# SoccerDiffusion: Modeling Football Possession Dynamics Using Fractional Brownian Motion

Samuel Mir Arribas, Joaquín Carrión Gil, Gonzalo Hurtado Sanhermelando, Marcos Gómez Soler, Jorge Durá Jiménez, Iván García Donderis

*Universitat Politècnica de València – Escola Tècnica Superior d'Enginyeria Informàtica*
*Data Science – Academic Year 2024/2025*

**Abstract**

This study explores how principles of fractional Brownian motion can be applied to football data analysis to better understand possession dynamics. The main objective is to model and interpret the behavior of ball movement through the Hurst exponent (H), a mathematical measure that captures the level of persistence or randomness in trajectories.

To achieve this, we generated over 50,000 synthetic football possessions using stochastic models simulating different playing styles. These were used to train machine learning models that predict the H value based on features such as displacement, duration, speed, and directionality. Once validated, the models were applied to real match data from StatsBomb to estimate H values for actual possessions.

Beyond measuring H, the core aim of the project is to explore how this parameter relates to concrete in-game situations—such as pressing intensity, goal events, possession structure, and match momentum. By combining H with a wide range of contextual variables, we provide new insights into how different phases of play reflect underlying possession dynamics.

Our results show that H can serve as a meaningful descriptor of team behavior and playing style, especially when interpreted in relation to game context. This interdisciplinary approach combines physical modeling, machine learning, and sports analytics to offer a novel framework for understanding football at a deeper level.

## 1 Introduction

We chose this project because it allowed us to work on a subject we are genuinely passionate about: football. From the beginning, our motivation was to explore how data science could be applied to the sport in a meaningful way. The idea of using data to better understand what happens on the pitch—beyond just goals or passes—was exciting and aligned perfectly with our academic interests.

Although we initially didn't know the project would involve mathematical models like fractional Brownian motion, we embraced the challenge once it was introduced. Learning to apply these concepts to football data pushed us out of our comfort zone and allowed us to connect ideas from physics, statistics, and machine learning in an innovative way.

Football has become a fertile ground for data science, largely due to the increasing availability of rich event datasets like those from StatsBomb. While traditional metrics offer valuable summaries of performance, they often fail to capture the underlying dynamics of ball movement during possessions. Understanding how the ball flows—not just where it ends up—is a more complex and subtle task.

In this context, we explore the application of fractional Brownian motion as a model for ball progression, focusing on the Hurst exponent (H) as a descriptor of movement style. By estimating H for both synthetic and real-world possessions, we aim to characterize playing patterns in a continuous and interpretable way.

Our approach goes beyond measuring H in isolation: we analyze how it correlates with specific match situations. Using event-level data from StatsBomb, we examine how H values vary with tactical scenarios such as high pressing, transitions, goal-scoring moments, and possession structure. This allows us to identify not only how the game evolves over time, but also how certain behaviors are reflected in the flow of the ball.

The goal of this work is to demonstrate that H is not just a theoretical construct but a practical metric for understanding team strategy and game momentum. Our findings highlight the value of combining physical modeling with contextual football data to produce actionable insights for coaches, analysts, and researchers.

# 2 Materials and Methods

## 2.1 Data Source

For this project, we used the **StatsBomb Open Data** repository, which provides detailed, event-based match data in a structured and standardized format. The data comes pre-cleaned and well-organized, meaning no additional preprocessing or cleaning was necessary before analysis. Each event is recorded with temporal, spatial, and categorical information, allowing for in-depth analysis of in-game behavior.

We based our study on different FC Barcelona matches from the **2019/2020 LaLiga season**:

These matches were chosen for their contrasting game dynamics and the high level of tactical complexity from the participating teams, having sense because this team in this season was a team that changed his dynamics and way of playing a lot, depending on the situations of each match.

Rather than relying on predefined possessions, we **constructed possessions manually** by grouping sequences of events of type `pass` and `carry`, both of which describe ball movement actions. A new possession was considered to start whenever control of the ball switched from one team to the other.

From the StatsBomb event data, we extracted only the information required to build and segment possessions. The primary event types used were:

- **Pass**: includes origin, end location, angle, and outcome
- **Carry**: includes start and end coordinates and duration

These two variables served as the foundation for reconstructing the ball movement over time. We ignored other event types at this stage, as our initial focus was purely on identifying continuous team possessions based on movement actions.

StatsBomb's Open Data is organized into several structured datasets, each serving a specific role in the match representation:

- **Matches**: general metadata for each match, such as competition, season, date, teams, final score, stadium, and referee.
- **Events**: the core dataset, containing all match actions in chronological order, including passes, carries, shots, pressures, duels, and more. Each event includes spatial coordinates, time stamps, and contextual attributes like `under_pressure` or `counterpress`.
- **Lineups**: provides information about starting players, jersey numbers, positions, and nationalities for both teams.
- **360 Frames**: adds spatial context by capturing player positions relative to key events, offering freeze-frame views of on-pitch situations.
- **Competitions**: describes the league and season structure, including gender and country of the competition.

Together, these datasets allow for comprehensive analysis of both isolated events and broader game dynamics. In our case, we primarily relied on the `events` dataset to reconstruct possessions and extract movement patterns, while the rest served to enrich contextual understanding and data validation during later phases.

## 2.2 Synthetic Data Generation

In order to simulate diverse football possession styles, we generated a large set of synthetic possessions using **fractional Brownian motion (fBm)**. This technique allowed us to model the spatial progression of the ball as a stochastic process, parameterized by the **Hurst exponent (H)**. The value of H determines the nature of the movement:

- $H > 0.5$: persistent and fluid movements
- $H < 0.5$: anti-persistent, irregular patterns
- $H \approx 0.5$: standard Brownian motion (random walk)

The motivation for generating synthetic data stems from a limitation in the dataset: the number of real possessions available from StatsBomb was insufficient to robustly train machine learning models. By producing artificial sequences with controlled variability, we ensured that the model would be exposed to a wide range of movement styles. Additionally, the synthetic data helped introduce noise and prevent overfitting to specific match situations. This approach reflects the concept of data augmentation discussed in class, where generating synthetic scenarios helps mitigate data sparsity and expand modeling capacity.

