applied sciences

MDPI

Article

Football Analytics: Assessing the Correlation between Workload, Injury and Performance of Football Players in the English Premier League

Victor Chang [1,*], Sreeram Sajeev [1], Qianwen Ariel Xu [1], Mengmeng Tan [2] and Hai Wang [3]

[1] Department of Operations and Information Management, Aston Business School, Aston University, Birmingham B4 7ET, UK; sreeram.sk.07@gmail.com (S.S.); qianwen.ariel.xu@gmail.com (Q.A.X.)

[2] Alliance Manchester Business School, The University of Manchester, Manchester M15 6PB, UK; mengmengtan0108@gmail.com

[3] School of Computer Science and Digital Technologies, Aston University, Birmingham B4 7ET, UK; h.wang10@aston.ac.uk

* Correspondence: v.chang1@aston.ac.uk or victorchang.research@gmail.com

Abstract: The aim of this research is to shed light on the complex interactions between player workload, traits, match-related factors, football performance, and injuries in the English Premier League. Using a range of statistical and machine learning techniques, this study analyzed a comprehensive dataset that included variables such as player workload, personal traits, and match statistics. The dataset comprises information on 532 players across 20 football clubs for the 2020–2021 English Premier League season. Key findings suggest that data, age, average minutes played per game, and club affiliations are significant indicators of both performance and injury incidence. The most effective model for predicting performance was Ridge Regression, whereas Extreme Gradient Boosting (XGBoost) was superior for predicting injuries. These insights are invaluable for data-driven decision-making in sports science and football teams, aiding in injury prevention and performance enhancement. The study's methodology and results have broad applications, extending beyond football to impact other areas of sports analytics and contributing to a flexible framework designed to enhance individual performance and fitness.

Keywords: football analytics; machine learning in sports; predictive modelling; sports data analysis; injury occurrence analysis

check for updates

Citation: Chang, V.; Sajeev, S.; Xu, Q.A.; Tan, M.; Wang, H. Football Analytics: Assessing the Correlation between Workload, Injury and Performance of Football Players in the English Premier League. *Appl. Sci.* **2024**, *14*, 7217. https://doi.org/10.3390/app14167217

Academic Editors: Joohyung Lee and Sae Yong Lee

Received: 27 June 2024
Revised: 4 August 2024
Accepted: 9 August 2024
Published: 16 August 2024

1. Introduction

1.1. Motivation

Football is a fascinating team sport that demands a wide range of skills to succeed both as a player and as a team. While it is easy to view a football game as a contest where the team with the greatest number of goals wins, football is much more than goals. It is a system constituted by complex interactions between players [1]. In such a multifaceted sport, data analytics are critical in assessing the significant variables associated with player performance, well-being, and fitness, fatigue, etc.

Numerous variables influence a player's performance on the field, and this has historically led to a lot of research being applied to find out the intricate relationship between these variables. Recently, the sports industry, including football, has experienced a lucrative rise in stature and has now become an important contributor to the global economy [2]. Substantial financial investments have led to an exponential rise in the amount of sports data generated. These data are being used by researchers and analytical departments of various teams to discover patterns, trends, and insights that would assist their respective teams to make data-driven decisions to obtain better performance and results.

Historically, research has primarily concentrated on analyzing player performance, with significant efforts directed towards creating predictive models that assess performance metrics and predict outcomes of football matches [3]. This focus stems from the fact that goals ultimately decide football matches and from everyone's interest in match results. Although predicting football results is a very complex task, the football betting industry has grown significantly. The unpredictability of football results and the growing betting business justify the need for prediction models to support gamblers [4].

While goal prediction remains a primary focus in football analytics, it is critical to recognize that football is more than just scoring; it is a sophisticated symphony of multifaceted elements. While goals are the ultimate aim in football, they represent only one note in a complicated orchestra of variables that affect a team's success. Football is a sport in which a plethora of forces collide and interact synergistically to shape the outcome of each match. These variables include, but are not limited to, player workload, physical and mental exhaustion, team tactics, player positioning, ball possession, passing accuracy, defensive techniques, and a plethora of others. Football is an ecosystem of interconnected measures that define a team's performance.

Football's complexity is further influenced by various elements, including a player's personal characteristics, such as position, age, and club, as well as workload metrics, performance indicators, injuries, and fitness levels. While many machine learning (ML) algorithms focus on goal prediction, the game itself is significantly more complex. This study is motivated by the need to investigate the intricate interplay of various factors. Its goal is to unearth hidden correlations, interpret patterns, and demonstrate how these variables intricately impact a player's performance and well-being in the English Premier League (EPL). Beyond merely analyzing scores, our research seeks to redefine football analysis by delving into the sport's rich complexity, providing valuable insights for player management, strategy optimization, and injury prevention, ultimately enhancing the beautiful game.

1.2. Background and Context

Understanding the structure of the English Premier League (EPL) is crucial, particularly its unique format of relegation and rewards. In this league, the top teams not only reap significant financial benefits but also gain opportunities to compete in prestigious tournaments such as the UEFA Champions League. Conversely, the bottom three teams are relegated to a lower division, impacting their status and financial stability.

The 2020–2021 EPL season saw Manchester City emerge as the champions, followed by Manchester United, Liverpool, and Chelsea, all securing spots in the UEFA Champions League. Key individual performances included Harry Kane as the top scorer, which underscores the correlation between playing time and scoring success. The analysis of player workload, especially among forwards, showed top scorers like Harry Kane and Mohammed Salah ranking high in minutes played, suggesting a link between on-field time and goal-scoring achievements.

This season was also notable for its injury trends, which varied significantly across teams. Liverpool, for instance, faced the highest number of injuries, which contributed to their drop to third place after their previous season's championship win. In contrast, teams like Arsenal and Wolves experienced different injury patterns, affecting their season's progress and final standings. Arsenal encountered numerous injuries at the season's start, whereas Wolves had fewer absentees in the initial weeks, possibly contributing to their stronger finish in the league. These injury statistics and their impact on team performance highlight the importance of player health and workload management in professional football.

1.3. Aims and Objectives

The primary objective of this research is to understand the complex relationships between the many variables that affect football players. These variables include player workload, personal traits, and performance in the English Premier League. This study also aims to investigate the significant effects of match- and game-related variables on player performance and injury occurrences. The main goal is to unravel the complex web of correlations within these variables to improve our comprehension of football dynamics. In the pursuit of comprehensive insights into the complex world of football dynamics, this study seeks to answer several research questions:

- What is the correlation between payer workload, player characteristics, and performance of football players in the English Premier League?
- How do game/match-related variables impact football players' performance in the English Premier League, and what are the essential variables contributing towards the prediction of player performance?
- How do game/match-related variables impact the occurrence of injuries among football players in the English Premier League, and what are the important variables contributing to injury prediction?
- How do various factors influence injury occurrence among players?

The study sets explicit aims to solve these research questions:

1. **Correlation Exploration**: Conduct detailed statistical analyses, such as descriptive analytics and correlation analysis, to identify and quantify the complex correlations between player workload, individual traits, and performance measures.
2. **Predictive Modelling**: Develop and integrate advanced ML models to assess the prediction capacity of diverse variables. Determine the essential characteristics that have a major impact on the forecast of player performance (goals and assists), providing insights into the drivers of football success.
3. **Injury Occurrence Analysis**: Examine the impact of game and match-related variables on the occurrence of injuries among English Premier League football players. Use rigorous statistical tools to determine the key variables contributing to injury incidence.

2. Literature Review

2.1. Introduction

Football analytics has become increasingly important in modern football for evaluating and predicting the performance of players and teams. Research in this field has predominantly focused on tactical outcomes and player evaluations through traditional statistics. However, there remains a significant gap in understanding how off-field factors such as workload, injuries, and overall player performance interact with and influence on-field success.

The existing literature highlights the development of football analytics, particularly through ML, and its key role in understanding player performance, fitness, and health. Despite this, few studies have holistically examined the interplay between player workload, the frequency and nature of injuries, and their cumulative impact on player and team performance. This oversight presents a critical research gap, particularly within the high-stakes environment of the English Premier League (EPL), where player performance directly correlates with club financial stability and success.

Within the context of the EPL, a professional football league consisting of the top 20 clubs in England where pursuing success is paramount, gaining knowledge and understanding the factors that influence players' performance, fitness, and well-being is of utmost significance to the clubs. By focusing on these specific areas, this research aims to bridge the existing knowledge gap, providing insights that could significantly enhance player management strategies, reduce injury rates, and optimize overall team performance. This approach aligns with the growing demand for a more comprehensive analysis that goes beyond traditional metrics, offering a nuanced understanding of what influences player output and team success in one of the world's most competitive football leagues.

2.2. Research and Findings

While ML models for football analytics have previously focused on injury prediction and prevention, skill or market value evaluation, and team or player performance prediction, Pantzalis and Tjortjis [3] take a broader approach. They extend their analysis beyond injury prediction and performance assessment to emphasize the prediction of long-term team and player performance. Their study demonstrates that by using historical data and sophisticated statistical methods, it is possible to accurately predict final league standings for specific leagues. Additionally, their study examines a team's performance to determine whether it will advance from a previous season.

Due to the game's constant fluidity, quantitative analysis in football faces significant challenges, making it difficult to establish a rigid framework. Despite the abundance of spatiotemporal data, practical methods for extracting useful insights are limited. Seidenschwarz et al. [5] propose a novel approach by presenting a method for extracting football-specific concepts from interviews, formalizing them within a performance model, and implementing data structures and algorithms in STREAM TEAM, a framework designed for the detection of complex team events. This paper provides a thorough examination of their approach, as well as insights into its potential applications.

For this study, a deep understanding of workload, injuries, and performance is vital to establish or discover correlations between these factors and to build a framework that would ultimately help teams optimize their performance. Workload refers to the physical or mental strain a player endures, which can have numerous implications. Each player is affected differently, depending on their physical attributes and the level of physical or mental strain they experience.

As a result, it is critical to investigate how workload affects a player, whether positively or negatively. The workload is a significant contributor to injuries. Windt and Gabbett's [6] proposed model emphasizes the importance of workloads. Internal risk factors are classified as modifiable or non-modifiable in this model, with workloads influencing injuries in three ways: (1) through exposure to external risk factors and potential inciting events, (2) through inducing fatigue and negative physiological effects, and (3) by promoting fitness and positive physiological adaptations. Exposure is solely determined by total load, whereas positive and negative adaptations are governed by both total workloads and changes in load, such as the acute to chronic workload ratio. This model explains the relationships between load and injuries in detail, encompassing total workloads, acute to chronic workload ratios, and the training load-injury paradox.

Performance analysis in football has gained giant strides over the past few years. With the aid of new technology and equipment, researchers around the globe have established new ways to accurately visualize and explain the performance and contribution of each player to the team. Various types of analysis typically used in football nowadays include:

- **Statistical Metrics**: Analytics involving the collection of statistical data such as goals scored, goals assisted, passes completed, etc., which reflects a player's overall contribution to the team's performance.
- **Position-Specific Analysis**: Analytics estimating the effectiveness of various players across different areas of the playing field, allowing them to assess their strong and weak areas, which can highlight potential improvements.
- **Physical performance**: Data related to player physical attributes, such as sprint speed, distance covered, and high-intensity runs, help gauge a player's fitness level and work rate during matches.
- **Video Analysis**: In addition to statistical data, video analysis is used to evaluate a player's decision-making, movement, positioning, and technical skills during matches.

Football performance evaluation heavily relies on player positions and their dominant areas on the field. By using clustering techniques similar to Key Performance Indicators (KPIs), Cefis and Carpita [7] establish composite indices for various performance areas (e.g., technical, mental, and physical) categorized by player roles. This strategy aims to give coaches and scouts an impartial player performance evaluation and objective tools for decision-making. The analysis helps coaches compare players with similar characteristics and positions by enabling a thorough understanding of player performance in particular roles and pitch areas [1].

Although the rules of football are quite straightforward, there are many factors to consider when assessing the quality of a player: how he/she plays on the ball, off the ball, the reactions, stamina, etc. However, when it comes to comparisons, attackers or forward players typically receive the most attention. This is because what matters the most in football are goals. Scoring more goals than the opponent is basically how a team wins a football match. Hence, players who contribute towards the team's total goals are vital to the team's success. Therefore, analyzing the performance of forwards to determine their goal-scoring ability and predicting their goals in a match/tournament/season can be a pivotal tool for teams when deciding whom to play and which opponent player to be cautious of.

