MODELLING PLAYER SIMILARITY AND VALUATION FOR ENHANCED SCOUTING AND RECRUITMENT IN FOOTBALL

M.Sc. Data Science and AI

Student Number: S267562

Supervisor: Anousheh Ramezani

Word Count: 11,981

This report is submitted in partial fulfilment of the requirement for the degree of **M.Sc. Data Science and Artificial Intelligence** by **S267562**.

Submission Date:

5th of August 2024

Abstract

This thesis endeavours to integrate advanced data analysis techniques with traditional football scouting and recruitment processes. The primary objectives are to develop a player similarity tool capable of identifying similar players to any given player and estimating the values of those players. This tool aims to streamline player identification, scouting, valuation, and recruitment processes, offering significant benefits to clubs with smaller financial budgets or those operating under UEFA Financial Fair Play (FFP) regulations.

This research investigates the effectiveness of K-Means clustering in grouping similar players based on performance metrics. By engineering features, conducting correlation analysis, and standardisation, the study identifies optimal clustering configurations and evaluates their quality using metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results indicate that a three-cluster solution offers a meaningful categorisation of players by position, attributes, and performance levels, balancing clustering quality and practical relevance. In parallel, Random Forest Regression (RFR) model is used to estimate player market values. This model's high R^2 score of approximately 0.93 demonstrates its strong predictive accuracy, providing a reliable tool for financial decision-making in player transfers. The findings highlight the potential of data-driven models to enhance traditional scouting methods by narrowing down player pools, saving costs, and time.

The research discusses limitations related to data accuracy, specificity, and the dynamic nature of transfer market. Finally, future work is suggested to further refine these models, including development of AI-generated scouting reports, a "Team Builder" feature for player recommendations, and an interactive web application for enhanced usability. Overall, this thesis contributes to the growing field of sports analytics by demonstrating how advanced data analysis can revolutionise football recruitment and scouting processes, providing clubs with a competitive edge in modern football landscape.

Acknowledgements

First and foremost, I am profoundly grateful to my family for their constant support and encouragement.

I would like to express my gratitude to my supervisor, Anousheh Ramezani. I am thankful for her patience and support throughout thesis writing process, her guidance in Python and SQL for Machine Learning, and MySQL & NoSQL databases modules. Her advice on writing technical reports for coding projects has significantly contributed to the successful completion of this work.

I would also like to extend my sincere thanks to the course leader, Dr. Adnane Ez-zizi, for Data Science and Artificial Intelligence.

I am thankful of all for their immense support and encouragement.

Table of Contents

Chapter 1: Introduction 1

1.1 Research Background 1

1.2 Scope of the Research 2

1.3 Research Aims and Objectives 3

1.4 Research Questions 3

1.5 Problem Statement 4

1.6 Significance of Study 4

1.7 Assumptions, Hypothesis and Challenges 4

1.8 Overview of Project 5

Chapter 2: Literature Review 6

2.1 Traditional Scouting and Recruitment Practices 6

2.2 Challenges of Traditional Scouting and Recruitment 8

2.3 Modern Scouting Practices in Football 9

2.3.1 Revolutionising Player Recruitment 9

2.3.2 Introduction of The Financial Fair Play (FFP) Rule 9

2.3.3 Role of Player Agents and Agencies 10

2.4 Related Work 10

2.4.1 Supervised Learning Techniques in Sports 10

2.4.2 Unsupervised Learning Techniques in Football 11

2.4.3 Candidate Prediction in Other Professions 12

2.5 Conclusion 14

Chapter 3: Research Methodology 15

3.1 Business Background and Data 15

3.1.1 Data Sources 15

3.1.2 Data Collection 15

3.1.3 Data Description 15

3.2 Architecture 18

3.3 Exploratory Data Analysis (EDA) 20

3.3.1 K-Means Clustering Model for Player Similarity 20

3.3.2 Random Forest Regression to Estimate Player Value 23

3.4 Data Cleaning & Feature Engineering 25

3.4.1 FM Data for K-Means 25

3.4.2 Transfermarkt Data for RFR Model 26
3.5 Feature Selection 27
3.5.1 Feature Selection for K-Means Model 27
3.5.2 Feature Selection for RFR Model 29
3.6 Model Building 30
3.6.1 K-means Model 30
3.6.2 Random Forest Regression Model 36
3.7 Model Evaluation 37
3.7.1 K-Means Model Evaluation 37
3.7.2 Random Forest Regression Model Evaluation 38
3.8 Ethical Considerations 39
3.9 Conclusion 40
Chapter 4: Result Interpretation 41
4.1 Correlation and Feature Engineering 41
4.2 Hyperparameter Tuning and Model Optimisation 42
4.2.1 K-Means Clustering for Player Similarity 42
4.2.2. Random Forest Regressor for Player Valuation 42
4.3 Model Performance Evaluation 43
4.3.1 K-Means Clustering Performance 43
4.3.2 Random Forest Regressor Performance 43
4.4 Conclusion 44
Chapter 5: Conclusion and Discussion 45
5.1 Interpretation of Outcomes in Relation to Research Questions 45
RQ1 45
RQ2 46
RQ3 46
5.2 Limitations and Challenges 46
5.2.1 Data Accuracy and Descriptiveness in Player Similarity Model 47
5.2.2 Challenges in Player Valuation and Data Availability 47
5.3 Future Work 47
5.3.1 AI-Generated Scouting Reports 47
5.4 Conclusion 48
Appendix 49

Appendix A 49
Appendix B 50
References 52

List of Figures and Tables

Figure 2.1: Early Days Practices. 7

Figure 2.2: Scouting and Recruitment Practices After Globalisation and Financial Success of the Sport 8

Table 2.1: Summary of Related Work 13

Figure 3.1: Number of Players in Different Leagues 16

Table 3.1: FM Data Frame 17

Figure 3.2: K-Means Model Architecture 18

Figure 3.3: RFR Model Architecture 18

Table 3.2: FM Data (First 5 Rows) 20

Table 3.3: Data Information Summary 21

Table 3.4: Descriptive Statistics 21

Figure 3.4: Distribution of Overall Rating 22

Table 3.5: Missing Data Counts 23

Table 3.6: Transfermarkt Data for First 5 Row 24

Table 3.7: Transfermarkt Data Descriptive Statistics for Key Variables 24

Figure 3.5: Distribution of Player Market Value 25

Figure 3.6: Correlation matrix for Positional Features 29

Figure 3.7: Elbow Graph Plot 31

Figure 3.8 Average Attacking Finishing Per Cluster 32

Figure 3.9: Average Defending Marking Awareness Per Cluster 32

Figure 3.10: Average Goalkeeping Reflexes Per Cluster 33

Figure 3.11: t-SNE Cluster Plot. 34

Figure 3.12: PCA Cluster Plot 35

Table 3.8: Similar Players Dataframe 35

Table 3.9: Estimated Player Value 36

Figure 3.13: RFR (comparison between Actual and Predicted Values) 39

List of Abbreviations

Fédération Internationale de Football Association – FIFA

Financial Fair Play – FFP

Mean Absolute Error - MAE

Random Forest Regression – RFR

R-Squared - **R^2**

Union of European Football Association – UEFA

CHAPTER 1: INTRODUCTION

This chapter introduces the background, scope, aims, and objectives of the research, which aims to integrate advanced data analysis with traditional scouting methods to streamline player scouting and recruitment in football. Section 1.1 provides the research background, emphasising football's global influence and the complexities of player recruitment in professional football. Section 1.2 outlines the scope of the research, focusing on developing a player similarity and estimated valuation model using statistical techniques and machine learning algorithms. Section 1.3 details the research aims and objectives, while Section 1.3.1 presents key research questions that guide the study. Furthermore, Section 1.4 outlines the problem statement, highlighting the challenges and risks inherent in traditional player recruitment methods. Section 1.5 discusses the significance of the study, explaining its potential impact on enhancing player recruitment practices. Additionally, Section 1.6 outlines the assumptions and challenges associated with the research. Finally, Section 1.7 provides an overview of the project and its expected outcomes, considering practical applications and financial constraints faced by smaller clubs.

1.1 Research Background

Football has evolved into a globally influential and economically significant sport, captivating millions of fans across more than 200 countries. Governed by the Fédération Internationale de Football Association (FIFA) (Pariath et al., 2018), football's transformation over the past three decades into a lucrative industry has amplified the importance of strategic player recruitment. Modern professional clubs, now sophisticated business entities, navigate complex financial landscapes involving sponsorships, endorsements, and broadcasting deals (Müller et al., 2017; Rohde and Breuer, 2017).

The importance of player recruitment cannot be overstated, as it directly influences a club's competitive edge and financial health. Effective recruitment underpins sporting success and revenue generation, making it a critical focus for clubs operating under financial constraints imposed by UEFA Financial Fair Play (FFP) regulations (Rossi et al., 2016; Baroncelli & Lago, 2016; Szymanski, 2015). These regulations necessitate a balance between spending and revenue, thereby intensifying the need for efficient and effective recruitment strategies (Franck, 2014; Plumley et al., 2017).

In this context, traditional scouting and recruitment methods, heavily reliant on subjective assessments by scouts and coaches, are increasingly being supplemented by data-driven approaches. Inspired by the "Moneyball" methodology popularised in baseball, football clubs are adopting advanced statistical analyses to uncover undervalued talent and optimise recruitment decisions (Lewis, 2004; Gerrard, 2016). This approach, pioneered in football by clubs like FC Brentford and FC Midtjylland, utilises comprehensive data analysis to identify players whose potential and performance may not be immediately apparent through conventional scouting (Biermann, 2016).

Despite these advancements, many clubs still predominantly rely on human observation and intuition, which are prone to biases and subjectivity (Kerr et al., 2022; Radicchi & Mozzachiodi, 2016). There is a growing advocacy for integrating scientific and technological methodologies to enhance decision-making in player recruitment, minimising human biases and improving objectivity (Gerrard, 2017; Schumaker et al., 2010; Hakes & Sauer, 2006).

The objective of this research is to streamline the recruitment process by enabling recruitment teams to identify players with similar attributes to a given player and provide an estimated valuation for the identified players, thereby saving time and resources. This project employs K-means clustering to group players based on their Football Manager (FM) statistics and uses Random Forest Regression (RFR) to estimate player value using Transfermarkt data. By integrating advanced data analysis with traditional scouting methods, this approach aims to provide a comprehensive model that streamlines player recruitment.

1.2 Scope of the Research

This research aims to integrate traditional scouting methods with advanced data analysis to streamline player scouting and recruitment practices in football. The study focuses on developing a player similarity model using K-means, a clustering machine learning algorithm to assist football clubs in identifying players with similar attributes and performance levels, as well as predicting player value using RFR a machine learning technique used in predicting continuous outcomes. The project utilises K-means clustering to group players with similar attributes based on their FM statistics and employs RFR to estimate player value using Transfermarkt data taking UEFA Financial Fair Play (FFP) regulations into consideration. This added valuation empowers clubs with low financial pull to discover and recruit cost-effective alternatives to top-tier players, allowing them to stay competitive on the pitch while adhering to their budgets and FFP regulations.

1.3 Research Aims and Objectives

The primary goal of this research is to create a comprehensive model that enables recruitment teams to accurately identify players who exhibit similar characteristics and performance levels to any given player. The specific aims and objectives are:

1. To develop a robust player similarity model using K-means clustering to group players with similar performance levels based on their performance metrics.
2. To develop a regression-based model using Transfermarkt data for accurately estimating player values.

This approach will integrate advanced data analysis with traditional scouting methods to enhance player scouting and recruitment practices.

The model's effectiveness in identifying similar players will be evaluated using several metrics, including Mean Absolute Error (MAE) and R-squared score for regression accuracy, Silhouette, Davies-Bouldin Index, and Calinski-Harabasz Index scores for clustering quality, and visual validation technique t-SNE plot to ensure meaningful clustering and accurate player similarity mapping. Ultimately, this will bring about a cost-effective and time-saving approach to player recruitment for clubs and give a competitive edge to smaller clubs with smaller budgets.

1.4 Research Questions

To achieve the outlined research objectives the following research questions will be addressed:

- RQ1. How can advanced data analysis be utilised to enhance traditional scouting methods and improve player scouting and recruitment in football?
- RQ2. How can a data analysis model enhance success in football recruitment?
- RQ3. How can the player similarity and valuation models help clubs with limited financial resources or those constrained by FFP?

An overview of the research background will be provided in Chapter 2 to understand the significance of these questions. My proposed methodology will be comprehensively outlined delivering answers to the research questions.

1.5 Problem Statement

In the highly competitive world of professional football, clubs constantly seek to gain a strategic advantage by scouting and acquiring players who seamlessly align with their team's composition, playing styles, and overall vision. Despite the resources invested in scouting and recruitment, the process is laden with difficulties. The task of securing the perfect player is not always successful, leading to significant risks and major financial implications.