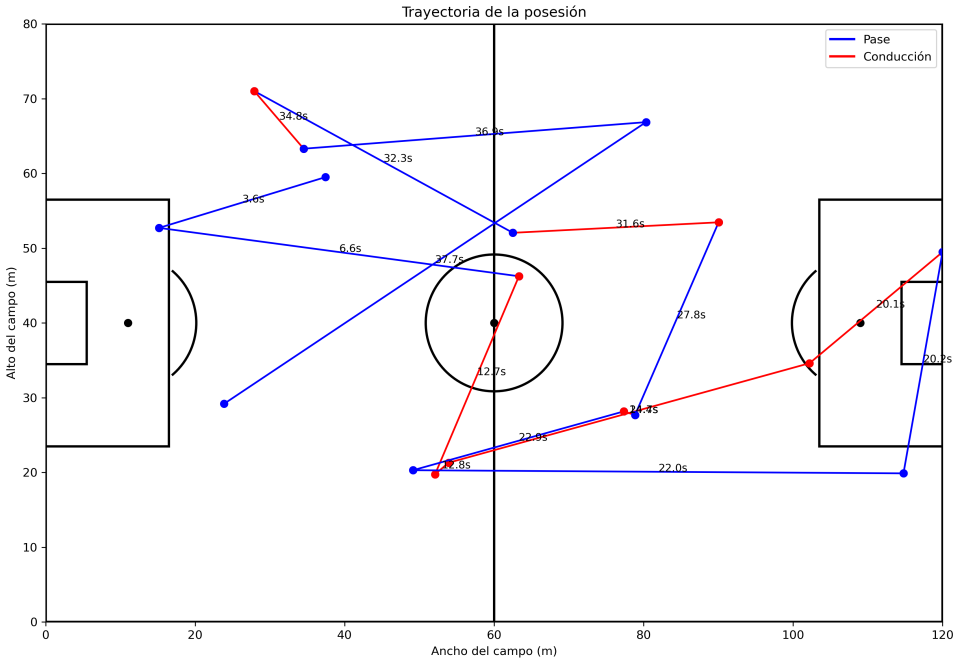



Figure 1: Low H: 0.23

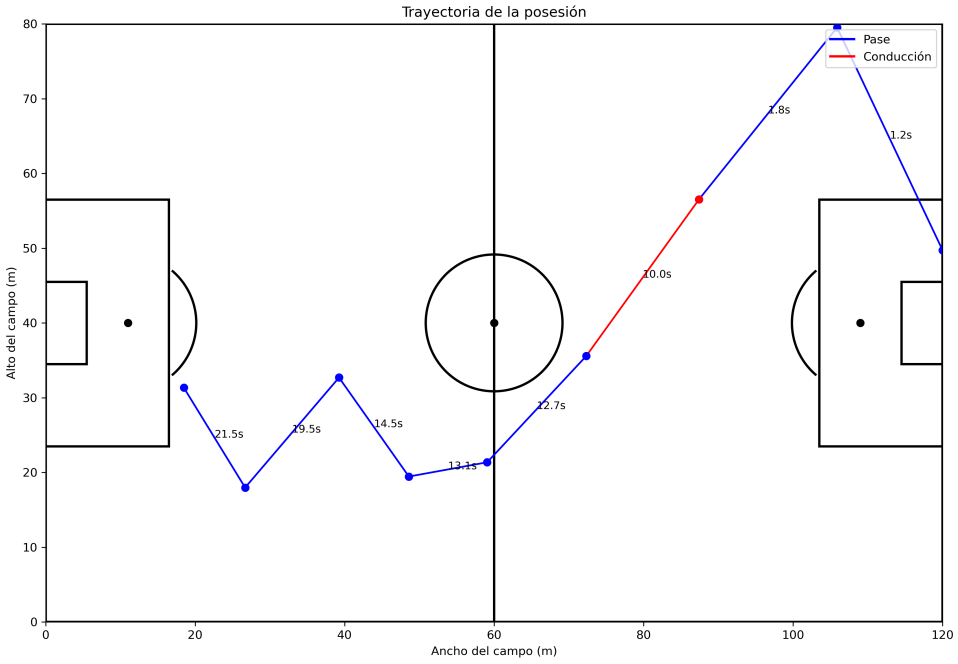



Figure 2: High H:0.96

We used the `fbm` Python package to generate fBm paths for both the **x** and **y** coordinates on a 120m × 80m pitch. Each possession simulates a sequence of ball movements composed of two action types: `passes` and `carries`, randomly assigned with probabilities of 70% and 30%, respectively—values chosen to reflect typical match dynamics.

The main steps of the generation process were:

- **Trajectory simulation**: fBm sequences were generated using the `FBM` class, with lengths adapted to the number of actions per possession.

- **Time modeling**: Action durations were sampled using exponential distributions with different scales for passes (shorter) and carries (longer).
- **Scaling and smoothing**: Movement deltas were rescaled to ensure realistic field coverage, and all coordinates were clipped to remain within pitch bounds.
- **Timestamps**: Cumulative time was computed for each possession to simulate real-time sequences.

We generated a total of **50,000 synthetic possessions**, with H values drawn uniformly from the interval $[0.2, 1.0]$. The number of actions per possession varied between 5 and 20 to simulate different lengths of ball control.

The generation pipeline was implemented using the following Python libraries:

- `numpy` for numerical operations
- `pandas` for data structuring
- `fbm` for stochastic path generation
- `tqdm` to display generation progress

All generated data was saved in a single `CSV` file for use in the later modeling stages.

This synthetic dataset provides a controlled and scalable framework for learning how the Hurst exponent relates to possession characteristics, enabling robust model training and evaluation despite limited real match data.

## 2.3 Feature Engineering

Once the synthetic possessions were generated, we extracted a set of descriptive features for each possession. These features capture the spatial and temporal dynamics of the ball's movement and serve as inputs for the machine learning model that predicts the Hurst exponent ($H$).

The features were computed by iterating over each synthetic possession and calculating a combination of statistical, geometric, and physical metrics. The most relevant features extracted were:

- **Number of passes and carries**: count of actions labeled as `Pass` or `Carry`.
- **Variation in position (X and Y)**: mean absolute difference between consecutive $x$ and $y$ coordinates.
- **Possession duration**: total time from the first to the last action in the possession.
- **Total distance covered**: sum of Euclidean distances between consecutive ball positions.
- **Average speed**: total distance divided by duration.
- **Average distance per movement**: mean length of all movement segments.
- **Average movement angle**: mean angle formed between movement vectors, using the arctangent function.

- **Speed change**: average change in movement speed across consecutive segments.
- **Acceleration**: average difference in speed changes (second derivative of position over time).

To compute these variables, we used common Python libraries including `numpy` and `pandas`. A custom function was used to calculate angles between movement vectors using `arctan2`, and all derived statistics were stored in a dictionary and then converted into a structured DataFrame.

The final result is a table where each row represents one synthetic possession, and each column corresponds to a specific feature. This table is referred to as our **minable view**, since it contains all the preprocessed information needed to train machine learning models. As covered in the seminars, a minable view is essential to ensure that each observation (in our case, a possession) is represented by relevant and well-structured features. In addition to the input features, we included the ground-truth value of $H$ for each possession, which was known in advance since it was used during the synthetic data generation process.

This minable view enables supervised learning by providing a clear and consistent input-output structure and was exported to CSV format for easy use in the modeling phase.

## 2.4 Modeling Approach

The objective of the modeling phase was to learn a function capable of predicting the Hurst exponent ($H$) of a possession based on the engineered features. To do this, we followed an iterative model development cycle that involved testing multiple algorithms, selecting suitable evaluation metrics, and using tools for systematic hyperparameter optimization.

**Data Splitting.** We split the dataset into training (80%) and test (20%) subsets using `train_test_split` with a fixed random seed to ensure reproducibility. The input features ($X$) came from the feature engineering stage, while the target variable ($y$) was the known value of $H$ used during synthetic data generation.

**Evaluation Metric.** The main evaluation metric used was the **Mean Absolute Error (MAE)**, which provides an intuitive and interpretable measure of average prediction error. We chose MAE over RMSE due to its robustness to outliers and direct interpretability in the context of H values. Moreover, since the Hurst exponent values lie within the bounded interval $[0, 1]$, the absolute error is naturally constrained, making MAE a suitable and meaningful metric for this task.

**Model Iteration.** We tested several regression algorithms in increasing order of complexity, starting from a baseline `Linear Regression`, moving to a `Random Forest Regressor`, and finally to more advanced models like `XGBoost`, a `Multi-Layer Perceptron (MLP)`, and a `Recurrent LSTM` network.

For each model, we evaluated performance using cross-validation and tracked the MAE across iterations. The process allowed us to incrementally improve results by tuning features and hyperparameters. This iterative strategy follows the model refinement cycles discussed in class, where diagnostic feedback and metric tracking guide the selection of optimal models.

**Hyperparameter Tuning.** For tree-based models like Random Forest and XGBoost, we used `GridSearchCV` from `scikit-learn` to explore combinations of parameters such as number of estimators, tree depth, and learning rate. The search used 5-fold cross-validation and was parallelized to speed up computation. This strategy allowed us to find optimal configurations and systematically reduce the error.

**Neural Networks.** We implemented a `Multi-Layer Perceptron (MLP)` and a `Recurrent LSTM` using `PyTorch`, training them over 50 epochs. These models were useful for exploring temporal or sequence-based learning, although the added complexity made them harder to fine-tune.

**Results.** Among all models tested, the best performance was obtained using `XGBoost`, with a final **MAE of 0.1171**. The LSTM model followed closely with a MAE of 0.1207, while Random Forest achieved 0.1179. The MLP performed less consistently, with an MAE of 0.3080.

**Model Selection.** Based on performance, efficiency, and interpretability, we selected **XGBoost as the final model**. It consistently outperformed the others and aligned well with the nature of our dataset, which is tabular and feature-rich rather than sequential.

**Best XGBoost Parameters:**

- `n_estimators`: 100
- `max_depth`: 10
- `learning_rate`: 0.05
- `subsample`: 0.8
- `colsample_bytree`: 0.9
- `gamma`: 0.1
- `min_child_weight`: 1

This modeling process reflects a rigorous approach to model development, combining experimentation, metric-driven evaluation, and informed model selection.

## 2.5 Matching Real Games to Synthetic H Values

To apply the model trained on synthetic data to real football matches, we first needed to estimate the Hurst exponent ($H$) for each possession from real event data. This process involved two main steps: reconstructing the possessions and computing the same features that were used for training.

**Possession Reconstruction.** Using the StatsBomb event data for the selected matches of FC Barcelona, we rebuilt each team possession by grouping consecutive events of type `pass` and `carry`, as we had done in the synthetic dataset. We considered a new possession to begin whenever control of the ball changed from one team to the other.

**Feature Extraction.** For each reconstructed possession, we computed the same set of features as described in the Feature Engineering section (Section 4.3), including total distance, duration, movement variation, angle, and acceleration. This ensured that the

format and scale of the real data matched the synthetic input on which the model was trained.

