2099177

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLOS ONE*, *19*(12), e0310839. https://doi.org/10.1371/journal.pone.0310839

Wu, J., Li, Y., Zhou, C., Xiong, X., Qin, X., & Cui, Y. (2024). Impact of substitutions on elite soccer team performance based on player evaluation system. *Journal of Human Sport and Exercise*, *20*(1), 316–327. https://doi.org/10.55860/sbh12g81

Xiao, Z., & Zhang, H. (2024). More substitutions changed team substitution strategy? An analysis of the FIFA World Cup 2002–2022. *BMC Sports Science, Medicine and Rehabilitation*, *16*, 165. https://doi.org/10.1186/s13102-024-00956-9

Appendices

Appendix 1: Code

The complete code used in the study has been uploaded to the following GitHub repository:

https://github.com/Yosef317/MasterThesis

The repository contains the 13 Jupyter notebooks used in the analyses together with a README file. The underlying datasets are not included due to a non-disclosure agreement with Divisionsforeningen. Consequently, the notebooks cannot be executed without access to the original data. However, all notebooks can be downloaded, and all outputs generated during the analyses have been preserved within the notebooks, allowing the results and workflow to be reviewed without further access to the underlying datasets.

Appendix 2: Detailed Match Minute Distribution for All Substitutions

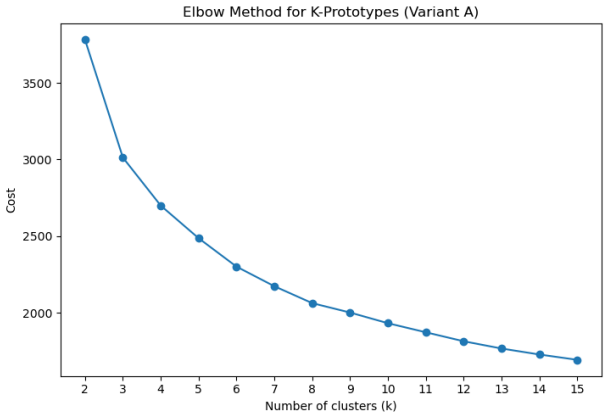
Match Minute	Number of Substitutions	Match Minute	Number of Substitutions
45	168	73	47
46	0	74	39
47	0	75	29
48	0	76	26
49	0	77	47
50	0	78	50
51	0	79	44
52	1	80	36
53	4	81	67
54	1	82	40
55	4	83	49
56	17	84	51
57	21	85	30
58	24	86	47
59	22	87	35
60	40	88	24
61	54	89	29
62	43	90	29
63	51	91	19
64	37	92	10
65	42	93	13
66	48	94	4
67	36	95	6
68	38	96	3
69	41	97	3
70	40	98	1
71	59	99	1
72	40		

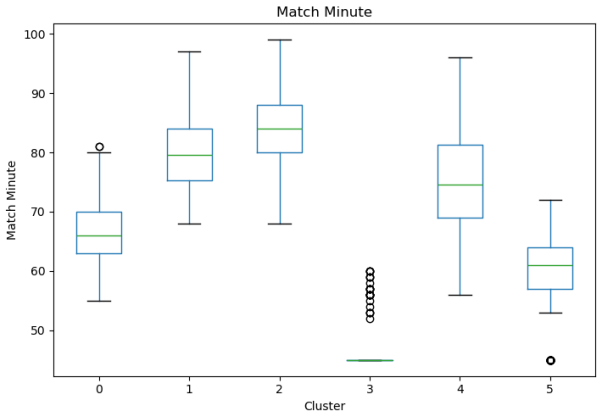
Appendix 3: Detailed Opponent Strength Distribution for All Substitutions

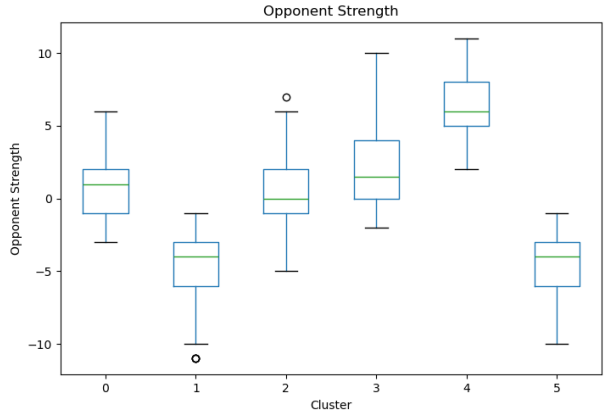
Opponent Strength	Number of Substitutions	Opponent Strength	Number of Substitutions
11	5	-1	207
10	14	-2	152
9	34	-3	87
8	22	-4	98
7	36	-5	78
6	57	-6	56
5	65	-7	33
4	97	-8	15
3	92	-9	33
2	150	-10	14
1	209	-11	4
0	52		

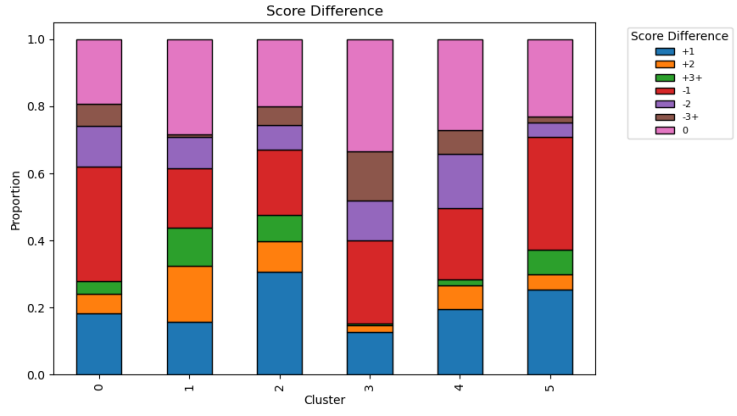
The following appendices present the detailed results for each dataset variant (A–FD) included in the study. Each appendix contains the elbow method results used to support the selection of the number of clusters, visualizations of the cluster distributions across the seven variables, and detailed clustering results including cluster descriptions, cluster sizes, and mean impact scores. In addition, each appendix includes the corresponding Adjusted Rand Index results, feature importance results, and detailed classification results including precision, recall, F1-score, macro F1-score, validation accuracy, and test accuracy.

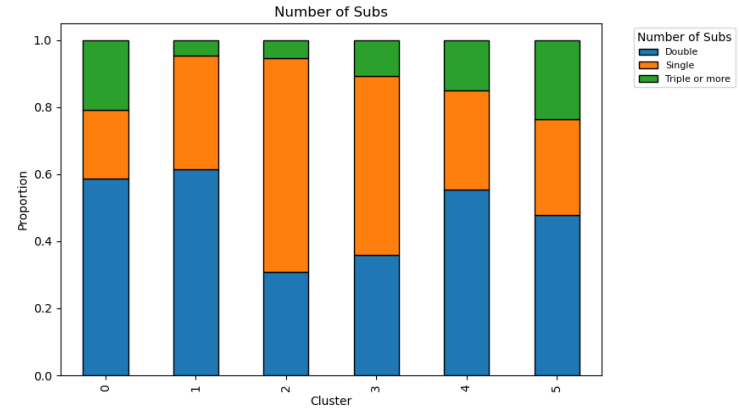
Appendix 4: Variant A Results

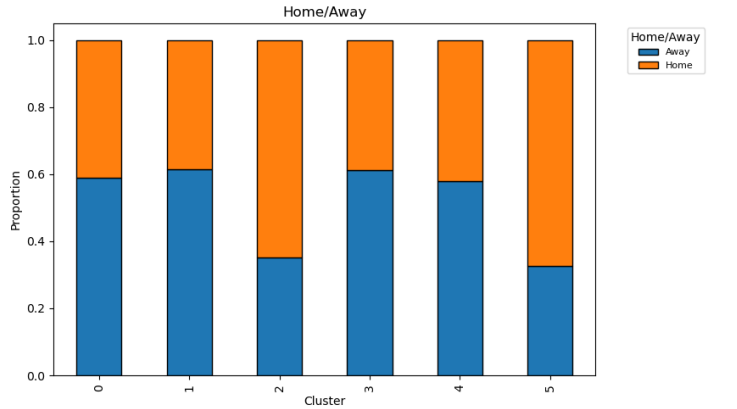


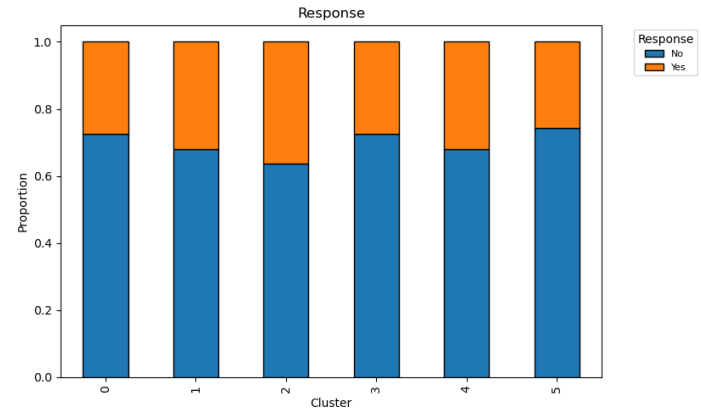


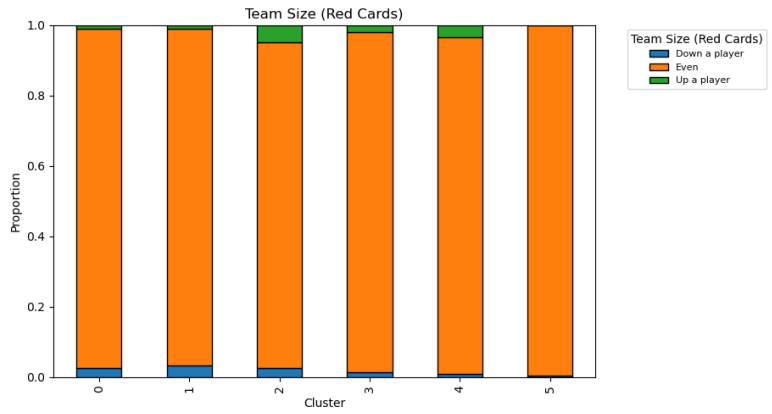










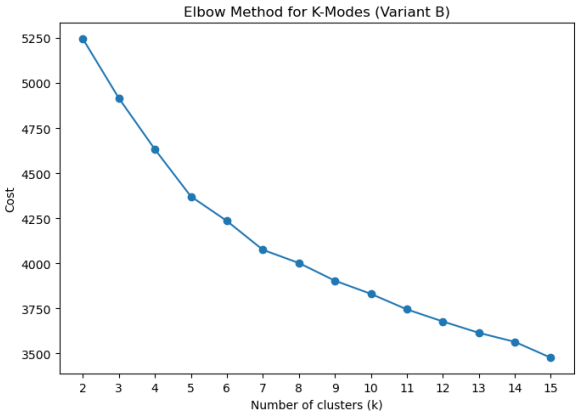


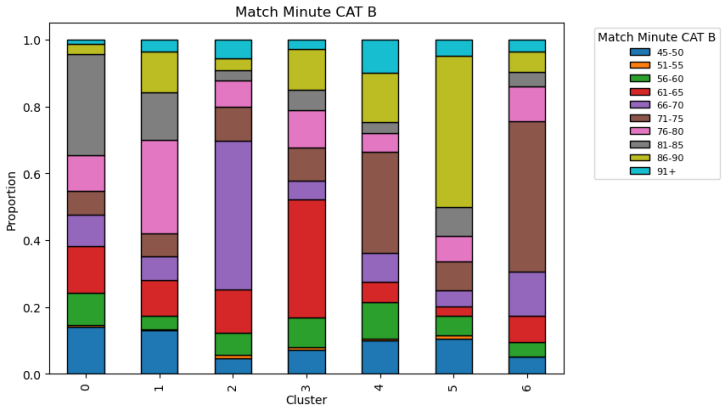
Variant A		
Number of Clusters = 6		
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**
0 (372)	3.040 (1.434)	Late Non-Reactive Double Substitutions in Drawing Away Games vs Weaker Opponent
1 (262)	3.019 (1.332)	Mid-Second Half Non-Reactive Double Substitutions in Drawing Home Games vs Weaker Opponent
2 (369)	2.873 (1.409)	Late Non-Reactive Single Substitutions in Winning Home Games vs Equal Opponent
3 (150)	3.180 (1.559)	Early Non-Reactive Single Substitutions in Drawing Away Games vs Equal Opponent
4 (240)	3.038 (1.385)	Mid-Second Half Non-Reactive Double Substitutions in Drawing Away Games vs Stronger Opponent
5 (217)	2.949 (1.412)	Mid-Second Half Non-Reactive Double Substitutions in Drawing Home Games vs Equal Opponent