In a recent study, a thorough football analytics approach explores goal prediction and player performance assessment [8]. By using historical football data and sophisticated analytics, this research makes use of a variety of attributes to build a Goal Prediction Model (GPM). The GPM allows for the evaluation of player and team performance while also providing reliable goal predictions. The study also examines and records the unique skills and statistical subgroups that set exceptional goal scorers apart from others. It is important to remember that the model's outcomes can be affected by variables like feature selection, data size, and parameters. The large training dataset contains information from 9074 games over five seasons in the top five leagues in Europe. In the end, this study has the potential to transform football analytics.

Another recent study [9] highlights a notable surge in the creation and application of predictive models in football analytics, which have greatly influenced various aspects of club operations. The Expected Goals (xG) metric, which quantifies the likelihood of a shot resulting in a goal, has gained prominence. However, traditional xG models often overlook essential factors like player and team ability and psychological effects, leading to limited trust in the model within the football community. This study addresses these concerns by leveraging ML techniques to enhance xG modelling. It introduces previously untested features and assesses their impact on predictive performance. Results indicate that the developed xG models perform competitively with optimal values from previous research. Moreover, xG is shown to be a superior predictor of a team's future success compared to traditional statistics, outperforming industry-standard metrics.

The difficulty of integrating predictive models in football is that it is a highly unpredictable game. Mere statistics cannot determine the outcome of a match. A historically stronger team, statistically speaking, cannot be 100% certain of winning the match, as football does not work this way, even if the winning percentage is higher. On the day of the match, anything could happen, and there are a lot of factors to consider when predicting a football match. One notable example of odds being dealt a heavy blow is when Leicester City won the English Premier League in the 2015/2016 season, despite finishing in the bottom three the previous season, which stat-wise does not make sense.

In their research [10], the authors explore ML-based predictive analysis and modeling of football match outcomes in the English Premier League. They use exploratory data analysis and feature engineering to carefully curate a feature set necessary for result prediction. They demonstrate the high accuracy of their ML-based predictive system, with the model using gradient boosting achieving a performance score of 0.2156 on the ranked probability score (RPS) metric. Over two English Premier League seasons (2014–2015 and 2015–2016), this performance evaluation covers game weeks 6 to 38. Despite the encouraging results, their model does not outperform those of betting companies Bet365 and Pinnacle Sports, which had an RPS value of 0.2012 for the same period. Their model, while promising, did not surpass the accuracy of the bookmakers' predictions, as shown by the fact that the lower the RPS value, the higher the predictive accuracy.

Assessing the impact and reasons for each type of injury has been of keen interest in the football research world and is of great importance for this study. Predicting injuries using statistical data is closely observed by many clubs around the world; any model that can accurately predict player injuries can be groundbreaking. However, not all kinds of injuries are predictable. For example, if a player sustains a leg fracture in a match due to a vicious tackle from an opponent, or if he/she gets a concussion or a fractured skull by accidentally hitting their head on the goalpost, these are one-off incidents. They cannot be predicted. However, some injuries, like medial collateral ligament (MCL), hamstring injuries, etc., are sometimes workload injuries, occurring due to the continuous strain on the respective muscle. Such injuries can be predicted as they have a direct relationship with the match statistics of a player, such as minutes played, matches played, and physical attributes such as age, strength, height, etc.

A systematic video analysis conducted by researchers [11] aimed to identify the factors contributing to acute hamstring injuries in professional male football (soccer) players. This analysis included video footage from the top two divisions of German male football from 2014 to 2019. The study focused on moderate to severe hamstring injuries caused by non-contact and indirect contact situations during matches, resulting in a seven-day time loss. The inciting events were classified, and two primary injury patterns emerged: sprint-related injuries (48%), occurring during linear acceleration or high-speed running, and stretch-related injuries (52%), associated with various closed and open chain movements. Despite the variety of inciting events, rapid movements with high eccentric demands on the posterior thigh appeared to be a common factor in hamstring injuries. These findings shed light on the mechanisms of hamstring injuries in professional male football players and highlight the importance of customized, multi-component risk reduction programs.

Though the deep investigation into workload, injury, and performance has laid a foundation for this study, the main crux lies in assessing the correlation between these aspects to derive insights and construct models. This can be impactful for several reasons:

- **Injury Prevention**: Understanding the relationship between workload and injury is crucial for developing effective injury prevention strategies. By identifying workload thresholds and patterns associated with higher injury risks, clubs and medical staff can implement targeted measures to reduce the likelihood of injuries.
- **Performance Optimization**: The correlation between workload and performance is a critical aspect of player management. Balancing the right level of workload can positively impact player performance, ensuring optimal physical and technical abilities on the pitch.
- **Player Management**: A deep understanding of the workload-injury-performance relationship allows for better player management.

There are many more advantages and use cases for football teams. Coaches have a model or framework that establishes a relationship between workload, injury, and performance and provides useful insights to make decisions.

It would be insightful to understand the extent to which players can handle workload, beyond which they become increasingly prone to injury, especially muscle injury. A lot of the top football teams worldwide play at least one match every week and three matches per week during some stages of the season. This takes a toll on the players and might lead to injuries and cramps.

Howle et al. [12] conducted an analysis that provides substantial insights into the issues faced by congested schedules and their implications for player well-being and performance in football. The study's findings are notable because they illustrate the critical impact of fixture congestion on football injury rates. It was discovered that there is a significant rise in injury incidence under crowded schedules, demonstrating a strong link between increased workloads and the risk of injury. The legitimacy of the study is enhanced by its robust methodology, which includes objective workload measures and injury tracking, making it a significant resource for the football community.

Comprehensive studies, such as [13], highlight the crucial role of analytics in elucidating the impacts of age, position, and injuries on NBA player performance and economic outcomes. Such research lays a robust groundwork for investigating comparable patterns in football, especially within the English Premier League (EPL), where the financial implications and pressures on player performance are similarly high.

The rapid growth of Artificial Intelligence (AI) has reaped benefits in almost every sector in the world over the past few years. The major advantage of AI is its ability to solve problems using algorithms quickly and with maximum accuracy. This technology is also being applied in the world of football. Machine learning (ML) and deep learning models, developed using complex algorithms, now have the capability to predict the outcomes of football matches, as observed in several studies reviewed above. It is essential to analyze the origins, growth, and current applications of complex AI algorithms and models in football. See Table 1 for the summary.

Table 1. All studies cited in this literature review along with author name, date, category of the topic, and its research aim.

Study Title	Author(s)	Category	Aim	Year
How do training and competition workloads relate to injury?	Windt and Gabbett [6]	Workload, Injuries and Performance in Football	Present a new framework for managing football injuries	2017
Injury Incidence and Workloads during congested Schedules in Football	Howle et al. [12]	Workload, Injuries and Performance in Football	Examine the relationship between injury incidence and workloads in congested football schedules	2019
Predictive analysis and modelling football results using machine learning approach for English Premier League	Baboota & Kaur [10]	Performance Analysis in Football	Une machine learning for predictive analysis of football match outcomes	2019
Football Analytics: Performance analysis differentiate by role	Cefis and Carpita [7]	Performance Analysis in Football	Develop composite indices for performance assessment in football	2020
Hamstring injury patterns in professional male football	Gronwald et al. [11]	Workload, Injuries, and Performance in Football	Identify factors contributing to acute hamstring injuries in professional male football players	2021
Expected goals in football: Improving model performance and demonstrating value	Mead et al. [9]	Performance Analysis in Football	Improve Expected Goals (xG) modeling for assessing team success in football	2023
Football Analytics for Goal Prediction to Assess Player Performance	Javed et al. [8]	Performance Analysis in Football	Greate a Goal Prediction Model and assess player performance in football	2023

3. Methodology

3.1. Objective

An extensive data-driven methodology was used in this research project to analyze the complex relationships in professional football. The primary objectives were to assess correlations, forecast player performance, and identify injuries. To achieve this, a sizable dataset comprising various player-related indicators was assembled. Through comprehensive data preprocessing and exploratory data analysis, machine learning (ML) models played a crucial role in providing predictive insights. Regression models were used to forecast performance measures like goals and assists, utilizing various player-related variables as predictive features. Conversely, classification models were used to predict injuries, contributing to a comprehensive understanding of player success on the field.

3.2. Dataset

The dataset consists of information on 532 players across 20 football clubs in the 2020–2021 season of the English Premier League. It includes 11 variables, all used for analysis in this study, categorized as follow:

- **Personal Information**: Player Name, Club Name, Age, Position etc.
- **Individual Workload Features**: Minutes Played, Matches Played, Matches Started, and 90 Minutes Played.
- **Individual Performance Features**: Goals Scored, Assists, Goals Plus Assists etc.
- **Individual Injury Occurrences**: Injured (Injured or Not Injured), Injury Reason, Injury Occurrences etc.

The data on Personal Information, Individual Workload Features, and Individual Performance Features were obtained from WhoScored.com, a well-reputed football analysis website that constantly updates its database regarding players and clubs across the top leagues in Europe. The data on Individual Injury Occurrences were obtained from an article published by Sky Sports, a popular television and broadcasting network based in the UK.

3.3. Methods and Structure

The methodological approach and structure followed in this research are represented in Figure 1 below.

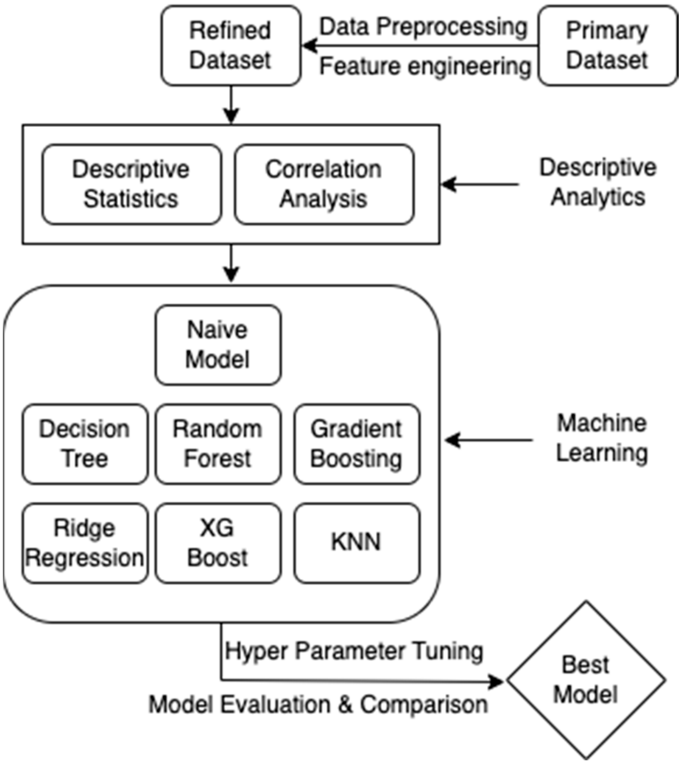


Figure 1. Methodology Approach and Structure.

3.4. Data Preprocessing

By resolving errors and deleting unnecessary variables, data cleaning is an essential step in maintaining dataset dependability. The original dataset for this study contained information on all participants and contained 34 variables. A criterion was defined to include only pertinent data, excluding records that had minimal impact on the study, and irrelevant variables were deleted to maintain analytical integrity.

(a) **Irrelevant variables**: The numerical data collected from the sources include several variables irrelevant to the analysis, such as expected goals (the number of goals a player is likely to score in the current season by analyzing his previous season's performance), nationality, and the number of yellow and red cards acquired by player throughout the season. These variables need to be removed before conducting the analysis. The process of feature selection in this study was guided by a combination of domain expertise and statistical analysis to ensure that only the most relevant features were included in our models. Initially, features were chosen based on their known impact on player performance and injury risk from existing sports science literature. This foundational selection was further refined through exploratory data analysis, where correlation matrices and preliminary regression models helped identify features with significant predictive power and minimal collinearity.

(b) **Minutes played**: The data include players who played as little as one minute of a match in the season. This will result in the data analysis being irrelevant and misleading. To address this, we set a parameter to include only those data that qualify for data analysis. We set the parameter to a minimum of 1140 min (setting an average of 30 min for each match played, so 38 matches imply a total of 38 multiplied by 30, which is 1140 min). This minimum threshold was chosen because it represents a substantial portion of the match time to contribute meaningfully to team dynamics while allowing for the inclusion of players who may not start every match but are regular contributors. This method not only reduces the risk of analyzing data that might not truly reflect a player's impact but also helps to eliminate incomplete records from the dataset.

$$\text{Minutes played} = \text{Total number of matches in the season} \times \text{Minimum standard set} = 38 \times 30 = 1140\ \text{min}$$

(c) **Player position**: Due to the different roles that midfielders and forwards play on the field, the research will focus on these two groups of players when predicting performance. The central planners of attacking moves are the midfielders (MF) and forwards (FW) players, who create opportunities for goals through assists and their own goals. They are the perfect subjects for this research because their performance measures are, by nature, focused on attack and creativity. Unlike goalkeepers or defenders, whose contributions are assessed differently, analyzing MF and FW players' offensive prowess offers deeper insights into the fundamental elements of football performance that directly impact a team's ability to win. This tactical decision enables us to focus on the players whose actions impact goal-scoring and offensive playmaking most.

