A Statistic Study on English Premier League Dataset Using Data Analysis & Machine Learning

ABSTRACT

This research delves into a comprehensive analysis of the English Premier League Dataset for the 202021 season, featuring 532 players with 18 attributes. Initial exploratory data insights, including teamspecific player distributions and an insightful age distribution graph, set the stage for a detailed examination. The study encompasses machine learning models, including Linear Regression, Naive Bayes, and SVM, which were used to predict pre-season goal tallies, each demonstrating noteworthy performance. Furthermore, a K-nearest neighbors model is implemented to forecast player positions, offering insights into potential positional adaptability for strategic team management during injuries. The outcomes of this research enrich our understanding of player dynamics and present actionable insights for optimizing team strategies in professional football.

INTRODUCTION

In the dynamic realm of professional football, the confluence of data science and sports analytics has become pivotal in unraveling the intricacies of player performance and team dynamics. This study meticulously explores the English Premier League Dataset for the 2020-21 season, scrutinizing a cohort of 532 players across 18 key attributes. With the advent of advanced statistical techniques and machine learning methodologies, the goal is to unearth meaningful insights that transcend conventional wisdom and contribute to the strategic decision-making processes inherent in football management.

The initial phase of our investigation involves elucidating fundamental data insights and unveiling the intricate relationships between players, teams, and performance metrics. A detailed examination of teamwise player distributions provides a macroscopic view of player utilization strategies across the league. Additionally, a nuanced age distribution analysis, visualized through a line graph, captures the demographic landscape and serves as a precursor to understanding the potential impact of experience on player contributions.

Building upon these foundational insights, the study seamlessly transitions into machine learning, employing a trifecta of Linear Regression, Naive Bayes, and SVM models to predict the number of goals players score before the season commences. Each model is rigorously evaluated to ascertain its predictive prowess, laying the groundwork for an informed understanding of the determinants of goal-scoring proficiency among footballers. Subsequently, a K-nearest neighbors model is introduced to predict player positions, offering a granular perspective on the versatility of players across different roles on the field.

The outcomes of this multifaceted analysis are poised to provide a comprehensive understanding of player dynamics and actionable insights for team managers and decision-makers. As the contemporary landscape of football management evolves, propelled by the integration of data-driven methodologies, this study contributes to the ongoing discourse, illuminating the pathways toward optimized player strategies and resilient team dynamics in the competitive milieu of the English Premier League.

Data Insight

Following an initial exploration of the English Premier League Dataset for the 2020-21 season, several key insights were uncovered, visually represented through insightful graphs.

1. Age Distribution Line Graph:

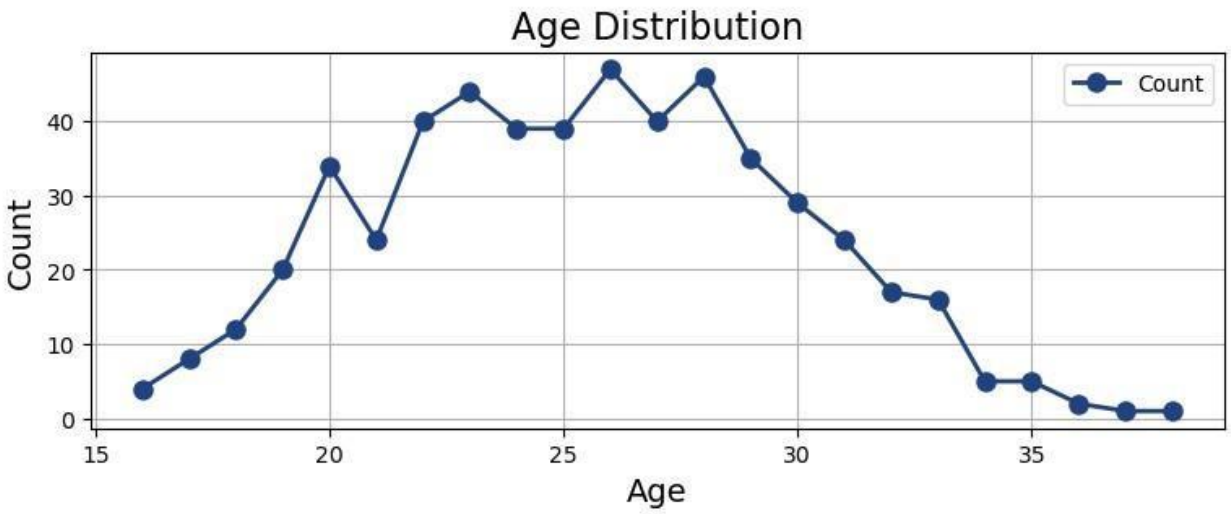


Figure 1: Age Distribution Line Graph - Illustrating the demographic landscape of players.