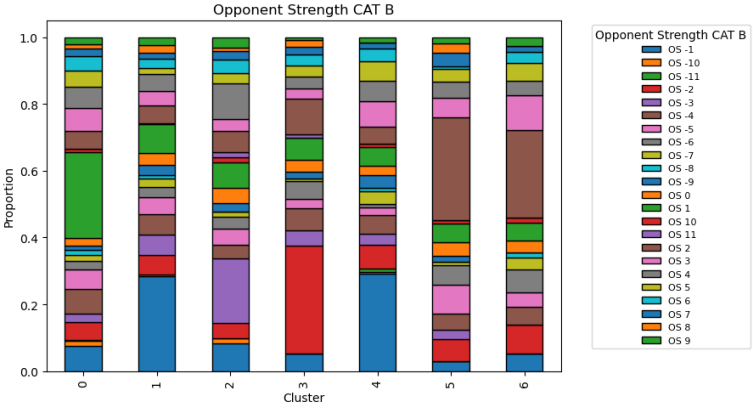
Variable	**Adjusted Rand Index**	**Feature Importance**
Opponent Strength (Num)	0.2073	0.3346
Match Minute (Num)	0.2211	0.3694
Number of Subs	0.4187	0.0711
Response	0.5145	0.0467
Home/Away	0.5177	0.0383
Team Size (Red Cards)	0.6736	0.0182
Score Difference	0.7732	0.1215

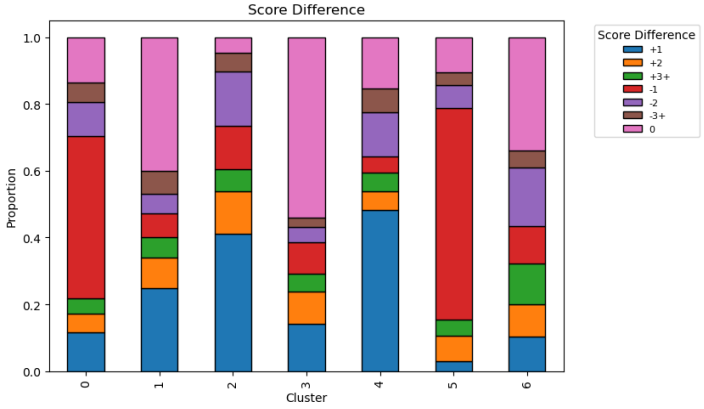
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.69	0.50	0.58	0.61	0.60	0.61
2	0.62	0.63	0.63			
3	0.54	0.60	0.60			
4	0.54	0.62	0.62			
5	0.70	0.67	0.67			

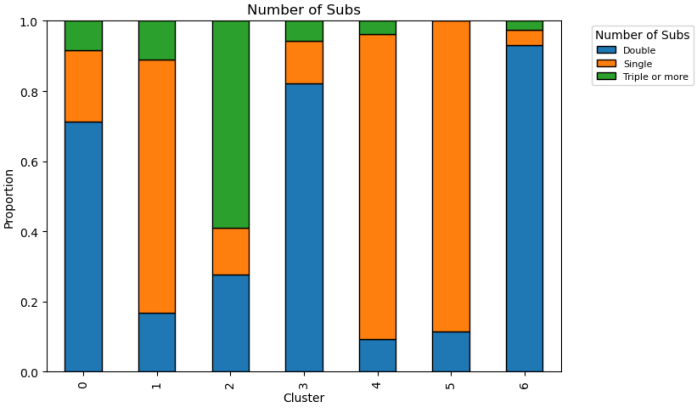
Appendix 5: Variant B Results

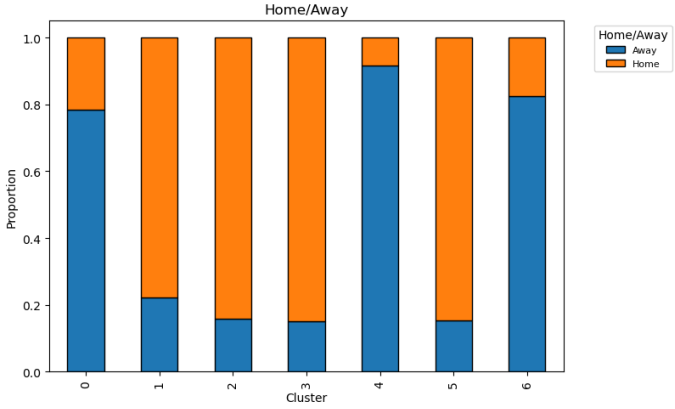


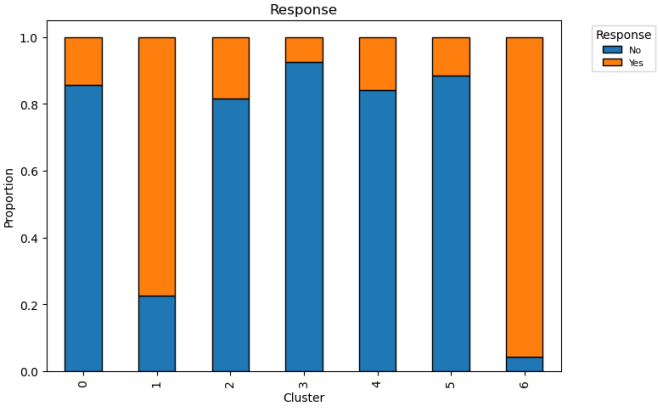


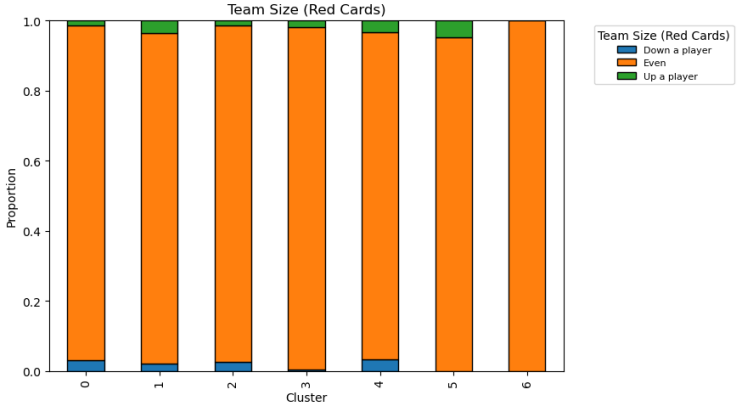










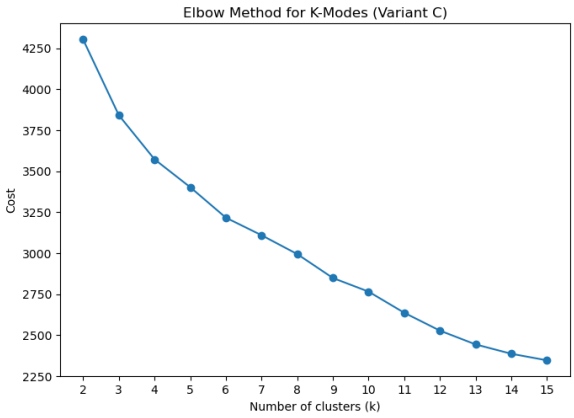


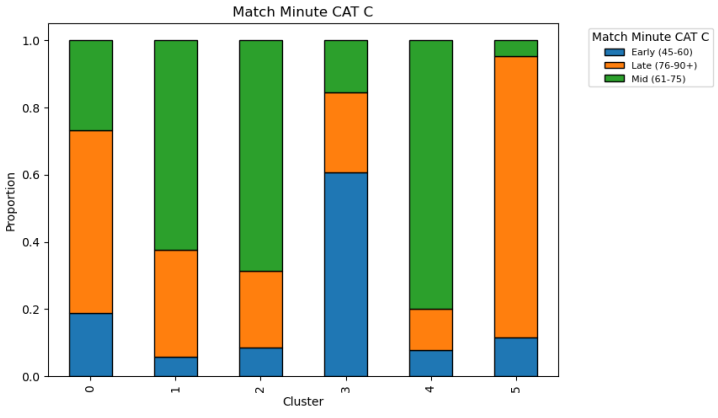
Variant B		
Number of Clusters = 7		
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**
0 (522)	3.079 (1.441)	Late Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent
1 (279)	2.914 (1.365)	Late Reactive Single Substitutions in Drawing Home Games vs Equal Opponent
2 (195)	2.964 (1.419)	Mid-Second Half Non-Reactive Triple Substitutions in Winning Home Games vs Weaker Opponent
3 (213)	3.141 (1.370)	Mid-Second Hal Non-Reactive Double Substitutions in Drawing Home Games vs Weaker Opponent
4 (182)	2.879 (1.467)	Late Non-Reactive Single Substitutions in Winning Away Games vs Equal Opponent
5 (104)	2.962 (1.441)	Late Non-Reactive Single Substitutions in Losing Home Games vs Stronger Opponent
6 (115)	2.861 (1.369)	Late Reactive Double Substitutions in Drawing Away Games vs Stronger Opponent

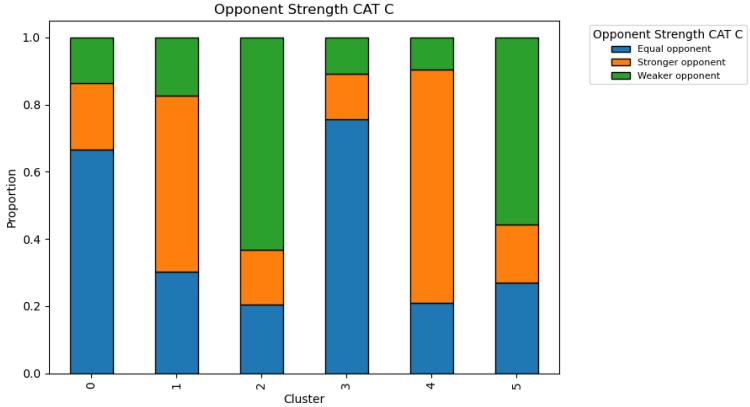
Variable	**Adjusted Rand Index**	**Feature Importance**
Opponent Strength (Cat B)	0.4397	0.3754
Match Minute (Cat B)	0.5438	0.2263
Number of Subs	0.3234	0.0922
Response	0.3078	0.0601
Home/Away	0.3152	0.0594
Team Size (Red Cards)	0.3991	0.0187
Score Difference	0.3531	0.1679