**H Estimation.** Once the features were extracted, we passed them through the final XGBoost regression model to obtain a predicted value of $H$ for each possession. This value represents an estimate of the persistence or randomness in the ball movement, derived from real in-game behavior.

This process allowed us to assign a continuous and interpretable measure of possession dynamics to each sequence in the real matches. These estimated H values form the basis for further analysis, where we aim to link them with contextual game events such as goals, pressure phases, and disciplinary actions.

**Linking H to Match Events.** Once we estimated the Hurst exponent ($H$) for each possession in real matches, the next step was to analyze how $H$ values evolved in relation to key contextual events within the game. Our objective was to assess whether specific match situations—such as goals, red cards, or substitutions—are associated with noticeable shifts in ball movement patterns, as captured by $H$.

To do this, we aligned the estimated $H$ values with a timeline of relevant events extracted from the StatsBomb dataset. For each target event (e.g., goal scored), we identified the possessions that occurred immediately before and after the event. We then computed the difference in average $H$ before and after each event type (denoted as $\Delta H$) to quantify the directional change in possession dynamics.

In addition to event-driven analysis, we studied the relationship between $H$ and pressure-related variables available in the dataset, such as:

- `acciones_bajo_presion`: number of actions performed under pressure
- `porcentaje_presion`: proportion of pressured actions per possession
- `eventos_Pressure`: count of pressure events during the match

We joined these variables with the corresponding $H$ values at the possession level, enabling us to compute statistical correlations. This integration allowed us to explore how possession behavior—persistent vs. erratic—is affected by the intensity and frequency of defensive pressure.

This enriched dataset, combining estimated $H$ with both categorical match events and continuous pressure metrics, serves as the foundation for the analyses and visualizations presented in the Results section. This method reflects the seminar approach to event-based pattern analysis, where we link model predictions to contextual in-game scenarios.

# 3 Results

## 3.1 Distribution of H in Real Match Possessions

Once we estimated the Hurst exponent ($H$) for all possessions in real matches played by FC Barcelona, we analyzed the overall distribution of values. We observed that most $H$ values—both for Barcelona and their opponents—ranged between 0.6 and 1.0. This pattern is consistent with previous work by Joan Castillo Esteve [3], whose thesis also reported $H > 0.5$ for most possession sequences.

The reason for this skewed distribution lies in how possessions are extracted. Our methodology grouped sequences of `pass` and `carry` actions, which naturally include ball circulation phases that last several seconds. However, very short possessions, such as a single-action corner or a one-pass penalty, are excluded by design. These sequences would likely produce $H$ values below 0.5 (anti-persistent), but are underrepresented in the data.

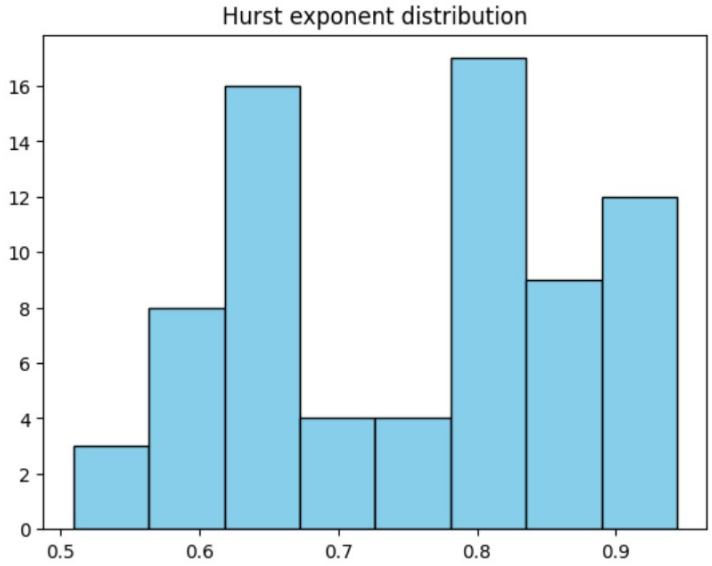



Figure 3: Distribution of Hurst Exponent (H) for FC Barcelona possessions in a LaLiga match.

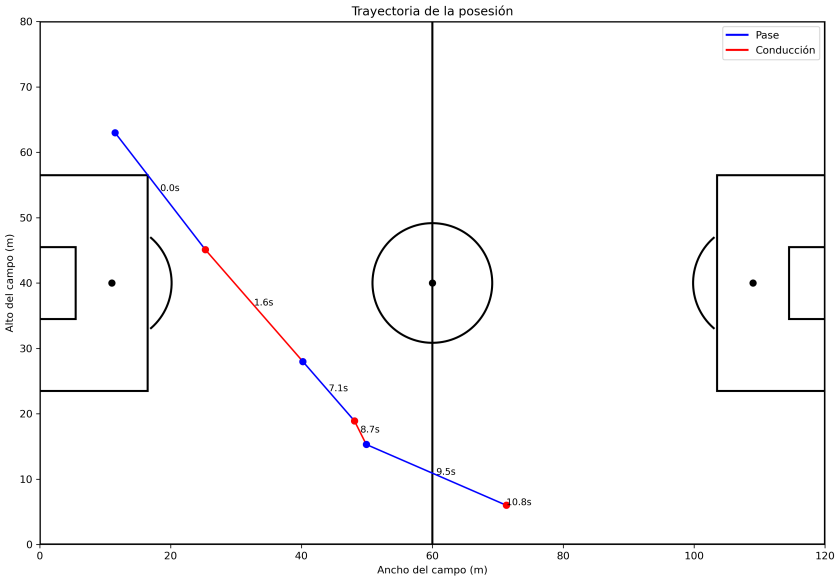



Figure 4: Example of a real FC Barcelona possession trajectory with estimated H = 0.84

## 3.2 Impact of Match Events on Hurst Exponent (H)

To explore how different types of in-game events affect possession dynamics, we analyzed the change in $H$ before and after key contextual events. Specifically, we focused on:

- Goals scored and conceded
- Red cards (for and against)
- Substitutions

For each of these events, we computed the average change in $H$ ($\Delta H$) by comparing the mean $H$ value of possessions before the event to those immediately after it. This allows us to quantify how disruptive or stabilizing each event is in terms of ball movement behavior.

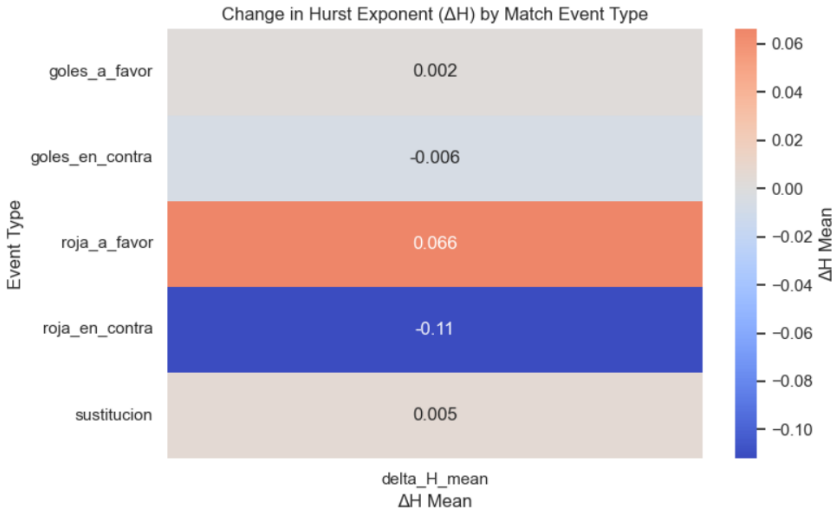



Figure 5: Average change in Hurst exponent ($\Delta H$) after key match events.

The heatmap reveals several key insights:

- **Goals in favor** ($\Delta H = 0.002$) and **goals conceded** ($\Delta H = -0.006$) showed minimal changes in $H$. These small shifts suggest that scoring events do not immediately alter possession dynamics in a consistent way, likely because the team resumes play in a relatively structured manner, or because the opponent regains control.
- **Red cards received** caused the most significant negative shift in $H$ ($\Delta H = -0.112$), indicating a breakdown in structured play and increased randomness, likely due to the tactical disruption and pressure from playing with one fewer player.
- **Red cards in favor** led to a noticeable increase in $H$ ($\Delta H = +0.066$), suggesting that teams in numerical superiority tend to retain possession more steadily and move the ball with greater control.
- **Substitutions** showed only a slight increase in $H$ ($\Delta H = 0.005$), which is too small to draw strong conclusions. Their impact on possession behavior appears negligible in the short term.