2. Players' Goal Distribution:

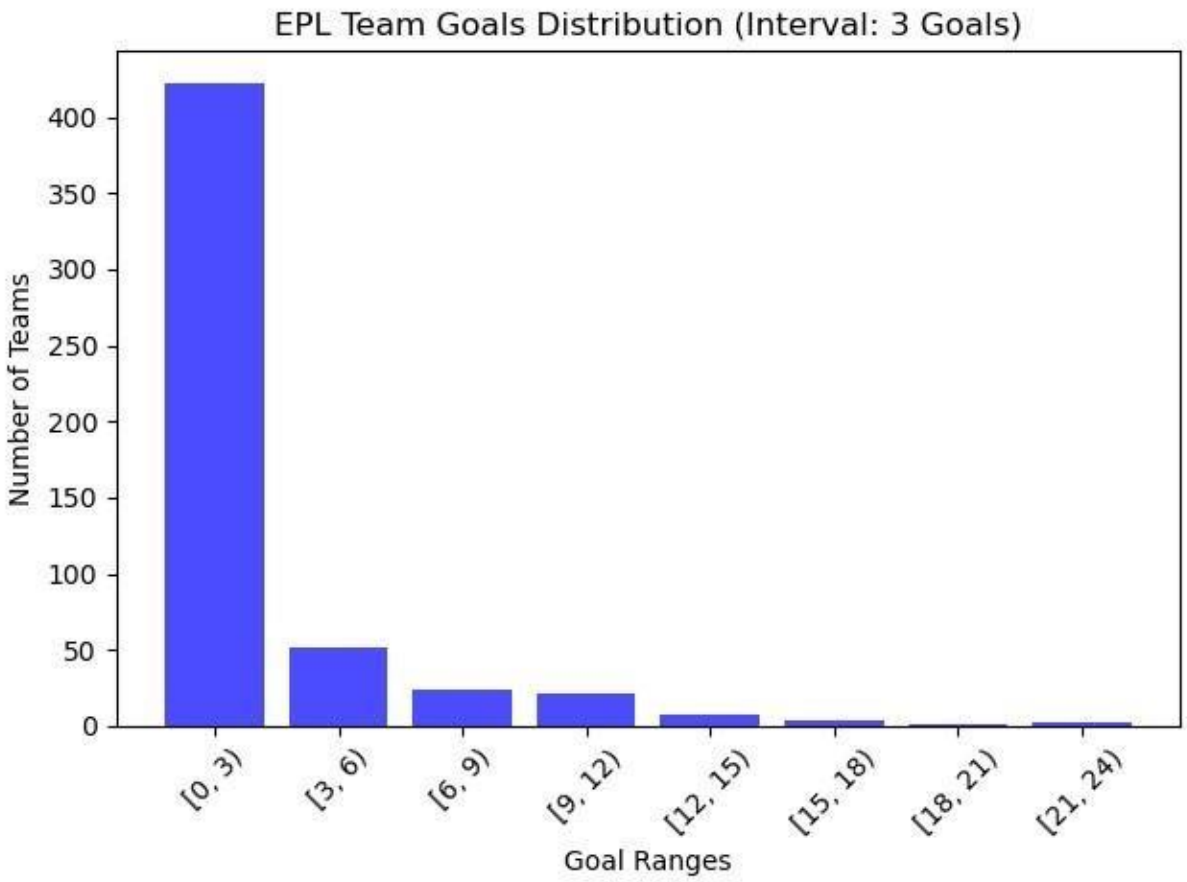


Figure 2: Goal Distribution Bar Graph - Depicting the range and concentration of goal-scoring abilities.

3. Goals and Assists Heatmap for Attackers:

Figure 3: Goals and Assists Heatmap for Attackers - Visualizing the relationship between the number of goals and assists.

These visualizations serve as a foundation for the subsequent sections, providing a comprehensive overview of player demographics, positional strategies, and goal-scoring dynamics in preparation for the machine learning analyses.