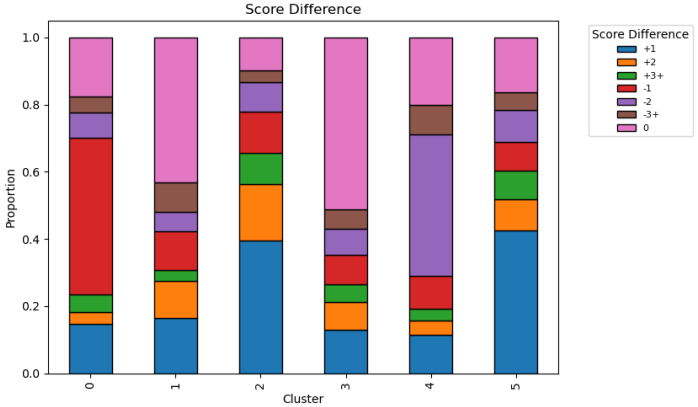
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.47	0.44	0.45	0.58	0.55	0.58
2	0.64	0.57	0.60			
3	0.56	0.62	0.59			
4	0.59	0.60	0.60			
5	0.64	0.65	0.65			

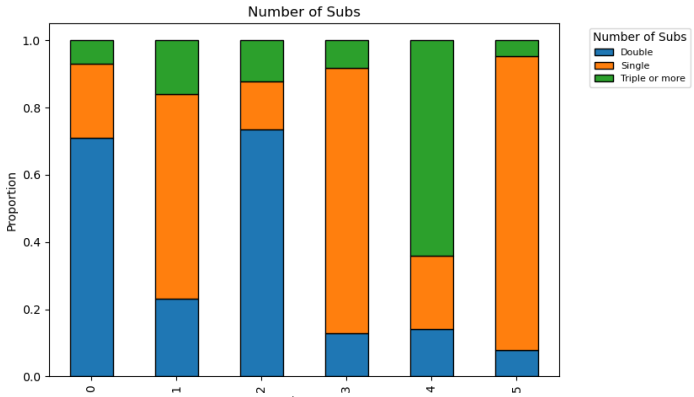
Appendix 6: Variant C Results

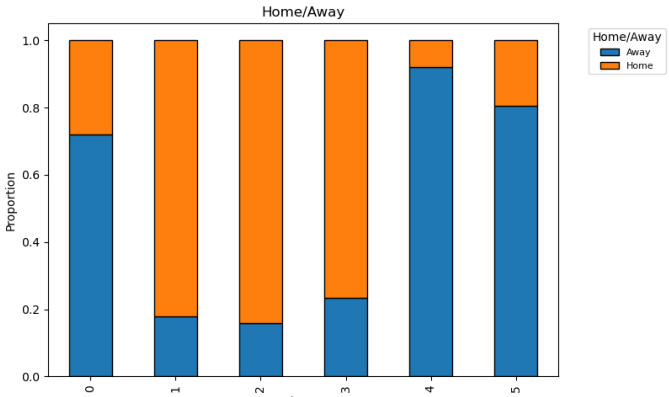


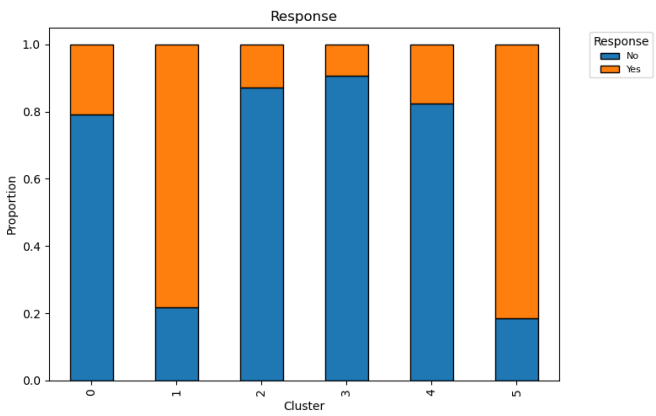


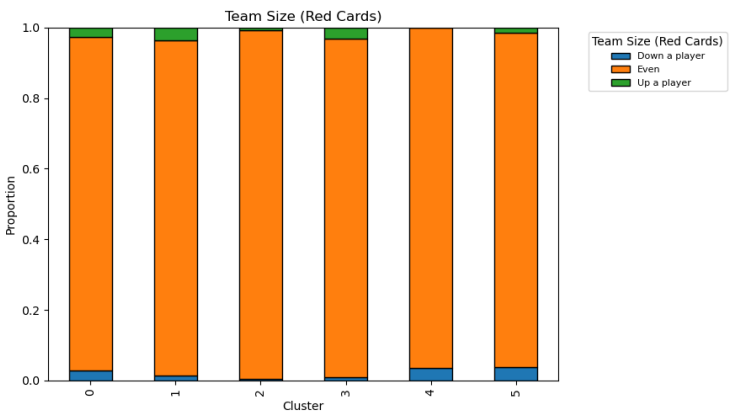










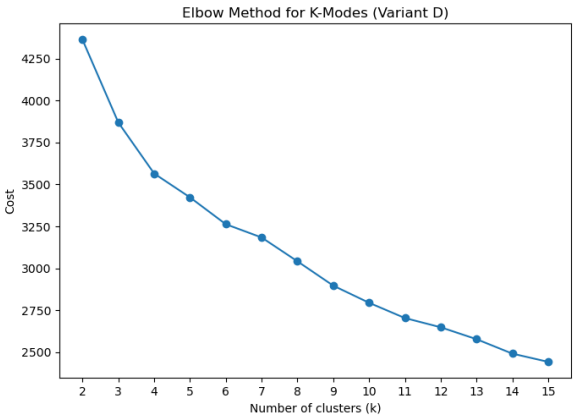


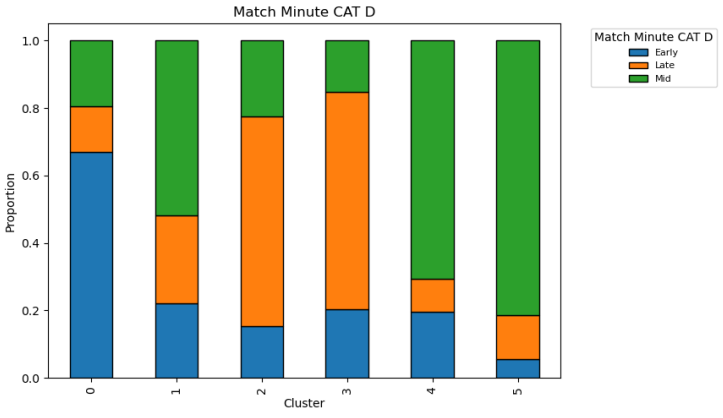
Variant C		
Number of Clusters = 6		
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**
0 (655)	3.066 (1.418)	Late Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent
1 (225)	2.898 (1.347)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Stronger Opponent
2 (294)	3.027 (1.365)	Mid-Second Half Non-Reactive Double Substitutions in Winning Home Games vs Weaker Opponent
3 (193)	2.917 (1.515)	Early Non-Reactive Single Substitutions in Drawing Home Games vs Equal Opponent
4 (114)	3.079 (1.512)	Mid-Second Half Non-Reactive Triple Substitutions in Losing Away Games vs Stronger Opponent
5 (129)	2.822 (1.378)	Late Reactive Single Substitutions in Winning Away Games vs Weaker Opponent

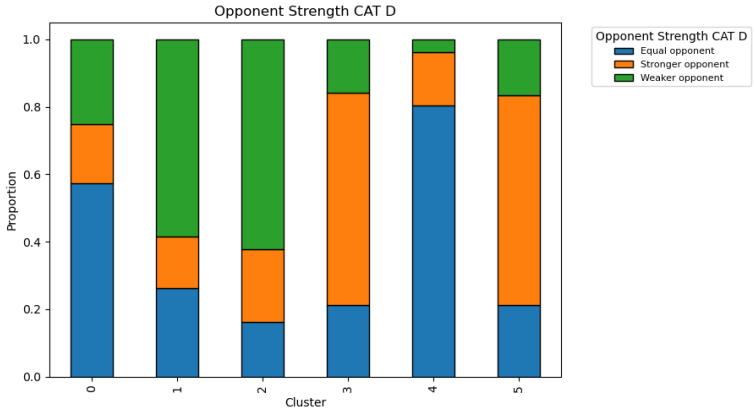
Variable	Adjusted Rand Index	Feature Importance
Opponent Strength (Cat C)	0.3346	0.1848
Match Minute (Cat C)	0.3832	0.1821
Number of Subs	0.3280	0.1452
Response	0.3087	0.1054
Home/Away	0.3449	0.0924
Team Size (Red Cards)	1.0000	0.0355
Score Difference	0.3364	0.2546

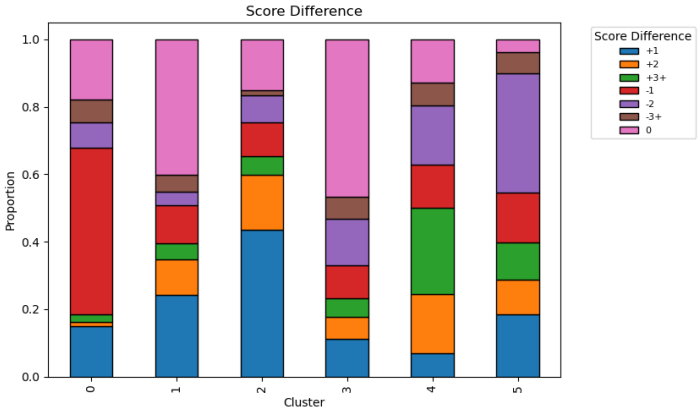
Impact Score	Precision	Recall	F1 Score	Macro F1	Validation Accuracy	Test Accuracy
1	0.37	0.35	0.36	0.40	0.41	0.40
2	0.42	0.41	0.41			
3	0.42	0.46	0.44			
4	0.37	0.35	0.36			
5	0.40	0.41	0.40			