In order to reinforce the previous interpretations, we considered it essential to also analyze the correlations between contextual game variables and the Hurst exponent (H). To this end, we calculated the Pearson correlation between H and the previously mentioned variables. The results are summarized in the table below:

| Variable | Correlation with H |
|---|---|
| acciones_bajo_presion | 0.67 |
| goles_a_favor_cerca | +0.026 |
| goles_en_contra_cerca | +0.002 |
| roja_a_favor_antes | +0.006 |
| roja_en_contra_antes | -0.001 |
| sustitucion_cerca | +0.021 |

A particularly noteworthy finding is the strong negative correlation between pressure and H. This result suggests that greater pressure on the team in possession tends to generate more chaotic and less persistent patterns of play, leading to less structured passing sequences.

In contrast, the remaining variables analyzed show no significant correlation with H, indicating that, at least in linear terms, their impact on possession dynamics is limited or negligible.

Based on these results, it is relevant to delve deeper into the most significant findings derived from the different analytical methods used.

In the case of red cards, we observed that they cause the greatest average change in H values. To illustrate this effect, we analyzed the evolution of the Hurst exponent throughout a match in which the analyzed team received a red card.

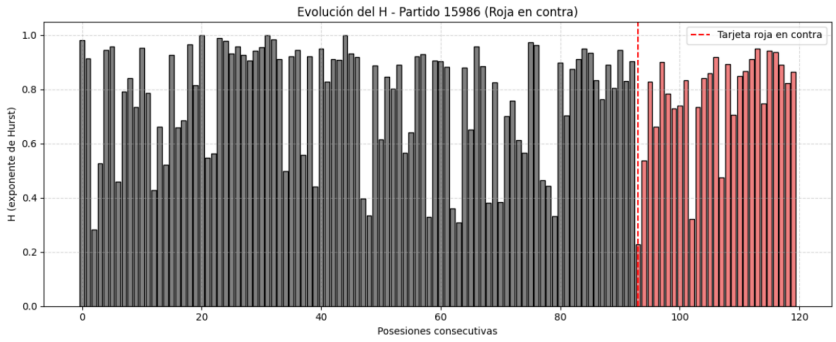



Figure 6: Evolution of the Hurst Exponent during a match

The corresponding graph shows that, immediately after the red card, there is a sharp drop in H, reflecting a period of disorganization and volatility in possession. However, as the match progresses, the H values tend to stabilize, indicating a recovery of structure in the team's play. This behavior suggests that the impact of the red card is primarily short-term. This conclusion is supported by the correlation between H and the variable "red card against before," which is nearly zero (r = -0.001). Tactically, this makes sense: a team playing with one fewer player will often seek to maintain long and structured possessions to slow the game down and compensate for their numerical disadvantage.

On the other hand, pressure emerges as the factor with the greatest structural influence on playing style. The strong negative correlation between actions under pressure and H indicates that defensive pressure directly contributes to the breakdown of order in possessions. In high-pressure scenarios, the team in possession is forced to accelerate play and make quicker decisions, leading to greater randomness in their sequences. Given this finding, we consider it worthwhile to further investigate the variables related to pressure to better understand their role in possession dynamics.

## 3.3 Correlation with Defensive Pressure Variables

We also analyzed how $H$ correlates with pressure-related metrics, such as:

- `acciones_bajo_presion` — actions executed under defensive pressure
- `porcentaje_presion` — percentage of actions under pressure
- `eventos_Pressure` — total pressure events during possessions

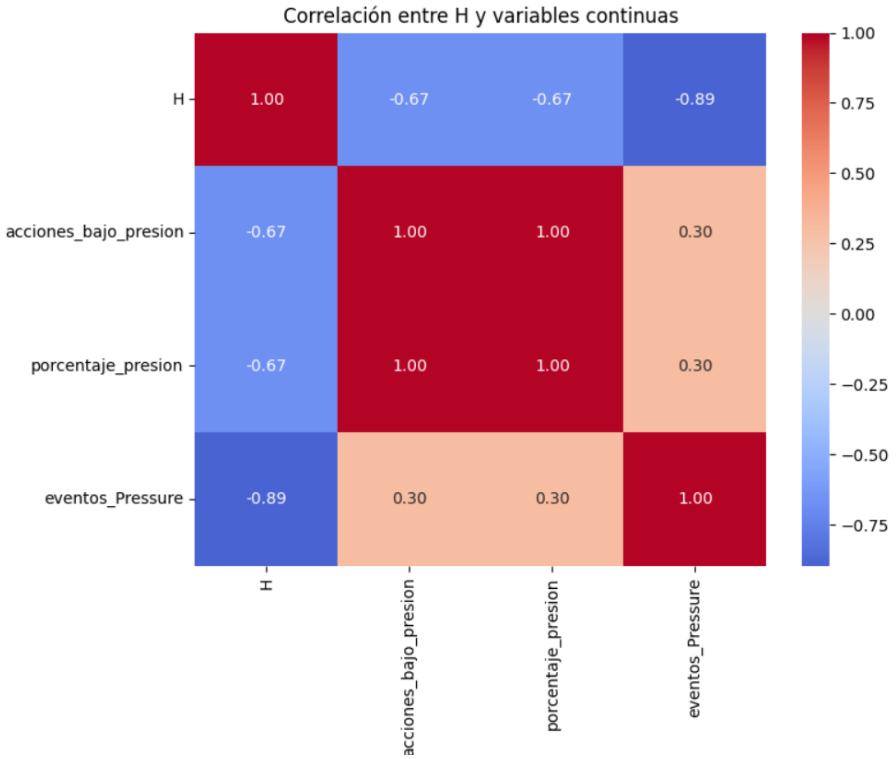



Figure 7: Correlation matrix between H and pressure-related variables.

The correlation matrix reveals strong negative relationships between $H$ and pressure metrics:

- A high number of `eventos_Pressure` correlates with a lower $H$ ($r = -0.89$), indicating that high-pressure situations force more erratic or reactive possessions.
- Both `acciones_bajo_presion` and `porcentaje_presion` also negatively correlate with $H$, further supporting this interpretation.

These results indicate that the Hurst exponent is sensitive to contextual pressure within a match. Specifically, the strong negative correlations observed between $H$ and the pressure-related variables suggest that when teams are subjected to defensive pressure, their possession dynamics become significantly more chaotic and less structured. A higher number of pressure events or a greater proportion of actions under pressure corresponds to lower $H$ values, meaning that the movement of the ball becomes more unpredictable and less persistent. This makes intuitive sense: under intense pressure, teams are forced

to react quickly, often with rushed decisions, vertical passes, or loss of spatial control—all of which reduce the continuity and coherence of possession. Therefore, $H$ serves not only as a measure of style or control but also as a reliable indicator of how external tactical factors, such as pressing intensity, directly impact ball circulation patterns.

These results support the idea, emphasized in several seminar examples, that domain-specific variables—like pressure—not only need to be quantified but interpreted within the behavioral structure they affect.

# 4 Discussion

## 4.1 Value and Applicability

This project introduces an innovative perspective to football analytics by modeling ball progression using the Hurst exponent, a concept drawn from fractional Brownian motion. Instead of relying solely on classic event-based statistics, we offer a dynamic and interpretable metric that captures the structural nature of possessions.

The novelty lies not only in the use of a mathematical model rarely applied to football, but also in how we trained it with synthetic data and validated it against real matches. By doing so, we demonstrate that complex tactical behavior can be quantified and analyzed through modern data science tools.

**Potential beneficiaries** of our work include:

- **Coaching staff and analysts**, who can detect whether their team is losing control or breaking structure during key match phases.
- **Performance departments**, which can use $H$ trends to design specific training routines aligned with desired possession styles.
- **Scouting teams**, by integrating $H$ as a compact descriptor of how rival teams build up and circulate the ball.
- **Sports tech companies**, which could integrate this metric into dashboards or tactical visualization software.

**Real-world integration** is straightforward: $H$ can be computed from event data (like StatsBomb or Opta) and visualized over time. This allows analysts to interpret possession quality across different match situations — e.g., after a goal, a red card, or a tactical change — and adjust strategy accordingly.

## 4.2 Limitations and Risks

While promising, the model has limitations. The precision of $H$ depends on possession segmentation: oversimplified or misclassified sequences could skew results. Also, the model focuses on ball dynamics, not player positioning, so some contextual insights might be lost.

