UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

MILESTONE 2: Soccer Diffusion

Data Science

Author: Samuel Mir Arribas
Joaquín Carrión Gil
Gonzalo Hurtado Sanhermelando
Marcos Gómez Soler
Jorge Durá Jimenez
Ivan García Donderis

Course 2024-2025

Contents

CHAPTER 1
Introduction

The objective of this project is to analyze ball diffusion in football using mathematical models and advanced data analysis techniques. The central idea is to investigate patterns in ball movement and their relationship with team playing strategies. To achieve this, we will apply concepts from Brownian motion theory and machine learning techniques to real match data from LaLiga.

Football data analysis and modeling is a growing discipline with a significant impact on decision-making within the sport. From improving player scouting to refining game strategies, the application of data science tools can provide competitive advantages to teams. Our project aims to generate valuable insights that can be used by coaches, analysts, and football teams to optimize performance and enhance game understanding.

This report covers data preparation, detailing the filtering and selection process of relevant information from the StatsBomb dataset, as well as the creation of new variables for analysis. Next, we will explain the modeling task, including variable selection and the methodology used. In the modeling and evaluation phase, we will analyze the performance of different machine learning algorithms and their applicability to our project's objective. Finally, we will discuss the obtained results, their impact on football, and the feasibility of implementation in analytical tools.

CHAPTER 2
Data Preparation and Minable View

2.1 Data Preparation

For this project, we worked with event data from StatsBomb, which was already cleaned and well-structured. Since the dataset did not require extensive cleaning, we focused on data selection, exploration, and feature engineering to ensure the data was suitable for analysis.

First, we filtered the original dataset to focus on possessions and passes to work with a controllable dataset. This subset became the foundation for both analyzing real data and generating synthetic data. Using the real data, we created visualizations and calculated summary statistics to explore how the ball moves during possessions, looking at both spatial and temporal patterns.

A major part of our data preparation involved generating synthetic possessions using fractional Brownian motion (fBm). By adjusting the Hurst exponent (H), we simulated different styles of ball movement, from more chaotic to smoother trajectories. These synthetic possessions included random sequences of passes and carries, with durations and spatial displacements modeled using stochastic processes. In total, we generated over 50,000 synthetic possessions, which were saved for later analysis.

To make both the real and synthetic data more useful for modeling, we engineered several new features. These included metrics like average displacement in x and y directions, total distance covered, possession duration, mean speed, distance per pass, average movement angle, speed variation, and acceleration. We calculated these features for each possession individually and stored them in a structured format that could be used directly in machine learning models.

Overall, this phase of the project was crucial for turning raw positional and temporal data into meaningful metrics that capture important aspects of ball movement. Both the real and synthetic datasets gave us a solid foundation to analyze patterns in football possessions and build predictive models later on.

2.2 Materialising Minable View

Building on our data preparation work, we created a structured dataset (minable view) specifically designed for training machine learning models to predict possession styles. Each row in our dataset represents one synthetic possession, containing all the engineered features we developed during data preparation.

For our modeling task, we focused on predicting the Hurst exponent (H) of each possession, which characterizes its movement pattern. We organized our minable view with:

Input features: The spatial and temporal metrics we calculated (displacement, distance covered, duration, speed, acceleration, etc.)

Target variable: The Hurst exponent value (H) that was used to generate each synthetic possession

The Hurst exponent serves as our key prediction target because it directly relates to possession behavior:

H > 0.5 indicates persistent movement (consistent direction)

H < 0.5 shows anti-persistent movement (frequent direction changes)

H = 0.5 represents completely random movement

We stored this structured dataset in CSV format, with each row containing:

- All the calculated movement features

- The generated Hurst exponent value

This minable view connects directly to our modeling phase by providing:

- Clean, preprocessed data ready for algorithm training

- Clear input-output relationships for supervised learning

- Consistent formatting for synthetic possessions

- All necessary features in numerical format for machine learning

This well-structured dataset enables our models to effectively learn the relationships between possession characteristics (input features) and movement patterns (Hurst exponent values). More importantly, this approach serves a dual purpose: first, it allows us to compare real football possessions against our synthetic data with known Hurst values to validate their similarity; second, it provides the foundation to study how Hurst exponent variations during matches correlate with actual game dynamics - including defensive pressure levels, goal-scoring events, and other critical match factors that influence possession behavior.

CHAPTER 3

Model Prototype and Evaluation

3.1 Model building

In this project, we developed and trained different machine learning models to predict the Hurst exponent (H) from football possession data. We implemented five different approaches to ensure a comprehensive comparison: Linear Regression (as a baseline), Random Forest, XGBoost, a Multi-Layer Perceptron (MLP), and an LSTM network.

For the Linear Regression model, we standardized the features using StandardScaler to ensure consistent scaling. While simple, this model provided a useful baseline for comparison. The Random Forest and XGBoost models were chosen for their ability to capture non-linear relationships in the data. We performed hyperparameter tuning for both using GridSearchCV with 5-fold cross-validation to optimize their performance. The best parameters for the Random Forest included 500 estimators with no maximum depth restriction. Besides these hyperparameters, we included others in the gridsearch like min_samples_split and min_samples_leaf. Focusing on XGBoost, its better performance was with a learning rate of 0.05, maximum depth of 10 and 100 estimators. We also included other hyperparameters like subsample, colsample_bytree or gamma.