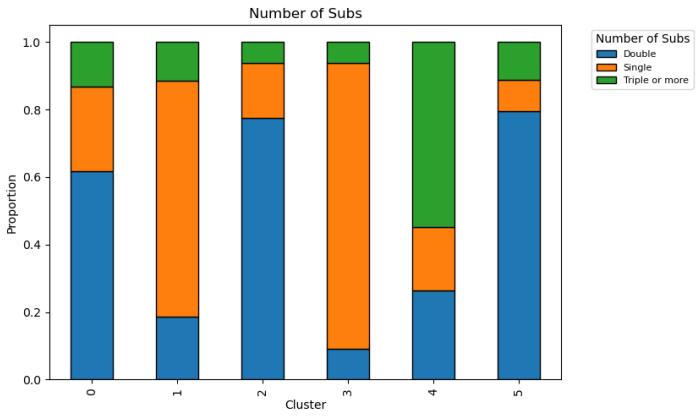
Appendix 7: Variant D Results

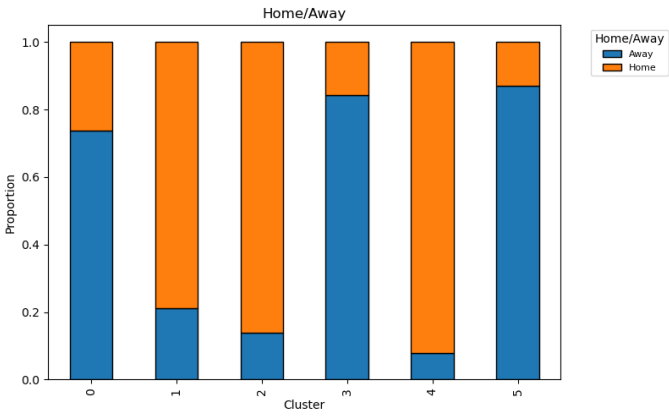


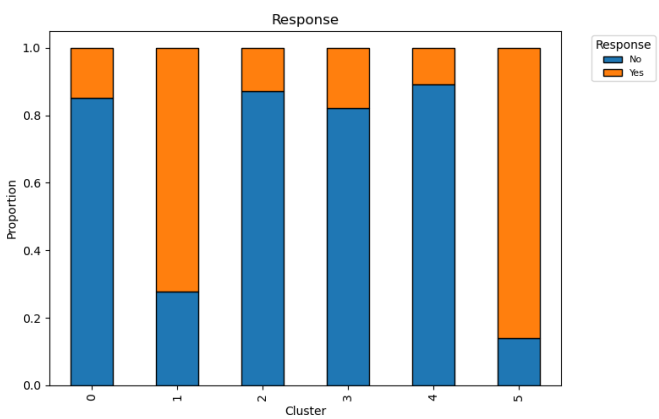


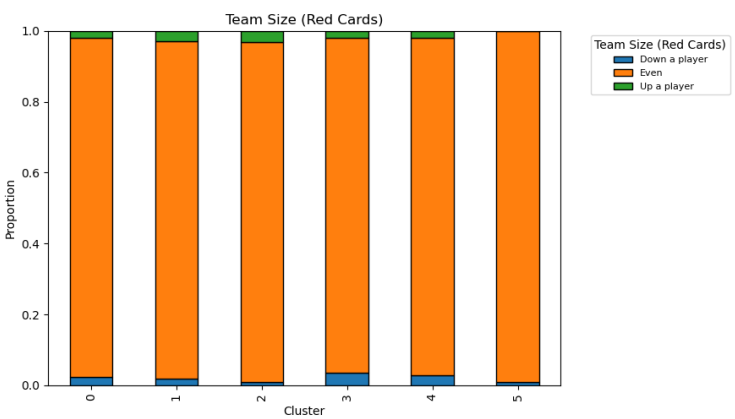










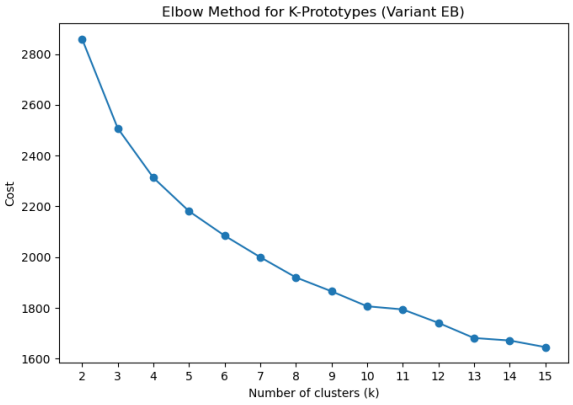


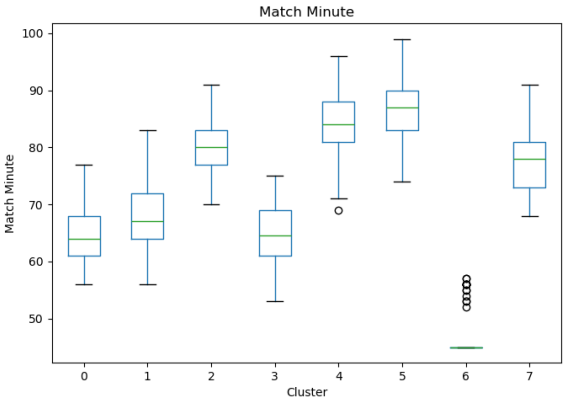
Variant D		
Number of Clusters = 6		
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**
0 (592)	3.176 (1.396)	Early Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent
1 (317)	2.890 (1.311)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Weaker Opponent
2 (294)	2.840 (1.447)	Late Non-Reactive Double Substitutions in Winning Home Games vs Weaker Opponent
3 (197)	2.990 (1.414)	Late Non-Reactive Single Substitutions in Drawing Away Games vs Stronger Opponent
4 (102)	2.912 (1.599)	Mid-Second Half Non-Reactive Triple Substitutions in Winning Home Games vs Equal Opponent
5 (108)	2.880 (1.464)	Mid-Second Half Reactive Double Substitutions in Losing Away Games vs Stronger Opponent

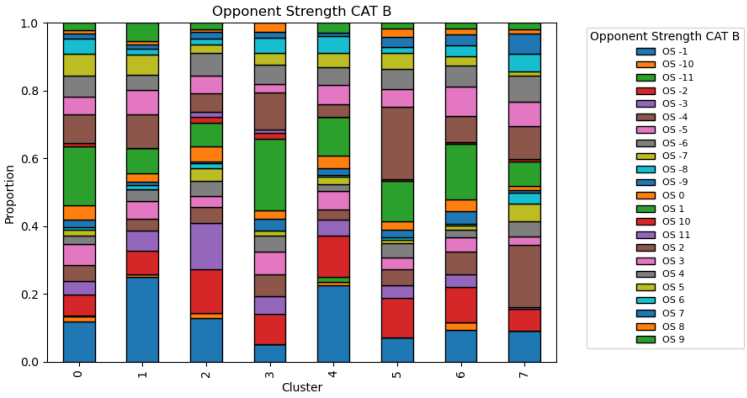
Variable	**Adjusted Rand Index**	**Feature Importance**
Opponent Strength (Cat D)	0.3349	0.1908
Match Minute (Cat D)	0.3011	0.1733
Number of Subs	0.3837	0.1508
Response	0.3487	0.0970
Home/Away	0.3875	0.0873
Team Size (Red Cards)	1.0000	0.0324
Score Difference	0.4510	0.2684

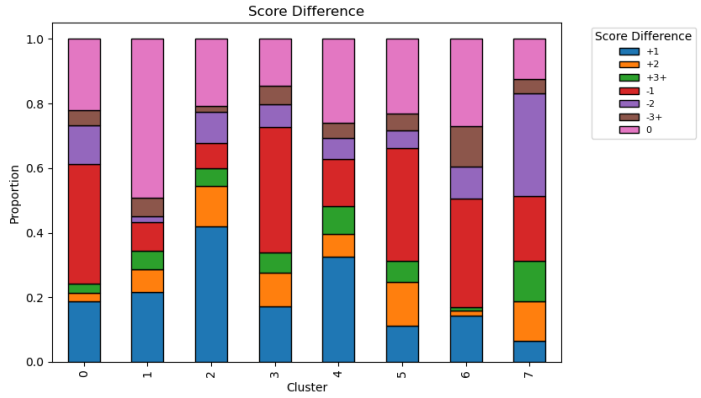
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.29	0.29	0.29	0.36	0.39	0.36
2	0.46	0.39	0.39			
3	0.40	0.44	0.44			
4	0.33	0.31	0.31			
5	0.36	0.39	0.39			