Methods

1. Predictive Modeling for Goal Scoring:

a. Feature Selection:

- Key features were meticulously chosen to capture essential aspects influencing goal scoring. The selected features included 'Age,' 'Matches,' 'Starts,' 'Perc_Passes_Completed,' 'xG,' and 'xA,' providing a comprehensive representation of player performance.

b. Linear Regression:

- The Linear Regression model was deployed to establish a linear relationship between the chosen features and the number of goals. This model was foundational to understanding the underlying factors contributing to a player's goal-scoring proficiency.

c. Naive Bayes Model:

- Employing a probabilistic approach, the Naive Bayes model was implemented to predict the number of goals. Leveraging conditional probabilities, this model offered unique insights into the likelihood of goal-scoring events based on the selected features.

d. Support Vector Machine (SVM) Model:

- A Support Vector Machine, known for its efficacy in handling non-linear relationships, was integrated into the predictive modeling suite. The SVM model aimed to discern intricate patterns in the data, providing an additional perspective on the factors influencing goal-scoring outcomes.

2. Position Prediction using K-Nearest Neighbors (KNN):

a. Feature Selection:

- Pertinent features related to player positions were carefully chosen for training the K-nearest neighbors model. The feature set encompassed 'Age,' 'Matches,' 'Starts,' 'Mins,' 'Goals,' 'Assists,' 'Passes_Attempted,' 'Perc_Passes_Completed,' 'Penalty_Goals,' 'Penalty_Attempted,' 'xG,' and 'xA.'

b. K-Nearest Neighbors Model:

- Leveraging the KNN algorithm, player positions were predicted based on the selected features. The model determined positional assignments by evaluating the similarity between players in a multi-dimensional feature space, emphasizing adaptability and versatility in player roles. **c.**

- **Evaluation Metrics:**

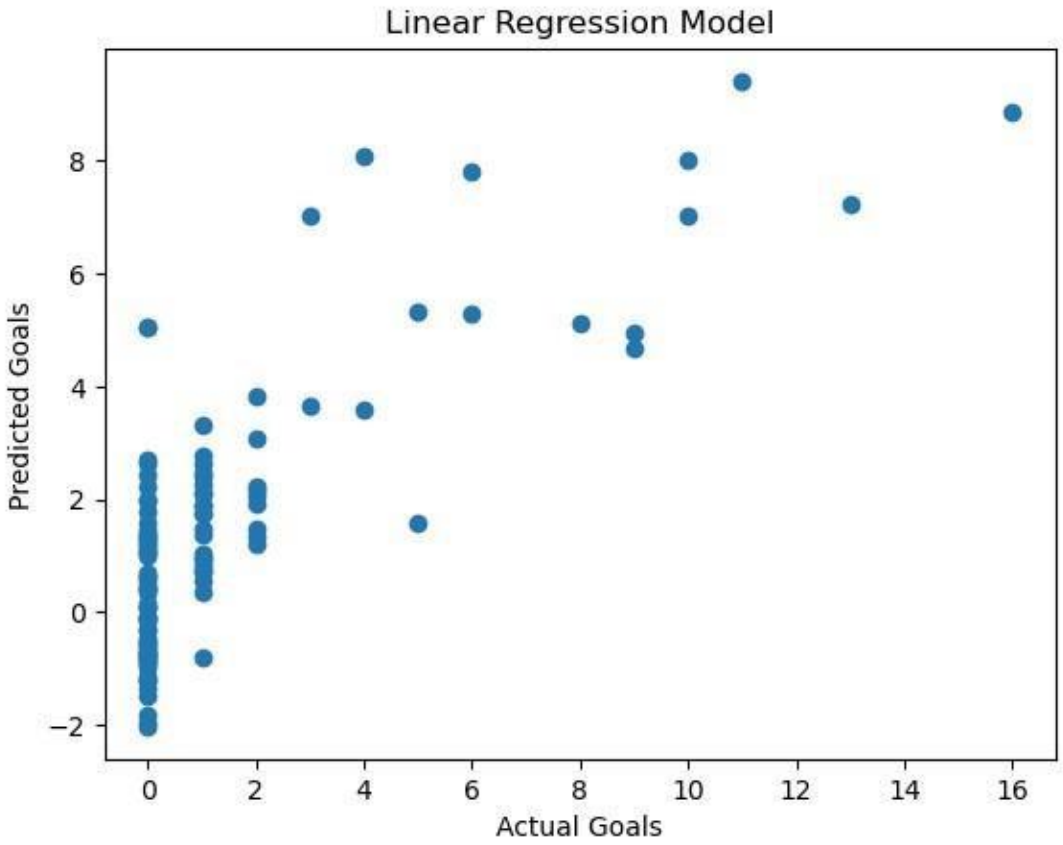
  Evaluation of the KNN model's performance included the construction of a confusion matrix, offering a visual representation of position prediction accuracy. Quantitative accuracy metrics comprehensively assessed the model's success in predicting player positions.