3.4.1. Feature Engineering

In the context of this study, feature engineering is an essential technique for gaining greater insight. The dataset includes a wide range of variables, but it is crucial to enhance these features through clever combinations and adjustments. These engineered variables are meticulously constructed to reveal complex correlations by utilizing domain knowledge relevant to football. This method greatly enhances the depth of analysis by enabling the discovery of hidden patterns and relationships.

(a) **Player Workload:**
The 'player workload' variable classifies football players as 'Rare Starter', 'Average Starter', and 'Frequent Starter', depending on how frequently they start matches. This variable provides insights into workload differences and their consequences on injuries and performance. These cut-off points were derived from a review of historical data and existing literature on player health and performance, providing a structured approach to monitor and manage player workload effectively.
Number of matches started:

- Less than 10 ⇒ Rare starter
- More than 10 but less than or equal to 25 ⇒ Average starter
- More than 25 ⇒ Frequent starter

(b) **Player Usage:**
Based on the number of minutes played, the 'Player Usage' variable classifies football players as 'Squad Rotation Players', 'Sporadic Players', or 'Crucial Players'. This emphasizes the importance of specific players to their teams and provides information about their usage's impact on performance, fitness, and injury risk.
Number of minutes played:

- Less than 1140 min ⇒ Sporadic Player
- Between 1140 and 2280 (inclusively) minutes ⇒ Squad Rotation Player
- More than 2280 min ⇒ Crucial Player

(c) **Age Category:**
Based on age, the 'Age Category' variable classifies football players as 'Youngster', 'Prime', or 'Veteran'. This sheds light on the impact of age on career duration, injury risk, and performance. It aids players in understanding how age and experience affect their roles and output.
Age falling between:

- 16 and 23⇒ Youngster
- 24 and 31 ⇒ Prime
- 32 and above ⇒ Veteran

(d) **Average Minutes per Match**
By determining a player's average minutes played per match, the 'Average Minutes per Match' variable offers a fair assessment of player playing time. This helps in determining the consistency of player participation in games, which may affect injury risk and performance. It serves as a starting point when examining how player workload, injury, and performance are related.

$$\text{Average Min/Match} = (\text{Minutes Played})/(\text{Matches Played})$$

3.4.2. Dummy Variables

Using OneHotEncoder, categorical variables such as player workload, player usage, age category, etc., were converted into numeric dummy variables. Dummy variables are binary (0 or 1) indicators that represent categorical data in numerical form. Each category within a variable is transformed into a separate column, with a value of 1 indicating the presence of that category and 0 indicating its absence. This conversion is necessary because many of the ML techniques used in this study require numerical inputs to produce predictions for the performance metrics and injuries of football players.

3.4.3. Train-Test Strategy

The dataset was randomly split into two sets: training and testing sets. The training set was used to train and optimize the machine learning models, while the testing set was reserved for evaluating their performance. The common practice of an 80–20 ratio was followed, where the larger portion was assigned to the training set to ensure adequate data for model learning [14].

3.4.4. Sampling Methods

The dataset shows an imbalance in terms of injury prediction, with just about 30% of players reporting injuries. If a random train-test split is performed, this can result in an unbalanced test set. As a result, the test set may contain significantly fewer injured players than the entire dataset, which could impair the ML model's capacity to forecast injuries accurately.

Stratified sampling is used to overcome this problem. This method ensures that the ratio of injured to healthy players in the training and test sets is the same as it was in the original dataset. This strategy ensures that both classes are fairly represented in each subset, resulting in a more useful dataset for training and evaluating models.

3.5. Descriptive Statistics

Prior to implementing ML models, descriptive statistics were utilized as an initial step. This method was used to thoroughly understand the dataset, including mean, median, standard deviation, and percentiles. These statistics facilitated the evaluation of central tendencies, dispersions, and data distributions. Descriptive statistics revealed valuable information about the dataset's characteristics, which aided in data preprocessing and model selection for the subsequent ML analysis.

3.6. Correlation Matrix

Correlation analysis was conducted as an initial data exploration step. This analysis aimed to reveal relationships and dependencies between different variables within the dataset. Key statistical measures, such as correlation coefficients, were used to assess the strength and direction of associations between pairs of variables. This preliminary correlation analysis was critical in identifying potential predictor variables for subsequent ML models, aligning with the study's goal of investigating the impact of various factors on player performance and injuries.

3.7. Machine Learning (ML)

ML is an essential tool in this study for understanding the underlying dynamics within football player data. The research deciphers the predictive potential of various factors by utilizing diverse algorithms and methodologies, ultimately predicting player performance in terms of goals and assists, as well as the likelihood of injuries. Recognizing the distinct nature of the predictive targets, regression models are used for goals and assists, leveraging their numerical nature. In contrast, classification models are invaluable for injury prediction and handling binary outcomes effectively. This tailored approach enables the study to delve into the dataset's nuanced patterns, shedding light on the multifaceted determinants of player success and well-being on the football field.

Since numerous factors in sports data are connected, a phenomenon known as multicollinearity, Ridge Regression was chosen for analysis. Ridge Regression is perfect for evaluating the relative relevance of various aspects of player performance because of its L2 regularization, which helps prevent overfitting while preserving the interpretability of linear models.

It was selected due to the Extreme Gradient Boosting's remarkable predictive capacity and aptitude for managing intricate, non-linear interactions.

3.7.1. Naive Model

A baseline or naive model was designed to offer a benchmark for measuring the performance of ML algorithms. In essence, a naive model provides a point of comparison to assess whether more sophisticated models are truly effective in making accurate predictions. For the research topic involving numerical 'Goals and Assists', the naive model computes the target variable's mean. However, a baseline model was not constructed for injury prediction due to the significant class imbalance. With the majority of players being injury-free (67%), a naive model would predict all players as non-injured, yielding high accuracy by chance rather than genuine predictive power. In such cases, the naive model would not accurately reflect the model's ability to identify injured players, as the prediction of 'non-injured' for all instances would not capture the complexity of the data. Consequently, given the disproportionate data distribution, establishing a naive model would not provide meaningful insights into injury prediction.

3.7.2. Decision Tree

A decision tree, originating in ML theory, is an efficient tool for solving classification and regression problems [15]. Unlike other classification approaches that use a set of features (or bands) jointly to perform classification in a single decision step, the decision tree is based on a multistage or hierarchical decision scheme or a tree-like structure.

These trees can divide data into subsets based on feature values, enabling them to predict continuous and categorical results. Their ability to capture complex feature-outcome relationships is made possible by their hierarchical decision structure, making them an adaptable option. Furthermore, decision trees offer interpretability, which helps understand the variables influencing predictions and is useful in the context of player performance and injury prediction.

3.7.3. Random Forests

In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed [16]. Each node is split in standard trees using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy performs very well compared to many other classifiers, including discriminant analysis, support vector machines, and neural network, and is robust against overfitting. In regression, the Random Forest algorithm combines the predictions of multiple decision trees to provide a robust estimate. It does so by averaging the outputs of individual trees. For classification, Random Forests use a majority voting mechanism. Each tree in the forest predicts the class, and the class with the most votes is the final prediction.

3.7.4. K-Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) method is a simple but effective method for classification [17]. KNN is used because it can spot certain patterns within collections of data. KNN offers localized insights for the study topics regarding injury differences across positions and clubs by categorizing positions or clubs with comparable injury patterns. This improves the analysis by highlighting groups of players and teams that frequently suffer injuries. KNN enhances the more general information gained from previous models by providing a more focused understanding of how certain positions and teams experience injury patterns.

KNN calculated distances between data points and identified the k-nearest neighbors for each data point in the dataset. The majority class among these neighbors was assigned as the predicted category for a given player, determining whether they were 'injured' or 'not injured'.

3.7.5. Gradient Boosting

Due to its proficiency in handling regression issues, Gradient Boosting, a potent ML technique, was chosen for predicting goals and assists. Gradient Boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current 'pseudo'-residuals by least squares at each iteration [18]. The pseudo-residuals are the gradient of the loss function being minimized with respect to the model values at each training data point evaluated at the current step. It sequentially creates a group of decision trees, each correcting the flaws in the previous one. Gradient Boosting minimizes a loss function mathematically by adding weak learners (in this case, decision trees) in a weighted manner. Gradient Boosting is a great option for goal and assistance prediction when the result is a continuous numerical value because it combines multiple decision trees to capture complex relationships within the data. It excels at data fitting, bias reduction, and prediction accuracy enhancement.

3.7.6. Ridge Regression

Ridge Regression is a regularization technique used in ML to handle regression problems, particularly when dealing with multi-collinearity (highly correlated predictors) and to prevent overfitting. It adds a regularization term to the equation for linear regression. Ridge Regression minimizes a loss function that includes the sum of squared regression residuals [19]. Ridge Regression lessens the effect of multi-collinearity and prevents coefficients from becoming too extreme, which helps combat the overfitting issue. It is especially helpful when dealing with datasets where predictors are highly correlated, as is frequently the case when predicting player performance.

3.7.7. XGBoost

Extreme Gradient Boosting, also known as XGBoost, is a potent ML algorithm used to address classification issues like injuries. Contrary to conventional gradient boosting, this ensemble method is designed to maximize the functionality and computational efficiency of these decision trees. It combines the predictions from multiple decision trees and has good scalability [20].

XGBoost builds an ensemble of decision trees sequentially, with each new tree attempting to fix the flaws of the ones that came before it. To make accurate predictions about a player's injury status, XGBoost can learn from the dataset's features and injury labels in the context of injury prediction. Its adaptability gives it a competitive edge because, when necessary, it can handle both binary classification (injured or not injured) and multi-class classification.

3.8. Hyperparameter Tuning

Grid Search CV was used for hyperparameter tuning to enhance the performance and accuracy of ML models. This method systematically explored various hyperparameter combinations to identify the optimal settings for improving model efficacy [21]. By exhaustively searching through the specified parameter ranges, it ensured the models were well-equipped for robust and insightful predictions in workload, injuries, and football performance indicators.

3.9. Evaluation Metric

The selection of evaluation metrics, namely Root Mean Squared Error (RMSE) for the second research question and accuracy for the third and fourth research questions, is based on the specific nature of the target variables and research aims.

- In the case of numerical 'Goals and Assists', RMSE is an ideal choice since it estimates the average size of forecast errors. It gives a comprehensive assessment of how well the ML model's predictions match the actual numerical outcomes. Because RMSE prioritizes avoiding both overestimation and underestimation, it is appropriate for evaluating the prediction accuracy of models attempting to estimate continuous variables such as goals and assists.
- Accuracy is a suitable choice for categorical 'Injured' and 'Not Injured' outcomes since it represents the proportion of accurately predicated cases. This metric is critical when it comes to appropriately classifying the occurrence or non-occurrence of an event. Given the significance of accurately detecting injuries, accuracy clearly indicates the model's ability to categorize these occurrences.

4. Results

4.1. Correlation Analysis

In this subsection, the statistical examination focused on understanding the relationships between player workload indicators and their performance outcomes, primarily goals and assists. This analysis directly addresses the first research question concerning the correlation between player workload and performance metrics.

The correlation analysis of English Premier League football players reveals strong positive correlations between workload metrics ('Min', 'MP', 'Starts', and '90s') and 'Gls_Ast' (Goals + Assists), indicating that players who spend more time on the field during games typically have higher goal-scoring and assisting performance. In particular, MP (total number of matches a player has participated in) has the strongest positive association with 'Gls_Ast', highlighting the significance of being included in the playing 11 and getting a chance to play every match irrespective of the number of minutes played. 'Age', on the other hand, shows relatively modest positive correlations, showing that it plays a minimal role in explaining variances in goals and assists. Match load measures are the main factors influencing player performance. See Figure 2.

