Project Title: Data-Driven Analysis of European Football: From Shot Patterns to Goal Prediction

Team 06: Maryamsadat Parpenchi, Mahsa Rajabi Nejad

Abstract

Modern football analytics moves beyond final scores by utilizing granular event-level data to understand the underlying factors of match dynamics.

The goal of this research work is to predict the probability of a shot resulting in a goal using supervised Machine Learning models and to identify team tactical profiles through unsupervised clustering.

The chosen dataset integrates detailed play-by-play events with match metadata, comprising approximately 230,000 shooting attempts across five major European leagues.

Keywords Football Analytics — Event Data — Goal Prediction — Machine Learning — KNIME

Contents The following table outlines the structure of the report:

Introduction

In recent years, football has evolved into a data-intensive sport where decisions are increasingly supported by analytical evidence rather than intuition alone. Beyond final scores and league standings, modern performance analysis focuses on granular event-level data that captures every on-field action, such as shots, passes, and fouls. Among these, shooting actions play a central role in determining match outcomes. However, not all shots carry the same probability of resulting in a goal; factors such as location, body part, and match context significantly affect scoring likelihood. Understanding these relationships is essential for developing advanced metrics like expected goals (xG).

The primary objective of this project is to study shooting behavior and model goal probability using machine learning techniques. Despite the availability of data, raw football datasets are often complex and high-dimensional. Therefore, this study implements a structured pipeline using Python for data preparation and the KNIME Analytics Platform for modeling.

The Dataset The final analytical dataset is derived from the integration of three key sources:

1. **Event Data (events.csv):** Contains detailed play-by-play information for each match.

2. **Match Metadata (ginf.csv):** Provides contextual information such as league, season, and date.
3. **Dictionary Data (dictionary.txt):** Used to translate numerical codes into meaningful categorical labels for variables like event_type, shot_place, bodypart, and location.

The cleaned dataset contains 229,135 shot events, each described by 19 variables.

- **Id_odsp :** Categorical(String)- Unique identifier for the match.
- **Id_event :** Categorical(String)- Unique identifier for the specific event within the match.
- **League :** Categorical- The league where the match was played (e.g., E0 for Premier League).
- **Season :** Categorical- The season year (e.g., 2016).
- **Time :** Numerical- The minute of the match when the event occurred (0-90+).
- **Sort_order :** Numerical- Sequential order of events within the same minute.
- **Event_type :** Categorical- The primary classification of the event (Selected 1 for Shot/Attempt).
- **Event_team :** Categorical- Name of the team performing the action.
- **Opponent :** Categorical- Name of the opposing team.
- **Player :** Categorical- Name of the player performing the action (the shooter).
- **Player2 :** Categorical- Name of the second player involved (usually the assist provider).
- **Shot_place :** Categorical- The placement of the shot relative to the goal (e.g., Bottom left corner).
- **Shot_outcome :** Categorical- The result of the shot (e.g., On target, Blocked, Missed).
- **Location :** Categorical- The zone on the pitch where the event took place (e.g., Centre of the box).
- **Bodypart :** Categorical- The body part used for the attempt (Right foot, Left foot, Head).
- **Assist_method :** Categorical- The type of pass or action leading to the shot (e.g., Cross, Through ball).
- **Situation :** Categorical- The play context (e.g., Open play, Set piece, Corner).
- **Fast_break :** Binary (0/1)- Indicator if the event was part of a fast break (Counter-attack).
- **Is_goal :** Binary (Target)- **Target Variable.** 1 indicates a Goal, 0 indicates No Goal.

Report Organization This report is organized as follows:

1. **Data Exploration:** We examine the characteristics of shots, focusing on spatial and temporal distributions.
2. **Preprocessing:** We describe the cleaning pipeline, including decoding categorical variables and handling missing values.
3. **Models:** We describe the Supervised (Logistic Regression, Decision Tree) and Unsupervised (K-Means) models used.
4. **Evaluation:** We evaluate the performance of the models using metrics like AUC and analyze the clustering results.
5. **Conclusion:** We summarize the findings and discuss future work.