This research aims to address these challenges by using data that considers all aspects of a player's performance to provide a list of similar players, saving the football club’s scouts time for watching loads of games, discovering players that the club were not even aware of while also minimising the challenges of recruiting the wrong players.

1.6 Significance of Study

The significance of this thesis lies in its potential to revolutionise player recruitment practices within the football industry. By leveraging intelligent data analysis, the proposed model offers football clubs a more informed, cost-effective, time-saving, and global approach to scouting and recruiting football players.

The proposed methodology can save scouts time and football clubs' expenses, by assisting clubs find players with desired traits that are outside of their scouting scope. By adding valuation to the provided similar players, the model ensures financial stability and compliance with UEFA FFP regulations, benefiting clubs with limited financial resources. The significance of this research aligns with the strategic "Moneyball" approach, enabling clubs to recruit talented players with low transfer fees and wages. The concept of "Moneyball," popularised by Michael Lewis in his 2004 book, revolutionised traditional scouting techniques in baseball by leveraging data analytics to identify undervalued players, leading to remarkable success for the Oakland Athletics despite their restricted budget (Lewis, 2004). This innovative approach involves utilising sophisticated statistical analysis to extract insights into player performance and potential, which may not be evident through conventional methods (McHale, 2018).

1.7 Assumptions, Hypothesis and Challenges

The research assumes that integrating advanced data analysis with traditional scouting methods will result in more efficient and effective player scouting and recruitment processes.

It also hypothesises that the player similarity model will significantly reduce scouting time and resources while maintaining or improving the quality of player acquisitions. Challenges anticipated in this research include the availability and quality of football data, the complexity of developing accurate machine-learning algorithms, and the need for clubs to adapt to new recruitment practices.

1.8 Overview of Project

The project aims to develop player similarity and valuation models that integrate traditional scouting methods with advanced data analysis to improve player recruitment in football. The outcome will be comprehensive models that enable football clubs to identify players with similar characteristics and performance levels efficiently. The model will be evaluated for its effectiveness in real-world scenarios, considering UEFA FFP regulations and financial constraints faced by smaller clubs. The project ultimately seeks to provide a cost-effective, efficient, and informed approach to player transfers, benefiting football clubs at all levels.

CHAPTER 2: LITERATURE REVIEW

In Chapter 2 of the thesis, the focus is on the evolution of football scouting and recruitment. The literature review discusses the transition from traditional methods to modern data-driven practices. It begins by exploring the historical background of scouting and recruitment in football, emphasising the impact of international competitions and legal changes global player transfers. The chapter also addresses the challenges inherent in traditional scouting. Furthermore, it examines modern advancements in scouting and recruitment. Finally, the chapter reviews related works, focusing on the application of machine learning algorithms for player selection and recruitment in football, and candidate selection in other industries, identifying gaps that the current project aims to address.

2.1 Traditional Scouting and Recruitment Practices

Scouting and recruitment in football have evolved significantly over the decades, influenced by technological advancements, globalisation, revenue, and changes in club management strategies.

Football scouting in its early years focused predominantly on youth acquisition through talent identification. The primary aim was to identify youngsters with exceptional athletic prowess to groom them into future elite footballers (Williams and Franks, 1998). During this period, it was not common practice for established players to transfer between clubs. Instead, the emphasis was on recruiting young, unattached players who could develop within the club and eventually secure their place in the first team. This period was primarily based on the experience and intuition of the scouts. From the 1960s to the 1980s, figures like Geoff Twentyman, Chief Scout at Liverpool, exemplified the traditional scouting role. Twentyman travelled extensively, attending numerous matches, compiling detailed reports based on repeated observations of promising players (Hughes, 2009). This approach emphasised the importance of extensive experience and personal judgment in identifying talent.

Figure 01 Early Days Practices.

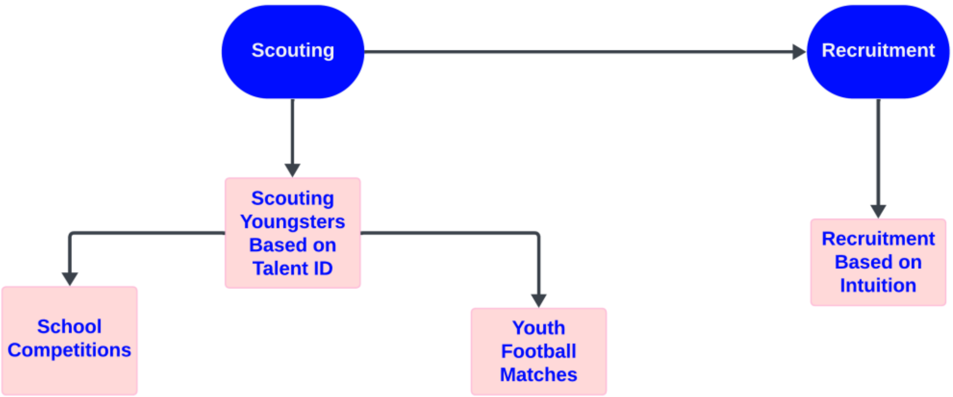

Scouting
Recruitment
Scouting Youngsters Based on Talent ID
Recruitment Based on Intuition
School Competitions
Youth Football Matches

Figure 2.1 illustrates traditional scouting practices where scouts focused on identifying young talent through talent identification in school competitions and youth football matches. Recruitment decisions during this period were heavily based on the scouts' intuition and personal judgment.

The establishment of international competitions, such as the Olympic football tournament in 1896 and the FIFA World Cup in 1930, played a crucial role in putting football on the world stage and increasing its popularity (Rookwood & Buckley, 2007; Taylor, 2006). This global exposure was further enhanced by the expansion of the World Cup. For instance, the 1982 World Cup hosted by Spain, which featured 24 teams, was the first expansion since 1934 and brought more international talent into the spotlight. Football clubs increased their global scouting efforts because of this expansion, which improved scouting and recruitment procedures in football. Scouts were now able to observe and evaluate a wider array of talent.

In response to this popularity and accompanying financial gains, European clubs pushed for laws that eased restrictions on the number of foreign players they could recruit, with that rule change coming in 1995, in the form of the 1995 Bosman ruling which further accelerated international player transfers by removing these restrictions, significantly increasing global player transfers (Schokkaert, 2016). This legal change, propelled clubs towards an increasingly globalised approach to player recruitment, transforming the landscape of football talent acquisition.

Figure 0.2 Scouting and Recruitment Practices After Globalisation and Financial Success of the Sport

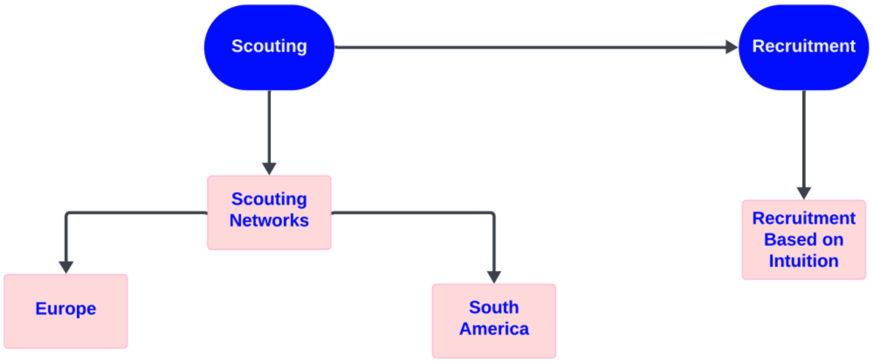


Figure 2.2 illustrates the evolution of scouting and recruitment practices following the globalisation and financial success of football. It depicts the expansion of scouting networks into Europe and South America. Despite this expansion, recruitment decisions remained largely based on intuition, lacking objectivity.

2.2 Challenges of Traditional Scouting and Recruitment

Traditional scouting in football faces challenges due to its reliance on subjective judgment and intuition, which can undermine the accuracy and effectiveness of talent identification.

The decision-making process is inherently uncertain and often influenced by hunches and gut feelings about a player's potential (Schumaker et al., 2010; Nash & Collins, 2006). This intuition-based approach can lead to inconsistent assessments, heavily dependent on scouts' personal experiences and cultural contexts (Christensen, 2009). The perceived ability to spot talent blends visual experience, intuition, and gut feeling, but is often influenced by previous identifications and personal interpretations of professional football (Lund & Söderström, 2017). Consequently, the process is shaped by social and cultural influences, leading to biases and oversights in talent evaluation (Tranckle & Cushion, 2006).

Considering the subjectivity involved in scouting, assessing the reliability of human observation, memory, and recollection is crucial. Although scouts' subjective experiences provide valuable insights, they also present significant challenges. Studies have revealed that human observation and recollection are prone to biases and inaccuracies (Franks and Miller, 1986). Effective scouts are required to process large amounts of information quickly and accurately, which can be challenging due to the vast volume of visual data available in a single football match (Reeves et al., 2019).

2.3 Modern Scouting Practices in Football

2.3.1 Revolutionising Player Recruitment

The landscape of player recruitment in football has evolved significantly since the 1990s. Disparities in financial capabilities among clubs have led to variations in resources allocated for talent scouting and recruitment. Major European clubs, such as Arsenal and Manchester United, began establishing large scouting networks in recent years (Elberse, 2013; Rivoire, 2011).

One new advancement in player recruitment is the emergence of video scouts. These scouts watch football games from leagues around the world using specialised platforms (Stats Perform, 2024; Wyscout, 2024). Implementing video scouting enables clubs to economise on logistics expenses while extending their reach across large geographical areas for player identification.

2.3.2 Introduction of The Financial Fair Play (FFP) Rule

In 1995, clubs advocated for the Bosman ruling to facilitate global player recruitment (Schokkaert, 2016). By 2009, however, the introduction of Financial Fair Play (FFP) regulations became necessary to maintain the sport's competitiveness and prevent financial doping that could lead many clubs to bankruptcy (Peeters & Szymanski, 2014). Implemented by UEFA in 2011, FFP marked a significant step towards addressing the financial instability afflicting European football clubs. A 2009 UEFA review revealed that over half of the 655 European clubs examined had incurred losses in the previous year. While some clubs managed these losses through wealthy owners, at least 20% were in severe financial distress.

FFP regulations mandate that a club's income must balance its player wage expenditures and net transfer fee spending over three years, thereby imposing budgetary restrictions aligned with revenue generation capacities (Franck, 2014). These measures address concerns about financial "doping," where external funds distort competitive balance (UEFA, 2011). Clubs that exceed spending limits face sanctions such as fines, withholding of prize money, player transfer bans, and even disqualification from European competitions (UEFA, 2011).

For the first time, football clubs' financial practices in recruitment came under scrutiny, with FFP introducing mandatory financial discipline.

2.3.3 Role of Player Agents and Agencies

Player agents and agencies have become stakeholders in the recruitment process. They provide consultation and representation for players in transfer negotiations, employment contracts, and commercial deals (Geey, 2020). According to Rossi et al. (2016), agents help clubs access information on potential recruits through their player portfolios and well-developed networks. However, in some cases, agents can wield significant influence over a club's transfer policy, potentially leading to conflicts of interest (Maguire, 2019).

2.4 Related Work

This section provides an overview of significant research efforts related to predicting and identifying suitable candidates in sports and other industries, with a particular focus on the application and effectiveness of machine learning algorithms and methodologies in sports, candidate prediction in various industries.

2.4.1 Supervised Learning Techniques in Sports

Recruiting the right players is essential for the success of any sports team. Numerous studies have been conducted to predict player performance and suitability in football and other sports.

AlShboul et al. (2017) conducted a study using a competitive neural network model to select players for a football club's squad. The model was then utilised to predict the squad's chances of victory against top opposition teams based on the individual player ratings of the selected players. By analysing the attributes of the selected team's players and those of the best opposition team, the neural network evaluated the likelihood of winning or losing. In an initial test with 11 players, representing the number of players in an official football match, the neural network achieved an accuracy rate of 54%. A subsequent experiment with 22 players, representing a full football squad, resulted in a 60% accuracy rate.

In 2018, Mathew et al. undertook a study to predict suitable replacements for football players who had been recruited by other clubs. They achieved this by predicting the player rating from the dataset and selecting players with similar ratings to the departed player. They also observed how the number of classes influenced the accuracy of the model. Their findings revealed that Linear Discriminant Analysis (LDA) outperformed other machine learning models.

Aalbers and Van Haaren (2018) conducted a study using 18 predefined player categories for outfield football players. They organised the players into the five common positional groups in football (as per Dellal et al., 2011) and further subdivided them into specific sub-groups.

In 2019, Bunker and Thabtah put forward the idea of using a neural network to predict football match results based on player attributes. The model utilised a dataset comprising player performance data, team performance data, and historical team results. Player attributes were classified into four categories: player resistance, speed, physical status, and technique. After implementing the neural network model, it was observed that its accuracy in predicting match results based on player attributes was relatively low.