To explore more complex architectures, we designed an MLP with two hidden layers (128 neurons each) and ReLU activation functions, trained for 50 epochs using the Adam optimizer. Despite its potential, the MLP showed unstable training and higher errors compared to tree-based models. We also experimented with an LSTM network, even though the data lacks a true temporal structure, to test if sequential modeling could offer any advantages. However, the LSTM's performance was similar to simpler models, confirming that the added complexity was unnecessary.

3.2 Evaluation

We evaluated all models using the Mean Absolute Error (MAE) on a held-out test set (20 % of the data). This metric is particularly appropriate for our task because the Hurst exponent is bounded between 0 and 1, with critical behavioral thresholds at 0.5. Values above 0.5 indicate superdiffusive (persistent) possessions with memory effects, while values below 0.5 correspond to subdiffusive (anti-persistent) behavior. Given this context, our achieved MAEs (0.11-0.12 for the best models) represent good performance - they correspond to average prediction errors of just 11-12% of the total scale, meaning our models distinguish between these different possession regimes.

The Linear Regression baseline achieved an MAE of 0.1179, setting a reference point for comparison. The Random Forest model matched this performance exactly (MAE: 0.1179), suggesting that the data may not contain strong non-linear patterns that Random Forests typically exploit. XGBoost performed slightly better with an MAE of 0.1171, making it the best-performing model.

The MLP, despite its theoretical capacity for modeling complex relationships, underperformed significantly (MAE: 0.3080). This was particularly surprising given our dataset contains 50,000 artificial possessions, which should theoretically be sufficient for neural networks. The poor performance suggests the underlying relationships may be fundamentally simple rather than data-limited. The LSTM achieved an MAE of 0.1207, demonstrating that its sequential processing capability did not provide any meaningful improvement for this task.

To ensure robust evaluation, we used a consistent train-test split (80-20) and employed cross-validation during hyperparameter tuning for the Random Forest and XGBoost models. This approach helped us avoid overfitting and provided reliable estimates of model performance.

In conclusion, XGBoost has been selected as the most suitable model for this task, though its advantage over simpler models was minimal. The consistent performance across models and the neural networks' struggles suggest the relationships between possession features and the Hurst exponent may be inherently simple enough to be captured by non-complex models.

CHAPTER 4
Discussion

4.1 Deployment Mockup

4.1.1. Practical Application for Coaching Staff

Our project provides coaches and technical staff with a powerful tool to analyze team possession patterns through the value of the Hurst exponent. By classifying possessions as persistent (controlled play), anti-persistent (disjointed circulation), or random (transition moments), we offer actionable insights into a team's playing style effectiveness. The model's output helps identify: when the team maintains optimal ball progression (high H values), when possession becomes unstable (low H values), and how these patterns correlate with match situations like high pressing or scoreline changes.

4.1.2. Mockup Demonstration

The accompanying comic strip illustrates how this analysis could be implemented in real coaching scenarios:

Comic view (click on it)

In addition, we attach a prototype of the website intended to present our project:

Mockup view (click on it)

4.2 Use of Technology

Throughout the project, the team effectively integrated various technological tools and Python libraries to approach sports data analysis from multiple perspectives. We used statsbombpy for football data collection, along with pandas, numpy, and scipy.stats for data processing and statistical analysis. Visualization and progress tracking were supported by matplotlib and tqdm, while the modeling phase included advanced techniques using xgboost, sklearn modules (model_selection, metrics), and neural networks developed with torch.nn, torch.optim, and torch.utils.data. Additionally, we implemented a physical model based on Brownian motion to simulate player movement. This presented an initial challenge due to our lack of familiarity with the concept. To understand how it worked, we consulted physics students and reviewed several academic articles. This pro-

cess reflects the team's ability to learn autonomously, adapt to new tools, and effectively apply interdisciplinary knowledge.

4.3 Seminar and AI Use

Throughout the project, we applied several approaches and tools introduced during the course seminars, adapting them to our specific needs. Clearly assigning roles within the team was essential for organizing the work, leveraging individual strengths, and maintaining smooth communication, which enabled effective collaboration and coordinated progress.

CRISP-DM, one of the core methodologies covered in class, served as our structural guide. Its iterative, business-focused approach allowed us to move systematically from understanding the problem to evaluating the model, always maintaining a practical perspective.

Feature engineering was another key stage. We learned that properly transforming data is crucial to improving prediction quality. This step helped ensure the model aligned more closely with our project goals.

We were also introduced to the use of the Lean Canvas Template, which we applied to define and validate the project's business vision. This tool helped us identify key elements of the business model in an agile and structured way, supporting informed decision-making in early stages.

Additionally, we used ChatGPT, Deepseek and Claude as support tools throughout the development process. They helped us better understand complex concepts such as the Brownian model, translate technical content into English, and optimize sections of code. Their use complemented our learning and allowed us to progress more efficiently.

Bibliography

[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, 2023. Bachelor's Thesis, Universitat Politècnica de València.

[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, 2024. Bachelor's Thesis, Universitat Politècnica de València.

[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.