

Club Football Match Data from 27 leagues worldwide, 2000-2025

Adam Gábor

March 20, 2025

1 Abstract

This paper presents a comprehensive association football dataset of match statistics collected from 2000 to 2025. The dataset includes approximately 230,000 matches across 27 countries and 42 leagues. It provides detailed match results along with statistical data from the match. It includes pre-match odds which reflect prediction market probabilities, as well as team-specific metrics such as recent Elo ratings and team form. The dataset features an advanced match classification system that categorizes matches into six distinct clusters, each representing different style of play. Provided in two CSV files, the dataset is readily accessible and designed for researchers interested in sports analytics, machine learning, and football outcome prediction.

2 Background & Summary

The growing global popularity of sports has been accompanied by an increased availability of sports data. However, progress in sports analytics has been hindered by the fragmentation and inconsistency of publicly available datasets, as well as the lack of standardized open-access resources. This paper introduces a dataset designed to facilitate the analysis and prediction of football matches, addressing the need for a large, up-to-date, and easily accessible resource for the football analytics community.

Inspired by open-science principles, this dataset is made publicly available to encourage further research and development in the field. While similar datasets exist, they often lack either scale, recency, or a combination of key features—such as Elo ratings, team descriptors, and match clusters—offered here. Although commercial data providers offer even more comprehensive statistics, these datasets remain prohibitively expensive for amateur analysts, researchers, and small sports organizations.

Beyond the raw data collection, which integrates match event data from Football-Data.co.uk with the well-established Elo rating system, this dataset contributes to the research community through the novel application of cluster analysis to identify distinct match types. Building upon the author's prior research in football analytics, matches have been categorized into six clusters representing different tactical approaches and game characteristics. These clusters introduce a new dimension for analysis, enabling researchers to examine team performance across various match types, explore the effectiveness of specific tactical systems in different contexts, and track the evolution of playing styles over time.

The dataset has broad applications across multiple disciplines. In sports science, researchers can analyze performance patterns, examine the impact of various factors on team success, and assess the effectiveness of different tactical strategies. In computer science, the dataset provides a rich and reliable testing ground for machine learning algorithms focused on result prediction, pattern recognition and anomaly detection. In economics, researchers can study market efficiency, the relationship between team quality and prediction market indicators, and the effects of information asymmetry on prediction markets. In the social sciences, the dataset facilitates research on competitive dynamics, home advantage phenomena, and winning streaks in sports performance.

A key motivation for making this dataset publicly available is to lower the barrier to entry for researchers and practitioners interested in football data analytics, thereby democratizing access to high-quality football data.

3 Methods

3.1 Data Acquisition

Our dataset compilation leveraged a comprehensive acquisition and processing workflow designed to ensure data consistency, accuracy, and analytical utility. The primary data collection process involved two distinct data sources. For match results and statistics, we systematically acquired data from Football-Data.co.uk, which aggregates match information from multiple leagues dating back to 2000. This platform hosts the largest publicly available football match results and statistics database on the internet. Simultaneously, we collected team strength data through the ClubElo API, capturing standardized Elo ratings for approximately 500 clubs across Europe. These ratings were sampled at regular intervals (specifically on the 1st and 15th of each month) to create a temporal record of team performance evolution over the dataset’s 25-year span. This dual-source approach provided complementary perspectives on team performance: objective match outcomes from Football-Data.co.uk and mathematically derived team strength metrics from ClubElo.

3.2 Data Processing and Integration

The data integration process presented significant technical challenges, particularly in harmonizing team and statistical identifiers across sources. We developed a custom Python pipeline using pandas and MySQL to standardize formats, resolve inconsistencies, and create a unified dataset. A critical component of this pipeline was a comprehensive team name mapping system that resolved variations in spelling, abbreviations, and linguistic differences. For instance, "Bayern" from ClubElo was consistently mapped to "Bayern Munich" in the match data, enabling accurate linkage across sources. The mapping system contains over 100 team name variants linked to canonical identifiers, ensuring that teams could be tracked consistently across their entire history despite naming changes or different conventions used by data providers. This mapping was implemented through a series of data transformation rules that were applied systematically across the entire dataset.