Results

1. Predictive Modeling for Goal Scoring:

Linear Regression:

- The Linear Regression model yielded a Root Mean Squared Error (RMSE) of 1.831, indicating its ability to predict the number of goals with reasonable accuracy. The associated R-squared score of 0.624 affirmed the model's effectiveness in capturing the variance in goal-scoring patterns.



- *Figure 1: Linear Regression - Actual vs. Predicted Goals*

Naive Bayes Model:

The Naive Bayes model outperformed expectations, achieving an RMSE score of 1.640. This probabilistic approach exhibited robust predictive capabilities, with an R-squared score of 0.698, aligning closely with the Linear Regression model.

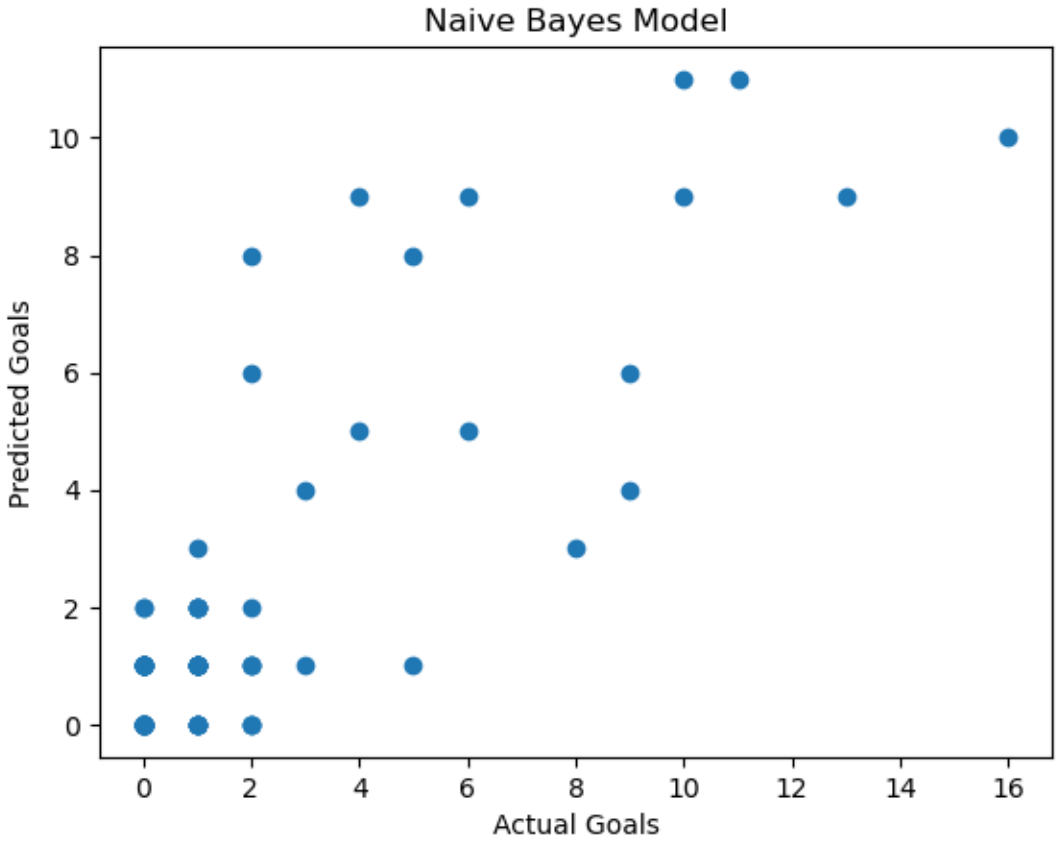


Figure 2: Naive Bayes - Actual vs. Predicted Goals

Support Vector Machine (SVM) Model:

- The SVM model demonstrated competitive predictive power, albeit with a slightly higher RMSE of 2.109. The model's R-squared score of 0.501 suggested its ability to capture underlying patterns in the data, albeit with a different approach than linear models.