1. Data Exploration

1.1 Dataset Overview The dataset covers five major European leagues: Serie A (I1), La Liga (SP1), Ligue 1 (F1), Bundesliga (D1), and the Premier League (E0). This broad coverage allows for a comparative analysis of different footballing styles and competitive environments. The final dataset consists of 229,135 shot events, providing a robust sample size for statistical modeling.

1.2 Target Variable Analysis (is_goal) Since the primary goal is to predict goal outcomes, we analyze the distribution of the target variable is_goal. The dataset is naturally imbalanced, with goals representing approximately **10.7%** of all shooting attempts. This reflects the real-world difficulty of scoring in professional football.

- **0 (No Goal):** ~89.3%
- **1 (Goal):** ~10.7% This class imbalance suggests that accuracy alone is not a sufficient metric for evaluation, motivating the use of the ROC Curve and AUC in the modeling phase.

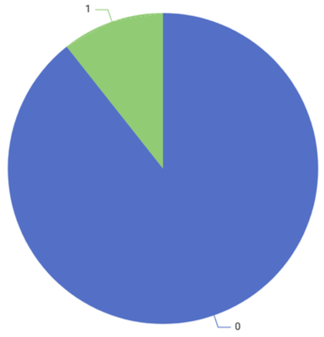

Figure 1: Distribution of the target variable is_goal, highlighting the class imbalance (10.7% Goals vs. 89.3% No Goals).

1.3 Shot Characteristics To understand the factors influencing scoring probability, we analyzed shot distribution across spatial and anatomical dimensions:

- **Shot Location:** Exploratory analysis reveals a clear structural pattern in shot locations. A large proportion of shots originates from areas **outside the penalty box**. However, despite their high frequency, these long-range attempts exhibit a relatively low conversion rate. In contrast, shots taken from the **centre of the box** occur less frequently but demonstrate a substantially higher likelihood of resulting in a goal. This confirms that spatial proximity is a critical factor in goal prediction.

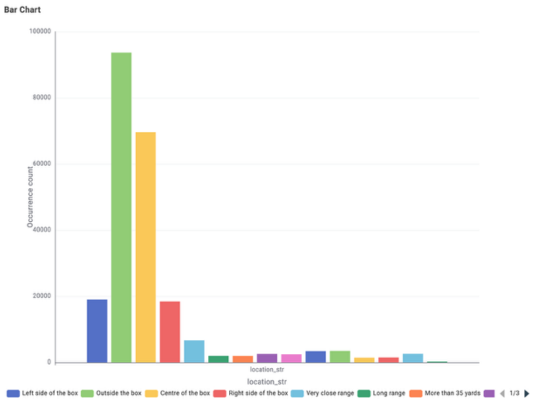

Figure 2: Distribution of shot frequency by pitch location

- **Body Part:** The analysis of body parts used in shooting shows a pronounced asymmetry. Approximately **60%** of shots are taken with the right foot, while the left foot accounts for around **25%**, and headers represent roughly **15%** of all attempts. This imbalance reflects both player dominance and tactical preferences.

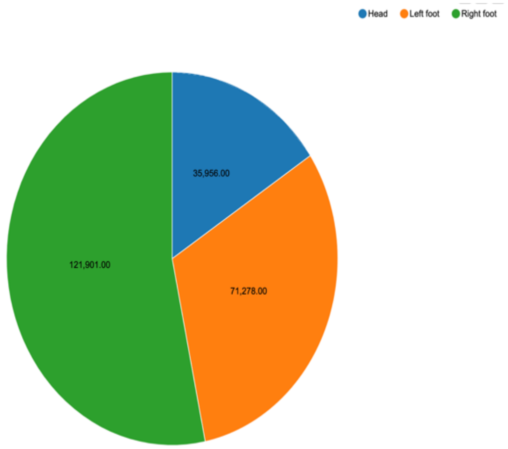

Figure 3: Share of body parts used in attempts

- **Temporal Distribution:** Goal frequency is not uniform throughout the match. Analysis shows a steady increase in goals as the match progresses, with a pronounced peak in the **final 10 minutes (80–90)**. This late-match surge can be attributed to physical fatigue and increased tactical risk-taking by trailing teams.