Form metrics, which capture the recent performance of the team, were calculated through a windowed aggregation process that considered the results of the match within specific time frames. The Form3 and Form5 variables represent points accumulated by teams in their previous three and five matches respectively, calculated using the standard football point allocation system (3 points for a win, 1 for a draw, 0 for a loss). This calculation was implemented using a temporally aware algorithm that ensured only matches chronologically preceding the current match were included, preserving the integrity of the predictive information available at the time of each match. For early-season matches where sufficient prior data within the same season was unavailable, the algorithm incorporated matches from the end of the previous season to maintain consistency in the form metrics. This approach ensured that form metrics remained meaningful throughout the dataset, including at season boundaries.

Elo rating integration followed a nearest-match temporal alignment strategy. For each match, the most recent available Elo rating prior to the match date was identified and associated with the participating teams. This process required careful handling of temporal relationships to ensure that no future information leaked into historical match records. When Elo ratings were unavailable for certain teams—particularly for lower-division teams or leagues outside ClubElo’s primary European focus—we did not implement fallback mechanisms, such as estimating Elo values based on betting odds, as this would have introduced redundancy and added no analytical value. Out of the 230,000 matches in the dataset, over 112,000 include Elo rating information.

3.3 Match Clustering

A distinctive feature of our dataset is the inclusion of match clustering, which categorizes games based on their statistical characteristics. This clustering was implemented using the K-means algorithm in scikit-learn, applied to a feature space constructed from key match attributes. The feature engineering process transformed raw match statistics into meaningful dimensions including shot efficiency differential (measuring the relative shooting accuracy of teams), possession dominance (derived from indicators such as shots and corners), game physicality (based on fouls), game tempo (based on total shots), and Elo differential (capturing pre-match team strength disparity). These features were scaled using a robust scaler to minimize the influence of outliers, then clustered into six distinct match types. The

resulting clusters represent fundamentally different match patterns: LTH (low tempo home-oriented), HTB (high tempo balanced contests), LTA (low tempo away-oriented), VAD (visibly away-dominated games), VHD (visibly home-dominated games), and PHB (physical balanced encounters). The cluster probabilities for each match were calculated using a softmax transformation of distances to cluster centroids, allowing for nuanced characterization of matches that exhibit characteristics of multiple clusters.

3.4 Data Validation and Outlier Handling

The entire data processing workflow was implemented as a fault-tolerant pipeline with extensive validation checks and error handling. Missing values were preserved rather than imputed to maintain data authenticity, enabling end-users to make informed decisions about handling incomplete information. The pipeline incorporated parallel processing techniques to efficiently handle the large volume of data, processing match files in batches to optimize computational resources. All transformations were applied consistently across the entire dataset, ensuring that analytical patterns detected in one period or league could be meaningfully compared with others. This methodological consistency is particularly valuable for longitudinal studies examining how football has evolved over the dataset’s quarter-century timespan. The resulting dataset structure provides clear documentation for researchers to develop additional derived features beyond those included in the base dataset, extending its utility for diverse analytical applications.

4 Data Records

Goal: Describe the dataset files, their formats, and contents. Explain the file structure and data organization. Reference the repository. Provide overview tables of column descriptions.

The dataset described in this paper can be accessed publicly via Kaggle, GitHub or HuggingFace under the MIT licence. The data is organized in two CSV files: `elo_ratings.csv` and `matches.csv`. Together, these files contain over 470,000 rows of data and occupy roughly 48MB of storage.

The `elo_ratings.csv` file contains semi-monthly snapshots of Elo ratings for the most prominent football clubs in europe, with ratings taken on the 1st and 15th of each month. The Elo ratings range from approximately 1200 to 2000, with higher-ranked teams occupying the upper echelon above 1900 points during their peak performance period.

The `matches.csv` file forms the core of the dataset, containing detailed information about individual football matches. The data includes match identifiers (division, date, time), team information, Elo ratings, form indicators, full-time and half-time results, and match statistics (shots, shots on target, fouls, corners, yellow and red cards). Additionally, the file includes extensive prediction market average odds data from approximately 17 European bookmakers. Data completeness varies by league and season, with top-tier competitions having the highest completion rates for statistical fields, with smaller leagues having more limited statistical coverage. All matches in the dataset are sorted chronologically.

Table 1: Column Descriptions for `elo_ratings.csv`

Column	Data Type	Description
Date	date	Date of the snapshot.
Club	string	Club name in English corresponding to Matches table.
Country	enum	Club country three-letter code.
Elo	float	Club’s current Elo rating, rounded to two decimal spots.