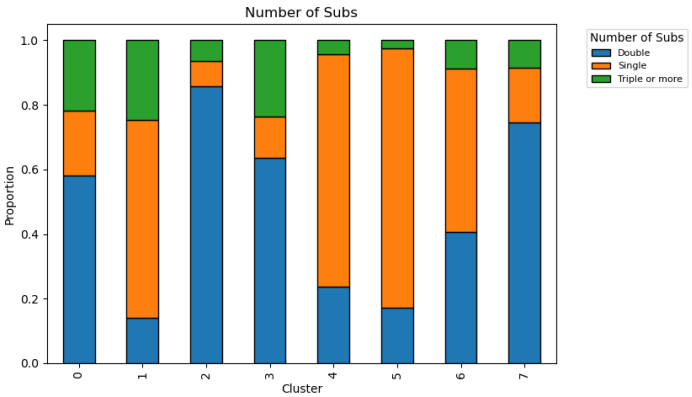
Appendix 8: Variant EB Results

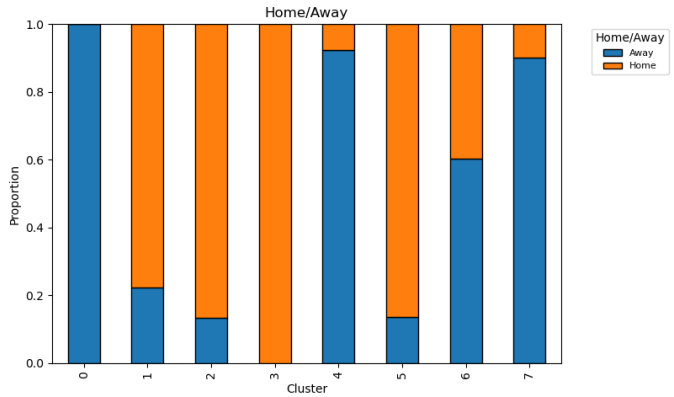


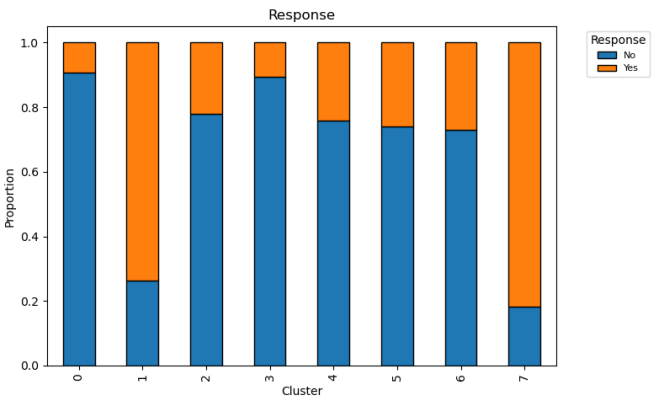


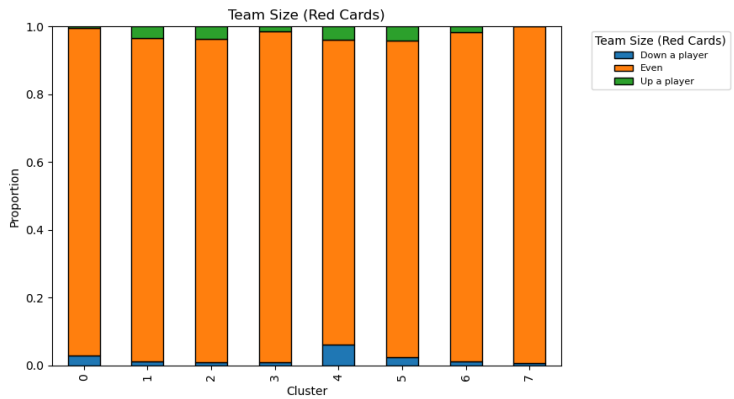






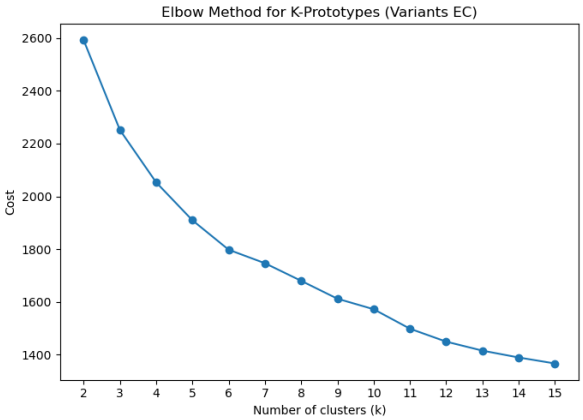


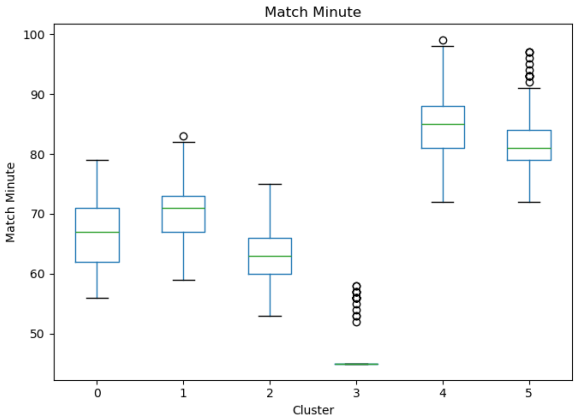


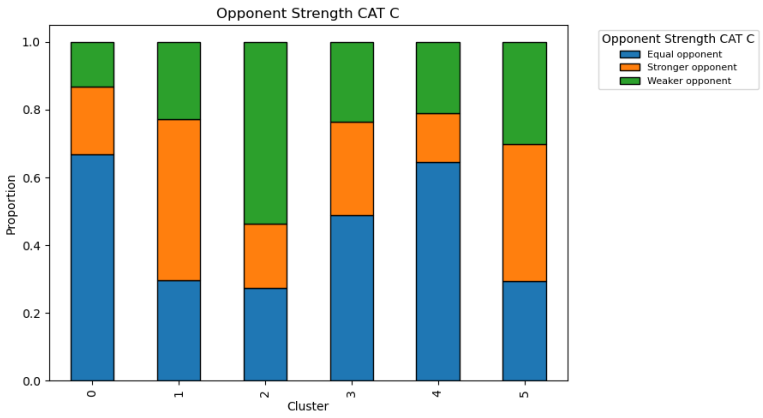


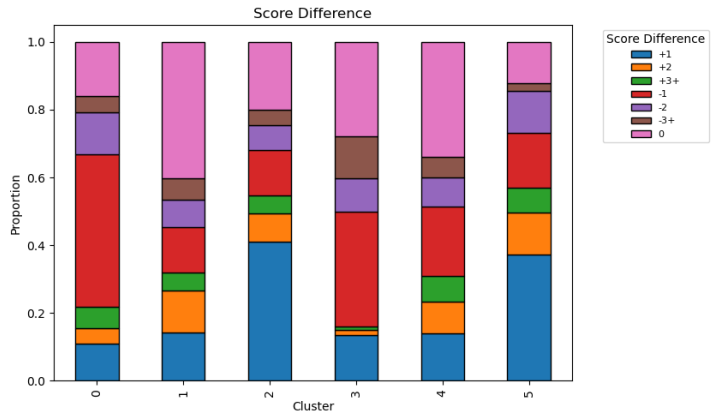
Variant EB						
Number of Clusters = 8						
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**				
0 (277)	2.924 (1.424)	Mid-Second Half Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent				
1 (171)	2.924 (1.328)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Equal Opponent				
2 (217)	2.968 (1.422)	Late Non-Reactive Double Substitutions in Winning Home Games vs Weaker Opponent				
3 (228)	3.268 (1.352)	Mid-Second Half Non-Reactive Double Substitutions in Losing Home Games vs Stronger Opponent				
4 (212)	2.896 (1.417)	Late Non-Reactive Single Substitutions in Winning Away Games vs Weaker Opponent				
5 (169)	2.929 (1.412)	Late Non-Reactive Single Substitutions in Losing Home Games vs Stronger Opponent				
6 (182)	3.060 (1.535)	Early Non-Reactive Single Substitutions in Losing Away Games vs Equal Opponent				
7 (154)	3.006 (1.403)	Late Reactive Double Substitutions in Losing Away Games vs Weaker Opponent				
Variable	**Adjusted Rand Index**	**Feature Importance**				
Opponent Strength (Cat B)	0.4721	0.3367				
Match Minute (Num)	0.1556	0.3086				
Number of Subs	0.3471	0.0845				
Response	0.3539	0.0540				
Home/Away	0.2726	0.0486				
Team Size (Red Cards)	0.3151	0.0178				
Score Difference	0.4553	0.1499				
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.62	0.48	0.54	0.59	0.58	0.60
2	0.67	0.59	0.63			
3	0.52	0.67	0.58			
4	0.55	0.56	0.56			
5	0.65	0.67	0.66			

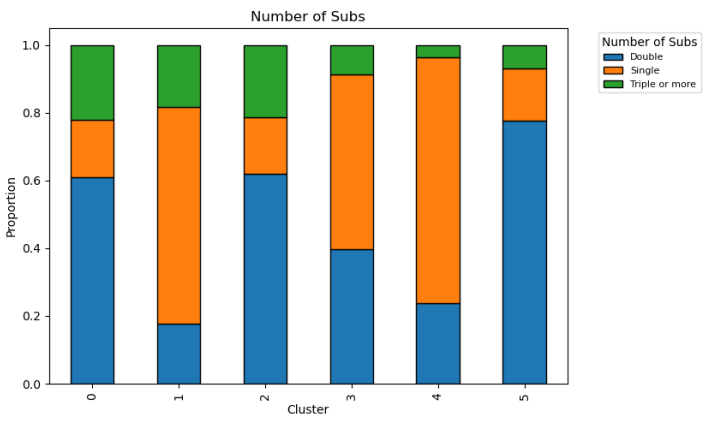
Appendix 9: Variant EC Results

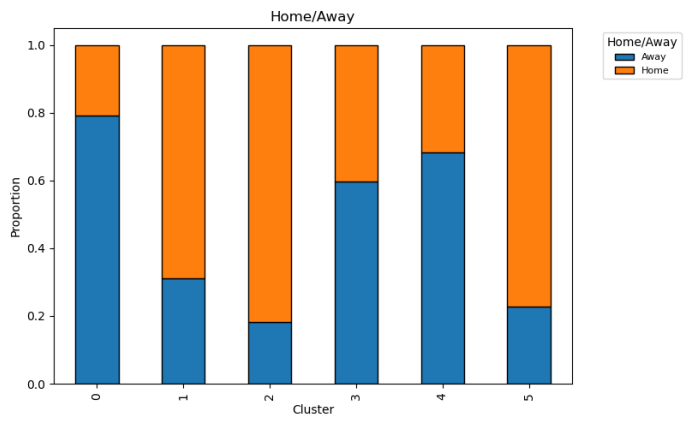


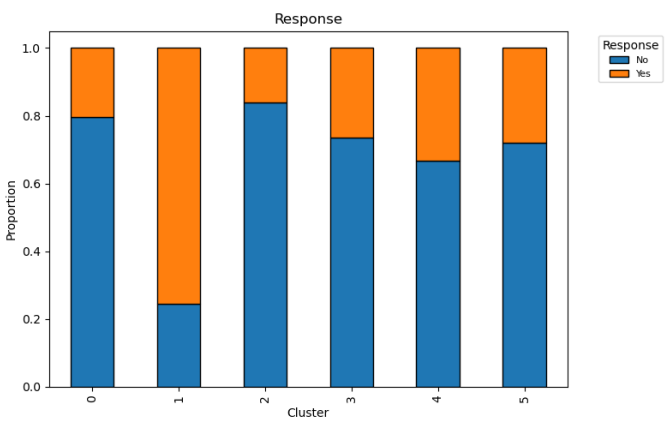


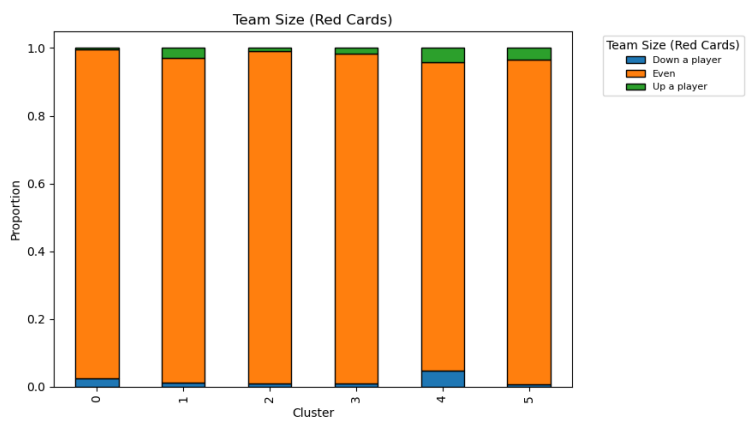










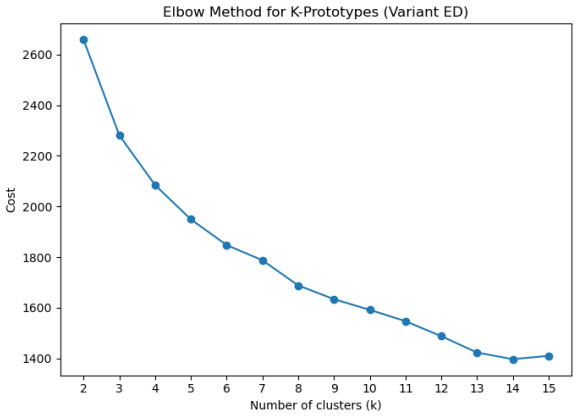


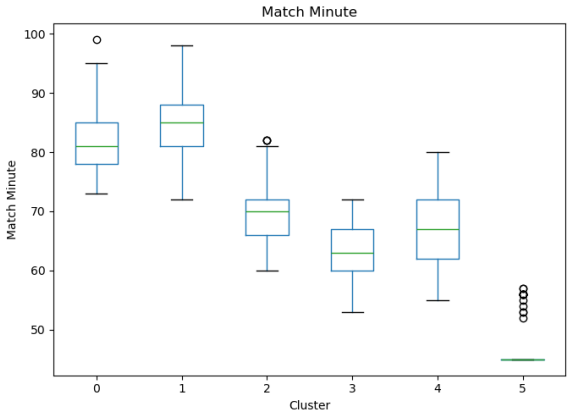
Variant EC		
Number of Clusters = 6		
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**
0 (373)	2.992 (1.400)	Mid-Second Half Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent
1 (176)	2.869 (1.305)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Stronger Opponent
2 (229)	3.227 (1.383)	Mid-Second Half Non-Reactive Double Substitutions in Winning Home Games vs Weaker Opponent
3 (186)	3.059 (1.532)	Early Non-Reactive Single Substitutions in Losing Away Games vs Equal Opponent
4 (356)	2.972 (1.394)	Late Non-Reactive Single Substitutions in Drawing Away Games vs Equal Opponent
5 (290)	2.900 (1.458)	Late Non-Reactive Double Substitutions in Winning Home Games vs Stronger Opponent

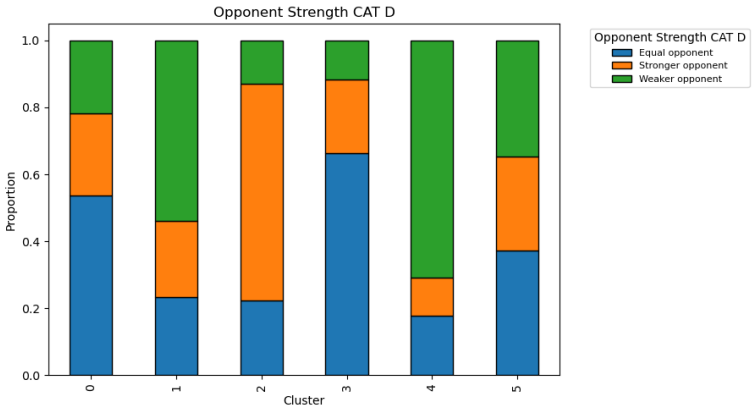
Variable	**Adjusted Rand Index**	**Feature Importance**
Opponent Strength (Cat C)	0.3697	0.1039
Match Minute (Num)	0.1431	0.5385
Number of Subs	0.4326	0.0890
Response	0.2894	0.0548
Home/Away	0.3182	0.0464
Team Size (Red Cards)	1.0000	0.0207
Score Difference	0.6572	0.1468