These studies highlighted the effectiveness of traditional algorithms such as LDA and decision trees for player recruitment and selection in football compared to neural networks. However, they primarily concentrated on predicting suitable replacements for individual players in a squad, rather than identifying the most similar player available on the market to the departed player. Moreover, they focused on predicting and analysing player performance metrics rather than accurately identifying similar players. Additionally, these studies neglected the financial aspects of the scouting and recruitment process in football by not estimating the associated costs of recruiting a player. This highlights a gap that this project aims to address.

2.4.2 Unsupervised Learning Techniques in Football

Unsupervised learning (such as clustering techniques) has been utilised to categorise players based on similar traits. This has helped in identifying player types and potential replacements.

Network motifs technique was applied to explore player categories based on their passing behaviour. Peña and Navarro (2015) employed affinity propagation to analyse players' pass motifs and found 37 clusters. Similarly, Bekkers and Dabadghao (2019) utilised pass motifs in a mean shift clustering phase and identified 25 clusters. Both studies showcase the potential of their models in aiding recruitment decisions, demonstrated by a case study on finding an alternative for former Barcelona midfielder Xavi (Bekkers & Dabadghao, 2019; Peña & Navarro, 2015).

2.4.3 Candidate Prediction in Other Professions

Machine learning algorithms have been applied to streamline job recruitment processes across various professions. In a study conducted by Jannat et al. (2016), a Naive Bayes classifier was employed to predict candidates for engineering companies. The study aimed to optimise the recruitment process and mitigate bias by ranking datasets based on knowledge level and GPA scores. Through calculations of probability and frequency, the model effectively identified the most prevalent class, aiding in the prediction of suitable job applicants. The Naive Bayes model demonstrated strong performance on both training and test datasets.

Jantan et al. (2010) employed a decision tree C4.5 classifier to predict employees deserving of promotion, achieving 95% accuracy using features such as skills, knowledge, and work outcomes.

In a study conducted by Apatean et al. (2017), the researchers aimed to predict suitable job candidates from a pool of applications using machine learning models. Their goal was to streamline the selection process by using these models to identify the best candidates for specific job positions. The study treated job positions as the target class and considered candidate information such as education, programming languages, and salary range as attributes. After data preprocessing, various data mining algorithms, including KNN, LDA, Naive Bayes, and decision trees, were applied. The results revealed that LDA and Naive Bayes outperformed the other models.

Li et al. (2011) developed a recruitment system using Support Vector Machines (SVM) and Multi-Criteria Decision Making (MCDM). They utilized e-questionnaires as the dataset, applying SVM to predict the best applicant and MCDM to evaluate the model's performance. The results indicated that the developed system is well-suited for recruitment purposes.

According to these studies, classification algorithms such as LDA, Naive Bayes, decision trees, and SVM show effectiveness in predicting job candidates. Apatean et al. (2017) acknowledged the financial aspects of recruitment by including salary range as an attribute. However, their primary focus was on non-sporting domain recruitment, making it clear that the findings may not directly apply to identifying similar players in sports or football. This underscores the novelty and specific relevance of this project.

Table 2.1 Summary of Related Work

Methodology	Reference	Advantages	Disadvantages or Limitations	Use Cases or Application Areas
Competitive Neural Network Model	AlShboul et al., 2017	Predicts team victory chances; improved accuracy with larger squad size.	Initial accuracy low (54% with 11 players); higher with 22 players (60%).	Football squad selection, match outcome prediction.
Linear Discriminant Analysis (LDA)	Mathew et al., 2018	Outperforms other models in predicting player ratings.	Accuracy is influenced by the number of classes.	Replacing recruited players, player rating prediction.
Stochastic Gradient Descent (SGD) Classifier	Aalbers & Van Haaren, 2018	Detailed classification of player positions.	Does not account for unknown player types; positional biases.	Categorising football players, scouting.
Neural Network Model	Bunker & Thabtah, 2019	Uses extensive player and team data.	Relatively low prediction accuracy.	Football match outcome prediction.
Network Motifs Technique	Bekkers & Dabadghao, 2019;	Identifies player types and replacements; demonstrates model potential in case studies.	Focuses on specific player behaviours (passing).	Categorising football players, and recruitment decisions.
Affinity Propagation	Peña & Navarro, 2015	Categorises based on pass motifs; identifies many clusters.	Limited to passing behaviour analysis.	Identifying player replacements, tactical analysis.
Naive Bayes classifier	Jannat et al., 2016	Optimises recruitment; mitigates bias.	Limited to engineering companies.	Engineering job recruitment.
Decision Tree C4.5 Classifier	Jantan et al., 2010	High accuracy (95%); uses relevant features.	Specific to promotion prediction.	Employee promotion decisions.
Various ML Models for Job Candidate Prediction	Apatean et al., 2017	Streamlines selection; LDA and Naive Bayes effective.	Focus on non-sporting domains.	General job recruitment.

Support Vector Machines (SVM) and Multi-Criteria Decision Making (MCDM)	Li et al., 2011	Effective for recruitment; incorporates salary range.	General recruitment, not specific to sports.	E-questionnaire-based recruitment, job candidate evaluation.

The table highlights a variety of methodologies employed in football and other industries for predicting and identifying suitable players and candidates. Each methodology comes with its own set of advantages and limitations, and they are applied in specific contexts. The insights from these studies underline the use of various machine learning algorithms (such as SVM, MCDM, and more) and the gaps in their application, particularly in the context of player similarity identification and value estimation, which this research aims to address.

2.5 Conclusion

The reviewed works of literature in this section indicate that, while football has made use of the massive amount of data it generates to enhance scouting and recruitment through Machine Learning, there is a gap in applying these methods to identify similar players and estimate their value. This project aims to address this gap by utilising K-Means clustering to categorise similar players based on their FM stats and employing RFR to estimate player value using Transfermarkt data. This approach not only simplifies the scouting process but also offers a global player search, presenting a unique contribution to the field.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Business Background and Data

3.1.1 Data Sources

For this project, data has been sourced from two reputable platforms, which are Football Manager (FM) and Transfermarkt. FM, which is a widely recognised football management simulation game, offers an extensive database of player statistics compiled by a network of around 1,300 researchers and 100 head researchers who gather intelligence on player attributes (Bleaney, 2014; Skandalis et al., 2015). This database is highly regarded for its depth and accuracy, often used by real-world football clubs for scouting purposes.

Transfermarkt, on the other hand, is a prominent football website that specialises in player market values and transfer information. The site relies on the input of its extensive online community, comprising approximately 190,000 members. Market values are determined through a hierarchical approach, where community members contribute to discussions, and designated 'judges' have the final say on the figures published (Herm et al., 2014; Müller et al., 2017). This community-driven model ensures a comprehensive and up-to-date reflection of the football market.

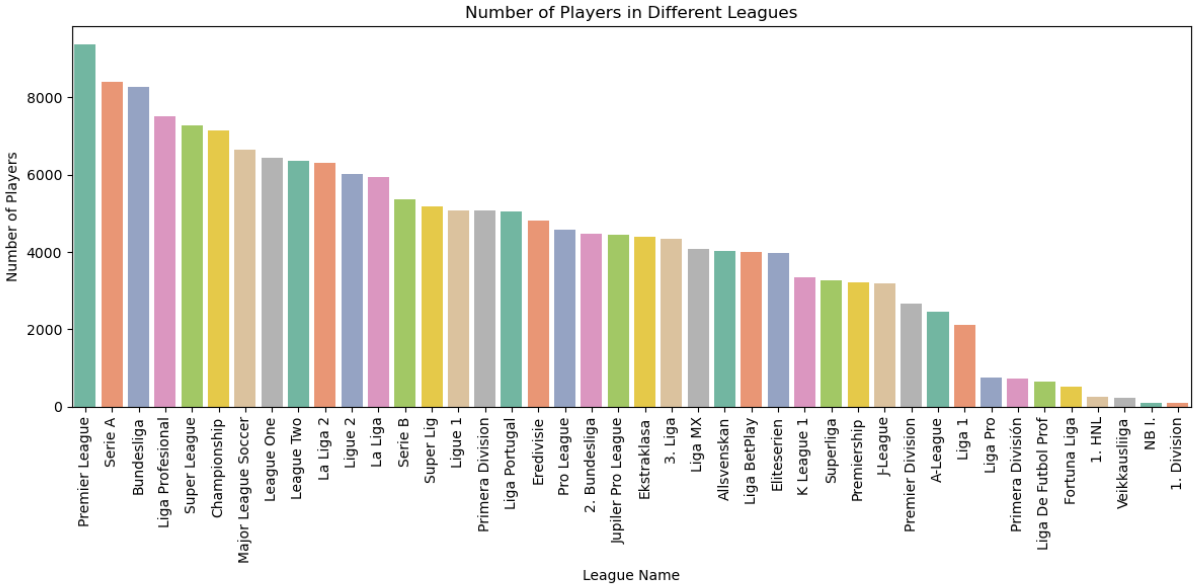
3.1.2 Data Collection

The data for this project was obtained by web scraping from FM and Transfermarkt. The rich datasets from FM and Transfermarkt are integral to the project's aims and objectives. This detailed collection supports rigorous statistical analyses and advanced model development to aid in scouting, recruitment and player valuation.

3.1.3 Data Description

The FM dataset offers an extensive and nuanced view of players and clubs from across the globe, making it a valuable tool for international scouting and analysis. The dataset includes players from various leagues, encompassing a diverse range of competitive environments and playing styles. This global scope covers major leagues such as the Premier League, La Liga, Serie A, and many others. It also includes data from lesser-known leagues, providing a comprehensive overview of football talent worldwide.

Figure 0 Number of Players in Different Leagues



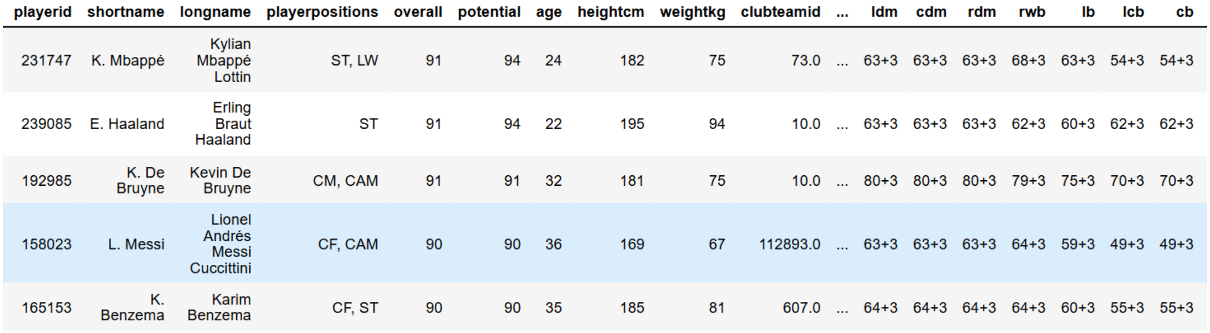
Each player's profile in the dataset contains detailed attributes ranging from technical, physical, and mental abilities, providing a comprehensive assessment of their skills and potential. The dataset includes players' overall, potential ratings, and specific skill attributes, which are numerically represented on a scale ranging from 1 to 100. This scale is crucial for evaluating a player's proficiency in various skills, such as shooting, passing, and tackling.

Key components of the dataset include:

- **Basic Information**: Unique identifiers (player id), names (short name, long name), and positions played (player positions).
- **Demographic Details**: Nationality (nationality), and Age (age).
- **Performance Metrics**: Overall skill ratings (overall), potential (potential), and specific skill attributes (e.g., attacking crossing, dribbling, defending standing tackle).
- **Position Data**: Detailed by the player's position within the club, which is captured in the club position column. The club position column specifies the player’s role in the team, such as ST (Striker), GK (Goalkeeper), CM (Central Midfielder), CB (Centre Back), and other positions. It also includes designations like SUB (Substitute) and RES (Reserve), which indicate players who are not regular starters, but are part of the squad rotation. This detailed categorisation allows for

a better understanding of a player’s specific role and expected contribution within their team setup.