There's also a risk of **overinterpretation**: a lower $H$ does not mean "worse", and a higher one is not inherently "better". Coaches must interpret $H$ in context — some phases require chaos (low $H$), while others benefit from structure (high $H$).

## 4.3 Legacy and Sustainability

To ensure that the work carried out in this project remains accessible, reusable, and understandable beyond the scope of the course, we have developed both a clear dissemination plan and a supporting website.

**Public repository structure.** All the code, data, and documentation is openly available in our GitHub repository: https://github.com/proyectsoccer/SoccerDiffusion. The structure of the repository is organized as follows:

- `data`: Contains all the CSV files used for training and evaluation.
- `jupyter_notebooks`: Includes all Jupyter notebooks developed during the project.
- `Information`: Gathers all reference documents such as:
  - StatsBomb PDFs,
  - Link to our first Prezi,
  - M2 report,
  - Final project report.
- `Pagina_web`: Contains the full source code of the website, including HTML, Python, and CSS files.
- `README.md`: Provides instructions to understand the project structure and replicate the results.

**Website development.** As part of the dissemination strategy, we have created a dedicated website to showcase the project in an intuitive and user-friendly manner. The website is implemented using `Flask`, a lightweight Python web framework. The backend logic is defined in the `app.py` file, which maps different routes to corresponding HTML templates via `render_template`. The structure is the following:

- `app.py`: Main application file that defines the website routes and initializes the Flask application.
- `templates`: Contains all HTML files structured with the Jinja2 templating engine:
  - `base.html`: Layout template used across all pages for consistency.
  - `index.html`: Homepage.
  - `about_us.html`, `use_cases.html`, `products.html`: Dedicated pages for project description, use cases, and further content.
- `static/style.css`: Custom CSS stylesheet to ensure a clean, modern, and responsive visual style.

The project uses the following key technologies and libraries:

- **Flask** for backend routing and server deployment.
- **HTML5 + Jinja2** for dynamic templating of content.

- **CSS3** for responsive styling and layout design.
- **Python** (v3.10) as the main programming language for the backend logic.

**Purpose and value of the web platform.** The website is not only a communication tool, but also a way to increase the project's accessibility and long-term impact. It allows a broader audience, including both technical and non-technical users to:

- Understand the concept of the Hurst exponent in football analytics.
- Explore visualizations of ball possession structures and how the $H$ index evolves during real matches.
- Examine how contextual factors such as goals, fouls, or high pressing affect the possession dynamics through changes in $H$.

This public platform serves as a bridge between academic research and real-world communication, reinforcing transparency, reproducibility, and educational dissemination. Screenshots of the different sections of the website are included in the annex to illustrate the interface and contents.

It is important to note that the platform is still in a preliminary stage, with a primary focus on presenting our results in a visual and accessible way. While it does not yet allow for interactive testing of our models or dynamic analysis of custom match data, this is a planned future extension of the project. We would like the website to become a tool in the future where users can explore the correlations between $H$ and the match variables in a more flexible and personalized way.

This approach aligns with the value-centric project strategies discussed during the course, where the goal is to create data products that go beyond academic demonstration and offer practical, communicative tools for end users.

## 4.4 Social and Educational Impact (SDGs)

This project supports several Sustainable Development Goals (SDGs):

- **SDG 4 – Quality Education**: This work showcases how data science and mathematics can be applied to real-world problems in sport, providing students and researchers with a concrete interdisciplinary case.
- **SDG 8 – Decent Work and Economic Growth**: By promoting innovation in the sports analytics industry, our tool contributes to the digital transformation of professional clubs and startups, offering potential economic value and employment opportunities in the growing field of AI in sports.

Overall, we see this not just as an academic exercise, but as a prototype for how advanced analytics can support better decision-making in modern football.

# 5 Conclusions

This project presents a novel approach to modeling football possessions using the Hurst exponent ($H$), a mathematical metric rooted in fractional Brownian motion. Through the combination of synthetic data generation, machine learning, and real match analysis, we developed a tool capable of quantifying possession behavior in terms of structural persistence and randomness.

One of the key contributions of this work is the demonstration that $H$ can serve as a dynamic indicator of how possessions evolve in response to match events. Our results show that while most events—like goals or substitutions—have negligible short-term impact on $H$, red cards produce sharp, temporary shifts in possession dynamics. Notably, defensive pressure emerged as the strongest and most consistent factor influencing $H$, revealing that pressing intensity is directly tied to how structured or chaotic a team's possession becomes.

By interpreting $H$ in relation to match context, analysts and coaches can identify when their team is gaining or losing control, or when specific disruptions are influencing playing style. This positions $H$ as a valuable, interpretable complement to traditional football metrics.

The methodology developed is reproducible, efficient, and adaptable. It offers an interdisciplinary bridge between physics-inspired modeling, machine learning, and applied sports analytics. Future directions could include integrating tracking data, expanding the event taxonomy considered, or exploring predictive use cases, such as forecasting shifts in possession dynamics based on live inputs.

In summary, the Hurst exponent proves to be a powerful descriptor for possession structure, especially when combined with contextual variables. Our work opens the door to more nuanced, data-driven evaluations of playing style, tactical adjustments, and match control.

# 6 Acknowledgments

We would like to thank the instructors of the Data Science course at the Universitat Politècnica de València for their guidance throughout this project, particularly for introducing us to advanced methodologies such as CRISP-DM and Lean Canvas. These frameworks were fundamental for organizing our workflow and aligning technical development with strategic goals.

We also acknowledge the creators of the StatsBomb Open Data project for providing high-quality football data that made our analysis possible.

Additionally, we benefited from several seminars and talks during the course, including the presentation by Tsuyoshi Osaki, which offered practical insights into football analysis from a coaching perspective.

Finally, we used tools like ChatGPT, Deepseek, and Claude to support our learning and development. These AI systems helped us clarify theoretical concepts, write and translate technical content, and optimize parts of our code during implementation.

# 7 References


[1] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.

[2] Deepika Dhingra and Shubhangi Bhardwaj. *Machine learning techniques in sports analytics: A review*. Materials Today: Proceedings, 47:6668–6673, 2021.

[3] Joan Castillo Esteve. *Characterization of trajectories in football matches using anomalous diffusion models*. Bachelor's Thesis, Universitat Politècnica de València, 2023.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Chapter on MLP networks and training with backpropagation.

[5] Bilel Guedri, Ahmed M. Elmisery, and Mohamed H. Aly. *Predictive analytics for football matches using machine learning techniques*. Procedia Computer Science, 170:1392–1397, 2020.

[6] Andy Liaw and Matthew Wiener. *Classification and regression by randomforest*. R News, 2(3):18–22, 2002.

[7] Benoît B. Mandelbrot and John W. Van Ness. *Fractional Brownian motions, fractional noises and applications*. SIAM Review, 10(4):422–437, 1968.

[8] Yusef Ahsini Ouariaghli. *Machine learning-based characterization of single-particle behavior with synthetic experiment videos*. Bachelor's Thesis, Universitat Politècnica de València, 2024.

[9] Ken Yamamoto, Seiya Uezu, Keiichiro Kagawa, Yoshihiro Yamazaki, and Takuma Narizuka. *Theory and data analysis of player and team ball possession time in football*. Physical Review E, 108(4), 2023.

# 8 Appendix – Web Interface Snapshots

This appendix contains snapshots from the website developed to visualize and communicate the results of our project. The web application provides an accessible interface for exploring the concept of the Hurst exponent in football, viewing real match possessions, and interpreting model predictions interactively.

