Artificial Intelligence Powered Heatmap Analysis of Pedri's Midfield Positioning and Spatial Behavior in FC Barcelona's 2024-25 La Liga Season: Tactical Roles, Zone Influence, and Impact on Team Strategy

Table of Contents

Introduction

Football has evolved into a highly complex and data-rich sport, where team strategies are increasingly influenced by granular player analysis. The role of the modern midfielder has become multifaceted, demanding a blend of creative playmaking, defensive resilience, and tactical intelligence. Players are no longer defined by singular positions but by their spatial behavior and influence across different phases of play. Traditional methods of performance analysis, such as counting goals or assists, often fail to capture this nuanced contribution. This gap highlights the need for advanced analytical tools that can translate raw positional and event data into actionable tactical insights.

This thesis presents the design and development of an AI-powered analytics dashboard focused on the performance of a single elite midfielder, Pedro González López (Pedri), during FC Barcelona's 2024-25 La Liga season. The project moves beyond simple statistics to provide a deep, contextual evaluation of his on-pitch behavior. It leverages unsupervised machine learning and advanced statistical models to analyze match-by-match event data, generate dynamic visualizations, and produce automated performance reports. The system is designed to identify a player's tactical role, quantify their creative impact, and benchmark their performance against a curated group of elite European midfielders.

The primary goal is to demonstrate how modern data science techniques can reveal hidden patterns in a player's spatial conduct. By transforming abstract coordinate data into intuitive heatmaps, passing networks, and comparative percentile profiles, the dashboard offers a powerful lens through which coaches, analysts, and fans can better understand a player's influence on team strategy.

Problem Context and Motivation

The analysis of an individual footballer's performance is often shallow, and narrative driven. Media commentary and fan discussions tend to focus on decisive moments like goals or assists, while overlooking the consistent, subtle actions that truly define a player's role. For a midfielder like Pedri, whose value lies in controlling tempo, progressing the ball, and finding space, these traditional metrics are insufficient. His contributions are distributed across the entire pitch and are

better measured through his positioning, movement, and the value of his actions rather than just their outcomes.

This creates an analytical challenge: how to objectively quantify and visualize the tactical contribution of a player whose impact is not always reflected on the scoresheet. Most existing analytics platforms present data in tables or simple charts, leaving the complex task of interpretation to the user. This research addresses that gap by building a system that not only presents data but also interprets it within a tactical and comparative context.

The motivation for this project stems from the need for more accessible and intelligent tools in sports analytics. By applying machine learning models to event data, this thesis seeks to automate the process of tactical analysis, making it possible to derive sophisticated insights from a single match file. The resulting dashboard serves as a proof-of-concept for a new generation of player analysis tools that are data-driven, evaluative, and context-aware.

Theoretical Framework and Analytical Models

The dashboard's analytical power is derived from the integration of several modern data science models and theories. These frameworks allow the system to move from description to interpretation.

Expected Threat (xT) and Valuing On-Ball Actions

A central challenge in football analytics is assigning value to actions that do not immediately lead to a shot. A simple pass in the defensive third is not as valuable as a through ball that breaks the opponent's defensive line. The **Expected Threat (xT)** model addresses this by assigning a goal-scoring probability to every zone on the pitch. An action generates value if it moves the ball from a low-xT zone to a high-xT zone. This project utilizes pre-calculated xT values as its core metric for measuring creative impact. By averaging the xT generated per action, the dashboard quantifies a player's ability to advance the ball into dangerous areas, providing a far more accurate measure of creative contribution than traditional metrics.

Unsupervised Machine Learning for Tactical Role Detection

A player's designated position on a team sheet often fails to capture their actual function during a match. To capture this tactical flexibility, the project employs an unsupervised machine learning

algorithm known as **K-Means Clustering**. The model works by grouping the (x, y) coordinates of a player's on-ball actions into clusters. By calculating the average pitch position of these clusters, the algorithm can objectively determine the player's "center of gravity" for that match. This quantitative output is then mapped to a tactical role classification (e.g., a deep-lying "#6", a box-to-box "#8", or an advanced "#10"), automating a task that would typically require hours of manual video analysis

Statistical Benchmarking and Percentile Ranking

A player's performance in a single match is difficult to evaluate in isolation. To provide context, the project uses a **statistical benchmarking model** based on **percentile ranking**. A comprehensive dataset of season-long performance metrics was compiled for a curated group of elite European midfielders. When a player's match data is analyzed, their KPIs are compared against this peer group. The system calculates the percentile rank for each metric, indicating where the player's performance stands relative to the elite standard. This transforms raw numbers into an intuitive and powerful comparative evaluation.

Research Aim

The primary aim of this thesis is to design and develop an AI-powered system that provides a deep and contextual analysis of a midfielder's match performance by integrating spatial data, machine learning, and statistical benchmarking.

Research Objectives

1. To collect and process real-world match event data for an elite midfielder, ensuring it is suitable for spatial and statistical analysis.
2. To implement and apply a K-Means clustering algorithm to automatically infer a player's tactical role based on the spatial distribution of their on-ball actions.
3. To integrate the Expected Threat (xT) model as a key performance indicator for quantifying a player's creative and territorial impact.
4. To develop a comparative framework using percentile ranking against a benchmark dataset of elite midfielders to contextualize and evaluate performance.