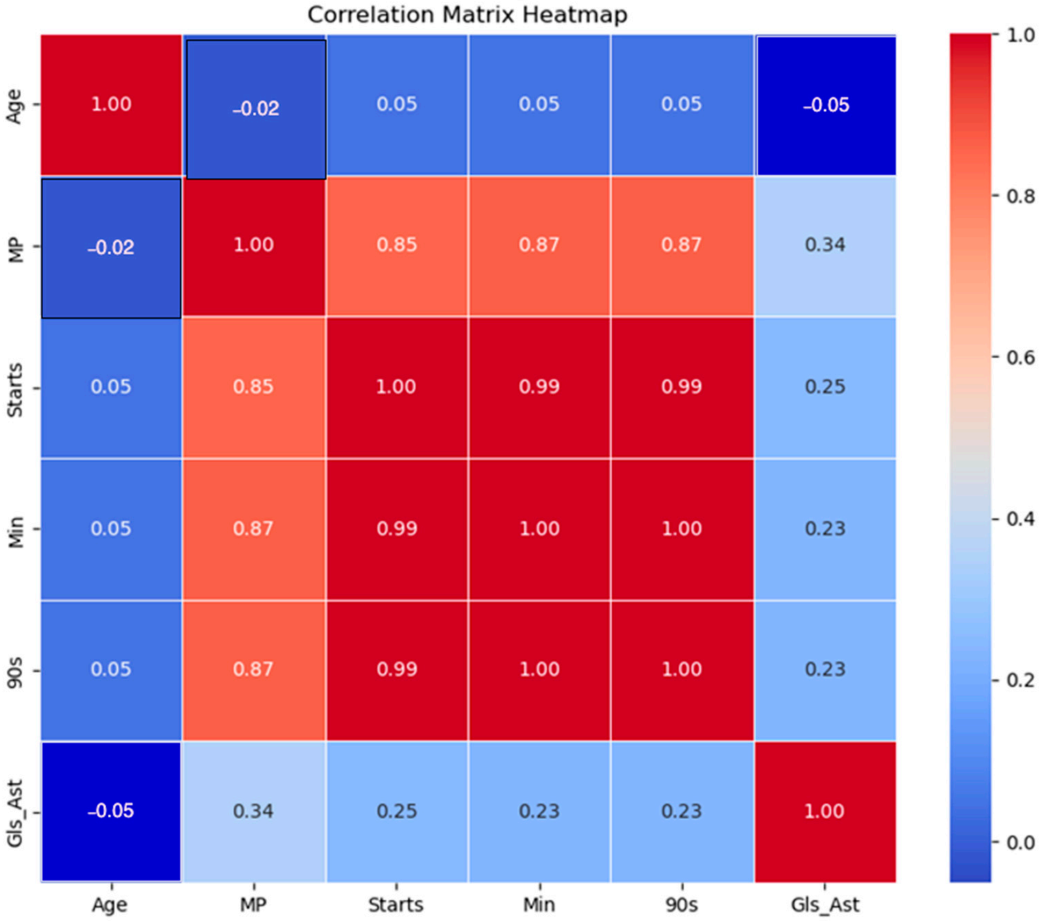


Figure 2. Correlation matrices. **Note**. Age: Player's age; MP: Matches Played; Starts: Number of matches started; Min: Total minutes played; 90s: Number of full 90-min matches played; Gls_Ast: Combined total of goals and assists.

4.2. Predicting Goals and Assists

The subsection on predicting goals and assists is dedicated to evaluating the effectiveness of various machine learning models in forecasting player performance metrics related to goals and assists. This aligns with the second research question, which aims to assess the predictive capacity of different variables affecting player performance outcomes.

4.2.1. Model Performance Assessment

Residual Analysis

Residual analysis involves examining the differences between the observed values and those predicted by the model. It helps identify any systematic deviations from the expected predictions, which can indicate model inadequacies. The residual analysis for the Decision Tree, Random Forest, Gradient Boosting, and Ridge Regression models reveals a consistent pattern across all models (Figures 3–6).

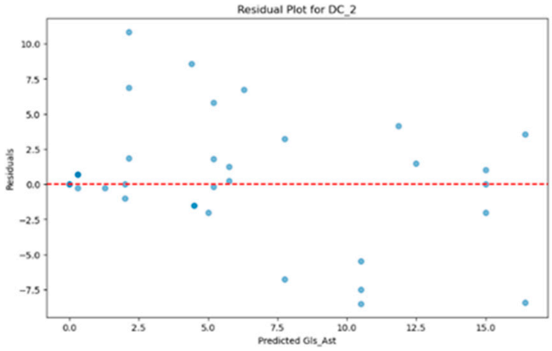


Figure 3. Residual plot: Decision Tree.

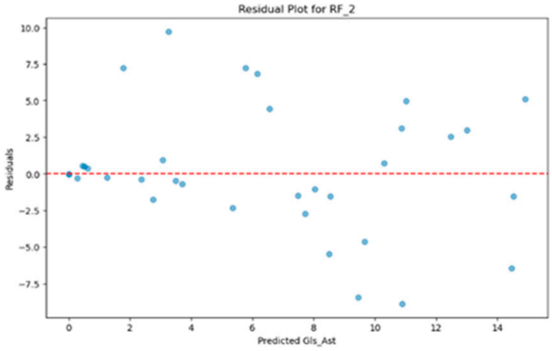


Figure 4. Residual plot: Random Forest.

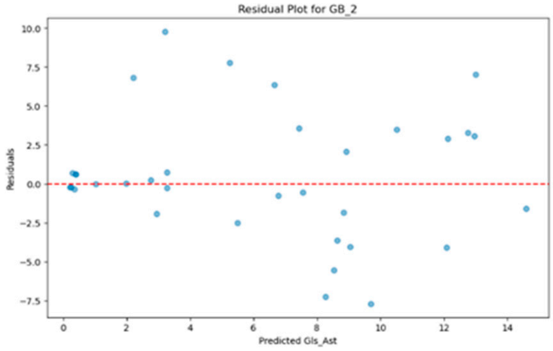


Figure 5. Residual plot: Gradient Boosting.

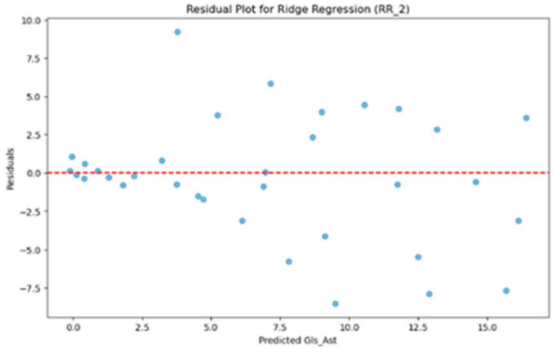


Figure 6. Residual plot: Ridge Regression.

Training and Testing Assessment

For lower predicted values, each model demonstrates a high level of accuracy, indicated by residuals closely scattered around the zero line. Specifically, this accurate prediction range is observed up to a predicted value of 5 for the Decision Tree model and Ridge Regression model and 4 for the Random Forest model and Gradient Boosting model. Within these ranges, the residuals indicate that the model effectively captures the underlying trends in the data, suggesting that the predictions are accurate. However, as predicted values increase, outliers and deviations from the zero line become more prevalent, and the variance of residuals fluctuates. Despite the absence of discernible patterns, these observations suggest that model performance may be affected when making predictions beyond certain thresholds. Their learning curves are shown in Figures 7–10.

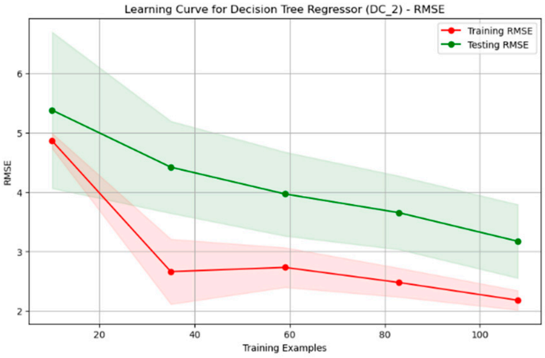


Figure 7. Learning curve: Decision Tree.

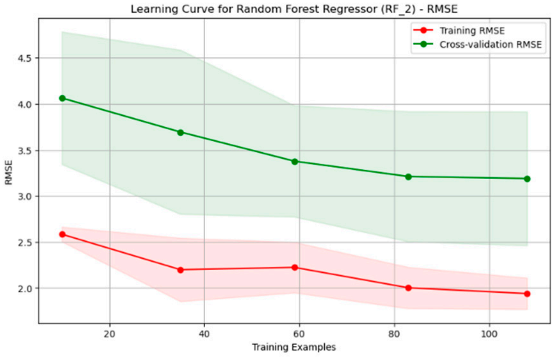


Figure 8. Learning curve: Random Forest.

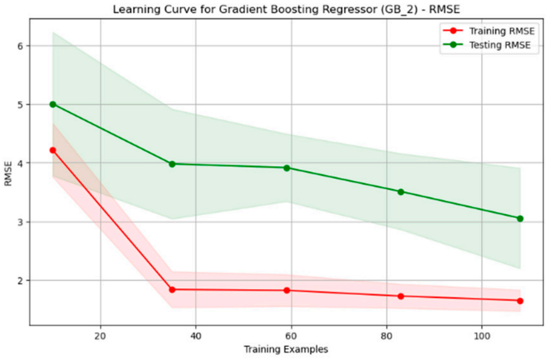


Figure 9. Learning curve: Gradient Boosting.

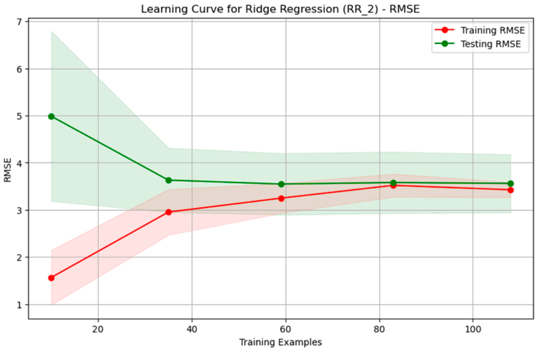


Figure 10. Learning curve: Ridge Regression.

- **Decision Tree versus Random Forests**: The Random Forest Regressor demonstrates superior initial performance, evidenced by its lower RMSE in both the training and testing sets, suggesting that it is a more effective model straight away when compared to the Decision Tree Regressor. Its smaller discrepancy between training and cross-validation RMSEs signals a reduced tendency toward overfitting, an advantage over the Decision Tree. For both models, adding more data improves the model's performance, but the rate of improvement slows down, which is typical as a model starts to reach its performance limit with the given features and model complexity. Notably, the Random Forest Regressor shows less variability in its testing RMSE, which is illustrated by a narrower confidence interval, indicating a more consistent performance regardless of the training set it encountered.
- **Gradient Boosting versus Ridge Regression**: The Gradient Boosting model has a considerable and continuous gap between training and testing RMSE. The training RMSE drops significantly at first, demonstrating the model's ability to fit the training data effectively. However, the testing RMSE improves more slowly, constantly remaining higher than the training RMSE. This difference indicates overfitting, as the model struggles to generalize to new data. The Ridge Regression learning curve, on the other hand, begins with a significant gap between training and testing RMSE. The gap narrows dramatically as training progresses, and the two RMSE curves converge. The training RMSE gradually rises while the testing RMSE falls, indicating better generalization. Ridge Regression finds a better balance between fitting the training data and generalizing to new data based on this behavior. While the RMSE values are not the lowest among the models, the model's ability to resist overfitting is a significant benefit.

In conclusion, the learning curves emphasize the trade-offs between model complexity and generalization. Random Forest and Gradient Boosting models provide a more balanced approach, whereas Decision Tree and Ridge Regression models struggle with overfitting. Fine-tuning and optimization efforts may be required to improve the generalization capabilities of the latter models.

4.2.2. Model Performance Comparison Based on RMSE

Referring to Table 2, the modelling process begins with a baseline comparison using the Naive Model to establish a starting point. Despite its simplicity, the Naive Model yields a relatively high RMSE of 6.62, highlighting its limitations and underscoring the need for more sophisticated modelling techniques. Subsequently, we explore various models, including Decision Tree (DC_1 and DC_2), Random Forest (RF_1 and RF_2), Gradient Boosting (GB_1 and GB_2), and Ridge Regression (RR_1 and RR_2).

Table 2. Comparison of default and fine-tuned ML models for predicting Goals and Assists.