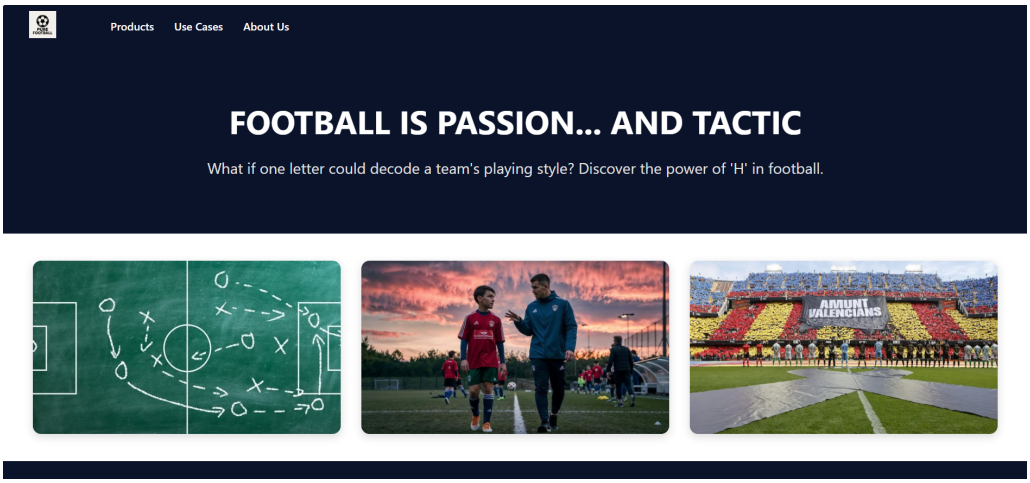



Figure 8

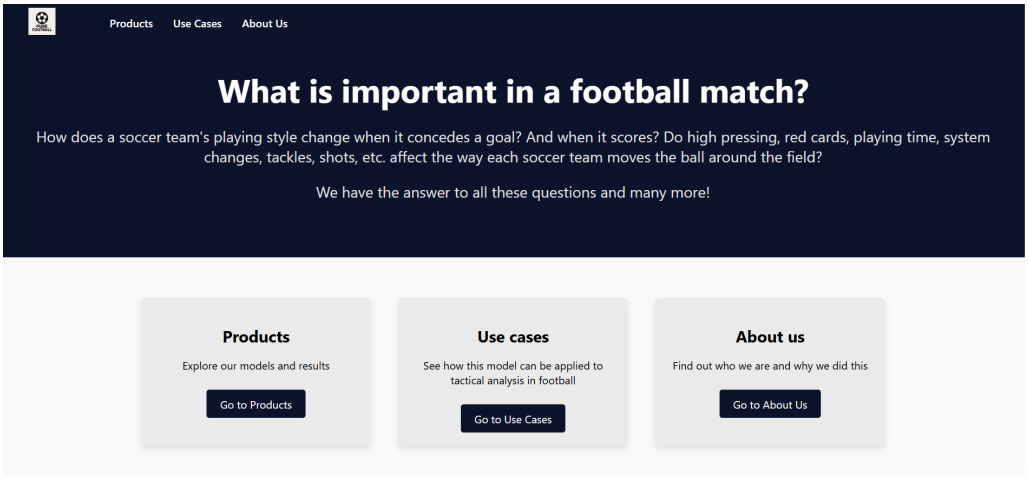



Figure 9


Figure 10

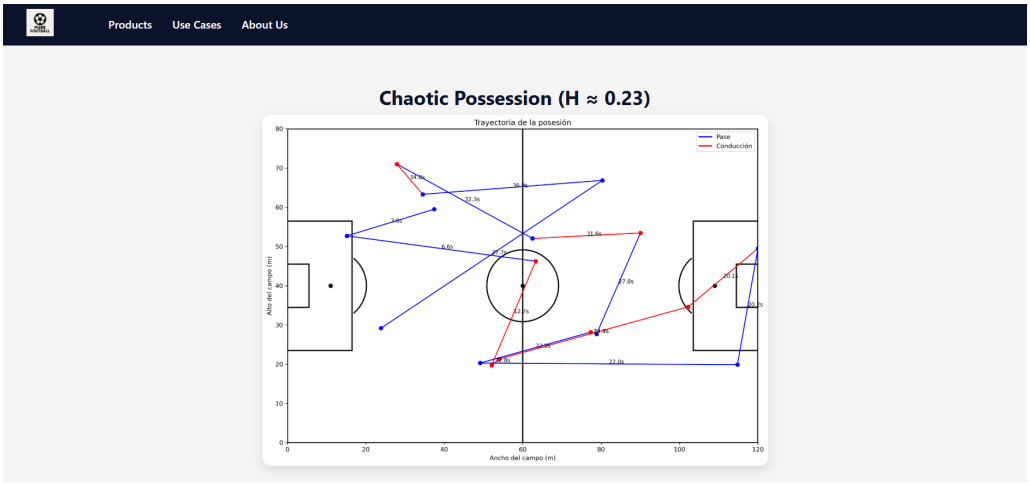



Figure 11

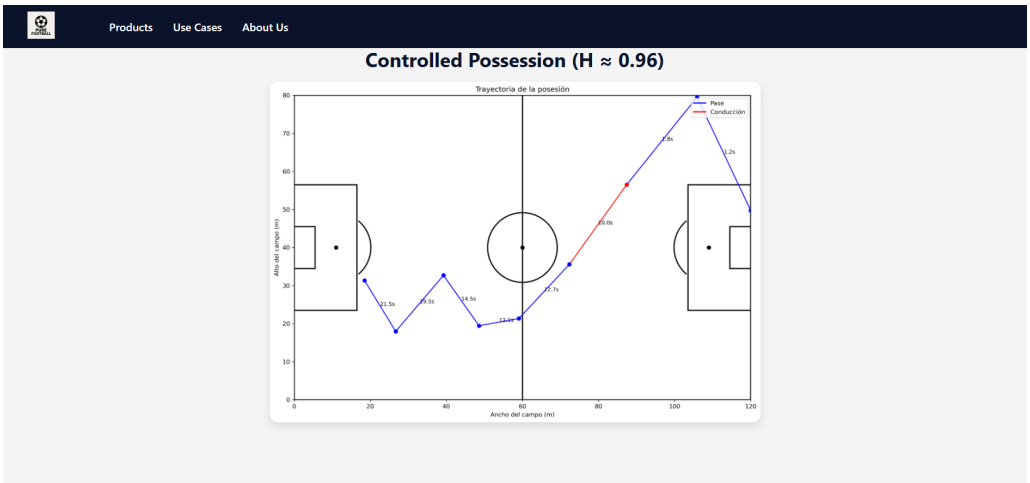



Figure 12

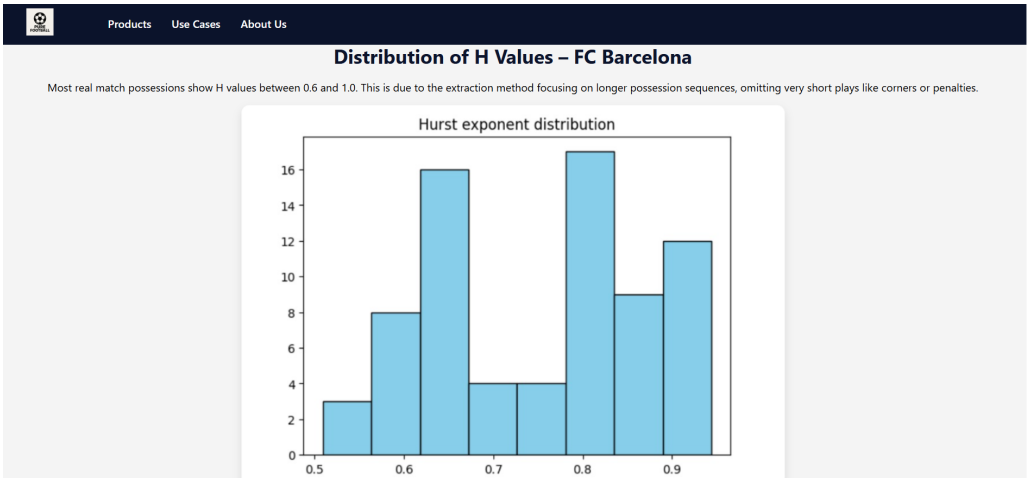



Figure 13

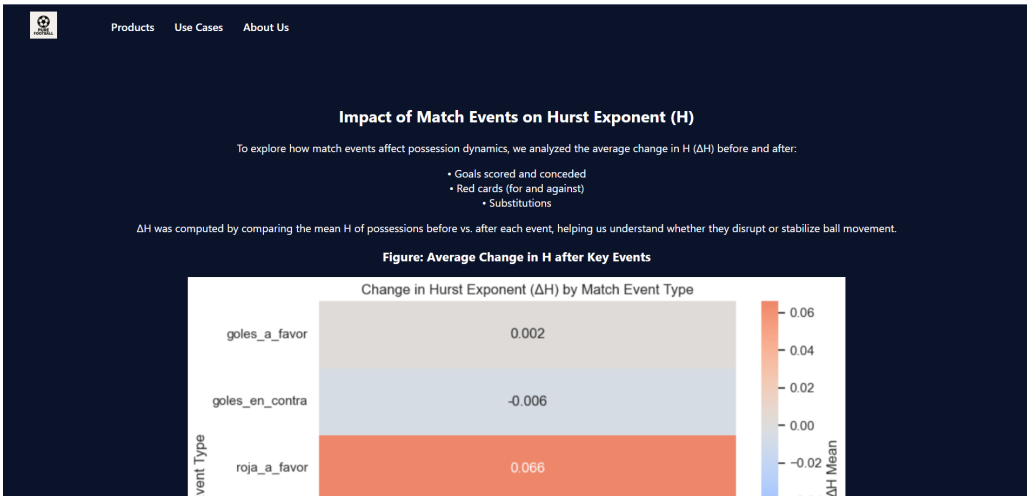