Table 3.1FM Data Frame



	playerid	shortname	longname	playerpositions	overall	potential	age	heightcm	weightkg	clubteamid	...	ldm	cdm	rdm	rwb	lb	lcb	cb
	231747	K. Mbappé	Kylian Mbappé Lottin	ST, LW	91	94	24	182	75	73.0	...	63+3	63+3	63+3	68+3	63+3	54+3	54+3
	239085	E. Haaland	Erling Braut Haaland	ST	91	94	22	195	94	10.0	...	63+3	63+3	63+3	62+3	60+3	62+3	62+3
	192985	K. De Bruyne	Kevin De Bruyne	CM, CAM	91	91	32	181	75	10.0	...	80+3	80+3	80+3	79+3	75+3	70+3	70+3
	158023	L. Messi	Lionel Andrés Messi Cuccittini	CF, CAM	90	90	36	169	67	112893.0	...	63+3	63+3	63+3	64+3	59+3	49+3	49+3
	165153	K. Benzema	Karim Benzema	CF, ST	90	90	35	185	81	607.0	...	64+3	64+3	64+3	64+3	60+3	55+3	55+3

This rich dataset enables a deep analysis of player attributes, supporting the identification and evaluation of talent across different contexts and competitive levels.

Additionally, the Transfermarkt dataset complements FM data by providing financial and contractual insights about the players. This dataset includes details on players' market values, wages, and contract durations, offering a comprehensive view of the economic aspects of football talent.

Key elements of the Transfermarkt dataset include:

- **Player Information**: This includes the player's full name (long name), positions played (player positions).
- **Financial Data**: It covers the player's market value (value eur), wages (wage eur), and, where applicable, the release clause amount (release clause eur).
- **Contractual Details**: Information on the player's club, including the club they are loaned from (club loaned from), date they joined their current club (club joined date), and the contract expiration year (club contract valid until year).

The Transfermarkt dataset enhances understanding of the economic dimensions in football and is invaluable for examining market trends and player movement patterns within the sport.

This combination of datasets allows for a holistic analysis of both on-field performance and off-field financial and contractual contexts, making it an important tool for achieving the research’s goal of integrating advanced data analysis with traditional scouting and recruitment processes.

3.2 Architecture

In this section, the architecture developed model is provided, focusing on the integration of data mining and machine learning processes to form a cohesive system aimed at analysing player performance statistics to identify similarities and predict market values. The architecture, as depicted in Figure 3.2 and Figure 3.3, outlines key components and their interactions, providing a functional blueprint for the project. It follows a structured flow, ensuring that each component contributes to the overall aim of developing player similarity and estimated valuation models.

Figure 3.2 K-Means Model Architecture

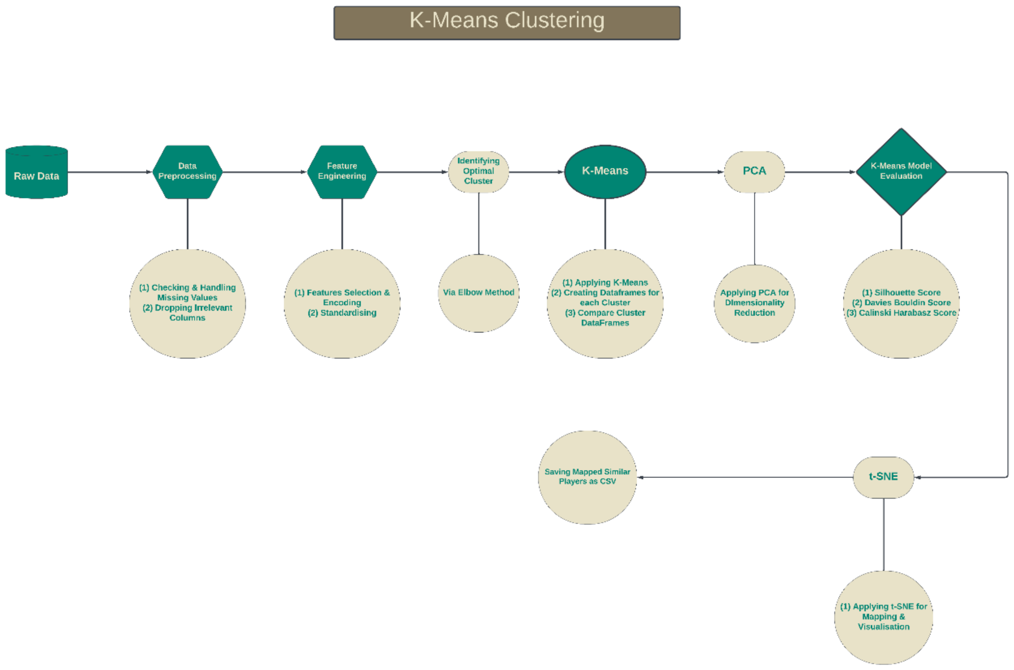
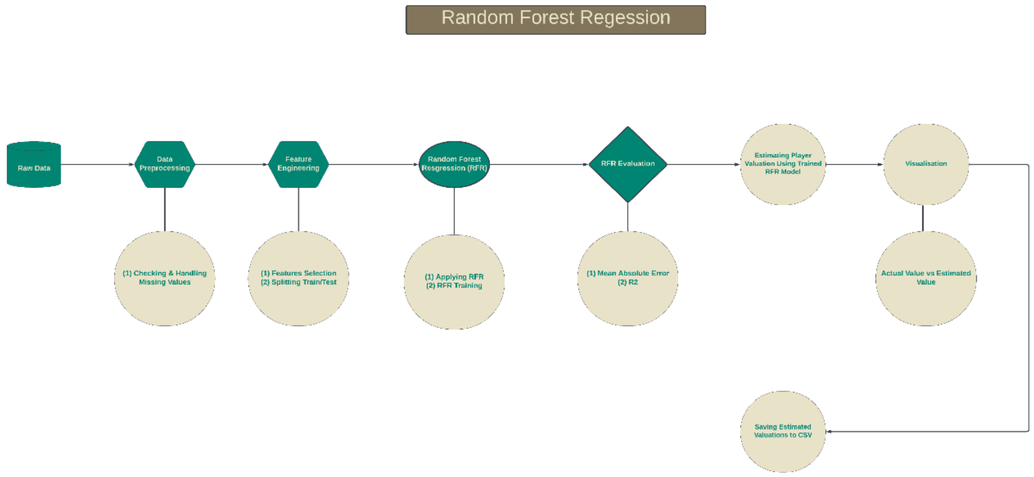

K-Means Clustering
Raw Data
Data Preprocessing
Feature Engineering
Identifying Optimal Cluster
K-Means
PCA
K-Means Model Evaluation
(1) Checking & Handling Missing Values
(2) Dropping Irrelevant Columns
(1) Features Selection & Encoding
(2) Standardising
Via Elbow Method
(1) Applying K-Means
(2) Creating Dataframes for each Cluster
(3) Compare Cluster DataFrames
Applying PCA for Dimensionality Reduction
(1) Silhouette Score
(2) Davies Bouldin Score
(3) Calinski Harabasz Score
Saving Mapped Similar Players as CSV
t-SNE
(1) Applying t-SNE for Mapping & Visualisation

Figure 3.3 RFR Model Architecture



The architecture includes two primary models, including K-Means Clustering and RFR, each consisting of several stages:

1. K-Means Clustering:
    - Data Preprocessing: This initial step involves handling missing values and dropping irrelevant columns, setting the stage for feature engineering.
    - Feature Engineering: Key player attributes are selected and encoded, followed by data standardisation.
    - Cluster Identification: The optimal number of clusters is determined using the Elbow Method, leading to the application of the K-Means algorithm.
    - Dimensionality Reduction: PCA is applied to reduce the data's dimensionality, facilitating easier visualisation and interpretation.
    - Model Evaluation: The performance of clustering model is evaluated using metrics like the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.
    - Visualisation: t-SNE is used for mapping and visualising clustered data, with results saved to a CSV file.
2. Random Forest Regression:
    - Data Preprocessing: Similar to the clustering process, this stage involves handling missing values and preparing data for analysis.
    - Feature Engineering: The data is split into training and test sets after selecting relevant features.
    - Model Application and Training: The Random Forest Regressor (RFR) model is applied and trained on the train set.
    - Model Evaluation: The model's performance is assessed using Mean Absolute Error and R^2 Score.
    - Player Valuation Estimation: The trained RFR model estimates player valuations on the test set, with the results visualised and saved.

This architecture serves as a comprehensive guide, detailing the essential machine-learning processes on the data involved, while also allowing flexibility for modifications as needed.

3.3 Exploratory Data Analysis (EDA)

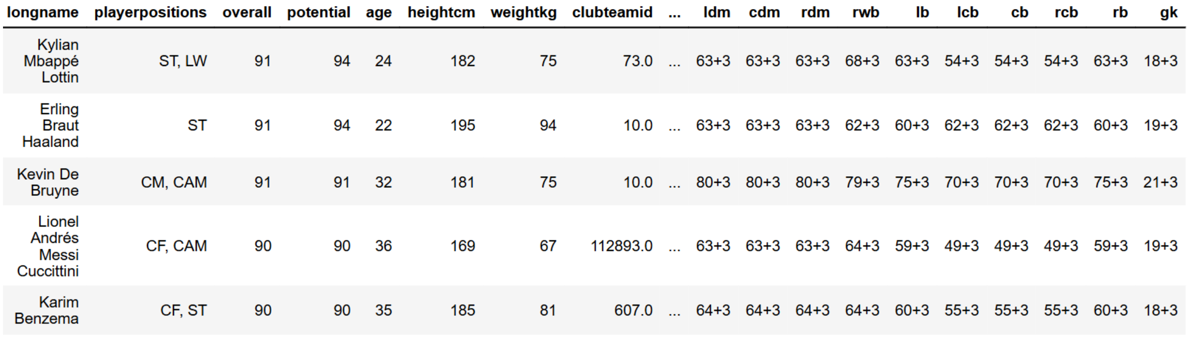
3.3.1 K-Means Clustering Model for Player Similarity

In this section, the Exploratory Data Analysis (EDA) of the FM dataset is discussed. EDA is crucial for understanding underlying patterns, distributions, and relationships within the data. This analysis provides insights into the central tendencies, variability, and data structure, setting the foundation for subsequent modelling, and analysis.

Data Overview: The dataset comprises various attributes of football players, including personal information (e.g., age, heightcm, weightkg), performance metrics (e.g., overall, potential), and specific skills and physical attributes such as pace, shooting, passing, dribbling, defending, and physic. The dataset consists of 180,021 rows and 97 columns, indicating a comprehensive set of features captured for each player.

Table 3.2 shows the first five rows of the dataset, providing a glimpse of the data structure and types of features available.

Table 0.2 FM Data (First 5 Rows)



longname	playerpositions	overall	potential	age	heightcm	weightkg	clubteamid	...	ldm	cdm	rdm	rwb	lb	lcb	cb	rcb	rb	gk
Kylian Mbappé Lottin	ST, LW	91	94	24	182	75	73.0	...	63+3	63+3	63+3	68+3	63+3	54+3	54+3	54+3	63+3	18+3
Erling Braut Haaland	ST	91	94	22	195	94	10.0	...	63+3	63+3	63+3	62+3	60+3	62+3	62+3	62+3	60+3	19+3
Kevin De Bruyne	CM, CAM	91	91	32	181	75	10.0	...	80+3	80+3	80+3	79+3	75+3	70+3	70+3	70+3	75+3	21+3
Lionel Andrés Messi Cuccittini	CF, CAM	90	90	36	169	67	112893.0	...	63+3	63+3	63+3	64+3	59+3	49+3	49+3	49+3	59+3	19+3
Karim Benzema	CF, ST	90	90	35	185	81	607.0	...	64+3	64+3	64+3	64+3	60+3	55+3	55+3	55+3	60+3	18+3

Data Structure and Types: The dataset contains a mixture of numerical, categorical, and textual data. An overview of data types and counts of non-null entries is provided in the summary:

Table 0.3 Data Information Summary

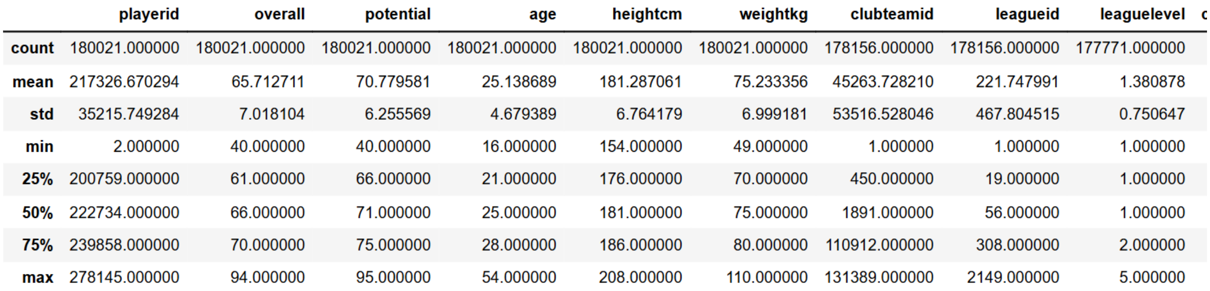
Column	Non-Null Count	Data Type
playerid	180021	Int64
shortname, longname	180021	object
overall, potential, age	180021	Int64
pace, shooting, passing, dribbling, defending, physic	159997	Float64
workrate	180021	Object
preferredfoot	180021	Object
attackingfinishing	180021	Int64
skilllongpassing	180021	Int64
movementsprintspeed	180021	Int64
powerstamina	180021	Int64
mentalityaggression	180021	Int64
defendingmarkingawareness	180021	Int64
goalkeepingreflexes	180021	Int64

The majority of the columns contain complete data, even if some columns have missing values.