Model	Hyper Parameter	RMSE
Naive Model		6.62
Decision Tree (DC_1)	Default	4.77
Decision Tree (DC_2)	Max depth: None Min samples leaf: 4 Min samples split: 2	4.41
Random Forest (RF_1)	Default	4.33
Random Forest (RF_2)	Max depth: None Min samples leaf: 2 Min samples split: 5 N estimators: 50	4.24
Gradient Boosting (GB_1)	Default	4.13
Gradient Boosting (GB_2)	Learning rate: 0.1 Max depth: 3 Min samples leaf: 4 Min samples split: 10 N estimators: 50	4.05
Ridge Regression (RR_1)	Default	3.91
Ridge Regression (RR_2)	**alpha: 10**	**3.90**

DC_1 represents the default configuration, while DC_2 incorporates fine-tuned hyperparameters. DC_2 outperforms its default counterpart with lower a RMSE. Transitioning to Random Forest models, RF_1, with default settings surpasses the predictive accuracy of Decision Tree models. The introduction of fine-tuned hyperparameters in RF_2 further enhances predictive accuracy, resulting in a lower RMSE. Similarly, Gradient Boosting models (GB_1 and GB_2) demonstrate their predictive power, with GB_1 showcasing a lower RMSE than previous models. Fine-tuning further improves performance in GB_2, leading to increased accuracy. Finally, two Ridge Regression models, RR_1 (default configuration) and RR_2 (fine-tuned with an alpha value of 10), are presented. RR_2 emerges as the best-performing model with the lowest RMSE among all models. This could be due to the fact that Ridge Regression models perform well for datasets where predictors are highly correlated with each other [21]. In this case, it has already been observed that predictor variables like minutes played, matches played, etc., are highly correlated with each other.

In conclusion, due to its low RMSE, the fine-tuned Ridge Regression model RR_2 stands out as the best-performing model, outperforming Decision Trees, Random Forests, Gradient Boosting, and the Naive Model. The selection of RR_2 as the favored model demonstrates its ability to produce accurate predictions about player performance regarding goals and assists.

4.2.3. Feature Importance

In Table 3, '90 Minutes Played (90 s)' is identified as the most important feature in both the Decision Tree model and Gradient Boosting models, with a feature importance of 0.874 and 0.552, respectively. According to these two models, the total time a player spends on the field during a match is the key element impacting the prediction of goals and assists. The Decision Tree model also highlights the contribution of the aspect 'Starts', but its impact is less pronounced. However, the Random Forest model considers it the most significant, with a feature importance of 0.412, followed by 'Minutes Played (Min)' at 0.397. These features align with the consensus that a player's playing time and match involvement are critical factors influencing performance predictions.

Table 3. Feature importance scores of predictors across models for Goals and Assists.

Feature Importance	Decision Tree	Random Forest	Gradient Boosting
Minutes Played (Min)	0.024	0.397	0.305
Matches Played (MP)	0.022	0.047	0.054
Starts	0.073	**0.412**	0.058
90 Minutes Played (90s)	**0.874**	0.114	**0.552**
Age	0.006	0.030	0.032

While different models assign different levels of importance to features, the analysis reveals a consensus across models regarding the central role of '90 min Played (90 s)' and the significance of playing time-related features in predicting goals and assists.

4.3. Predicting Injuries

The subsection on predicting injuries addresses the third research question by evaluating the effectiveness of machine learning models in forecasting injury occurrences among English Premier League football players. The analysis employs various ML models with a focus on evaluating their performance through accuracy, precision, recall, and AUC scores.

4.3.1. Model Evaluation

We compared several ML models to predict the target variable in the model comparison in Table 4. The fine-tuned Decision Tree model (DTC_2) stood out with an accuracy of 0.72, reflecting the significant impact of hyperparameter optimization. Similarly, the adjusted Random Forest model (RFC_2) saw an increase in accuracy to 0.69 after fine-tuning its parameters. XGBoost (XGB_2) performed admirably from the outset, maintaining a high accuracy of 0.72 following optimization, affirming its suitability for the predictive task at hand. The K-Nearest Neighbors model (KNN_2) also benefited from hyperparameter adjustments, achieving an improved accuracy of 0.63.

Table 4. Comparison of fine-tuned ML models for predicting injury.

Model	Accuracy	Precision	Recall	AUC Score
Decision Tree (DTC_2)	**0.72**	0.58	0.35	0.66
Random Forest (RFC_2)	0.69	0.50	0.50	0.70
XGBoost (XGB_2)	**0.72**	0.55	0.60	0.65
K-Nearest Neighbors (KNN_2)	0.63	0.405	0.50	0.54

However, accuracy alone is not the sole indicator of a model's effectiveness, especially when predicting injuries where the cost of false negatives can be significant. Therefore, it is crucial to shift the attention to precision and recall ratings, especially for cases involving injuries. Precision measures the rate at which models correctly predict injuries, whereas recall assesses their ability to identify all actual injuries, thus reducing missed cases. In predicting player injuries, the Random Forest model (RFC_2) achieves a precision of 0.5 and a recall of 0.5, indicating that it correctly identifies half of the injured players. The Decision Tree Model (DTC_2) has a higher precision of 0.58 but a lower recall of 0.35, indicating that it misses a large number of injured players. XGBoost (XGB_2) comes out on top with a balanced precision of 0.55 and the highest recall of 0.6, identifying 60% of the actual injuries, making it the most suitable model for this application.

Moreover, the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) values provided additional insights into the models' classification abilities. Among the models, the ROC curve for Random Forest (RFC_2) shows the most promise. With an AUC value of 0.70, it clearly outperforms random guessing in its ability to discern between injured and healthy players. The Decision Tree model (DTC_2) has an AUC of 0.66, which is better than random chance but less powerful than RFC_2. In contrast, the K-Nearest Neighbors model (KNN_2) has an AUC of 0.54, which is only slightly better than random chance and suggests the model struggles with class separation in this context. XGBoost (XGB_2) has an AUC of 0.65, placing it between the Random Forest and Decision Tree models in terms of performance, indicating a moderate ability to differentiate between the classes.

With an accuracy of 0.63 and an AUC score of 0.54, the K-Nearest Neighbors (KNN_2) model appears to marginally outperform random guessing. The recalculated values show that the model achieves a precision of around 0.405 and a recall of 0.5, implying that although the model has a high probability of false positives, it may detect 50% of actual injury cases. The below-average precision of the model, which yields many false positive predictions, clearly illustrates its incapacity to anticipate injuries. Similarly, the recall value shows that half of the real injuries can be identified by the model. Based on this performance, the KNN_2 model needs regular improvement to anticipate football players' injuries, even though it is easy to use and effective.

In conclusion, XGBoost (XGB_2) is the most successful model for predicting player injuries, achieving the highest accuracy and a critical balance between precision and recall. Its high recall rate is crucial for reducing the risk of overlooking actual injury cases. The AUC scores further reinforce the models' standings, with Random Forest (RFC_2) leading in terms of ROC performance, followed by Decision Tree (DTC_2) and XGBoost (XGB_2), which both offer reasonable performance in injury classification.

A thorough summary of the performance measures for each model examined in this study is given in Table 5. We report the Root Mean Square Error (RMSE) for goals and assist prediction; lower numbers denote better performance. We present accuracy, precision, recall, and Area Under the Curve (AUC) ratings for injury prediction. To give a thorough overview, we have included numerous measures for injury prediction and RMSE for goals/assists prediction. Certain findings are excluded for particular reasons:

- Injury measurements are not relevant for the Naive Model because it uses goals/assists prediction as a baseline exclusively.
- Since Gradient Boosting and Ridge Regression are suitable for managing continuous outcomes, they were only applied to objectives and assistance prediction. For these models, injury measurements are therefore irrelevant.
- As XGBoost and K-Nearest Neighbors are appropriate at classification tasks, they were especially used for injury prediction. Consequently, RMSE values for objectives and aids are not relevant.
- Decision-Tree and Random Forest can be comprehensive models, but they may sometimes underperform in any of metrics.

Table 5. Summary of Model Performance Metrics.

Model (Baseline)	**RMSE (Goals/Assists)**	**Accuracy (Injury)**	**Precision (Injury)**	**Recall (Injury)**	**AUC (Injury)**
Naive Model	6.62	NA	NA	NA	NA
Gradient Boosting	4.05	NA	NA	NA	NA
Ridge Regression	3.90	NA	NA	NA	NA
Model (Injury investigations)	RMSE (Goals/Assists)	Accuracy (Injury)	Precision (Injury)	Recall (Injury)	AUC (Injury)
XGBoost	NA	0.72	0.55	0.60	0.65
K-Nearest Neighbors	NA	0.63	0.405	0.50	0.54
Model (Comprehensive)	RMSE (Goals/Assists)	Accuracy (Injury)	Precision (Injury)	Recall (Injury)	AUC (Injury)
Decision Tree	4.41	0.72	0.58	0.35	0.66
Random Forest	4.24	0.69	0.50	0.50	0.70

We can immediately compare model performance inside each prediction task by presenting the data in Table 5. By using suitable methods for analysis, we can determine which models work best for a given analytical problem in football performance and injury risk assessment.

4.3.2. Feature Importance

In all three models in Table 6, 'Squad', which denotes a player's team allegiance/club at which he is playing, consistently stands out as the most significant predictor. This widespread agreement highlights the crucial role that a player's team plays in injury prediction, indicating that team dynamics and environmental factors significantly impact a player's vulnerability to injuries.

Table 6. Feature importance scores of predictors across models for injury.

Feature	Random Forest (RFC_2)	Decision Tree (DTC_2)	XGBoost (XGB_2)
Position	0.06	0.11	0.03
Squad	**0.16**	**0.26**	**0.27**
Age	0.11	0.12	0.20
Matches Played	0.12	0.08	0.12
Starts	0.10	0.04	0.11
Minutes Played	0.12	0.04	0.08
Average Minutes/Match	0.12	0.25	0.11
90s	0.16	0.09	0.07
Player Workload	0.00	0.00	0.00
Player Usage	0.01	0.00	0.01
Age Category	0.03	0.00	0.01

Second, 'Age' plays a key role in the models, especially in XGBoost (XGB_2), which assigns it the highest relevance score. This demonstrates the significant influence a player's age has on their risk of injury, with XGBoost giving special attention to this aspect. Age is undoubtedly an essential factor in injury prediction, highlighting the potential advantages of developing injury-preventive tactics according to a player's age range.

Additionally, the models identify the number of 'Starts' and 'Matches Played' as fairly relevant variables, with RFC_2 and XGB_2 giving relatively greater importance to these attributes. This suggests that when determining injury likelihood, it is important to consider a player's match experience and inclusion in starting lineups. In contrast, 'Player Workload' is consistently assigned a low relevance rating by all models, indicating that it might not substantially impact injury prediction. This emphasizes how complicated injury determinants are, with certain aspects being less significant in this situation.

Finally, 'Squad' appears as the key predictor across all models, highlighting the significance of this factor in injury prediction. Age also has considerable influence, and XGBoost is particularly aware of this. In addition to providing a greater understanding of the factors driving injury prediction, these insights establish the groundwork for injury prevention and management measures catered explicitly to the various player profiles.

5. Discussion and Findings

5.1. Predictors for Player Performance (Goals and Assists)

5.1.1. Matches Played

- **Strong Positive Correlations with Workload Metrics**: It became clear that player performance metrics, particularly goals and assists, exhibit strong positive correlations with workload metrics, including elements like minutes played, matches played, and 90s (minutes played divided by 90). This result is consistent with common sense because more time spent on the field naturally gives players more chances to assist their teammates and score goals.
- **Matches Played (MP) with emphasis**: Within the subgroup of match load measures, matches played (MP) stood out as especially significant. Compared to other measures like minutes played (Min) and 90s (minutes played per match), which had correlations of 0.23, MP had a higher correlation coefficient of 0.34. This suggests that a player's ability to score goals and provide assists is most significantly influenced by the sheer number of games in which they take part, regardless of how much time they spend on the field or whether they are a starter.

The data for 2021/22 and 2022/23 were collected from the official Premier League website. This is to further validate whether the top three performers could maintain their high performance in the succeeding seasons and to check the significance of matches played (MP) as a contributing factor towards goals and assists.

Referring to Table 7, as forwards, Harry Kane and Mohamed Salah averaged around 30 goals and assists, with their forward positions allowing numerous scoring opportunities. Bruno Fernandes, a midfielder, averaged 20 goals and assists, reflecting his playmaking role. It is worth noting that Bruno Fernandes had an outstanding season in 2020/21, with 30 goals and assists. This outstanding performance highlights his essential role in the team, his influence on team performance, and his ability to positively affect matches. In conclusion, the data emphasize the importance of these top performers' regular field presence (high MP) and their contributions relative to their positions and tactical roles.