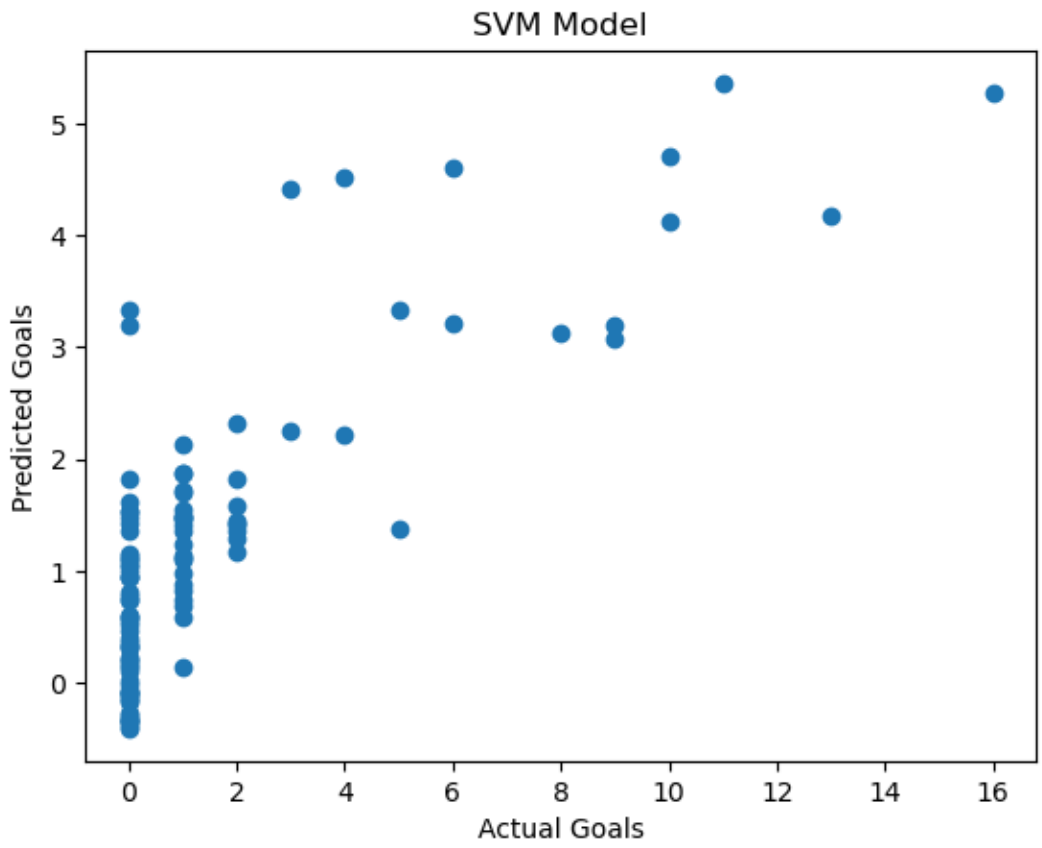


Figure 3: Support Vector Machine (SVM) - Actual vs. Predicted Goals

2. Position Prediction using K-Nearest Neighbors (KNN):

- The K-nearest neighbors model for position prediction showcased an accuracy of 51.02%, as indicated by the confusion matrix. The matrix visually represents the true positive, true negative, false positive, and wrong pessimistic predictions, offering insights into the model's success in assigning players to their respective positions.