Descriptive Statistics: To understand the central tendencies and dispersion of the numerical data in the dataset, descriptive statistics were calculated.

Table 3.4 presents descriptive statistics for selected numerical columns, including key performance metrics:

Table 0.4 Descriptive Statistics



	playerid	overall	potential	age	heightcm	weightkg	clubteamid	leagueid	leaguelevel	c
count	180021.000000	180021.000000	180021.000000	180021.000000	180021.000000	180021.000000	178156.000000	178156.000000	177771.000000	
mean	217326.670294	65.712711	70.779581	25.138689	181.287061	75.233356	45263.728210	221.747991	1.380878	
std	35215.749284	7.018104	6.255569	4.679389	6.764179	6.999181	53516.528046	467.804515	0.750647	
min	2.000000	40.000000	40.000000	16.000000	154.000000	49.000000	1.000000	1.000000	1.000000	
25%	200759.000000	61.000000	66.000000	21.000000	176.000000	70.000000	450.000000	19.000000	1.000000	
50%	222734.000000	66.000000	71.000000	25.000000	181.000000	75.000000	1891.000000	56.000000	1.000000	
75%	239858.000000	70.000000	75.000000	28.000000	186.000000	80.000000	110912.000000	308.000000	2.000000	
max	278145.000000	94.000000	95.000000	54.000000	208.000000	110.000000	131389.000000	2149.000000	5.000000	

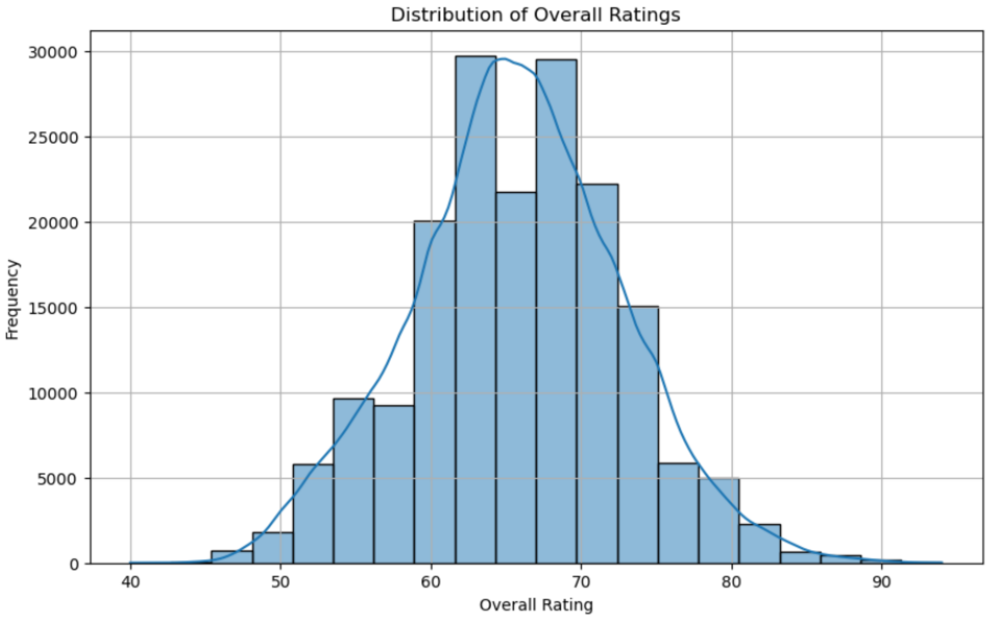
Key observations from the descriptive statistics include:

- The average 'overall' rating is approximately 65.71, with a standard deviation of 7.02, highlighting a diverse range of player skill levels in the dataset.

- The ‘potential’ rating, which indicates a player's possible peak performance, averages around 70.78, suggesting that many players are expected to develop further.
- The ‘pace’ attribute, a critical factor in a player's speed, shows a wide range from 1 to 99, with an average score around mid-range values, reflecting varying player roles and styles.
- Attributes like ‘shooting’, ‘passing’, ‘dribbling’, ‘defending’, and ‘physic’ provide a detailed view of players' technical and physical capabilities, each with their unique distribution patterns. For instance, ‘shooting’ and ‘passing’ are crucial for attacking players, while ‘defending’ and ‘physic’ are more relevant for defensive roles.

Data Distribution and Variability: Understanding the distribution of key attributes is essential for identifying patterns and potential outliers. The presence of exceptionally skilled players as well as lower-rated players demonstrates a wide range of talent levels within the dataset.

Figure 3.4 Distribution of Overall Rating



The distribution of ‘overall’ ratings shows a normal distribution centred around 65 to 70, indicating that most players are rated within this range. This suggests that most players have average ratings, with fewer players receiving extremely high or low ratings.

Distributions of specific skills such as ‘pace’ and ‘dribbling’ reveal significant variation, which is expected given the specialised nature of these attributes across different player positions. For instance, forwards typically have higher ‘shooting’ and ‘dribbling’ skills, while defenders might score higher in ‘defending’ and ‘physic’.

Missing Data: An important part of EDA is identifying missing data and duplicates that could affect the analysis. The dataset contains some missing values in columns such as ‘clubteamid’, ‘clubname’, and ‘nationteamid’, which will be excluded as the analysis will depend on performance metrics columns to group similar players. Table 3.5 shows missing data counts.

Table 0.5 Missing Data Counts

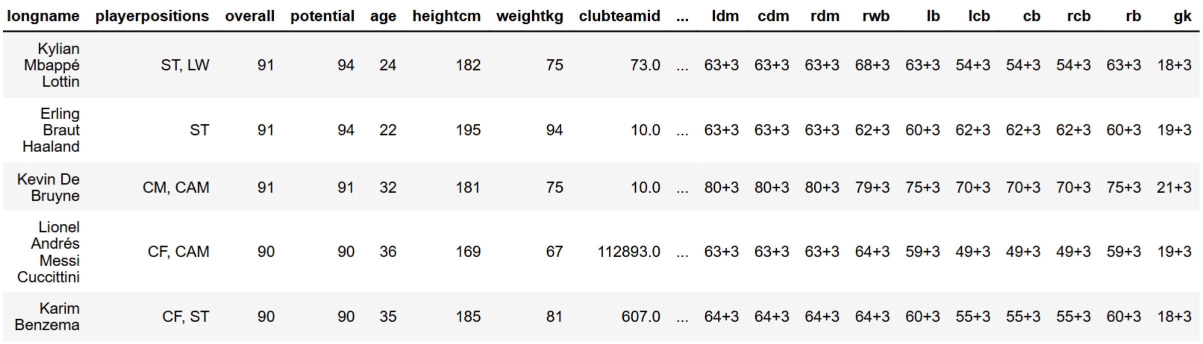
Column	Missing Data Count
pace	20024
shooting	20024
passing	20024
dribbling	20024
defending	20024
physic	20024
Mentality composure	32888

In conclusion, EDA provides a comprehensive overview of the dataset structure, central tendencies, variability, and data quality. This foundational understanding is critical for subsequent K means clustering, as it informs selection of features, handling of missing data, and interpretation of results.

3.3.2 Random Forest Regression to Estimate Player Value

Data Overview: The dataset for the RFR model includes detailed information on football players, such as personal data (age, date of birth), career data (club name, league name), and financial data (player market value, wage, release clause). The dataset comprises 180,021 rows and 16 columns, offering a comprehensive snapshot of players' professional and financial statuses.

Table 3.6: Transfermarkt Data for First 5 Row



longname	playerpositions	overall	potential	age	heightcm	weightkg	clubteamid	...	ldm	cdm	rdm	rwb	lb	lcb	cb	rcb	rb	gk
Kylian Mbappé Lottin	ST, LW	91	94	24	182	75	73.0	...	63+3	63+3	63+3	68+3	63+3	54+3	54+3	54+3	63+3	18+3
Erling Braut Haaland	ST	91	94	22	195	94	10.0	...	63+3	63+3	63+3	62+3	60+3	62+3	62+3	62+3	60+3	19+3
Kevin De Bruyne	CM, CAM	91	91	32	181	75	10.0	...	80+3	80+3	80+3	79+3	75+3	70+3	70+3	70+3	75+3	21+3
Lionel Andrés Messi Cuccittini	CF, CAM	90	90	36	169	67	112893.0	...	63+3	63+3	63+3	64+3	59+3	49+3	49+3	49+3	59+3	19+3
Karim Benzema	CF, ST	90	90	35	185	81	607.0	...	64+3	64+3	64+3	64+3	60+3	55+3	55+3	55+3	60+3	18+3

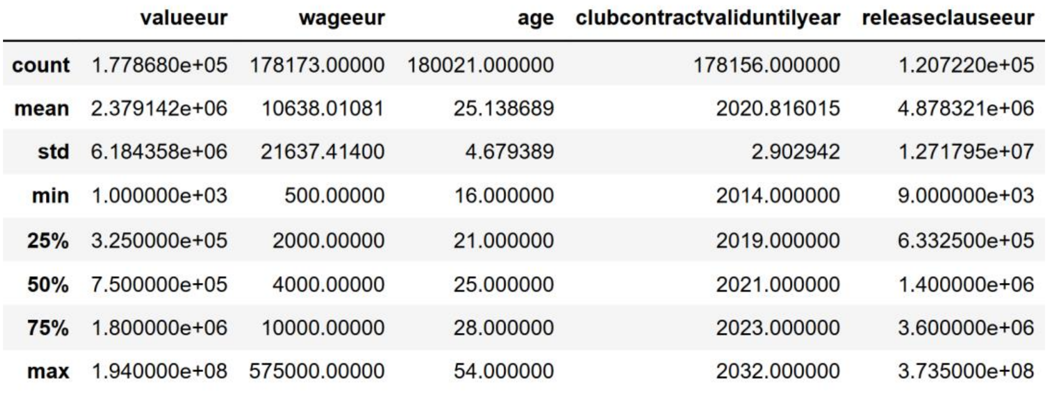
Data Structure and Type: The dataset consists of numerical, categorical, and textual data types.

A summary of the data structure reveals the following:

- Numerical Data: Includes attributes like player market value (value eur), wage (wage eur), and age.
- Categorical Data: Includes attributes like player positions, nationality, club name, and league name.
- Textual Data: Includes player URLs and names.

Data Descriptive Statistics: Descriptive statistics offer insights into the dataset's characteristics.

Table 3.7 Transfermarkt Data Descriptive Statistics for Key Variables



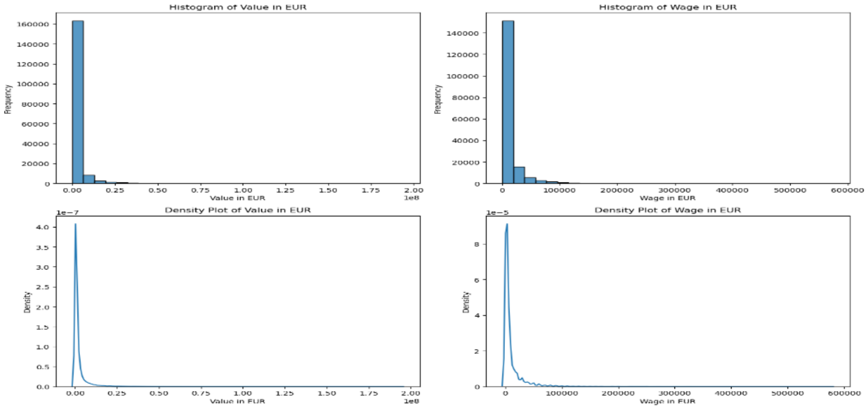
	valueeur	wageeur	age	clubcontractvaliduntilyear	releaseclauseeur
count	1.778680e+05	178173.00000	180021.000000	178156.000000	1.207220e+05
mean	2.379142e+06	10638.01081	25.138689	2020.816015	4.878321e+06
std	6.184358e+06	21637.41400	4.679389	2.902942	1.271795e+07
min	1.000000e+03	500.00000	16.000000	2014.000000	9.000000e+03
25%	3.250000e+05	2000.00000	21.000000	2019.000000	6.332500e+05
50%	7.500000e+05	4000.00000	25.000000	2021.000000	1.400000e+06
75%	1.800000e+06	10000.00000	28.000000	2023.000000	3.600000e+06
max	1.940000e+08	575000.00000	54.000000	2032.000000	3.735000e+08

Key observations include:

- **Player Value**: The average market value ‘valueeur’ is approximately €2.38 million, with a high standard deviation, indicating a wide range in player market values.
- **Wage**: The average wage is about €10,638, reflecting diverse player salary scales.
- **Age**: The dataset includes players from ages 16 to 54, with an average age of around 25, highlighting a range of career stages.