Table 7. Comparison of top three performers.

Player	2020/21		2021/22		2022/23	
	Matches Played	Gls_Ast	Matches Played	Gls_Ast	Matches Played	Gbs_Ast
Harry Kane	35	37	37	26	38	33
Mohammed Salah	37	27	35	36	38	31
Bruno Fernandes	37	30	36	16	37	16

Clubs can use these findings as a guide for evidence-based player management. While using the "matches played" knowledge, rotation methods can be improved. For example, the "85% rule" [22] could be implemented, requiring important attackers to participate in that percentage of games to strike a balance between exhaustion and sharpness. The association between age and performance supports adjusting training loads; veterans could benefit from fewer, lower-intensity sessions in between games. Real-time substitution decisions might be informed by injury prediction models, and players who pose a danger could be closely observed during games. Ultimately, these resources allow coaches to make data-driven decisions that protect players' well-being while pursuing awards.

5.1.2. Squad

Table 8 shows the top 20 goal scorers from each team; observations are as follows.

- **The prominence of Top Teams**: Players from the season's top four teams (Manchester City, Manchester United, Liverpool, and Leicester City) are marked in bold. This distinction is critical because it emphasizes that being a part of a high-performing team has a major impact on a player's ability to accumulate goals and assists. In this case, these four teams account for half of the top 20 performers (10 out of 20, as emphasized in bold in Table 8), demonstrating their dominance in player performance metrics.

- **Variety of Positions**: The table includes players from numerous positions, with forwards contributing the most goals and assists. Midfielders like Kevin De Bruyne, Bruno Fernandez, and Jack Harrison are also prominently featured in the top 20, highlighting their versatile responsibilities in both scoring and creating goals.
- **Individual Brilliance**: The list features some of the league's most prolific goal scorers and playmakers, including Harry Kane and Bruno Fernandes, at the top. These players are recognized for their extraordinary abilities and consistent impact on matches, routinely scoring goals and making assists.
- **Balance and Competitiveness**: The presence of players such as Patrick Bamford and Ollie Watkins, who play for clubs other than the conventional top tier, demonstrates that the Premier League retains a competitive and diverse player environment. This type of balance brings interest to the league by allowing developing talent to flourish.

Table 8. Squad Comparison of Top 20 Goal Scorers from Each Team.

Player	Goals + Assists	Team	Position
Harry Kane	37	Tottenham	Forward
Bruno Fernandes	30	Manchester United	Midfielder
Son Heung-min	27	Tottenham	Forward
Mohamed Salah	27	Liverpool	Forward
Jamie Vardy	24	Leicester City	Forward
Patrick Bamford	24	Leeds United	Forward
Marcus Rashford	20	Manchester United	Forward
Ollie Watkins	19	Aston Villa	Forward
Kevin De Bruyne	18	Manchester City	Midfielder
Sadio Mané	18	Liverpool	Forward
Matheus Pereira	17	West Brom	Forward
Raheem Sterling	17	Manchester City	Forward
Callum Wilson	17	Newcastle Utd	Forward
Jack Grealish	16	Manchester City	Forward
Roberto Firmino	16	Liverpool	Forward
Jack Harrison	16	Leeds United	Midfielder
Danny Ings	16	Burnley	Forward
Dominic Calvert-Lewin	16	Everton	Forward
Riyad Mahrez	15	Manchester City	Forward
Chris Wood	15	Burnley	Forward

Figure 11 unequivocally demonstrates that, compared to the other top clubs in the 2020–2021 Premier League season, Manchester City had significantly more goals and assists, which is one of the reasons for their success in the league. The collective performance of top teams like Manchester City, Manchester United, Liverpool, and Chelsea contributed to their players' success in scoring and assisting. However, this relationship is bidirectional. Specifically, individual performances were equally crucial in enhancing a team's league position. The consistent output from key players often translated into a higher finish in the league standings.

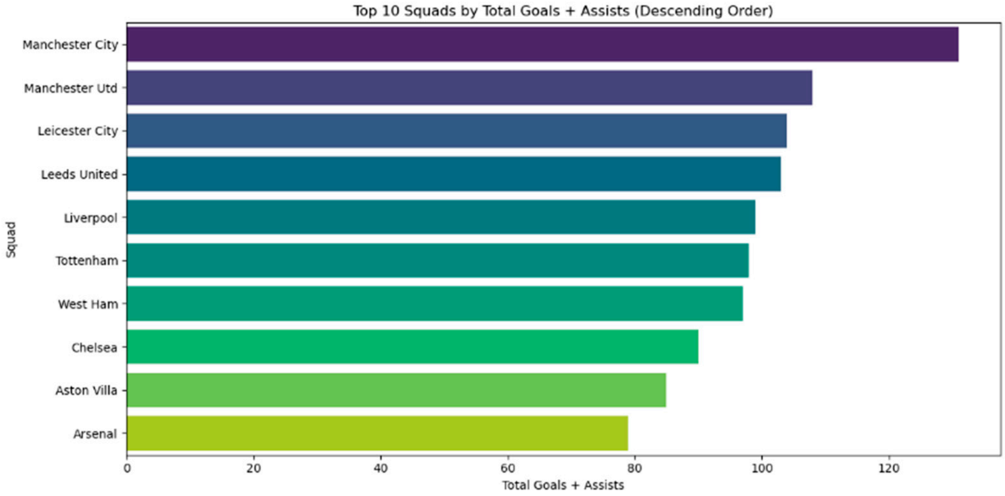


Figure 11. Goals and Assists by Teams (Top 10).

This comprehensive viewpoint emphasizes the dynamic and symbiotic relationship between individual player performance and team success. In essence, the squad's collective effort in continuously producing goal-scoring opportunities directly influenced the individual player's capacity to shine.

5.1.3. Age Category

In the field of football performance analysis, age is frequently seen as a relevant factor that may influence player performance, as shown in Figure 12. This research thoroughly examined player attributes and their relationship to performance indicators. Initial findings from correlation analysis and feature importance scores obtained by ML models revealed that age had a relatively minor impact on player performance, with correlation scores of −0.05 indicating a weak negative correlation, and feature importance scores ranging from −0.05 to 0.032, indicating that it is a weak predictor when it comes to predicting performance.

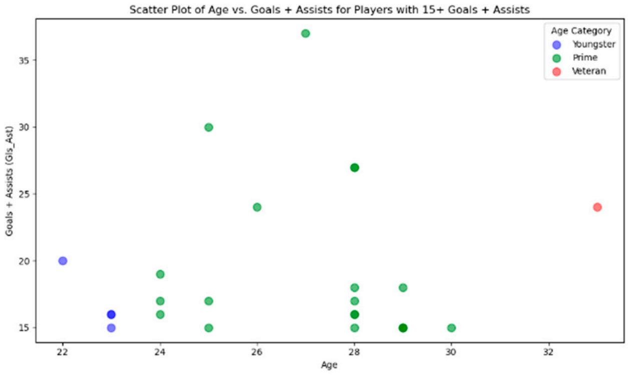


Figure 12. Age vs. Goals and Assists (Top 20).

However, a closer look at the top performers (Figure 12) revealed a dominance of players within the 'Prime' age group (23 to 31 years), suggesting that the prime years are likely the most productive due to a mix of experience, tactical understanding, and physical fitness. The 'Youngster' group, aged 16 to 22, was less represented but notably impactful, showcasing their potential. 'Veteran' players over 32 were scarcely found among the top ranks, indicating that while experience is beneficial, it might not be crucial for peak performance.

The study also observed that age, while not a strong standalone predictor, is indeed influential during a player's prime years, as shown in Table 9. For instance, elite players like Harry Kane (27), Mohammed Salah (28), and Bruno Fernandes (26) not only played almost every match but also ranked highly in performance, aligning with the observation that prime-aged players in top teams tend to exhibit higher levels of performance.

Table 9. Age Comparison; Top 3 Players.

Player	Squad	Age
Harry Kane	Tottenham	27
Mohammed Salah	Liverpool	28
Bruno Fernandes	Manchester United	26

In conclusion, while individual variables such as 'Matches Played', 'Age Category', and 'Squad/Team' may not stand out as dominating predictors of goals and assists, they highlight a discernible trend. Exceptional players who consistently offer top-tier statistics tend to fall into specific age groups and are frequently associated with high-performing clubs. This insight emphasizes the intricate interplay of these variables, underlining their overall impact on a player's output in terms of goals and assists.

5.2. Predicting Injuries

The ML algorithms deployed to predict injuries showed considerable success, with the best-performing model achieving a commendable recall score of 0.60. This means that the algorithm correctly identified 60% of the injuries among injured players, demonstrating its usefulness in predicting such occurrences. Notably, the analysis revealed that 'Squad' emerged as the most crucial feature in predicting injury, with an average feature importance of 0.23 across models and an even more pronounced significance of 0.29 in the best model, a fine-tuned XGBoost model. 'Average Minutes per Match', with an average feature importance score of 0.16, and 'Age', with an average importance rating of 0.14, were close behind. These findings shed light on the critical importance of squad dynamics and player age in comprehending and forecasting injuries in professional football.

5.2.1. Squad

Based on injury occurrences, the bar graph (Figure 13) depicts the injury distribution among the top ten teams in terms of the number of injury occurrences. Burnley has the most injuries on the list, totaling 12. Newcastle United is a close second with ten injuries. Manchester City, Manchester United, Crystal Palace, and Everton are tied for third position with eight injuries each. Southampton and Leicester City follow with seven injuries each, while Sheffield United is last with six. Arsenal has the fewest injuries among these top ten teams, with only five.

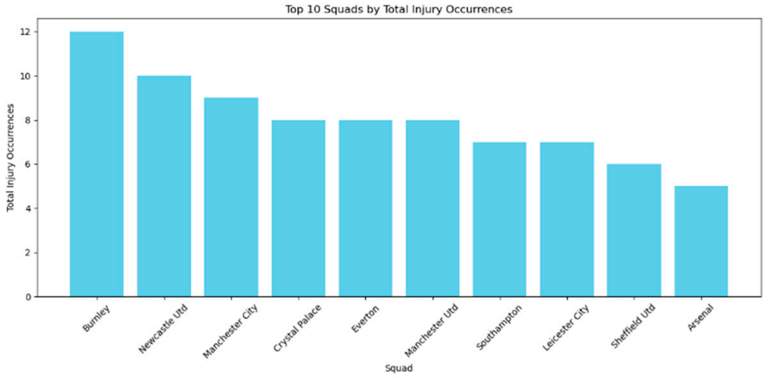


Figure 13. Top 10 Squads by Injury Occurrences.

Figure 14 clearly distinguishes between the squad classifications and their average injuries. The 'Most Injury Affected' category, which includes the top five clubs with the most injuries, stands out with an average of nine injuries per team. In other words, on average, these heavily afflicted teams face a significant injury load, approaching double digits. In sharp contrast, the 'Rest' group, which includes the remaining teams not in the top five, reports an average of only four injuries per squad. This considerable gap highlights the importance of squad categorization in injury occurrences. It supports the study's conclusion that squad composition is vital in forecasting and managing injuries in football teams.

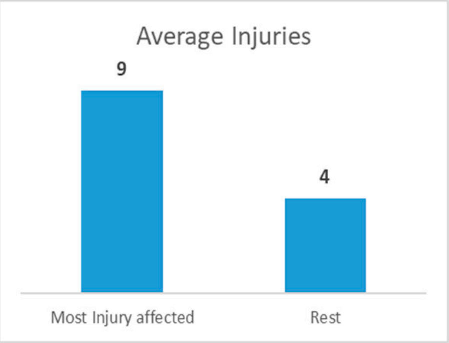


Figure 14. Average Injuries; Most affected Squads vs. Rest.

The pie chart in Figure 15 provides a revealing perspective on injury occurrences, categorizing them as 'Most Injury Affected' and 'Rest'. Notably, despite accounting for only a quarter (25%) of all teams, the 'Most Injury Affected' category accounts for 42% of all injury occurrences. This notable disparity highlights the considerable impact of squad makeup on injury rates. A concentrated set of teams, albeit a minority, clearly shoulders a disproportionate injury burden. This image significantly confirms the study's fundamental finding: squad composition plays a critical role in forecasting and affecting injury occurrences, with a few teams causing a substantial amount of the league's injuries.