Figure 14

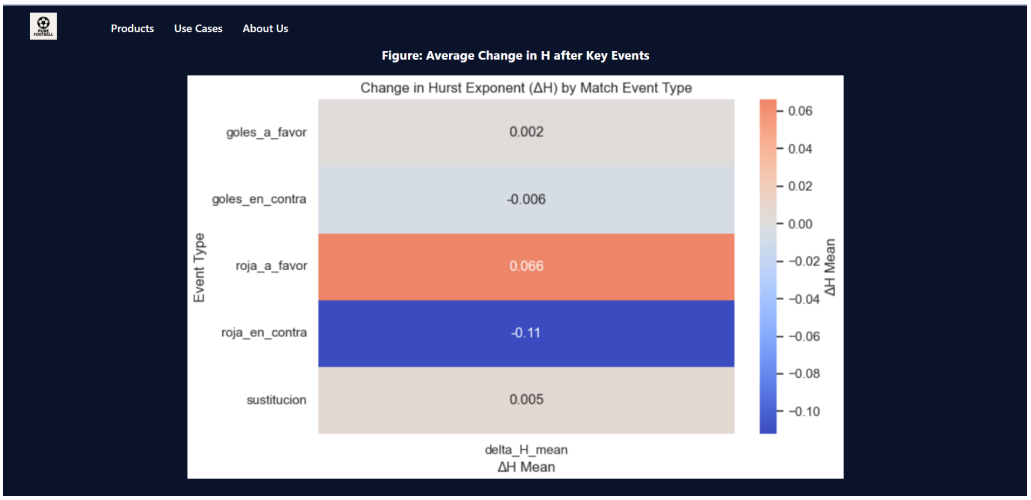



Figure 15

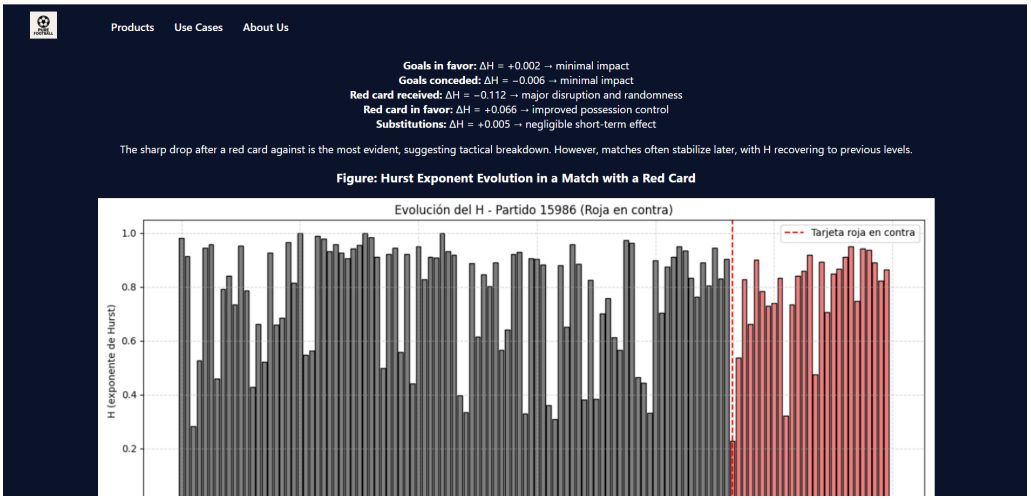



Figure 16

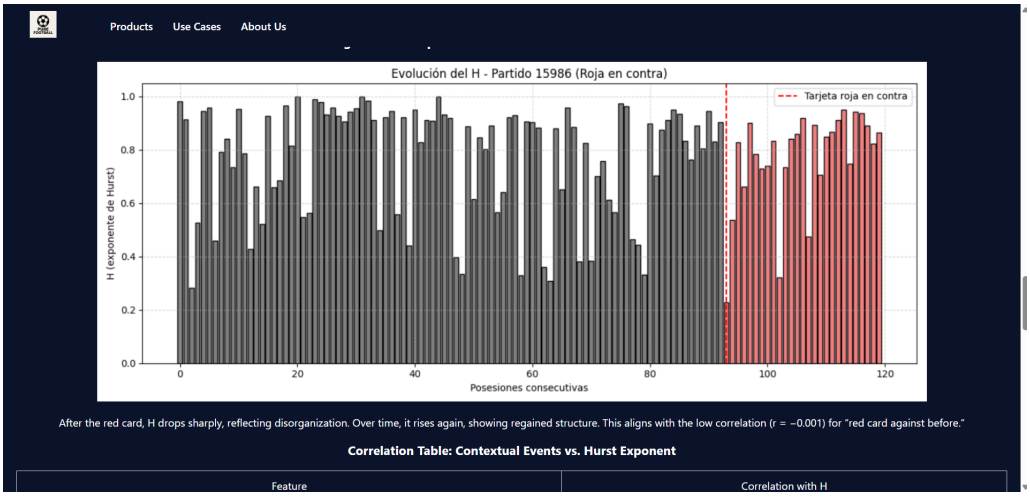



Figure 17

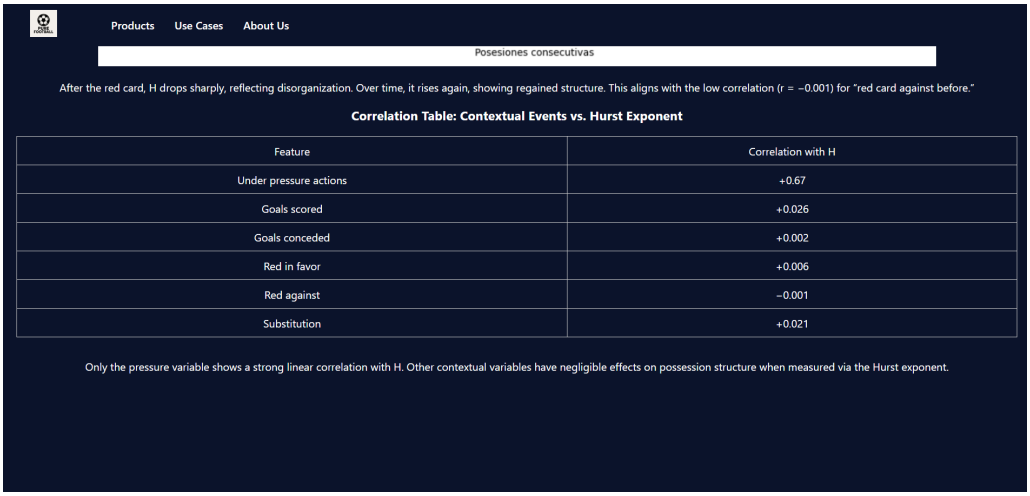



Figure 18

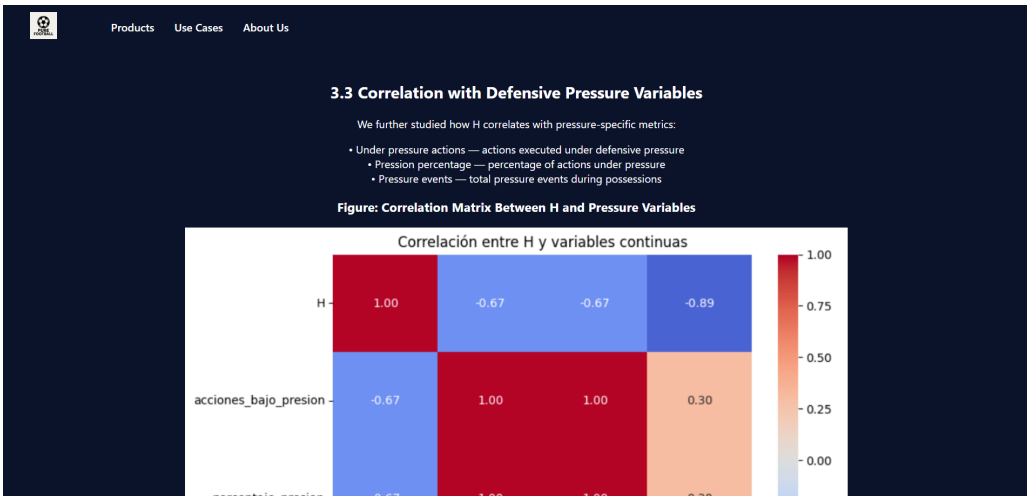



Figure 19

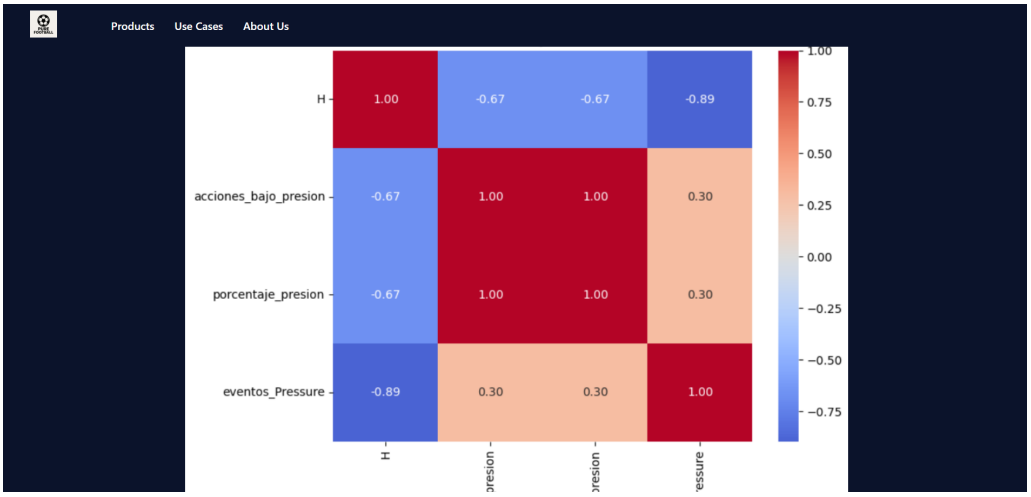