Data Distribution and Variability: Examining distributions of key variables reveals significant skewness, with a few players having exceptionally high values. This is visualised using histograms and density plots in Figure 3.5, which help identify potential outliers and inform feature engineering process.

Figure 05 Distribution of Player Market Value



3.4 Data Cleaning & Feature Engineering

3.4.1 FM Data for K-Means

In this section, the processes involved in data preparation are described, which are data cleaning and feature engineering. Proper data preparation is crucial for ensuring quality and reliability of any analysis or model built on the data. This involves handling missing values, standardisation, and transforming categorical variables.

Missing Values: The dataset contains missing values. Handling these missing values is essential to prevent biases in the analysis and ensure that K-Means algorithms can process data effectively.

- **Numerical Columns:** For numerical columns, missing values were imputed using the median of the respective columns. The median is a measure of central tendency, especially in cases where the data may contain outliers. This method was chosen because it does not get affected by extreme values, ensuring a more stable and representative imputation. This method is effective in maintaining the distribution of the data and preventing the introduction of any new biases that could arise from using the mean or mode.

- **Categorical Columns:** For categorical columns missing values were filled with a placeholder value 'Unknown'. This approach helps in retaining all data entries without the need for deletion, which could potentially reduce the size and hence the diversity of the data. This approach prevents the loss of data and ensures that every record is preserved. Using a placeholder also helps in later stages, such as model training, where it is essential to encode these categorical variables into numerical format. This ensures that all categories, including the 'Unknown' placeholder, are accounted for, thus preventing errors during K means processing.

Data Standardisation: Standardisation is a key step in preprocessing, particularly when different features are on varying scales. This step is crucial for the choice of algorithm (K-Means clustering), as it relies on distance metrics. Standardisation ensures that each feature contributes equally to analysis, rather than allowing some features to dominate. The ‘StandardScaler’ from ‘scikit-learn’ was used to transform the data. It is particularly crucial for clustering, where the scale of the data can significantly influence the results.

By addressing missing values and standardising the data, the foundation is built for applying K-means. These steps are crucial as they enhance the reliability and accuracy of the entire process.

3.4.2 Transfermarkt Data for RFR Model

Missing Values: The dataset contained missing values in several key columns.

- **‘valueeur’**: 2,153 missing values
- **‘wageeur’**: 1,848 missing values
- **‘clubjoineddate’**: 12,588 missing values
- ‘clubcontractvaliduntilyear’: 1,865 missing values
- **‘releaseclauseeur’**: 59,299 missing values

For numerical columns, missing values were imputed using the median of the respective columns. Median imputation was chosen again because it is resilient against outliers, providing a more reliable central tendency compared to the mean. For categorical and date columns such as the ‘club joined date’ column, missing values were filled with a placeholder date '2018-01-01'. This placeholder was used to maintain data continuity without making assumptions about missing dates. Similarly, for the ‘club contract valid until year’ column, missing values were filled with '0' to indicate unknown or non-applicable contract years.

3.5 Feature Selection

Feature selection is another step in the data preparation process, particularly for clustering tasks such as K-means and RFR. This step involves identifying and selecting the most relevant features that contribute to the analysis, ensuring the model's efficiency and accuracy. In this section, steps taken to refine the dataset for optimal clustering performance are outlined. The focus is on player performance metrics and discarding non-essential or redundant data.

3.5.1 Feature Selection for K-Means Model

Dropping Off-Field Features: Initially, a list of columns deemed irrelevant to the analysis was identified and removed from the dataset. The exclusion of these columns is based on their lack of contribution to evaluating on-field performance, which is critical for scouting and recruitment. While some features provide context or nuanced details about players, they contained too many missing values. Although variables could potentially illustrate a player's versatility, they primarily contain subjective data, as they include metrics for every position on the field for each player, even if certain players have never actually played in those positions.

Selection of Key Features: After data cleaning, a comprehensive set of key features was selected for K-means clustering, aligning with the goal of identifying players with similar performance levels. These features cover a broad spectrum of player attributes that are important for evaluating individual and positional strengths, making them highly relevant to the objective.

- Player Identifiers and Basic Info: ‘playerid’, ‘shortname’, ‘longname’. These identifiers are crucial for distinguishing players within the dataset and referencing specific individuals in the results.

- Performance Metrics: Metrics such as ‘overall’, ‘potential’, ‘pace’, ‘shooting’, ‘passing’, ‘dribbling’, ‘defending’, and ‘physic’ are core indicators of a player's on-field abilities. They provide an overview of players’ skills and potential, which is essential for comparing players across different positions.
- Player Attributes: Attributes like ‘weakfoot’, ‘skillmoves’, and ‘workrate’ offer additional context that can influence players’ performance and role within a team. These factors are important for understanding their physical profile and technical skills.

These features were carefully chosen for their relevance to various aspects of player performance, ensuring a holistic analysis:

- Attackers: Metrics such as ‘shooting’ and ‘attackingfinishing’ highlight offensive capabilities and goal-scoring potential.
- Midfielders: Attributes like ‘passing’, ‘dribbling’, ‘skill dribbling’, ‘skilllong passing’, and ‘mentality vision’ focus on creativity, control, and play-making abilities.
- Defenders: Key defensive metrics include ‘defending marking awareness’, ‘defending standing tackle’, and ‘defending sliding tackle’, which are essential for assessing defensive prowess.
- Goalkeepers: Specific attributes such as ‘goalkeeping diving’, ‘goalkeeping handling’, and ‘goalkeeping reflexes’ are essential for evaluating goalkeeping skills.

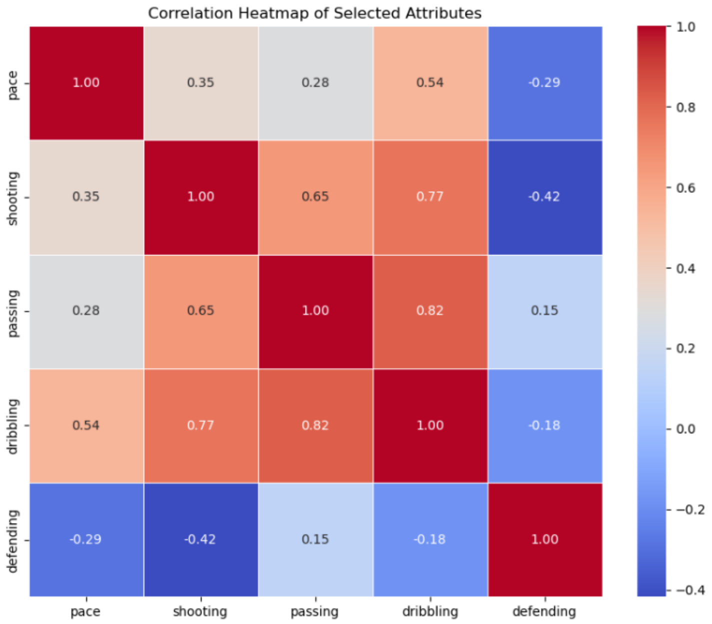
Additionally, universal attributes are included. These attributes provide an overview of each player's capabilities, regardless of their primary position, and are crucial for understanding the overall versatility and adaptability of players.

This extensive selection of features ensures that all key aspects of players’ performance and potential are captured, allowing for a detailed and nuanced clustering analysis. By incorporating metrics that cover all areas of the pitch and various types of player performance, the model can effectively identify similar players, facilitating targeted recruitment and talent identification.

To prepare the dataset for clustering, categorical features such as preferred foot were converted into numerical format using one-hot encoding. This process was essential to include these variables in clustering analysis effectively, allowing the model to recognise different categories as distinct features.

Finally, to understand relationships between selected features, a correlation matrix was computed for the most obvious positional metrics like defending. The resulting correlation heatmap provides insights into how these attributes interrelate, which helps in understanding redundancies in the data.

Figure 3.6 Correlation matrix for Positional Features



This feature selection and preparation phases builds on earlier steps of data cleaning and EDA, as discussed in Sections 3.3 and 3.4.

3.5.2 Feature Selection for RFR Model

Selection of Key Features: The first step in this process was to focus on attributes directly impacting player valuation while excluding redundant data. The following features were chosen for this analysis.

1. **Age**: This feature is significant as it influences a player's market value.
2. **Wage**: Wage is an indicator of players' perceived weekly value by clubs and is often correlated with market value.
3. **Club Contract Valid Until Year**: This indicates the remaining duration of a player's contract.

4. **Release Clause**: This specifies a set price for which a player can leave their club, providing a direct measure of their market valuation.

These features were selected for their direct relevance to assessing player value, ensuring that the model could make accurate predictions based on the most critical aspects of a player's profile.

The target variable for the model will be ‘value eur’, representing the estimated market value of the players. This value will be used in the RFR to predict market values based on the selected features.

Feature selection focused on retaining attributes that are most indicative of players’ market value. By carefully choosing these features, the model can accurately estimate player values, providing financial in recruitment and scouting.

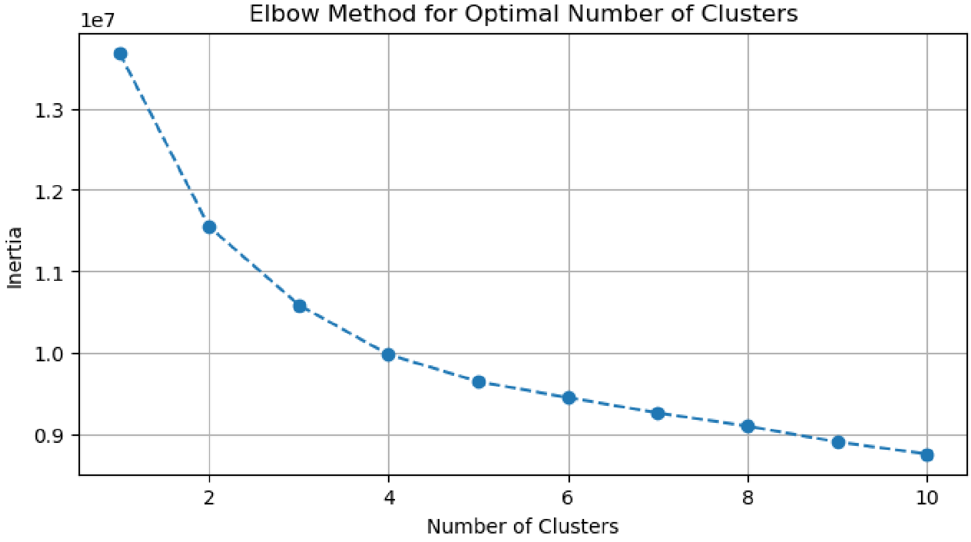
3.6 Model Building

In this section, the process of constructing and training K-means to categorise players into distinct clusters based on their performance metric and RFR to estimate their values is outlined.

3.6.1 K-means Model

Optimal Number of Clusters: The first step in the clustering process was to determine the optimal number of clusters. This was achieved using the Elbow Method. The method involves plotting the inertia (a measure of how internally coherent the clusters are) against the number of clusters. The point at which the decrease in inertia begins to level off, known as the "elbow," indicates the optimal number of clusters. This balance between minimising inertia and preventing overfitting allows for more meaningful segmentation of the data.

Figure 3.7: Elbow Graph Plot



The plot shows the elbow appears to be 3. This is because after 3 clusters, the decrease in inertia starts to become less pronounced, indicating diminishing returns on the added clusters.

Application of K-Means Clustering: With the optimal number of clusters identified as 3, the K-Means clustering algorithm was applied to the dataset. K-means clustering is widely used for partitioning a dataset into K-distinct subsets. Each data point is assigned to a cluster with the nearest mean. This method is particularly suitable for the dataset as it effectively handles diversity and range of player performance metrics.

These clusters represent different archetypes of players, as determined by the features considered in the model. The clustering resulted in a new column in the dataset, 'cluster_optimal', indicating an assigned cluster for each player.

Cluster Categorisation: To understand the nature of each cluster, a cluster summary analysis was conducted. This involved calculating average values of key numeric attributes within each cluster. The summary provided insights into typical characteristics of players in each group, allowing for a deeper understanding of common traits within each cluster. This analysis helps in identifying what differentiates each cluster, such as variations in average pace.