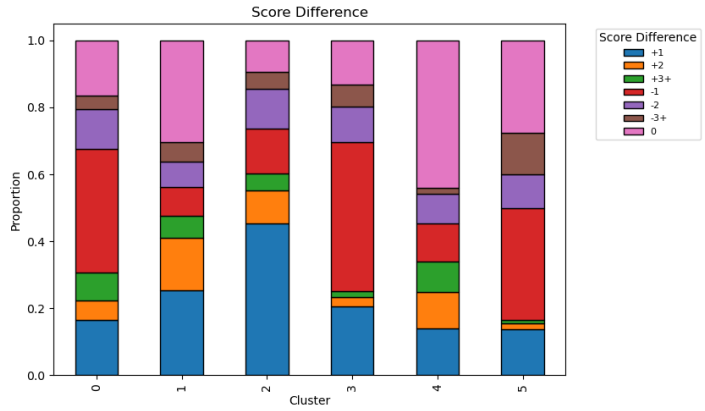
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.60	0.44	0.51	0.55	0.56	0.55
2	0.57	0.57	0.57			
3	0.50	0.56	0.53			
4	0.54	0.65	0.59			
5	0.57	0.55	0.56			

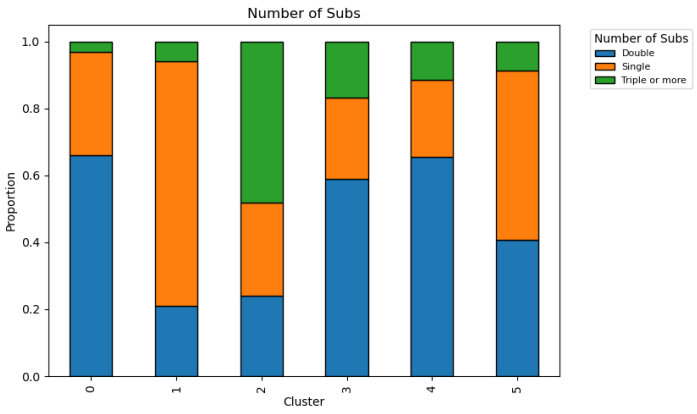
Appendix 10: Variant ED Results

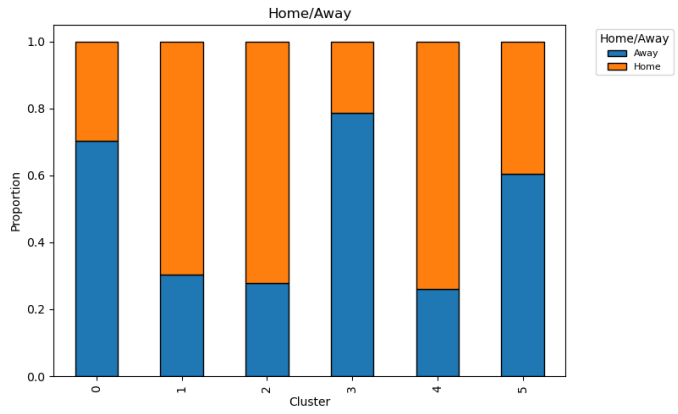


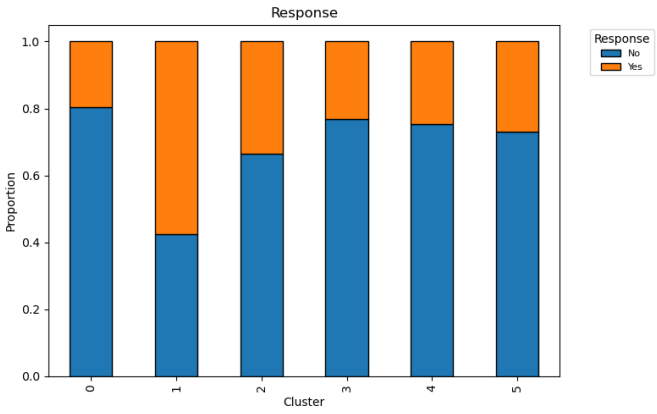


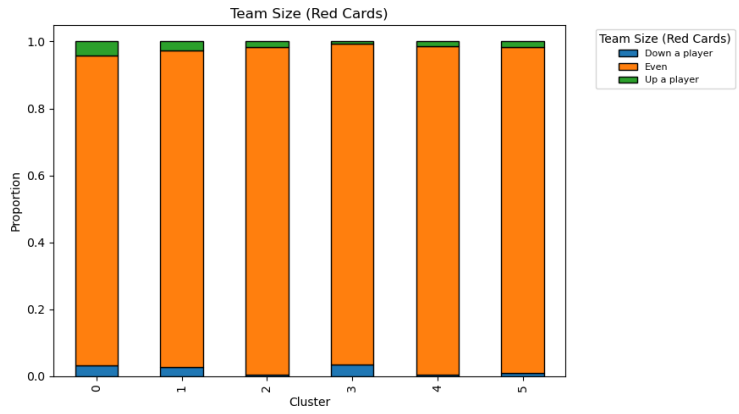






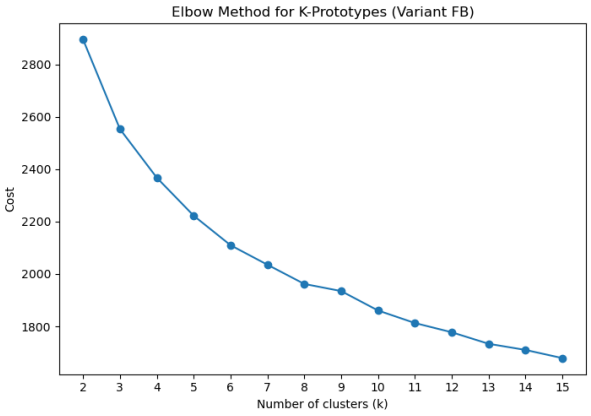


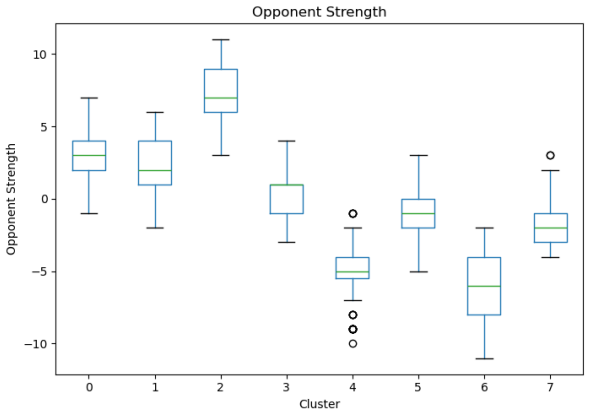


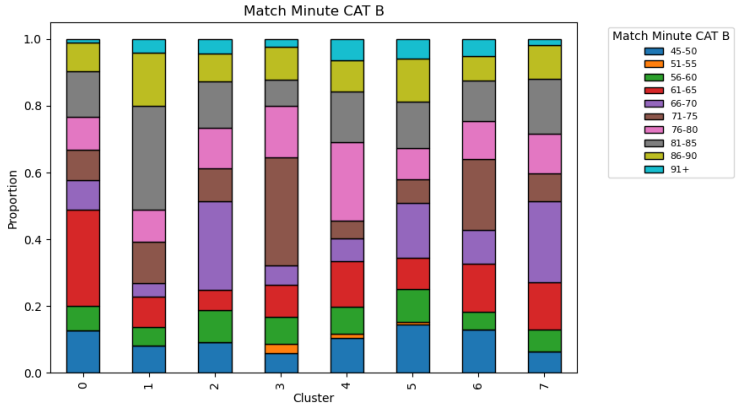


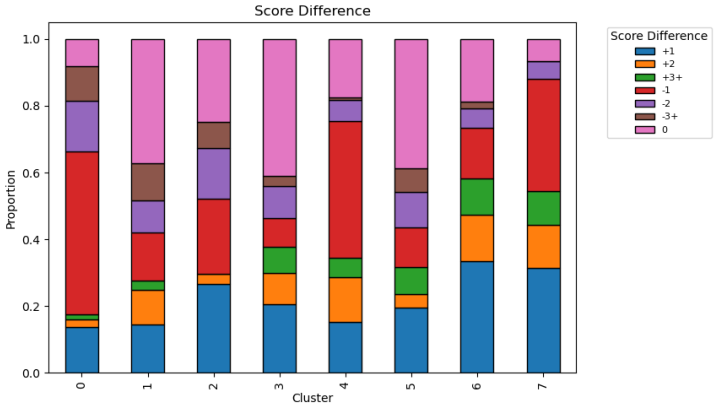
Variant ED						
Number of Clusters = 6						
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**				
0 (372)	2.970 (1.447)	Late Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent				
1 (304)	2.819 (1.319)	Late Reactive Single Substitutions in Drawing Home Games vs Weaker Opponent				
2 (179)	2.804 (1.378)	Mid-Second Half Non-Reactive Triple Substitutions in Winning Home Games vs Stronger Opponent				
3 (282)	3.135 (1.460)	Mid-Second Half Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent				
4 (291)	3.165 (1.337)	Mid-Second Half Non-Reactive Double Substitutions in Drawing Home Games vs Weaker Opponent				
5 (182)	3.071 (1.541)	Early Non-Reactive Single Substitutions in Losing Away Games vs Equal Opponent				
Variable	**Adjusted Rand Index**	**Feature Importance**				
Opponent Strength (Cat D)	0.4637	0.1076				
Match Minute (Num)	0.1855	0.5328				
Number of Subs	0.4386	0.0882				
Response	0.3392	0.0544				
Home/Away	0.3689	0.0473				
Team Size (Red Cards)	1.0000	0.0203				
Score Difference	0.3580	0.1496				
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.61	0.56	0.59	0.57	0.55	0.57
2	0.64	0.57	0.60			
3	0.49	0.50	0.49			
4	0.52	0.60	0.56			
5	0.59	0.59	0.59			

