Classifying Player Roles at the 2022 FIFA World Cup Using Machine Learning

Introduction

This project uses performance data from the 2022 FIFA World Cup to identify natural player roles based on their style of play. Rather than labeling players manually by position or using predefined categories, we applied machine learning to let the data reveal the groupings.

Objective

The objective was to group players by their match behavior — such as passing frequency, shooting rate, and defensive actions — and to discover distinct **player archetypes** without pre-labeling or assumptions. The goal was not to evaluate who was better, but rather to **understand how players contribute** differently on the field.

Methodology

3.1 Data Collection

We used a public dataset containing aggregated player statistics from the 2022 World Cup. This included metrics like:

- Total passes, shots, and tackles
- Expected goals (xG)
- Minutes played

To ensure reliability, only players who participated in at least **two full matches** (≥180 minutes) were included. The final dataset included **335 players**.

3.2 Feature Engineering

We normalized each stat per 90 minutes played (standard practice in sports analytics) to make fair comparisons across players with varying game time.

The features were grouped into:

- **Finishing metrics** (shots, xG)
- **Creative metrics** (passes, assists, progressive passes)
- **Defensive metrics** (tackles, interceptions, clearances)

3.3 Model Selection

We used **K-Means Clustering**, a simple but powerful algorithm that groups similar players based on their numeric features.

Other models were considered:

- **Hierarchical clustering**: good for small datasets, but less intuitive for visual interpretation
- **DBSCAN**: useful for noisy, unstructured data — not ideal for our clean numerical format
- **Gaussian Mixture Models**: handles overlapping roles well, but more complex and less explainable

K-Means was chosen for its speed, simplicity, and clear group boundaries — ideal for a project that prioritizes interpretability and stakeholder communication.

Results

After testing various group sizes using the Elbow Method and Silhouette Score, we selected **4 clusters** as the optimal number of player groups. These clusters revealed four clear player roles:

1. Finishers

These players had the highest number of shots per match and frequently got into scoring positions. They were less involved in buildup play or defense. Typical examples would include strikers and goal-hungry wingers.

2. Playmakers

This group featured players who completed the most passes, particularly forward or progressive passes into dangerous areas. They were the architects of the attack — midfielders who shaped the rhythm and direction of play.

3. Ball-Winning Defenders

Players in this cluster were highly active defensively, recording high numbers of tackles, interceptions, and clearances. They are critical in disrupting opposition attacks and protecting the defensive third.

4. Limited Involvement

This group had the lowest activity across all metrics. Players in this cluster may have played in limited roles (e.g., substitutions, non-possession-focused defenders), or had minimal impact across both attack and defense.

These roles emerged naturally from the data — no prior labels were used.

Visualization

To help stakeholders and non-technical audiences interpret the findings, we used **Principal Component Analysis (PCA)** to reduce the multi-dimensional data into a two-dimensional plot.

This visualization revealed clear grouping patterns, with each cluster forming distinct regions — visually confirming the validity of the roles assigned. Each player is represented as a dot, color-coded by role, showing how similar players tend to gravitate together based on their stats.

Why This Work Matters

Understanding player roles through data — instead of assumptions or labels — is a powerful tool for modern football analysis.

For instance, a scout evaluating players can use this model to identify talent who matches a team's tactical needs. A coach can use these insights to ensure the right balance of roles is present in the starting eleven. Even fans and broadcasters can use this information to better understand the dynamics behind team performance.

Moreover, organizations like **Sportable**, which collect real-time player and ball tracking data, can apply the same modeling techniques using even richer movement features. This

can lead to **live role classification**, on-the-fly tactical analysis, and data-driven substitution decisions during matches.

This project is a proof-of-concept for how meaningful roles can be uncovered and explained — without any subjective bias.

Future Extensions

This analysis opens the door to several exciting future directions:

- **Role prediction models**: Now that we've labeled players with roles using clustering, we could train a supervised machine learning model to predict a player's role given their stats — useful for scouting players in other tournaments or leagues.

- **Time-based role evolution**: With match-by-match data, we could track how players shift roles depending on opponent, formation, or tournament phase.

- **Team-level strategy insights**: By summarizing the role distribution within each national squad, we could compare tactical strategies across teams — e.g., whether a team is defense-heavy, balanced, or attack-oriented.

- **In-game tracking data**: Using tracking coordinates (speed, direction, position), this exact method could be used to classify roles and actions in real time — which is the kind of product Sportable offers to elite sports teams.

Conclusion

This project demonstrates the power of machine learning in sports analytics — not to predict the future, but to **understand the present more clearly**. By clustering players based on per-90-minute performance metrics, we were able to uncover **natural playing styles** without any labels.

The result is a data-driven method to understand player behavior, assess team dynamics, and support smarter, more objective decision-making in football. It's a model that could be deployed in scouting, tactical planning, or even live-match commentary.

With access to richer real-time data, such as those collected by Sportable's Smart Ball and player tracking technology, these same methods can evolve into powerful **real-time performance insight tools** used by elite teams around the world.