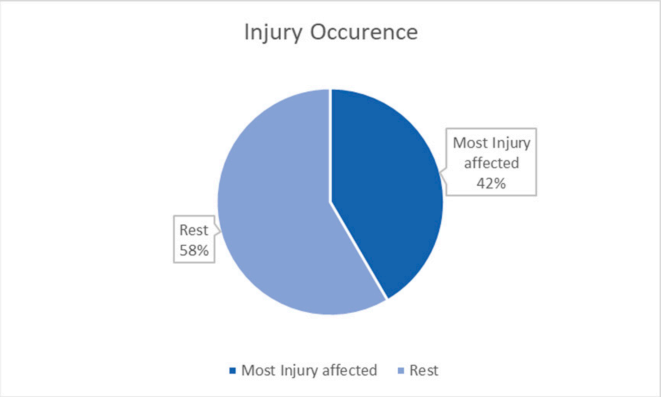


Figure 15. Injury Occurrence %; Most affected Squads vs. Rest.

5.2.2. Average Minutes per Match

Figure 16 divides players into three groups based on their 'Average Min/Match': 'Low', 'Medium', and 'High'. 'Low' refers to players who play less than 60 min per match, 'Medium' refers to players who play between 60 and 75 min per match, and 'High' refers to players who play more than 75 min each match.

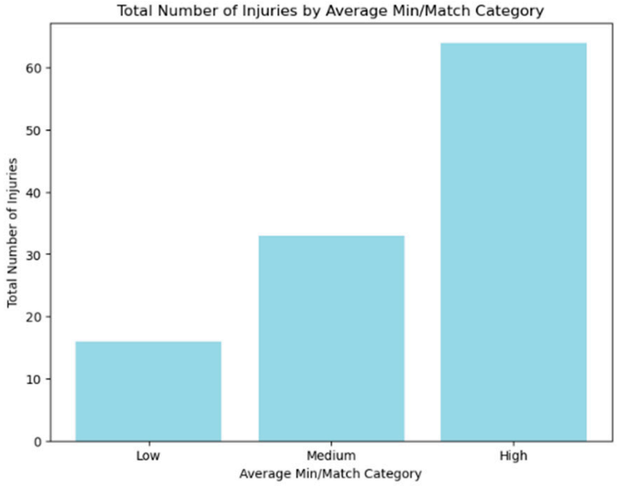


Figure 16. Total Injuries by Average Minute per Match Category.

According to Figure 16, a substantial proportion of injuries, roughly 57%, occur in the 'High' group, showing that players who routinely play more than 75 min per match are more likely to get injured. In contrast, the 'Medium' category, which includes players who average 60 to 75 min every match, accounts for approximately 30% of all injuries. Only 16 injuries occur in the 'Low' category, which includes players who play fewer than 60 min per match. This concept emphasizes the need for calculated player substitutions during games. Coaches and team managers should think carefully about when and how to replace players who have been on the field for extended periods of time. Timely substitutions allow players to rest and recover, lowering the chance of injury due to exhaustion or overexertion.

Furthermore, it underlines the significance of effectively controlling a player's playing time during the season. This includes substitutions during games and preparing for proper rest intervals between games and throughout training. Teams can limit the likelihood of injuries and preserve the health and performance levels of their athletes by maximizing playing time and giving adequate recovery. In essence, these findings reaffirm that 'Average Min/Match' is a reliable predictor of injuries due to the close relationship between playing time and injury occurrence.

5.2.3. Age

Figure 17 provides useful information on injury rates among different age groups of players. Notably, the 'Veteran' category, which includes athletes aged 23 to 31, has an incredibly high injury rate of nearly 60%. This means that six out of every ten players in this group were injured during the season, underscoring the importance of age as a reliable predictor of injuries. Despite its low feature importance score of 0.14, this research highlights the importance of age in injury prediction. It emphasizes the importance of managing veteran players carefully, balancing their game time with strategies to minimize overexertion and injury.

In contrast, the 'Prime' and 'Youngster' groups have injury rates that are roughly 30% and 25%, respectively. This implies that age may not be the most critical factor impacting injuries in these age groups. Instead, other variables or events may contribute to the occurrence of injuries.

In conclusion, while age is not a strong predictor of injuries for all players, it is a critical component for the 'Veteran' category, where approximately 60% of players have been injured. This finding supports the study's conclusions about the role of age in injury prediction and emphasizes the importance of cautious management of senior players to reduce injury risks.

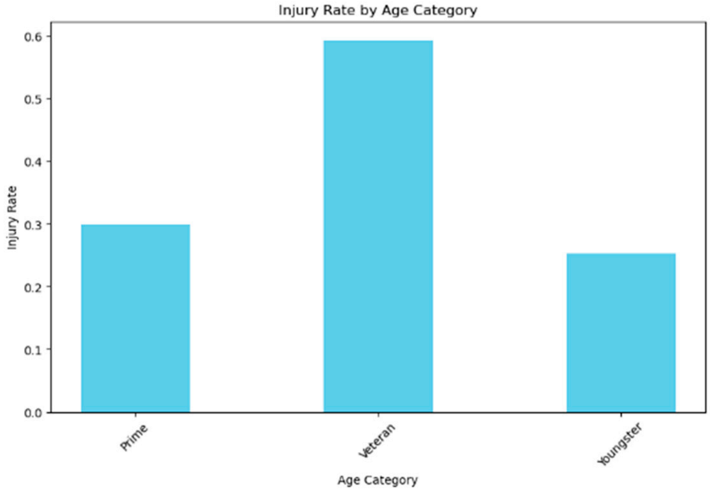


Figure 17. Injury Rate by Age Category.

This research identifies three key indicators of injuries in football teams: squad composition, average minutes per match, and age category. These are important predictors of injuries in football clubs. These findings highlight their significance and are consistent with previous research, underlining their vital roles in injury prediction and management. Effective squad management, playing time allocation, and age category consideration are all critical techniques for injury prevention and player well-being.

6. Conclusions

6.1. Key Findings

This study's most significant finding is the impact of matches played (MP) on player performance. MP, with the highest correlation value of 0.24, stands out among other workload indicators like minutes played, starts, and full games. This emphasizes the central role of MP in influencing player performance. The study also reveals that for players with exceptional offensive skills, the key to success lies not just in playing more minutes but in participating in as many games as possible. This insight suggests a strategic approach for managers and coaches: systematically rotating top players. By doing so, they can ensure these players' participation in successive games, balancing the need for rest and the goal of consistent, high-level performance while reducing injury risks. Furthermore, the research points to the importance of the squad's quality in player performance. Remarkably, half of the top 20 players belonged to the top 4 teams in the league, illustrating a clear link between player excellence and team strength. This finding highlights the necessity for aspiring athletes to join top teams and for such teams to focus on recruiting consistent, talented players.

In terms of injury prediction, squad composition emerges as a significant factor, with a feature importance score of 0.29. The study found that the 'Most Injury Affected' clubs, among the top 5 in injury incidents, accounted for a substantial portion of all injuries. This highlights the importance of squad balance and fatigue management in preventing injuries. Age also appears as a critical factor, showing a modest yet significant correlation with overall player performance and a stronger predictive power for injuries. Particularly, players in their prime (23–31 years) contribute to goals and assists. Conversely, players over 32 are more prone to injuries, suggesting the need to manage older players carefully to minimize their injury risks. Lastly, average minutes per match is identified as a key predictor of injuries. Players who play longer periods per game (>75 min on average) are more likely to sustain injuries. This finding is crucial for team management, highlighting the need for a balanced approach to player usage, emphasizing rotation and rest, especially for older players, to reduce injury risks.

6.2. Contribution to the Field of Sports Science and Analytical Research

This project significantly advances sports science and analytics by offering an in-depth analysis of factors influencing football players' performance and well-being. It achieves this through several key contributions:

The research provides a holistic view by integrating and scrutinizing a variety of football-related variables such as player workload, fatigue, performance indicators, and personal attributes. This comprehensive approach yields a deeper understanding of the complex interactions among these factors. It directly impacts decision-making in sports, offering valuable insights for athletes, coaches, and teams. The study guides informed, data-driven strategies in team management, training, and injury prevention by elucidating the relationships between workload, player characteristics, and performance outcomes.

This research adds depth to injury management by exploring variations in injury occurrences across different clubs and age groups. The insights gained enable the formulation of targeted injury prevention and management strategies, potentially reducing player downtime and enhancing overall team performance. For instance, the study's findings about the heightened injury risk in older players ('Veteran' category) suggest that teams should tailor their fatigue management approaches for this specific age group.

Moreover, this study paves the way for future research in football analytics, especially in injury prediction and prevention. It encourages further exploration into multidimensional factors affecting player injuries, both on and off the pitch, such as medical support, squad rotation strategies, and mental fatigue.

6.3. Impact on Football Clubs

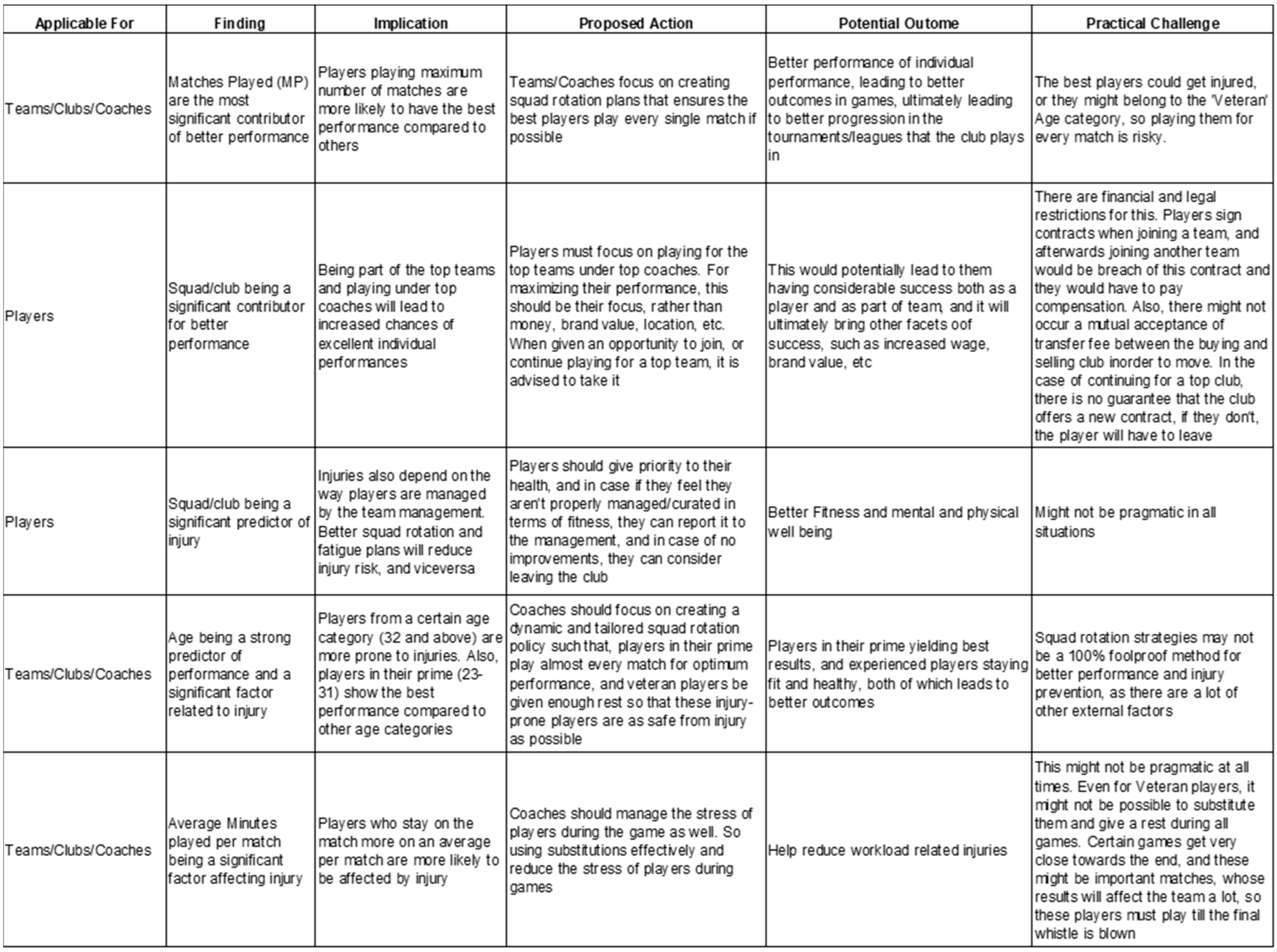
Football clubs operate in an extremely competitive environment, where even minor advantages can greatly impact performance, tournament advancement, and overall success. The results of this study have significant ramifications for football teams, trainers, and players alike.