5. To design and build an interactive web-based dashboard that visualizes these analyses through heatmaps, passing networks, and dynamic performance reports, making complex insights accessible to a non-technical audience.

Contribution and Significance

This research makes a practical and academic contribution to the field of sports analytics. Academically, it presents a cohesive framework for integrating multiple data science models (K-Means, xT, percentile ranking) into a single, unified system for player evaluation. It provides a replicable methodology for moving beyond descriptive statistics to generate automated, context-aware tactical insights.

Practically, the project serves as a prototype for a new generation of sports analytics tools. It demonstrates that complex machine learning models can be deployed in an accessible and intuitive dashboard, empowering users to conduct their own sophisticated analysis without needing a background in programming or data science. By focusing on a single player, it highlights the depth of insight that can be gained from granular, event-level data.

In a broader sense, this work illustrates the power of applied artificial intelligence to augment human expertise. The dashboard does not replace the analyst but equips them with a powerful tool to accelerate their workflow, identify patterns that might be missed by the human eye, and support their observations with quantitative evidence.

Justification of the Study

The field of sports analytics is undergoing a profound transformation. While the availability of granular player data has grown exponentially, the tools and methodologies for deriving meaningful tactical insights from this data have not kept pace. Most public-facing analysis remains descriptive, focusing on outcome-based metrics like goals and assists, or simple volume metrics like pass counts. This study is justified by its direct attempt to address this analytical gap by moving from descriptive to diagnostic and evaluative analysis. It focuses on the modern midfielder, arguably the most tactically complex position in football, whose contributions are often spatially and contextually dependent, making them difficult to capture with traditional statistics.

The selection of Pedri as a case study is deliberate. As a player renowned for his exceptional spatial awareness, press resistance, and ability to control game tempo, he embodies the qualities that simple metrics fail to quantify. His performance is an ideal test case for demonstrating the power of advanced models like Expected Threat (xT) and K-Means clustering to reveal a player's true influence on the game's structure and flow. By focusing on a single, high-profile player, this research can achieve a depth of analysis that broader, team-level studies often miss.

Technologically, this project is both timely and highly relevant. The maturation of open-source Python libraries for data science (such as Scikit-learn, Pandas, and Matplotlib) and for web application development (Streamlit) has democratized access to powerful analytical tools. It is now feasible for independent researchers to build and deploy sophisticated, interactive dashboards that were once the exclusive domain of elite professional clubs with dedicated analytics departments. This thesis capitalizes on this technological readiness to create a tangible, functional artifact that serves as a proof-of-concept for accessible, high-level sports analysis.

Furthermore, this study is justified by its contribution to a more evidence-based and nuanced discourse surrounding player performance. In a media landscape often dominated by subjective narratives and reactionary hot takes, there is a pressing need for objective, data-driven tools that can support analysis with quantitative evidence. The dashboard developed in this project provides such a tool. It is designed not to replace human expertise but to augment it, offering analysts, coaches, and educated fans a way to test their hypotheses, identify subtle performance trends, and understand a player’s tactical flexibility on a match-by-match basis.

The interdisciplinary nature of the research provides additional justification. The project does not reside purely within computer science but integrates concepts from sports science, statistical modeling, and data visualization. It demonstrates how abstract algorithms can be applied to solve concrete, real-world problems in a domain with massive public interest. By presenting its findings in an intuitive, visual format, the study also addresses the critical "last mile" problem in data analytics: translating complex quantitative outputs into insights that are easily understandable and actionable for a non-technical audience. This makes the project not just a technical exercise but also a study in effective data communication.

Finally, the study operates with a clear understanding of its scope and limitations. It acknowledges that quantitative models cannot capture every aspect of a player's performance, such as leadership

or off-ball intelligence that does not result in a recorded event. However, it is justified on the principle that providing an objective, data-driven layer of analysis is a significant improvement over relying on purely subjective assessments. By offering a replicable framework for single-player analysis, this research provides a valuable contribution that can be built upon, critiqued, and extended by future work in the growing field of football analytics.

Research Questions

- How can a specialized analytics dashboard be designed using machine learning (K-Means Clustering) and advanced statistical models (Expected Threat, Percentile Ranking) to provide a deep, contextual, and evaluative analysis of an elite midfielder's match-by-match spatial and tactical performance?
- What are the analytical limitations and ethical considerations inherent in using an AI-driven system to evaluate complex human performance in professional sport, and what principles must guide the responsible interpretation and deployment of such a tool to avoid oversimplification and potential bias?

Research Hypotheses

Hypothesis 1 (for Research Question 1):

The integration of K-Means clustering for role detection, Expected Threat for creative valuation, and percentile benchmarking for performance context will yield a multi-dimensional player profile that is significantly more insightful and tactically relevant than analysis based on traditional descriptive statistics (e.g., goals, assists, pass completion percentage) alone.

Hypothesis 2 (for Research Question 2):

It is hypothesized that while the AI-driven dashboard can provide objective and rapid performance analysis, its ethical and responsible application is fundamentally dependent on its ability to provide contextual benchmarks (i.e., percentile rankings), transparently acknowledge its methodological limitations (e.g., the assumptions of the xT model), and be presented as a supplemental tool designed to augment, rather than replace, expert human judgment.