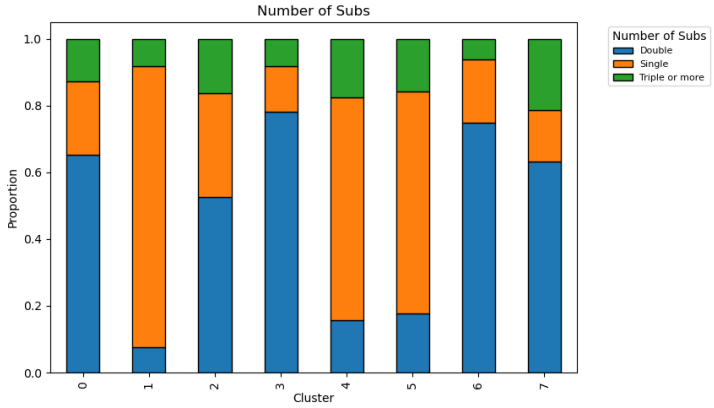
Appendix 11: Variant FB Results

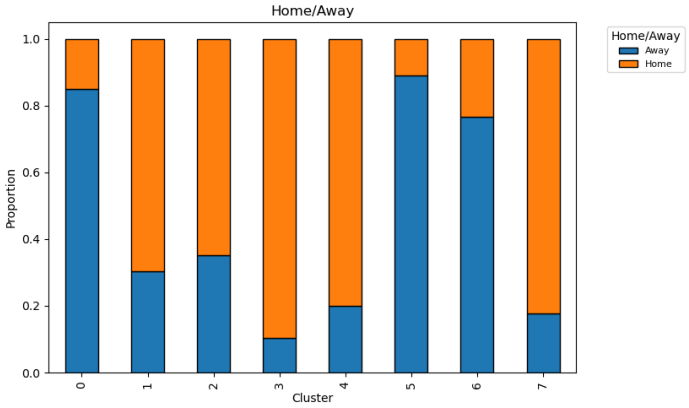


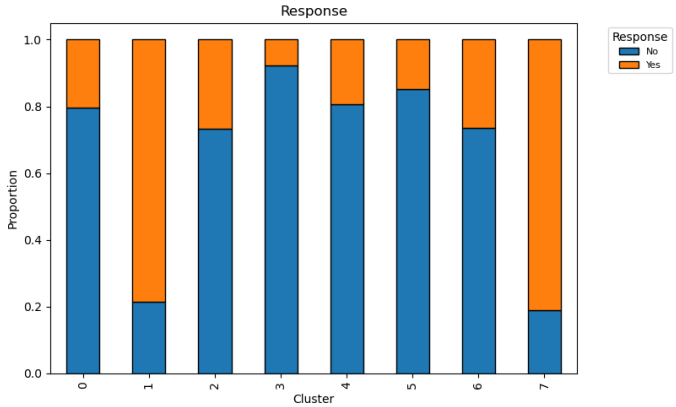


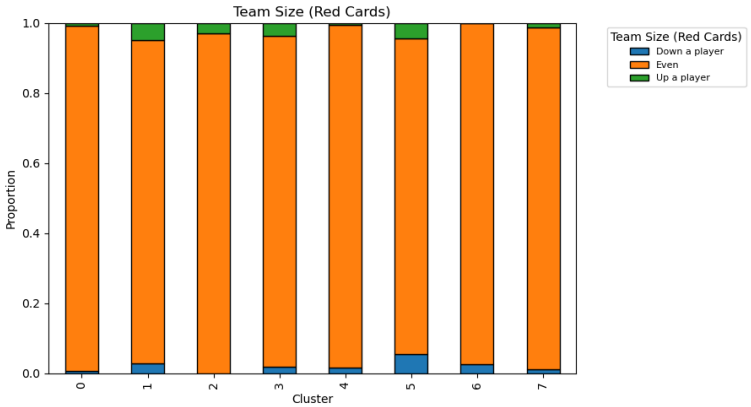






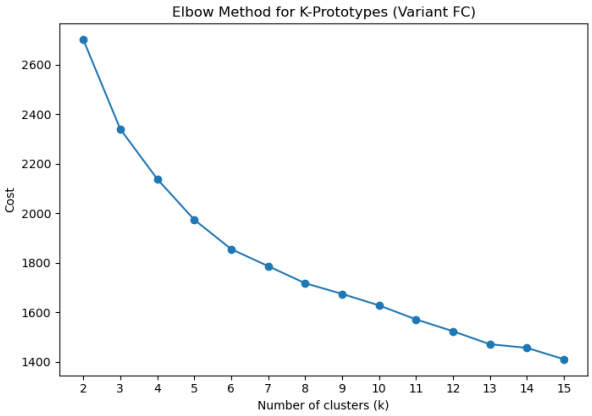


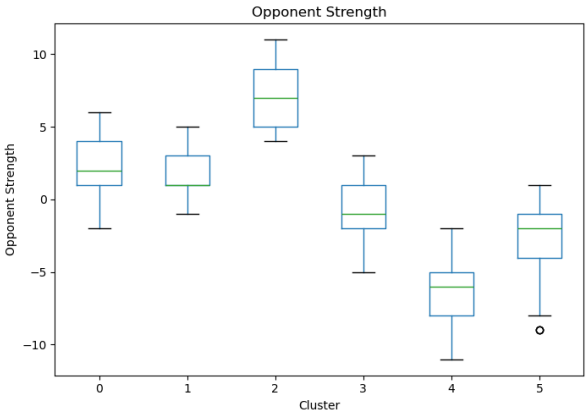


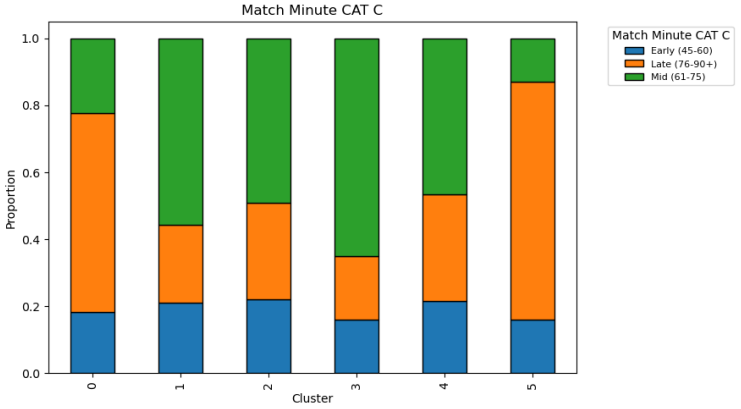


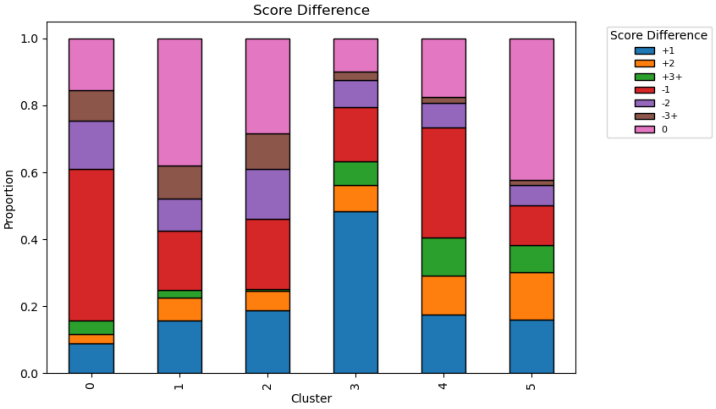
Variant FB						
Number of Clusters = 8						
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**				
0 (293)	3.007 (1.436)	Mid-Second Half Non-Reactive Double Substitutions in Losing Away Games vs Stronger Opponent				
1 (145)	2.910 (1.424)	Late Reactive Single Substitutions in Drawing Home Games vs Equal Opponent				
2 (165)	3.091 (1.431)	Mid-Second Half Non-Reactive Double Substitutions in Winning Home Games vs Stronger Opponent				
3 (220)	3.086 (1.416)	Late Non-Reactive Double Substitutions in Drawing Home Games vs Equal Opponent				
4 (171)	2.813 (1.398)	Late Non-Reactive Single Substitutions in Losing Home Games vs Weaker Opponent				
5 (255)	3.055 (1.399)	Mid-Second Half Non-Reactive Single Substitutions in Drawing Away Games vs Equal Opponent				
6 (192)	3.328 (1.396)	Mid-Second Half Non-Reactive Double Substitutions in Winning Away Games vs Weaker Opponent				
7 (169)	2.586 (1.312)	Mid-Second Half Reactive Double Substitutions in Losing Home Games vs Equal Opponent				
Variable	**Adjusted Rand Index**	**Feature Importance**				
Opponent Strength (Num)	0.1585	0.3892				
Match Minute (Cat B)	0.4084	0.2231				
Number of Subs	0.2873	0.0938				
Response	0.2798	0.0555				
Home/Away	0.2700	0.0542				
Team Size (Red Cards)	0.2907	0.0215				
Score Difference	0.2722	0.1628				
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.59	0.48	0.53	0.59	0.56	0.59
2	0.61	0.51	0.56			
3	0.48	0.56	0.52			
4	0.64	0.71	0.67			
5	0.64	0.69	0.67			