Football clubs can leverage data-driven management strategies to optimize player rotation, ensuring that players are neither overused nor underutilized. This balanced approach helps maintain player fitness throughout the season, enhancing performance during critical matches. Furthermore, clubs can implement customized training regimens informed by predictive insights on injury risks, reducing the likelihood of injuries and extending player careers.

The insights into performance metrics across different player roles enable clubs to make informed recruitment decisions, identifying players who best fit their tactical needs. Understanding which attributes correlate strongly with success in specific positions allows for more strategic player acquisitions. Additionally, by analyzing trends and performance outcomes of younger players, clubs can tailor development programs to nurture potential talent effectively, focusing on skills that contribute significantly to match outcomes.

Clubs can also make informed match day decisions with a deeper understanding of how various factors such as player workload and position influence game outcomes. Managers can optimize team performance through strategic player selection and substitutions during matches. Regular application of the analytical methods used in this study can assist in the continuous assessment and improvement of team strategies and player contributions. By integrating these insights into existing strategies, football clubs can not only enhance their competitive edge but also foster a healthier and more sustainable environment for their players. The recommendations provided are designed to offer a data-driven foundation for enhancing decision-making processes in football club management.

These implications and potential actions are summarized in Figure 18 below.



Applicable For	Finding	Implication	Proposed Action	Potential Outome	Practical Challenge
Teams/Clubs/Coaches	Matches Played (MP) are the most significant contributor of better performance	Players playing maximum number of matches are more likely to have the best performance compared to others	Teams/Coaches focus on creating squad rotation plans that ensures the best players play every single match if possible	Better performance of individual performance, leading to better outcomes in games, ultimately leading to better progression in the tournaments/leagues that the club plays in	The best players could get injured, or they might belong to the 'Veteran' Age category, so playing them for every match is risky.
Players	Squad/club being a significant contributor for better performance	Being part of the top teams and playing under top coaches will lead to increased chances of excellent individual performances	Players must focus on playing for the top teams under top coaches. For maximizing their performance, this should be their focus, rather than money, brand value, location, etc. When given an opportunity to join, or continue playing for a top team, it is advised to take it	This would potentially lead to them having considerable success both as a player and as part of team, and it will ultimately bring other facets oof success, such as increased wage, brand value, etc	There are financial and legal restrictions for this. Players sign contracts when joining a team, and afterwards joining another team would be breach of this contract and they would have to pay compensation. Also, there might not occur a mutual acceptance of transfer fee between the buying and selling club inorder to move. In the case of continuing for a top club, there is no guarantee that the club offers a new contract, if they don't, the player will have to leave
Players	Squad/club being a significant predictor of injury	Injuries also depend on the way players are managed by the team management. Better squad rotation and fatigue plans will reduce injury risk, and viceversa	Players should give priority to their health, and in case if they feel they aren't properly managed/curated in terms of fitness, they can report it to the management, and in case of no improvements, they can consider leaving the club	Better Fitness and mental and physical well being	Might not be pragmatic in all situations
Teams/Clubs/Coaches	Age being a strong predictor of performance and a significant factor related to injury	Players from a certain age category (32 and above) are more prone to injuries. Also, players in their prime (23-31) show the best performance compared to other age categories	Coaches should focus on creating a dynamic and tailored squad rotation policy such that, players in their prime play almost every match for optimum performance, and veteran players be given enough rest so that these injury-prone players are as safe from injury as possible	Players in their prime yielding best results, and experienced players staying fit and healthy, both of which leads to better outcomes	Squad rotation strategies may not be a 100% foolproof method for better performance and injury prevention, as there are a lot of other external factors
Teams/Clubs/Coaches	Average Minutes played per match being a significant factor affecting injury	Players who stay on the match more on an average per match are more likely to be affected by injury	Coaches should manage the stress of players during the game as well. So using substitutions effectively and reduce the stress of players during games	Help reduce workload related injuries	This might not be pragmatic at all times. Even for Veteran players, it might not be possible to substitute them and give a rest during all games. Certain games get very close towards the end, and these might be important matches, whose results will affect the team a lot, so these players must play till the final whistle is blown

Figure 18. Significant ramifications for football teams, trainers, and players.

6.4. Applicability to Other Domains

Beyond football, the methodologies and findings of this study have wider applicability and offer valuable contributions to various fields. In sports, the principles of data-driven decision-making highlighted here can be adapted to other team sports like basketball and hockey, where similar analytics can improve strategies and results.

In healthcare and injury prevention, the study's focus on predictive models and fatigue management can be applied to reduce injury risks in diverse sports and physically demanding professions. These methods can help in designing injury prevention strategies not only for athletes but also for individuals in occupations that entail physical exertion.

Additionally, the study's insights into workload management can inform strategies to optimize workforce performance in the corporate sector. By applying these principles, organizations can enhance productivity while mitigating burnout risks, paralleling the balance between training and recovery necessary for athletes.

6.5. Limitations

The study, while comprehensive, faces several limitations that must be acknowledged. The primary constraint is the dataset's limited scope, encompassing only player characteristics, performance indicators, and injury data from a single season. This limitation arises from the dynamic nature of football, where player transfers, team promotions, and relegations significantly alter team dynamics and player performance across seasons. The 2020–2021 EPL season was impacted by the COVID-19 pandemic, affecting various aspects including a delayed start and restricted fan attendance during the period for which the data was collected. Consequently, expanding the dataset to include multiple seasons or data from different football leagues was not feasible due to the variability in league characteristics, such as match frequency, playing styles, and other contextual factors influencing performance and injury rates.

Another significant limitation is the inherent unpredictability of injuries in football. Despite the study's ML models forecasting injuries with notable accuracy, the spontaneous nature of injuries in football, often resulting from unforeseen on-field events, presents a challenge. This unpredictability means that, regardless of the precision of predictive models, there will always be inherent limitations in forecasting injuries in a sport as unpredictable and complex as football.

Additionally, the analysis was constrained by the availability of certain variables. While diverse in player-related characteristics and performance indicators, the dataset lacked more intricate elements related to player fitness and injuries. These variables, often held confidential by league officials and team management, such as detailed medical records and specific injury-related data, were not accessible for this study, thereby limiting the breadth of the analysis.

Furthermore, it is essential to recognize potential biases introduced by the study's design and assumptions inherent in the modeling process. The representativeness of the sample is primarily constrained to a single, pandemic-affected season, which may not adequately reflect normal competitive conditions. This could influence the generalizability of the findings across different seasons with varying conditions and intensity. Additionally, assumptions made during the modeling, such as the linear relationship presumed between certain player attributes and performance outcomes, may not capture more complex interactions or nonlinear dynamics present in the actual data. Acknowledging these biases and assumptions is crucial for interpreting the study's findings accurately and understanding the scope of their applicability.

6.6. Recommendations for Future Work

Building on this study's findings, future research should delve deeper into the realms of youth and academy player analysis and the investigation of external variables influencing player performance. Analyzing data on younger players would identify potential early indicators of success and injury risks and guide the development of bespoke strategies aimed at nurturing talent while safeguarding against injuries. This approach promises a more proactive and nuanced method of player development and management within clubs.

Additionally, thoroughly examining how external factors such as weather, travel, and match schedules impact players could significantly enhance our understanding of performance and injury dynamics. Such research might lead to the creation of specialized training regimes and match strategies tailored to mitigate the adverse effects of these variables. Understanding the nuances of how these factors influence player fatigue and recovery could revolutionize the way teams manage their players, optimizing performance while prioritizing health and well-being.

Author Contributions: Conceptualization, V.C. and S.S.; methodology, V.C. and S.S.; software, S.S.; validation, V.C. and S.S.; formal analysis, V.C., S.S., Q.A.X. and H.W.; investigation, V.C. and S.S.; resources, V.C.; data curation, S.S.; writing—original draft preparation, V.C. and S.S.; writing—review and editing, V.C., Q.A.X., M.T. and H.W.; visualization, S.S.; supervision, V.C.; project administration, V.C.; funding acquisition, V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partly supported by VC Research (VCR 0000223) for Chang.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available upon the Chief Data Officer's (the second author) review.

Conflicts of Interest: There are no conflicts of interest.

References

1. Clemente, F.M.; Martins, F.M.L.; Mendes, R.S.; Figueiredo, A.J. A systemic overview of football game: The principles behind the game. *J. Hum. Sport Exerc.* **2015**, *9*, 656–667. [CrossRef]
2. Asif, R.; Zaheer, M.T.; Haque, S.I.; Hasan, M.A. Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research. *Int. J. Comput. Sci. Inf. Secur.* **2016**, *14*, 516.
3. Chazan-Pantzalis, V.; Tjortjis, C. Sports Analytics for Football League Table and Player Performance Prediction. In Proceedings of the 2020 11th International Conference on Information, Intelligence, Systems and Applications, Piraeus, Greece, 15–17 July 2020; pp. 1–8.
4. Rodrigues, F.; Pinto, Â. Prediction of football match results with Machine Learning. *Procedia Comput. Sci.* **2022**, *204*, 463–470. [CrossRef]
5. Seidenschwarz, P.; Rumo, M.; Probst, L.; Schuldt, H. A Flexible Approach to Football Analytics: Assessment, Modeling and Implementation. In *Proceedings of the 12th International Symposium on Computer Science in Sport (IACSS 2019)*; Springer International Publishing: Cham, Switzerland, 2020.
6. Windt, J.; Gabbett, T.J. How do training and competition workloads relate to injury? The workload-injury aetiology model. *Br. J. Sports Med.* **2017**, *51*, 428–435. [CrossRef] [PubMed]
7. Cefis, M.; Carpita, M. Football Analytics: Performance analysis differentiate by role. In *Third International Conference on Data Science & Social Research Book of Abstracts*; CIRPAS and University of Bari Aldo Moro: Bari, Italy, 2020; p. 22.
8. Javed, D.; Jhanjhi, N.Z.; Khan, N.A. Football Analytics for Goal Prediction to Assess Player Performance. In *Proceedings of Innovation and Technology in Sports*; Springer Nature: Singapore, 2023; pp. 245–257.
9. Mead, J.; O'Hare, A.; McMenemy, P. Expected goals in football: Improving model performance and demonstrating value. *PLoS ONE* **2023**, *18*, e0282295. [CrossRef]
10. Baboota, R.; Kaur, H. Predictive analysis and modelling football results using machine learning approach for English Premier League. *Int. J. Forecast.* **2019**, *35*, 741–755. [CrossRef]
11. Gronwald, T.; Klein, C.; Hoenig, T.; Pietzonka, M.; Bloch, H.; Edouard, P.; Hollander, K. Hamstring injury patterns in professional male football (soccer): A systematic video analysis of 52 cases. *Br. J. Sports Med.* **2021**, *56*, 165–171. [CrossRef] [PubMed]
12. Howle, K.; Waterson, A.; Duffield, R. Injury Incidence and Workloads during congested Schedules in Football. *Int. J. Sports Med.* **2019**, *41*, 75–81. [CrossRef] [PubMed]
13. Sarlis, V.; Tjortjis, C. Sports Analytics: Data Mining to Uncover NBA Player Position, Age, and Injury Impact on Performance and Economics. *Information* **2024**, *15*, 242. [CrossRef]
14. Alayón, S.; Hernández, J.; Fumero, F.J.; Sigut, J.F.; Díaz-Alemán, T. Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. *Appl. Sci.* **2023**, *13*, 12722. [CrossRef]
15. Xu, M.; Watanachaturaporn, P.; Varshney, P.K.; Arora, M.K. Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.* **2005**, *97*, 322–336. [CrossRef]
16. Liaw, A.; Wiener, M. Classification and Regression by Randomforest. *R News* **2002**, *2*, 18–22.
17. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. *KNN Model-Based Approach in Classification*; Springer: Cham, Switzerland, 2003; pp. 986–996.
18. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
19. de Vlaming, R.; Groenen, P.J. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. *Biomed. Res. Int.* **2015**, *2015*, 143712. [CrossRef] [PubMed]
20. Abdurrahman, G.; Sintawati, M. Implementation of xgboost for classification of parkinson's disease. *J. Phys. Conf. Ser.* **2020**, *1538*, 012024. [CrossRef]
21. Belete, D.M.; Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [CrossRef]
22. McKeown, G. To Build a Top Performing Team, Ask for 85% Effort. Available online: https://hbr.org/2023/06/to-build-a-top-performing-team-ask-for-85-effort (accessed on 14 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.