Figure 3.8 Average Attacking Finishing Per Cluster

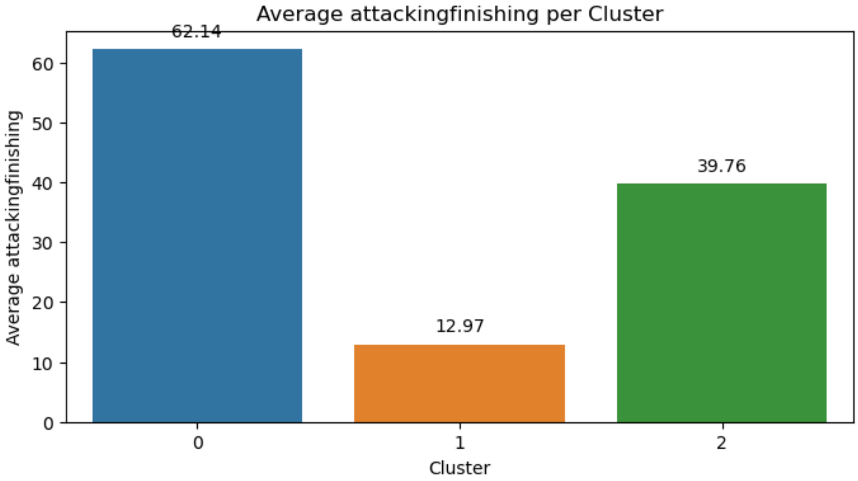


Figure 3.7 depicts the average values of "attacking finishing" metric for three clusters. Here’s a summary of the chart.

- **1st Cluster:** It has the highest average attacking finishing score, approximately 62.14.
- **2nd Cluster:** It has the lowest average score, around 12.97.
- **3rd Cluster:** It has moderate average score, approximately 39.76.

This visualisation is useful for identifying groups of players with different levels of attacking finishing abilities, which can help recruitment teams focus on players based on their scoring potential.

Figure 3.9: Average Defending Marking Awareness Per Cluster

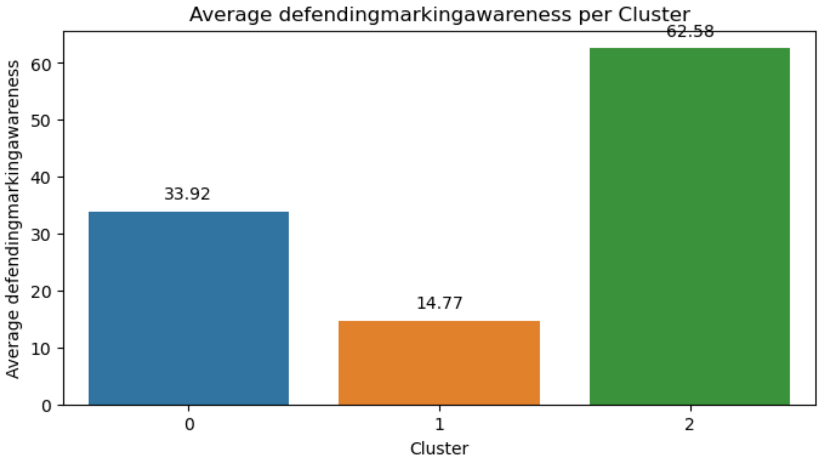


Figure 3.8 indicates the average values of the "defending marking awareness" metric for three different clusters as follows.

- **1st Cluster:** This group has an average defending marking awareness score of approximately 33.92.
- **2nd Cluster:** The lowest average score among the clusters, with a value of around 14.77.
- **3rd Cluster:** The highest average score, approximately 62.58.

This chart helps to categorise players based on their defensive skills, particularly in marking and awareness. Such insights are valuable for recruitment teams to identify players with specific defensive strengths or weaknesses.

Figure 3.10: Average Goalkeeping Reflexes Per Cluster

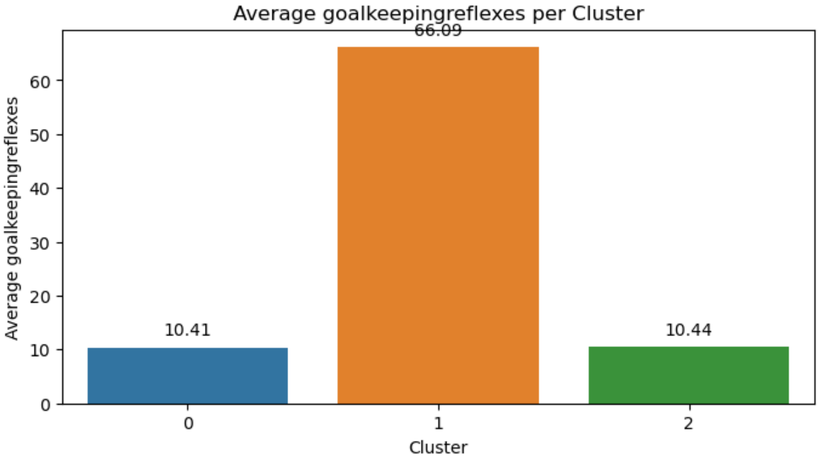


Figure 3.9 shows the average values of "goal keeping reflexes" metric across three clusters provided as follows.

- **1st Cluster:** This cluster has an average goalkeeping reflex score of approximately 10.41.
- **2nd Cluster:** Significantly higher than other clusters, with an average score of around 66.09.
- **3rd Cluster:** Similar to 1st Cluster, with an average score of approximately 10.44.

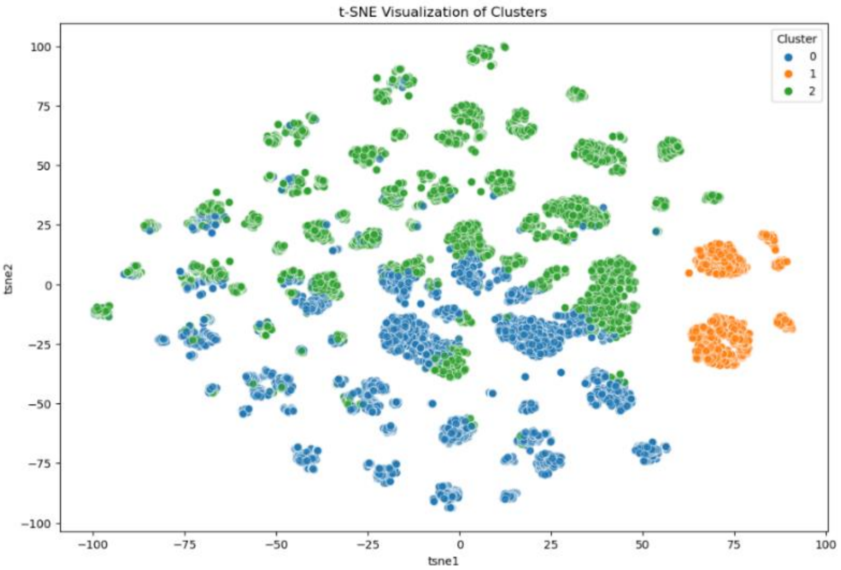
The 2nd Cluster stands out as having players with significantly higher reflexes, likely indicating specialist goalkeepers. Recruitment teams can use this information to identify key players with strong goalkeeping reflexes or potential areas for development in other clusters.

The charts reveal that clustering algorithm effectively categorises players into distinct groups:

- **1st Cluster:** It represents attacking players, with a high average score of 62.14 in attacking finishing.
- **2nd Cluster**: It clearly indicates goalkeepers, shown by a high average score of 66.09 in goalkeeping reflexes.
- **3rd Cluster**: It identifies defensive players, with an average score of 62.58 in defending marking awareness.

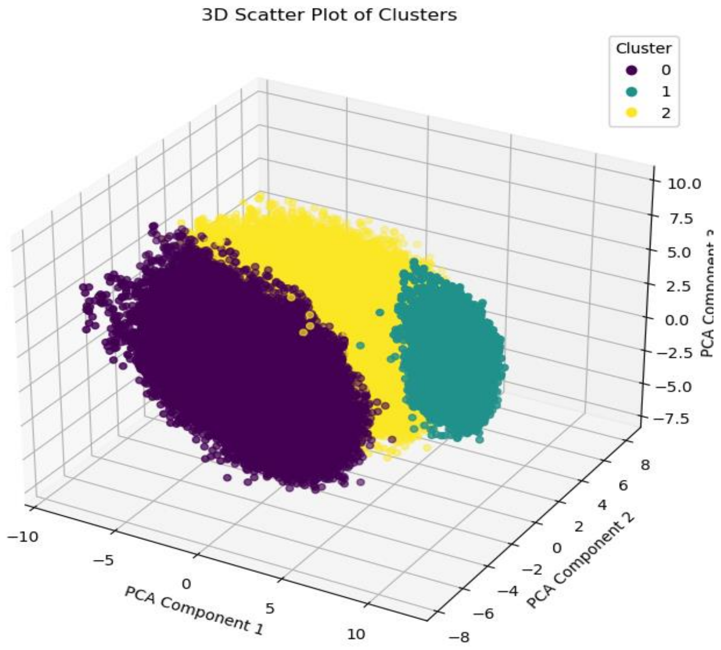
Dimensionality Reduction and Visualisation: For better visualisation of clusters, dimensionality reduction techniques were employed. Principal Component Analysis (PCA) was first used to reduce the data to three dimensions, facilitating easier handling and interpretation. Following PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to further reduce dimensions to two, making it possible to visualise clustering results in two-dimensional space. This step visualises the distribution and separation of clusters.

Figure 3.11: t-SNE Cluster Plot.



To further enhance understanding of cluster distributions, 3D scatter plot was generated using the first three principal components. This plot provides a comprehensive visual representation of player clusters in 3D space, highlighting distinct grouping of players based on their attributes. The clusters were color-coded for clarity, with each colour representing different cluster. The 3D plot not only confirms the separation observed in 2D t-SNE visualisation, but also offers deeper insights into spatial relationships between clusters.

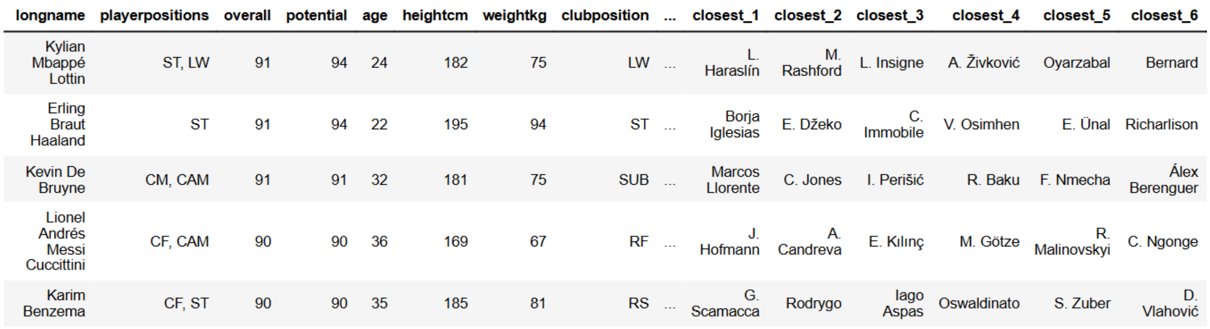
Figure 3.12 PCA Cluster Plot



Identification of Similar Players: The final part of the analysis focused on identifying similar players within the dataset. Using t-SNE transformed data, players were grouped based on their proximity in reduced-dimensional space. The Euclidean distance between players was used to find those with similar profiles, identifying clusters with comparable attributes and performance metrics.

This similarity analysis provides practical applications, such as finding potential recruits with similar profiles to targeted players, scouting for talent, and understanding player development paths. The identification of the ten closest players for each individual, were saved for further exploration and utilisation.

Table 3.8 Similar Players Dataframe



longname	playerpositions	overall	potential	age	heightcm	weightkg	clubposition	...	closest_1	closest_2	closest_3	closest_4	closest_5	closest_6
Kylian Mbappé Lottin	ST, LW	91	94	24	182	75	LW	...	L. Haraslín	M. Rashford	L. Insigne	A. Živković	Oyarzabal	Bernard
Erling Braut Haaland	ST	91	94	22	195	94	ST	...	Borja Iglesias	E. Džeko	C. Immobile	V. Osimhen	E. Ünal	Richarlison
Kevin De Bruyne	CM, CAM	91	91	32	181	75	SUB	...	Marcos Llorente	C. Jones	I. Perišić	R. Baku	F. Nmecha	Álex Berenguer
Lionel Andrés Messi Cuccittini	CF, CAM	90	90	36	169	67	RF	...	J. Hofmann	A. Candreva	E. Kılınç	M. Götze	R. Malinovskyi	C. Ngonge
Karim Benzema	CF, ST	90	90	35	185	81	RS	...	G. Scamacca	Rodrygo	Iago Aspas	Oswaldinato	S. Zuber	D. Vlahović

The K-Means building phase concluded with effective identification of clusters and implementation of methods to find similar players within those clusters.

3.6.2 Random Forest Regression Model

This section details the construction and training of the RFR model to estimate the market value of football players. Utilising the selected features to predict player valuations accurately.