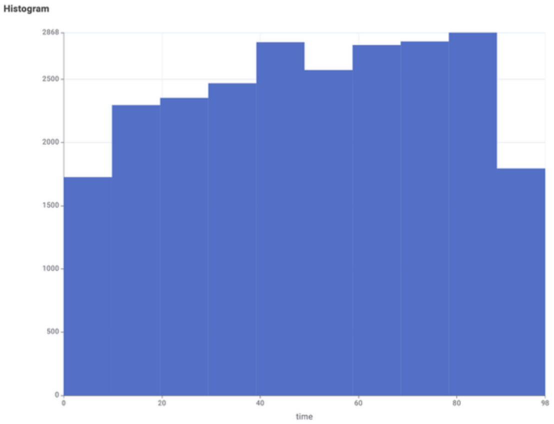

Figure 4: Temporal distribution of goals throughout the match

2. Preprocessing

Due to the complexity and scale of raw football event data, a dedicated preprocessing pipeline was required before any modeling tasks could be performed. This phase focused on transforming raw, code-based data into a clean, interpretable dataset using **Python**, while subsequent analysis was carried out in **KNIME**.

2.1 Data Integration and Decoding

The original event data encoded most categorical variables as numerical values (e.g., shot_place=3), which are unsuitable for meaningful analysis. Using the dictionary.txt file, we mapped these codes to descriptive textual labels (e.g., "Bottom left corner"). Simultaneously, the event-level data was merged with match metadata (ginf.csv)

using the unique match identifier id_odsp. This enriched each shot with context such as league, season, and country.

2.2 Feature Filtering and Target Definition

The raw dataset contained various event types, including fouls, cards, and substitutions. To align with the project's objective of **goal prediction**, we applied a strict filter to retain only records where event_type = 1 (Attempt/Shot). The target variable was defined as is_goal, a binary indicator where:

- **1:** Represents a Goal.
- **0:** Represents a miss or save .

2.3 Handling Missing Values

Missing values were addressed using domain-informed imputation strategies to avoid data loss:

- **Assist Method:** Missing values were imputed with **"Solo Run"**, reflecting shots taken without a recorded assist.
- **Player Names:** Missing entries were filled with "Unknown" to preserve dataset integrity.

3. Models : The modeling phase was conducted in the **KNIME Analytics Platform**, ensuring a transparent and reproducible workflow. The analytical framework was divided into two parallel branches: Supervised Learning for prediction and Unsupervised Learning for profiling.

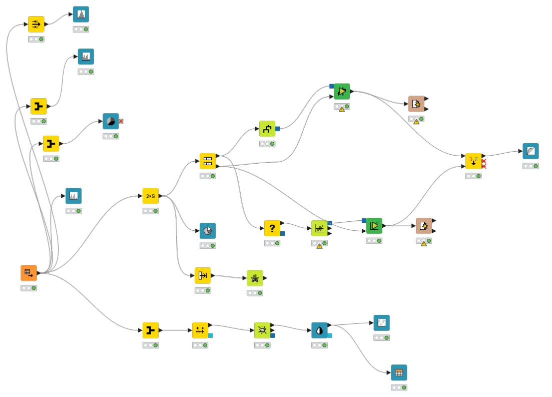

Figure 5: The KNIME analytical workflow

3.1 Supervised Learning: Goal Prediction

The primary task was to predict the is_goal variable based on shot characteristics (e.g., location, body part, situation).

- **Partitioning:** The dataset was split into **70% Training** and **30% Testing** sets. **Stratified Sampling** was applied to preserve the class balance of the target variable (10.7% goals) in both partitions.
- **Algorithms:** Two classifiers were implemented and compared:
    1. **Decision Tree:** Selected for its interpretability and rule-based structure, allowing us to visualize decision paths.
    2. **Logistic Regression:** Selected for its probabilistic nature. Unlike Decision Trees which classify based on hard rules, this model estimates the probability $P(y=1)$ that a shot will result in a goal using the sigmoid function:

$$P(is - goal = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$

This explicitly models the relationship between input features (location, body part, etc.) and the likelihood of scoring, making it the mathematical foundation for Expected Goals (xG) metrics. To prevent overfitting, high-cardinality variables such as player and date were excluded from this model.

3.2 Unsupervised Learning: Team Profiling In parallel, an unsupervised approach was used to group teams based on playing style.