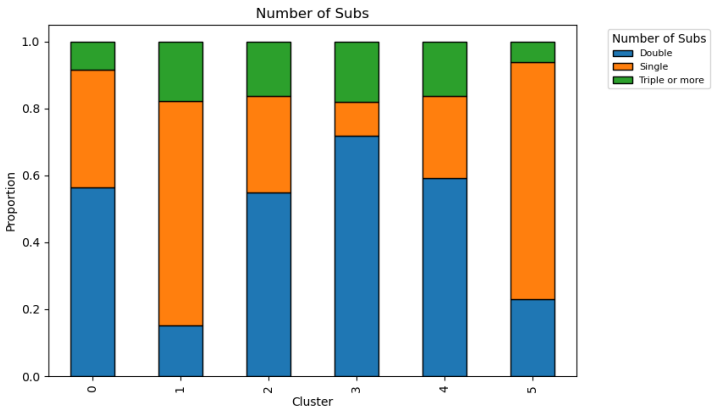
Appendix 12: Variant FC Results

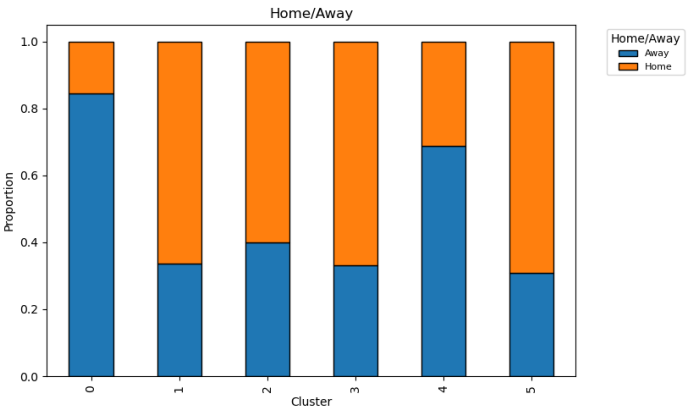


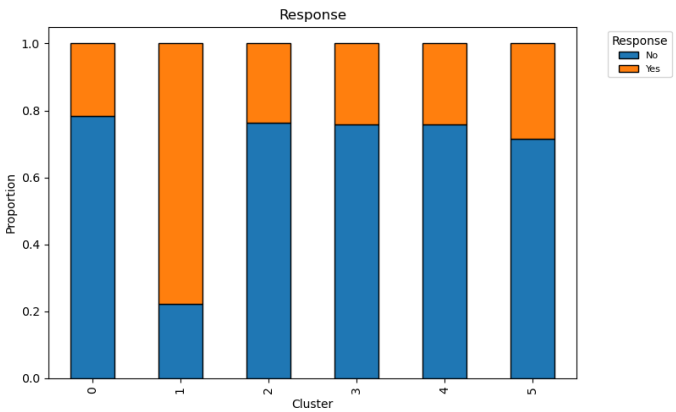


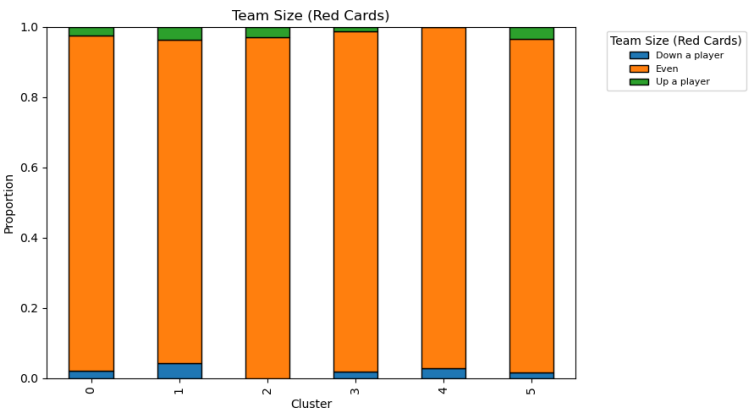






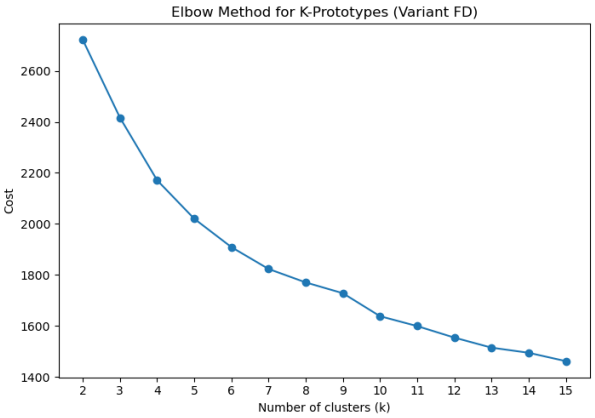


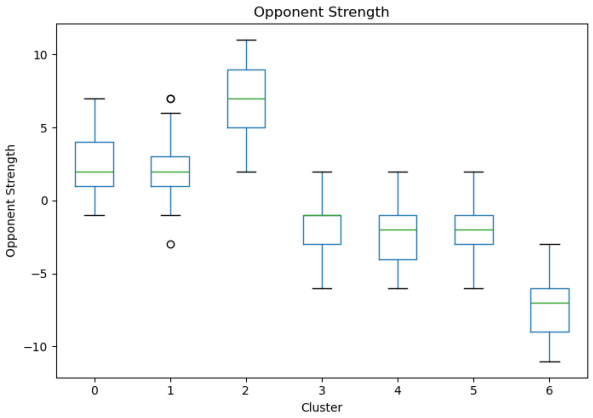


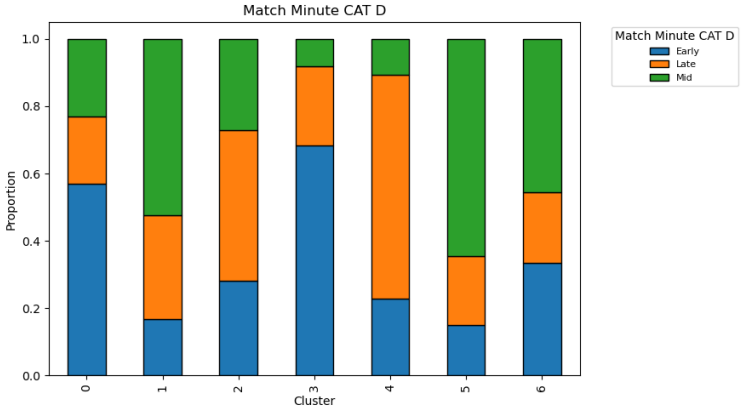


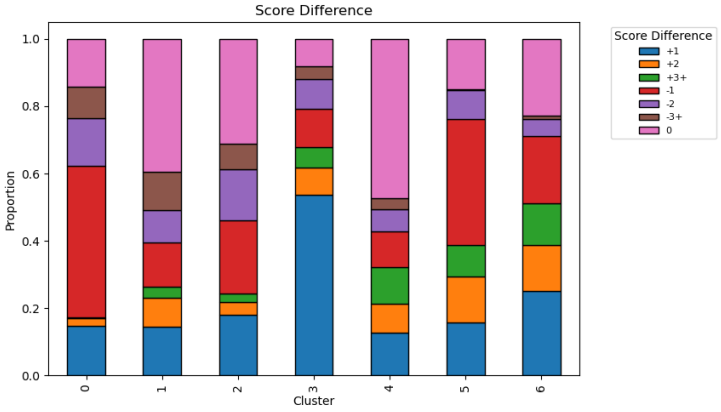
Variant FC						
Number of Clusters = 6						
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**				
0 (360)	2.967 (1.435)	Late Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent				
1 (190)	2.963 (1.423)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Equal Opponent				
2 (208)	3.043 (1.459)	Mid-Second Half Non-Reactive Double Substitutions in Drawing Home Games vs Stronger Opponent				
3 (317)	3.025 (1.421)	Mid-Second Half Non-Reactive Double Substitutions in Winning Home Games vs Equal Opponent				
4 (240)	3.138 (1.424)	Mid-second half Non-Reactive Double Substitutions in Losing Away Games vs Weaker Opponent				
5 (295)	2.888 (1.337)	Late Non-Reactive Single Substitutions in Drawing Home Games vs Equal Opponent				
Variable	**Adjusted Rand Index**	**Feature Importance**				
Opponent Strength (Num)	0.1898	0.5336				
Match Minute (Cat C)	0.4216	0.1008				
Number of Subs	0.4214	0.0910				
Response	0.3596	0.0581				
Home/Away	0.3571	0.0436				
Team Size (Red Cards)	0.6442	0.0228				
Score Difference	0.2145	0.1502				
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.56	0.40	0.46	0.51	0.52	0.51
2	0.52	0.49	0.51			
3	0.47	0.54	0.50			
4	0.46	0.54	0.50			
5	0.58	0.59	0.59			

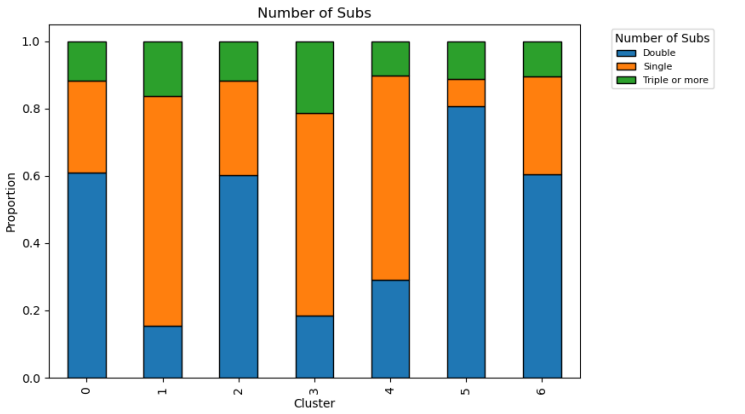
Appendix 13: Variant FD Results

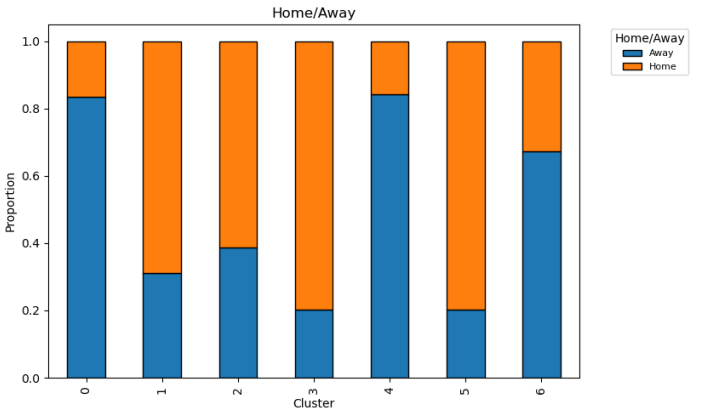


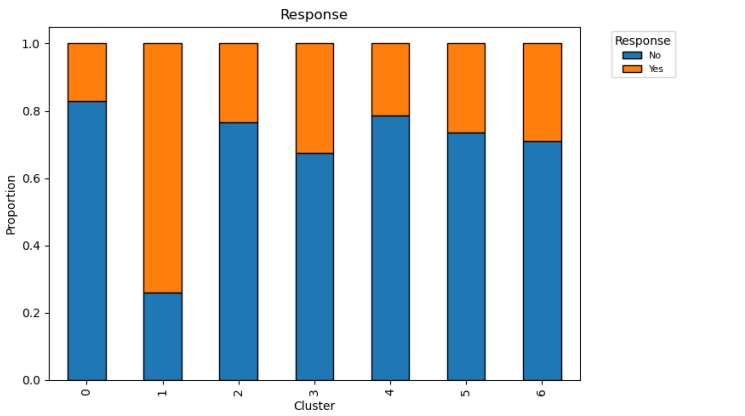


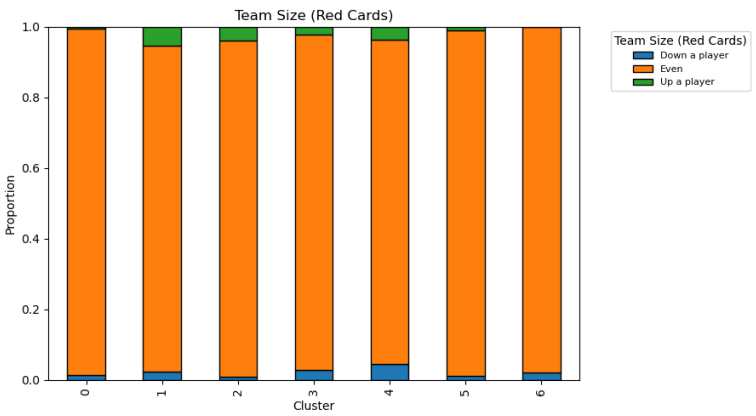












Variant FD						
Number of Clusters = 7						
Cluster (Number of Subs)	**Mean Impact Score (Std)**	**Cluster Description**				
0 (340)	2.915 (1.417)	Early Non-Reactive Double Substitutions in Losing Away Games vs Equal Opponent				
1 (208)	3.087 (1.384)	Mid-Second Half Reactive Single Substitutions in Drawing Home Games vs Equal Opponent				
2 (206)	3.146 (1.427)	Late Non-Reactive Double Substitutions in Drawing Home Games vs Stronger Opponent				
3 (212)	2.962 (1.420)	Early Reactive Single Substitutions in Winning Home Games vs Equal Opponent				
4 (196)	2.995 (1.430)	Late Non-Reactive Single Substitutions in Drawing Away Games vs Equal Opponent				
5 (268)	2.802 (1.391)	Mid-Second Half Non-Reactive Double Substitutions in Losing Home Games vs Equal Opponent				
6 (180)	3.228 (1.410)	Mid-Second Half Non-Reactive Double Substitutions in Winning Away Games vs Weaker Opponent				
Variable	**Adjusted Rand Index**	**Feature Importance**				
Opponent Strength (Num)	0.1556	0.5296				
Match Minute (Cat D)	0.4653	0.1013				
Number of Subs	0.4162	0.0933				
Response	0.3552	0.0558				
Home/Away	0.4201	0.0445				
Team Size (Red Cards)	1.0000	0.0227				
Score Difference	0.2946	0.1528				
Impact Score	**Precision**	**Recall**	**F1 Score**	**Macro F1**	**Validation Accuracy**	**Test Accuracy**
1	0.49	0.35	0.41	0.52	0.49	0.52
2	0.64	0.59	0.62			
3	0.48	0.52	0.50			
4	0.43	0.50	0.46			
5	0.57	0.63	0.60			

Appendix 14: Sensitivity Analysis for Cluster Separation

Variant	Number of Clusters	Min Mean Impact Score	Max Mean Impact Score	Difference	Cluster Difference
Variant A	6	2.873	3.180	0.307	-0.074
Variant A	7	2.848	3.081	0.233	
Variant B	7	2.861	3.141	0.280	-0.033
Variant B	5	2.888	3.135	0.247	
Variant C	6	2.822	3.079	0.257	+0.259
Variant C	9	2.767	3.283	0.516	
Variant D	6	2.840	3.176	0.336	+0.018
Variant D	4	2.805	3.160	0.354	
Variant EB	8	2.896	3.268	0.371	+0.221
Variant EB	10	2.560	3.152	0.592	
Variant EC	6	2.896	3.227	0.358	+0.162
Variant EC	9	2.618	3.138	0.520	
Variant ED	6	2.804	3.165	0.360	+0.070
Variant ED	8	2.769	3.199	0.430	
Variant FB	8	2.586	3.328	0.742	-0.311
Variant FB	10	2.799	3.230	0.431	
Variant FC	6	2.888	3.138	0.249	+0.069
Variant FC	8	2.842	3.160	0.318	
Variant FD	7	2.802	3.228	0.426	-0.217
Variant FD	8	2.929	3.139	0.209	