Table 2: Column Descriptions for matches.csv

Column	Data Type	Description
Division	enum	League that the match was played in - country code + division number (I1 for Italian First Division). For countries where we only have one league, we use 3-letter country code (ARG for Argentina).
MatchDate	date	Match date in the classic YYYY-MM-DD format.
MatchTime	time	Match time in the HH:MM:SS format. CET-1 timezone.
HomeTeam	string	Home team's club name in English, abbreviated if needed.
AwayTeam	string	Away team's club name in English, abbreviated if needed.
HomeElo	float	Home team's most recent Elo rating.
AwayElo	float	Away team's most recent Elo rating.
Form3Home	int	Number of points gathered by home team in the last 3 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 9).
Form5Home	int	Number of points gathered by home team in the last 5 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 15).
Form3Away	int	Number of points gathered by away team in the last 3 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 9).
Form5Away	int	Number of points gathered by away team in the last 5 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 15).
FTHome	int	Full-time goals scored by home team.
FTAway	int	Full-time goals scored by away team.
FTResult	enum	Full-time result (H for Home win, D for Draw and A for Away win).
HTHome	int	Half-time goals scored by home team.
HTAway	int	Half-time goals scored by away team.
HTResult	enum	Half-time result (H for Home win, D for Draw and A for Away win).
HomeShots	int	Total shots (goal, saved, blocked, off-target) by home team.
AwayShots	int	Total shots (goal, saved, blocked, off-target) by away team.
HomeTarget	int	Total shots on target (goal, saved) by home team.
AwayTarget	int	Total shots on target (goal, saved) by away team.
HomeFouls	int	Total fouls by home team.
AwayFouls	int	Total fouls by away team.
HomeCorners	int	Total corners taken by home team.
AwayCorners	int	Total corners taken by away team.
HomeYellow	int	Total yellow cards awarded to home team players (excl. staff).
AwayYellow	int	Total yellow cards awarded to away team players (excl. staff).
HomeRed	int	Total red cards awarded to home team players (excl. staff).
AwayRed	int	Total red cards awarded to away team players (excl. staff).
OddHome	float	Bet365's Home Team Win Odd.
OddDraw	float	Bet365's Draw Odd.
OddAway	float	Bet365's Away Team Win Odd.
MaxHome	float	Maximum Home Team Win Odd from 17 European bookmakers.
MaxDraw	float	Maximum Draw Odd from 17 European bookmakers.
MaxAway	float	Maximum Away Team Win Odd from 17 European bookmakers.
Over25	float	Bet365's Over 2.5 Total Goals Scored Odd.
Under25	float	Bet365's Under 2.5 Total Goals Scored Odd.
MaxOver25	float	Maximum Over 2.5 Total Goals Scored Odd from 17 European bookmakers.

Continued on next page

Table 7 continued

Column	Data Type	Description
MaxUnder25	float	Maximum Under 2.5 Total Goals Scored Odd from 17 European book-makers.
HandiSize	float	Asian handicap size for home team (negative number indicating stronger home team) .
HandiHome	float	Bet365's Home Team Win Odd with the given handicap size for Home team.
HandiAway	float	Bet365's Away Team Win Odd with the given handicap size for Home team.
C_LTH	float	Likelihood of falling into C_LTH cluster.
C_LTA	float	Likelihood of falling into C_LTA cluster.
C_VHD	float	Likelihood of falling into C_VHD cluster.
C_VAD	float	Likelihood of falling into C_VAD cluster.
C_HTB	float	Likelihood of falling into C_HTB cluster.
C_PHB	float	Likelihood of falling into C_PHB cluster.

Our dataset introduces a novel approach to football match classification through unsupervised machine learning. K-means clustering identified 6 clusters representing fundamentally different game styles and team matchups that have the ability to enhance predictive modeling:

Table 3: Key Characteristics of Football Match Clusters

Characteristic	LTH	LTA	VHD	VAD	HTB	PHB
Dataset %	20	18	13	14	19	16
Shot Efficiency	+23	-24	+6	-4	0	+1
Home Possession %	51	39	70	38	55	54
Total Fouls	24.90	24.93	24.81	25.30	25.27	30.40
Shot Count	22.46	22.53	22.76	22.15	24.32	23.49
Elo Diff (H-A)	+7	0	+200	-200	0	-4
Avg. Result (H-A)	1.6-0.7	1.0-1.3	2.3-0.7	1.0-1.8	1.6-1.4	1.4-1.1
Percentage Differences from Dataset Average						
Home Shots	-4.1	-2.9	+3.2	-3.1	+5.2	+2.9
Away Shots	-3.1	-3.5	-1.3	+3.0	+5.6	+1.3
Home Fouls	-4.1	-3.9	-3.4	-0.8	-1.6	+18.4
Away Fouls	-4.3	-3.5	-3.3	-1.4	-2.1	+18.0
Home Yellow Cards	-3.9	-4.1	-2.2	+4.3	-0.4	+13.8
Away Yellow Cards	-3.5	-2.8	-0.5	+2.5	+0.2	+12.2
Home Red Cards	-1.0	-5.1	-2.2	+5.1	-2.4	+17.5
Away Red Cards	-3.0	-1.4	-4.8	-5.5	-3.1	+25.8
Home Goals	-0.9	-2.6	+2.1	-8.1	+1.0	-0.7
Away Goals	-2.6	+2.6	+2.6	+3.7	+4.1	-2.0