- **Aggregation:** Shot-level data were aggregated by event_team to calculate **Total Shots** (Volume) and **Average Goals** (Efficiency).
- **Normalization:** Features were scaled using **Min-Max Normalization** (0–1 range) to ensure equal weight during clustering.
- **Clustering:** The k-Means algorithm was applied to group teams into distinct tactical profiles. We selected k=3 primarily based on domain knowledge. Alternative values of k were tested informally, but k=3 provided the most interpretable clusters, capturing three classic tiers of team performance: High-performing (Elite), Average (Mid-table), and Low-performing (Defensive/Struggling)

3.3 Association Analysis (Pattern Mining) To capture the sequential nature of football, we extended the analysis using **Association Rule Mining** (Apriori Algorithm). Unlike clustering which groups static features, this method identifies frequent patterns in event sequences. We treated consecutive match events as "transactions" to discover rules such as `{Corner, Header} $\rightarrow$ {Goal}`. We utilized **Lift** and **Confidence** metrics to filter out random occurrences and identify high-value tactical combinations .

4. Evaluation

4.1 Classification Performance

To rigorously evaluate the predictive power of our models, we utilized both Accuracy (threshold-dependent) and AUC - Area Under Curve (threshold-independent). Given that "Goals" are rare events (representing only ~10.7% of the data), Accuracy alone can be misleading. Therefore, the ROC Curve serves as our primary performance indicator to assess discrimination power.

Comparison of Models:

The table below summarizes the performance on the test set:

Model	**Accuracy**	**AUC**	**Observation**
Decision Tree	93.1%	0.941	Strong baseline, stable performance.
Logistic Regression	**94.3%**	**0.967**	Superior discrimination power.

ROC Curve Analysis:

As illustrated in Figure 6, both models perform significantly better than a random classifier. However, the Logistic Regression curve (Green line) consistently dominates the Decision Tree (Blue line).

- **Stability:** Unlike typical decision trees which often suffer from overfitting (memorizing training data), our tree model showed robust generalization on the test set.
- **Probabilistic Precision:** The Logistic Regression achieves a near-perfect AUC of 0.967. This indicates that the model is exceptionally good at ranking a "Goal" shot higher than a "No Goal" shot. This probabilistic output is crucial for **xG (Expected Goals)** calculations, confirming that linear relationships map well to shooting probabilities in this domain.

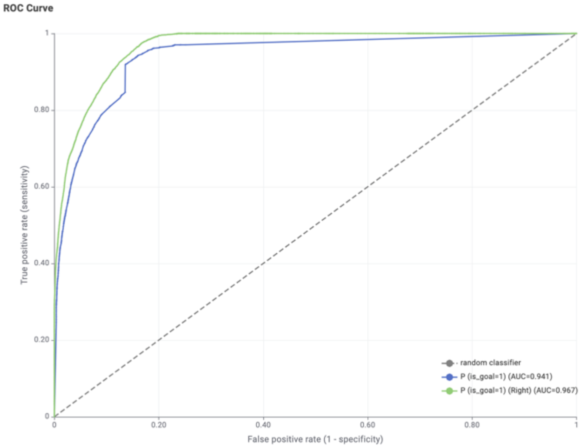

Figure 6: ROC Curves comparing Logistic Regression (Green line, AUC=0.967) and Decision Tree (Blue line, AUC=0.941).

4.2 Clustering Results (Unsupervised Learning)

In addition to prediction, we evaluated the tactical profiles of teams using k-Means Clustering ($k=3$). We analyzed the relationship between Shot Volume (Quantity) and Average Goals (Quality/Efficiency).

Figure 7 reveals three distinct clusters that align with real-world football knowledge:

1. **The Elite Cluster (Top-Right):** Teams in this group exhibit both high shot frequency and high conversion rates. These represent dominant league leaders who create many chances and finish them clinically.
2. **The Defensive/Low-Block Cluster (Bottom-Left):** Teams with low shot volume and low efficiency. These teams likely adopt a defensive strategy or struggle to convert the few chances they create.
3. **The Mid-Table Cluster (Center):** Teams with average performance in both metrics, representing the balanced majority of the leagues.