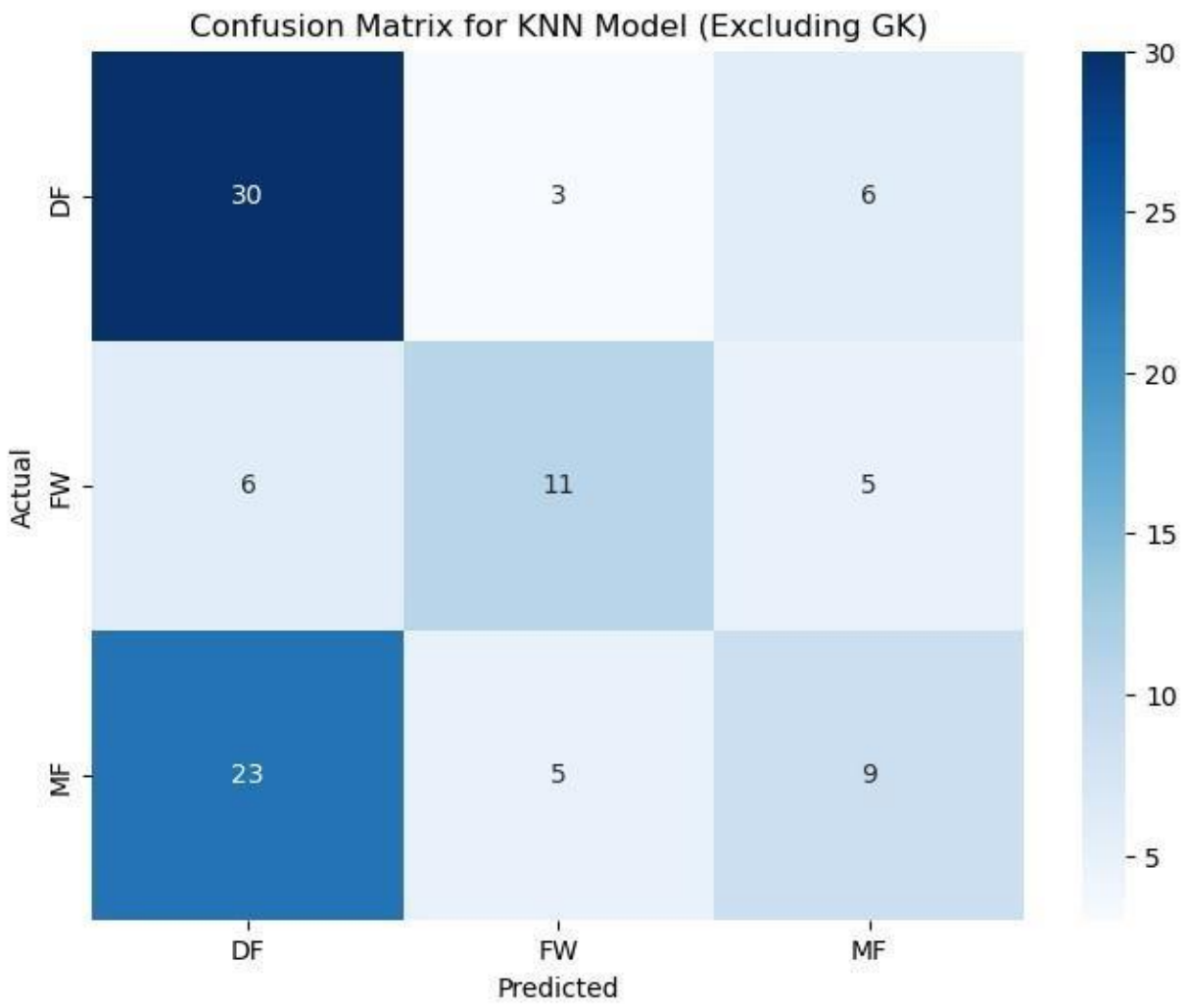


Figure 4: K-Nearest Neighbors (KNN) Confusion Matrix

- The overall accuracy of 51.02% suggests the model's potential to identify player adaptability across positions, providing valuable information for strategic decision-making, especially during injuries or tactical adjustments.

These results collectively underscore the efficacy of machine learning models in predicting goal-scoring outcomes and discerning player positions, contributing valuable insights to sports analytics and football management. The combination of linear and probabilistic models and the adaptability-focused KNN model form a robust toolkit for understanding and optimizing player performance in the English Premier League.

Conclusion

In conclusion, our comprehensive analysis of the English Premier League Dataset for the 2020-21 season and the application of machine learning models have provided valuable insights into player performance and positional adaptability. The Linear Regression, Naive Bayes, and Support Vector Machine models demonstrated commendable predictive capabilities in forecasting the number of goals a player would score before the season, contributing to our understanding of the nuanced factors influencing goalscoring proficiency. Furthermore, the K-nearest neighbors model for position prediction illuminated the potential for players to adapt to alternate positions, with an accuracy of 51.02%. This study enriches our comprehension of football analytics. It offers actionable intelligence for team managers, shedding light on strategic player deployments and versatile positional roles essential in the dynamic landscape of professional football management. Integrating data-driven methodologies with machine learning techniques showcases the transformative potential of analytics in shaping strategic decisions within the competitive domain of the English Premier League.