The dataset structure facilitates various types of analysis through its comprehensive set of pre-match, in-match, and post-match indicators. This multi-dimensional approach enables researchers to investigate not only what happened in each match but also the context in which it occurred and how it compared to expectations. The dataset is particularly valuable for developing predictive models that can estimate match outcomes, goal totals, or specific statistical events based on pre-match indicators. The longitudinal nature of the data, spanning 25 years, also enables investigation of how the game has evolved over time in terms of playing styles, competitive balance, and statistical patterns.

5 Technical Validation

To validate the reliability and utility of the dataset, we conducted a series of predictive modeling experiments. These experiments serve two purposes: first, to verify the internal consistency and predictive

power of the included features, and second, to establish performance benchmarks for different analytical approaches that researchers might apply to the dataset. We divided our validation process into three main areas: match outcome prediction (win, draw, loss), exact score prediction, and assessment of the added value provided by the match clustering.

The foundational validation involved predicting match outcomes using various methods, ranging from simple heuristics to sophisticated machine learning algorithms. As shown in Table , naive approaches such as random guessing achieved approximately 33% accuracy, while always predicting a home win (the most common outcome) yielded 44% accuracy. Market odds-based predictions demonstrated significantly better performance (48% accuracy), comparable to predictions based on pre-match Elo ratings (47%). This validates the dataset’s representation of team strength through both the included Elo rating system and the prediction market odds. More sophisticated machine learning approaches substantially outperformed these baselines, with the best models achieving in-play accuracies above 65%. The superior performance of these models underscores the richness of the dataset’s feature set and its potential for predictive applications. The table below shows accuracy for 3-way prediction (home, draw, away).

Method	Accuracy	F1 Score
Random Prediction	0.334	0.328
Always Predict Home Win	0.444	0.205
Odds-Weighted Prediction	0.480	0.446
Elo-Weighted Prediction	0.473	0.447
Form-Weighted Prediction	0.410	0.369
Half-time Result + Odds	0.617	0.558
Machine Learning Models		
Logistic Regression	0.646	0.567
LightGBM	0.648	0.610
Neural Network	0.653	0.614

Table 4: Performance of Match Outcome Prediction Methods

We extended our validation to the more challenging task of exact score prediction. This assessment not only tested the dataset’s capability to support detailed outcome predictions but also validated the consistency of goal-scoring patterns across the included leagues and seasons. Table below presents the performance metrics for score prediction models. The simple Poisson distribution model, which assumes that goals follow a Poisson process, achieved baseline performance ($R^2 = 0.426$), confirming that goal-scoring patterns in the dataset conform to established statistical distributions to some extent. However, more sophisticated models incorporating the dataset’s rich feature set substantially outperformed this baseline. The best (yet still fairly simple) ensemble model, combining neural networks with gradient-boosted trees, achieved an R^2 of 0.515, demonstrating the dataset’s capacity to support nuanced modeling of match dynamics beyond simple statistical assumptions.

Model	Exact Score Acc.	R^2	MAE	Brier
Poisson Distribution	0.190	0.426	0.642	0.190
Linear Regression	0.238	0.492	0.575	0.174
LightGBM	0.237	0.493	0.575	0.177
Neural Network	0.244	0.489	0.573	0.174
Ensemble (NN+XGB+LGB)	0.257	0.515	0.552	0.167

Table 5: Performance of Match Outcome Prediction Methods

A key innovation in our dataset is the inclusion of match clustering, which categorizes games into distinct types based on their statistical profiles. To validate the value of this approach, we have taken a subset of the dataset and compared models trained on the entire subset (global models) against cluster-specific models and meta-models that weighted combined both approaches. As shown in table below, we found that the meta-model approach consistently outperformed both global and cluster-specific models, validating the additional analytical dimension provided by the clustering. The improvement was modest but consistent across different model architectures, with meta-models showing an average accuracy improvement of 0.28% and F1 score improvement of 0.37% compared to global models.