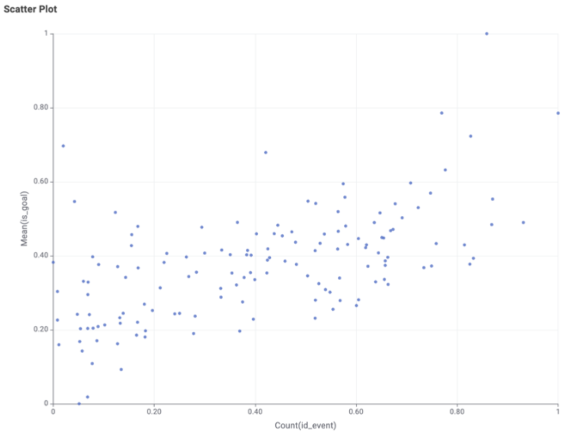

Figure 7: Team clustering based on Shot Volume (Quantity) vs. Average Goals (Efficiency).

This unsupervised evaluation proves that even simple aggregated features can successfully capture complex tactical differences between teams.

4.3 Association Rules Discovery The application of the Apriori algorithm revealed meaningful tactical patterns, filtering for rules with a Lift > 1. The strongest discovered rule was `{Assist: Cross} $\rightarrow$ {Body Part: Head}` with a significant Lift of 4.40 and a Confidence of 69%. Overall, a limited number of high-lift rules were identified, indicating that meaningful tactical patterns are relatively rare. This statistically validates the classic tactical reliance on crosses for aerial threats. In contrast, general play patterns like `{Open Play} $\rightarrow$ {Pass}` showed a much lower Lift (~1.2), confirming that the model can distinguish between specific tactical setups and generic game events.

5. Conclusion

As was argued in the introduction, predicting the outcome of a shooting attempt in a dynamic sport like football is not a straightforward task. This is reflected in the inherent imbalance of our dataset, where only about 10.7% of shots result in a goal. This statistic supports the intuitive idea that scoring is a rare event influenced by high uncertainty and complex variables, not all of which (such as defensive pressure or goalkeeper positioning) were captured in our event-level data.

On the other hand, the impact of the measurable features we *did* analyze—specifically shot location, body part, and assist method—is not negligible. As witnessed by the ROC and Accuracy measures of our models, these spatial and contextual factors are strong predictors of success. After the evaluation considerations, the **Logistic Regression** model proved to be the superior approach, achieving an AUC of 0.967 compared to 0.941 for the Decision Tree. The exceptionally high AUC suggests that the selected features—particularly the spatial attributes and shot placement—possess strong discriminative power. This indicates that the model successfully captured the clear boundary between high-probability scoring zones (e.g., inside the box) and low-probability attempts, effectively minimizing false positives.

This confirms that goal scoring is best modeled as a probabilistic event, aligning perfectly with the modern "Expected Goals" (xG) framework used in professional analytics.

Even though the results are encouraging, it is possible to improve them. Our unsupervised clustering successfully grouped teams into tactical profiles (Elite vs. Defensive), but this analysis could be deepened by adopting more complex models, such as Artificial Neural Networks, or by incorporating player-tracking data to quantify defensive pressure. Furthermore, while the current high AUC reflects strong spatial features, future validation using cross-season or cross-league splits could provide a more conservative and robust estimate of performance.

Finally, future iterations should address potential biases introduced by our imputation strategy. In this study, we imputed missing assist methods as "Solo Run" and missing player names as "Unknown." While practical, this approach assumes that missingness is informative, which might not always hold true. As noted in statistical literature (e.g., regarding Simpson's Paradox), improper imputation can sometimes distort relationships. Therefore, applying advanced techniques like Multiple Imputation or leveraging external data sources would provide a more robust handling of missing values, ensuring that the analysis remains unbiased and granular .

References

[1] Secareanu, A. (2016). Football Events: European Soccer Events Dataset. Kaggle. Retrieved from: https://www.kaggle.com/secareanualin/football-events

[2] Pappalardo, L., Cintia, P., Rossi, A. et al. (2019). A public data set of spatio temporal match events in soccer competitions. Nature Scientific Data 6, 236.

[3] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. (1984). Classification and Regression Trees. New York: Routledge.

[4] KNIME AG. KNIME Analytics Platform User Guide. Retrieved from https://www.knime.com/documentation.