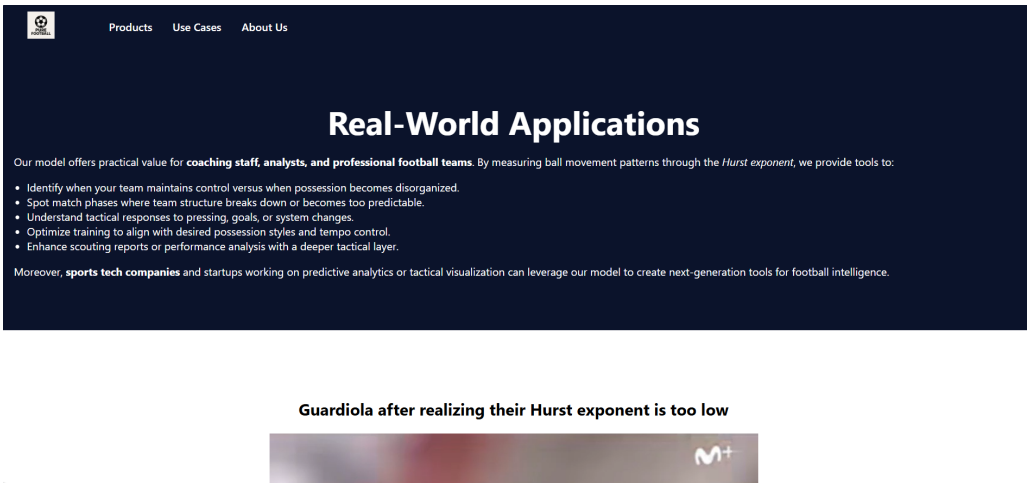



Figure 20


Figure 21




**Guardiola after realizing their Hurst exponent is too low**



Figure 22

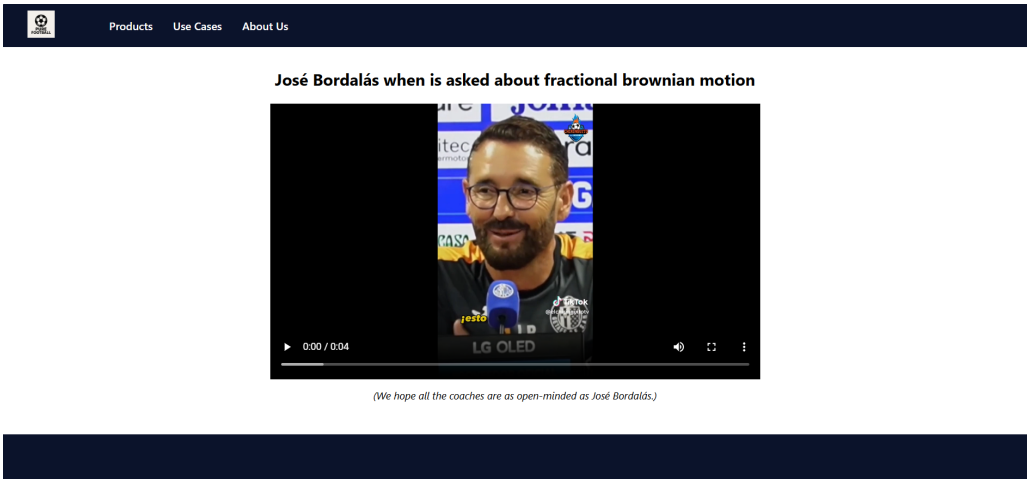



Figure 23

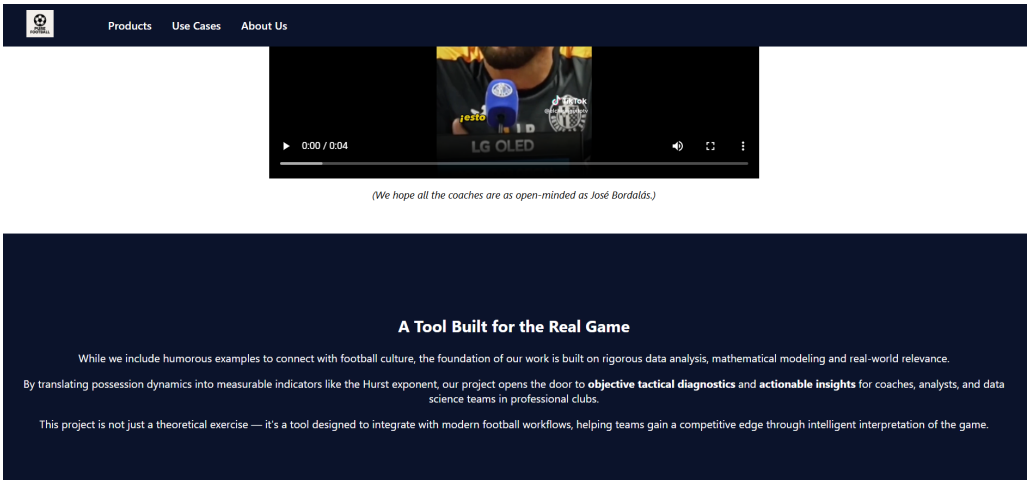



Figure 24

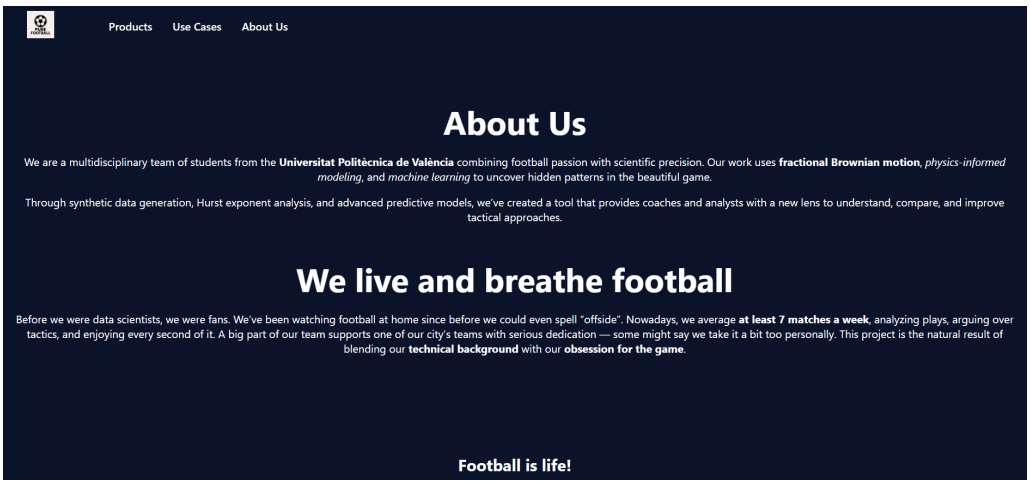



Figure 25