While these gains may appear incremental, they represent a significant enhancement in a domain where prediction improvements of even fractions of a percent are valuable.

Model/Accuracy	Global	Avg-Cluster	Avg-Duo	Meta
LightGBM	0.624	0.632	0.638	0.638
Logistic Regression	0.641	0.638	0.637	0.637
Neural Network	0.638	0.638	0.637	0.638
Random Forest	0.622	0.629	0.631	0.632

Table 6: Performance Comparison of Global, Cluster-Specific, and Meta Models

Our validation tests also revealed that different match clusters benefit from different prediction approaches. For instance, we found that visibly home-dominated (VHD) matches showed the highest predictability (accuracy up to 0.692), while low-tempo away-oriented (LTA) matches were the most challenging to predict (lowest accuracy at 0.606). This pattern is consistent with intuitive expectations about match dynamics, where clear dominance by one team creates more predictable outcomes. These findings validate the cluster categorizations as meaningful distinctions that capture genuine differences in match characteristics. The dataset allows researchers to explore these differences systematically, potentially leading to more nuanced understanding of how match dynamics influence outcomes.

Our technical validation confirms that the dataset provides a robust foundation for football analytics research, is consistent and can be reliably used for designing statistical models that align with theoretical expectations about goal-scoring and match outcome patterns. The value of novel match clustering approach is validated by the consistent, albeit modest, improvements shown by meta-models leveraging this additional dimension of information. These results establish performance benchmarks for future research using this dataset and confirm its suitability for a wide range of analytical applications in sports science and related fields.

6 Usage Notes

This dataset is designed to support a wide range of analytical applications in football research. For match outcome prediction, we recommend beginning with pre-match features (Elo ratings, form metrics, and market odds) before incorporating in-play statistics. When handling missing values, which are more common in statistical fields for lower leagues, researchers should consider league-specific or division-level imputation rather than global approaches, as statistical patterns vary significantly across competition levels, or refrain from using imputation at all, even to the extent of losing some data.

Time-series analyses should account for the seasonal structure of football competitions, with careful handling of season boundaries where team compositions change due to promotion and relegation. For researchers interested in exploring temporal patterns, we suggest implementing a sliding window approach when calculating team form and momentum indicators. This provides a more dynamic representation of team performance trajectories than the static Form3 and Form5 metrics included in the base dataset. Examples for that include number of goals scored in last N matches, number of points gained home/away in last N matches, win streak for home and away matches, recent Elo change or number of rest days between matches.

When working with odds data, users should note that these represent market expectations rather than ground truth probabilities; the implicit overround (bookmaker margin) should be accounted for in probability-based analyses. The match clusters provide an additional analytical dimension that can be leveraged through ensemble approaches or segment-specific modeling strategies.

The match clusters are based on a specific clustering method and feature set. Users should carefully consider the cluster descriptions and potentially explore alternative clustering approaches or feature sets depending on their research questions.

The table below outlines recommended derived features that researchers can implement to extend the dataset’s analytical capabilities. These features have demonstrated value in previous research and can be readily calculated from the base dataset without requiring external information:

Table 7: Recommended Derived Features

Column	Data Type	Description
Match Date-Time	datetime	Combination of match date and time into a single datetime data type.
Total Goals	int	Total goals scored by both the Home and Away teams.
1XOdd	float	Combined odds for a Double Chance bet: Home Team Win or Draw. (Maximum value 'Max1XOdd' is also possible).
X2Odd	float	Combined odds for a Double Chance bet: Away Team Win or Draw. (Maximum value 'MaxX2Odd' is also possible).
12Odd	float	Combined odds for a Double Chance bet: Home Team Win or Away Team Win. (Maximum value 'Max12Odd' is also possible).
Elo Difference	float	The difference between the Home Team's Elo rating and the Away Team's Elo rating.
Elo Total	float	The sum of the Home Team's Elo rating and the Away Team's Elo rating.
Elo Advantage	float	The EloDifference divided by EloTotal, normalizing the Elo difference to a percentage.
Form3 Difference	int	The Home Team's Form3 score minus the Away Team's Form3 score.
Form5 Difference	int	The Home Team's Form5 score minus the Away Team's Form5 score.
Form Momentum Home	int	Difference between Home Team's recent and older form, derived from the formula $\text{Form3Home} - (\text{Form5Home} - \text{Form3Home})$. Values range from -15 (worst momentum) to 18 (best momentum).
Form Momentum Away	int	Difference between Away Team's recent and older form, derived from the formula $\text{Form3Away} - (\text{Form5Away} - \text{Form3Away})$. Values range from -15 (worst momentum) to 18 (best momentum).
Implied Probability Home	float	Probability of a Home Team Win, derived from OddHome using the formula $1/\text{OddHome}$.
Implied Probability Draw	float	Probability of a Draw, derived from OddDraw using the formula $1/\text{OddDraw}$.
Implied Probability Away	float	Probability of an Away Team Win, derived from OddAway using the formula $1/\text{OddAway}$.
Implied Probability Total	float	The sum of ImpliedProbHome, ImpliedProbDraw, and ImpliedProbAway.
Bookmaker Margin	float	$\text{ImpliedProbTotal} - 1$. Understood as "market uncertainty," and can help distinguish between clear-favorite matches and matches with unpredictable outcomes.
Shots Difference	int	The Home Team's total shots minus the Away Team's total shots.
Shots Total	int	The sum of Home Team's total shots and the Away Team's total shots.
Shot Accuracy Home	float	Home Team's shots on target divided by Home Team's total shots, representing shot accuracy as a percentage. This tends to increase as the match progresses. Higher accuracy doesn't always mean a better team.
Shot Accuracy Away	float	Away Team's shots on target divided by Away Team's total shots, representing shot accuracy as a percentage. This tends to increase as the match progresses. Higher accuracy doesn't always mean a better team.
Shot Accuracy Diff	float	The Home Team's shot accuracy minus the Away Team's shot accuracy.
Corners Difference	int	The Home Team's total corners minus the Away Team's total corners.
Corners Total	int	The sum of the Home Team's total corners and the Away Team's total corners.

Continued on next page

Table 7 continued

Column	Data Type	Description
Game Dominance Index	float	A custom made-up metric that reflects match ball possession and the number of attacking chances, approximated by $((\text{CornersDifference} + \text{ShotsDifference})/2)$, and can be fine-tuned.
Card Points Home	int	$\text{HomeYellow} + 2 * \text{HomeRed}$. The red card multiplier (2) can be adjusted (e.g. to 3 or 4) for improved Machine Learning results.
Card Points Away	int	$\text{AwayYellow} + 2 * \text{AwayRed}$. The red card multiplier (2) can be adjusted (e.g. to 3 or 4) for improved Machine Learning results.
Card Diff	int	The Home Team's card points minus the Away Team's card points.
Draw Likelihood	float	The weighted percentage likelihood of a draw, derived from EloDifference , Form5Difference , and ImpliedProbDraw . This helps counter-weight machine learning methods biased towards clear winners.

For those new to football analytics, we recommend several starting approaches to familiarize themselves with the dataset. Beginning with exploratory data analysis of key distributions and relationships, researchers can develop intuition for typical patterns and anomalies in football statistics. Basic predictive models using only Elo differentials and home advantage can establish a performance baseline before introducing more complex features. Analyzing a single league over multiple seasons provides insight into competition-specific patterns while controlling for league structure and quality. These exploratory approaches provide entry points that can lead to more sophisticated analyses as researchers develop domain knowledge and technical expertise in football analytics.

7 Code Availability

The dataset is available on GitHub at <https://github.com/xgabora/Club-Football-Match-Data-2000-2025> under the MIT license. The repository contains both data files.

The pipeline responsible for data acquisition, cleaning, integration, feature engineering, and match clustering implementation was developed in Python 3.9 and primarily utilizes pandas (1.5.3) for data manipulation, scikit-learn (1.2.2) for preprocessing, and requests (2.28.1) for API interactions. Additional dependencies include numpy (1.24.3), matplotlib (3.7.1) for visualization, and statsmodels (0.13.5) for statistical analysis. The match clustering component uses scikit-learn's KMeans implementation with $k=6$ and a random state of 42 for reproducibility.

The pipeline includes separate modules for each processing stage: data acquisition, data cleaning, feature calculation and clustering. The codebase is updated on a bi-monthly basis to incorporate new matches into the dataset.

No restrictions are placed on the use of this code beyond those specified in the MIT license, and users are encouraged to extend the dataset for their own research purposes. All other information about the dataset are documented in the repository's README file, along with contact information for reporting issues or requesting enhancements.

References