Train and Test Split: To ensure the model's reliability and generalisability, the dataset was split into training and testing sets. The split was configured to allocate 80% of the data for training and 20% for testing. This 80-20 split is a common practice that provides a balanced approach by allocating sufficient data for training while reserving a reasonable portion for evaluating its performance. This split helps in assessing how well the model generalises to new, unseen data, which is crucial for its practical applicability.

Model Training: The RFR model was chosen for its ability to handle both linear and non-linear relationships. The training process involved the model learning from provided features to predict the target variable (player value). The Random Forest algorithm constructs multiple decision trees during training and outputs mean prediction of individual trees, enhancing prediction accuracy and reducing overfitting.

Estimating Player Valuation: The trained model was then used to estimate market value for all players, including those not part of the test set. The results were added to the dataset as a new column. This column provides a calculated estimation of each player's market value based on the learned patterns.

Table 3.9: Estimated Player Value

Long name	Value euro	Estimated value euro
Kylian Mbappé Lottin	181500000	174720000
Erling Braut Haaland	185000000	175270000
Kevin De Bruyne	103000000	100965000
Lionel Andrés Messi Cuccittini	41000000	37735000
Karim Benzema	51000000	48900000

Table 3.13 shows the estimated and actual values. The application can support strategic decisions in player acquisition, valuation, and overall team management.

3.7 Model Evaluation

3.7.1 K-Means Model Evaluation

Model evaluation is a crucial step in validating the effectiveness of clustering results. Given the unsupervised nature of the K-Means algorithm, supervised evaluation metrics are not applicable. Instead, three metrics were tailored to perform clustering tasks, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Each metric offers unique insights into the quality and characteristics of clustering, helping to assess the coherence and separation of clusters formed.

Silhouette Score: The Silhouette Score measures the average distance between points within the same cluster compared to the average distance between points in different clusters. A higher score indicates better-defined clusters. In this study, the sampled Silhouette Score of 0.14 indicates a moderate clustering quality, suggesting that while there is some degree of cluster separation, there may still be overlaps among certain players. This result is critical in understanding the clarity of the clustering and indicates potential areas for improvement in the clustering process or data preprocessing.

Davies-Bouldin Index: This index evaluates the average similarity ratio of each cluster with the cluster most similar to it, based on a ratio of within-cluster distances to between-cluster distances. A lower Davies-Bouldin Index indicates better clustering performance. The score of 2.33 obtained in my evaluation suggests moderate clustering quality, where clusters are not ideally distinct, but still relatively well-separated. This metric is particularly important as it helps in identifying how distinct the clusters are, which is vital for identifying similar players.

Calinski-Harabasz Index: This index considers the ratio of the sum of between-cluster dispersion to within-cluster dispersion. A higher score generally indicates better-defined clusters. In this evaluation, it yielded a score of 1460.60, indicating a reasonable degree of separation within clusters. This metric is important as it provides a balance between the separation of clusters, ensuring that the model does not create overly complex groupings.

3.7.2 Random Forest Regression Model Evaluation

The evaluation involves assessing the RFR model's predictive accuracy using two key metrics Mean Absolute Error (MAE) and the R-squared (R^2) score. These metrics provide a comprehensive view of how well the model's predictions match the actual vaue.

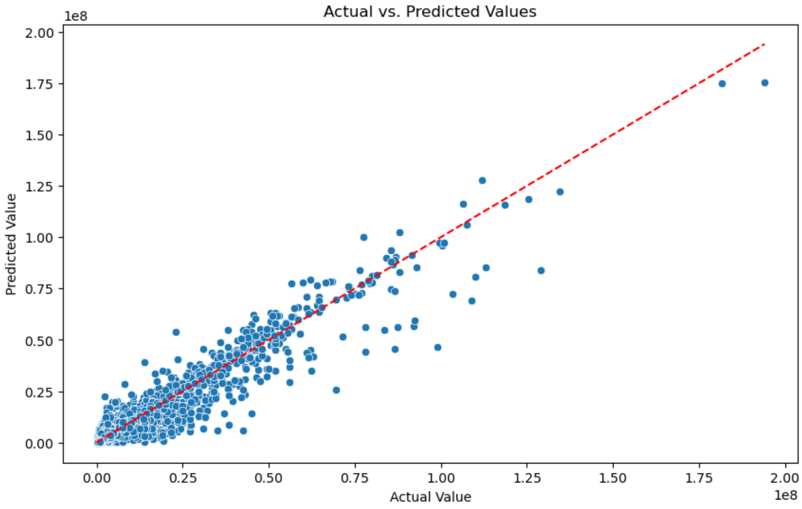
MAE: MAE is a measure of the average magnitude of errors between predicted and actual values. It is calculated as the average of the absolute differences between predicted and actual values. In this context, MAE helps quantify how far, on average, the estimated player values deviate from their actual market values. A lower MAE indicates better predictive accuracy. For RFR, the MAE was calculated to be approximately 480,858.25 euros. This value represents the deviation of estimated player values from their true market values. While this error margin may seem large, it is important to consider the scale of player valuations, where values can vary significantly.

R^2 Score: The R^2 score measures the proportion of variance in the dependent variable (player value) that is predictable from independent variables (features such as age). It indicates the model's explanatory power, with values ranging from 0 to 1. An R^2 score closer to 1 suggests that a larger proportion of variance in the target variable is accounted for by the model.

The R^2 score for this model was found to be approximately 0.93. This high R^2 score indicates that the model explains about 93.4% of the variance in player valuations. This suggests that selected features and model are highly effective in capturing factors that influence a player's market value. A high R^2 score is particularly valuable in financial and market prediction contexts, as it implies that the model is well-calibrated to the underlying data.

To visualise the Model’s accuracy, a scatter plot was created comparing the actual values against the predicted values.

Figure 3.13 RFR (comparison between Actual and Predicted Values)



This evaluation highlights the effectiveness of the model and underscores its potential applications in sports analytics, particularly in player valuation and scouting. The results provide a clear indication of the model's strengths and areas for potential refinement, guiding future efforts in model improvement and feature selection.

3.8 Ethical Considerations

In any data-driven project, particularly those involving personal or sensitive data, it is important to address confidentiality, security, and responsible data usage. This project aims to find similar football players to any given player by clustering players based on various performance metrics and attributes as well as estimating their values, which entails handling sensitive information that may relate to individuals' personal and professional lives.

- **Data Security:** Protecting data from unauthorised access and breaches is a critical ethical responsibility. During the project, data storage and processing were conducted in secure environments with restricted access. Passwords were used where necessary to ensure data at rest and in transit were safeguarded against potential security threats. Regular access controls were put in place to monitor and prevent unauthorised use or data leakage.

- **Responsible Data Usage:** Beyond confidentiality and security, ethical data usage involves using data in a manner that is fair, transparent, and aligned with intended purposes. In this project, data was solely used for developing player similarity and valuation tools. Any insights or outcomes from the analysis were communicated transparently, ensuring that interpretations of data were accurate and not misleading. The project adhered to ethical guidelines and standards set by relevant authorities, ensuring that data was not used for any discriminatory practices or decisions that could negatively impact individuals or groups.

- **Compliance:** It is also important to consider legal and ethical frameworks governing data usage. While data was sourced from publicly available datasets, the project complied with data protection regulations including GDPR.

These measures are essential not only for protecting individuals represented in the dataset but also for maintaining the integrity and credibility of research outcomes.

3.9 Conclusion

In this chapter, I have detailed in-depth the methodological process for the research to ensure that the results are firmly grounded in a well-structured approach. The comprehensive steps taken not only bolster the credibility of the research but also successfully achieve the project's goal of developing a robust player similarity model using K-means and estimating player valuation using RFR.

CHAPTER 4: RESULT INTERPRETATION

This chapter provides an in-depth interpretation of the results from K-Means clustering and RFR models. The focus is on understanding data correlations, feature significance, model effectiveness, and accuracy. Each sub-section addresses specific findings and their implications for broader objectives of streamlining scouting and recruitment processes.

4.1 Correlation and Feature Engineering

In the preprocessing phase, correlation analysis was conducted to identify and understand relationships between various player performance metrics. Key features such as ‘shooting’, ‘passing’, ‘dribbling’, and ‘defending’ were central to the analysis. It was crucial to examine potential for multicollinearity, particularly among features that could capture overlapping aspects of player abilities. The correlation heatmap showed that 'dribbling' was strongly correlated with both 'passing' and 'shooting', while 'defending' had moderate negative correlations with 'shooting' and 'pace'. These correlations suggest potential multicollinearity, especially among 'dribbling', 'shooting', and 'passing', which could affect linear models. Despite these correlations, selected features provide valuable information about player performance and are crucial for analysis. This was expected, as 'dribbling', 'shooting', and 'passing' are attacking traits. So, midfielders and strikers should possess a bit of these skills. According to the heatmap, specific performance metrics are tailored to each position on the field. Consequently, positional metrics like ‘attacking finishing’ were selected for strikers, ‘skill long passing’ for midfielders, ‘defending sliding tackle’ for defenders, and ‘goalkeeping reflexes’ to ensure players were clustered according to their distinct abilities and attributes.

For K-Means clustering analysis, feature engineering involved several key steps. Missing values in numerical features like ‘shooting’, ‘physic’, ‘defending’, ‘dribbling’, and ‘passing’, were imputed using median to ensure robustness against outliers. Categorical columns such as ‘club position’, ‘preferred foot’, and ‘work rate’ had missing values replaced with the placeholder 'Unknown' to maintain data completeness.

Standardisation was performed using ‘StandardScaler’, which removed mean and scaled features to unit variance. A necessary step for K-Means clustering, which relies on distance metrics, to ensure that all features contributed equally to the clustering process.

Irrelevant features, including ‘clubjersey number’, and ‘nation team id’, were excluded. Key features retained for clustering included ‘overall’, ‘potential’, ‘pace’, ‘mentality vision’, ‘defending marking awareness’, ‘goalkeeping diving’, ‘age’, and ‘work rate’.

In Transfermarkt dataset for RFR, missing values in important columns such as ‘value eur’, ‘wage eur’, ‘club joined date’, ‘club contract valid until year’, and ‘release clause eur’ were addressed using median imputation or placeholders. This ensured that dataset remained intact and useful for regression analysis.

4.2 Hyperparameter Tuning and Model Optimisation

4.2.1 K-Means Clustering for Player Similarity

In determining optimal number of clusters for K-Means clustering, Elbow method was utilised. The graph suggested that optimal number of clusters could be 2, 3, or 4. First clustering was performed with four clusters, which resulted in Silhouette Score of 0.12, Davies-Bouldin Index of 2.53, and Calinski-Harabasz Index of 891.26.

Next, a clustering with three clusters was performed, which yielded improved metrics, with Silhouette Score of 0.14, Davies-Bouldin Index of 2.33, and Calinski-Harabasz Index of 1460.60. This configuration appeared promising. Finally, a two-cluster solution was performed, which provided a higher Silhouette Score of 0.2. However, this configuration was less effective. The two-cluster solution grouped all goalkeepers into one cluster (Cluster 0), which was appropriate, but it lumped all outfield players into a single cluster (Cluster 1) without distinguishing between defensive, attacking, and midfield roles.

Therefore, despite slightly lower Silhouette Score, three-cluster solution is preferred. This configuration provided a more meaningful categorisation of players based on their positions and attributes, making it a better choice for our analysis.

4.2.2. Random Forest Regressor for Player Valuation

For estimating player values, RFR was employed using its default hyperparameters. The default settings include 100 trees in the forest and a maximum depth of None, allowing the trees to grow until all leaves are pure or contain fewer than the minimum samples required to split. The number of features considered for each split is set to auto, which in this case means the square root of the total number of features ($\sqrt{4} = 2$). The model uses bootstrapping by default and does not parallelise computation. It was trained using four specific features, which are age, wage in euros, club contract validity until year, and release clause in euros. The evaluation of this model yielded Mean Absolute Error of 480,858.25 and R² Score of approximately 0.93, indicating a high level of predictive accuracy and a strong fit to the data.

Given these promising results, it is decided to retain default hyperparameters for Random Forest Regressor. The model's effectiveness with its default configuration suggests that it is well-suited for given dataset and task, obviating the need for further tuning.

4.3 Model Performance Evaluation

4.3.1 K-Means Clustering Performance

The evaluation of the clusters utilised three main metrics to determine the quality and clarity of the clustering.

The Silhouette Score achieved was 0.14, indicating moderate clustering quality. The score also reflects that players within the same cluster are reasonably similar to each other in terms of features considered, but there are some players whose attributes could make them fit into more than one cluster.

Davies-Bouldin Index for clustering was 2.33. This value indicates that average similarity ratio within clusters is moderate, with room for clearer separation between clusters.

Calinski-Harabasz score of 1460.009 indicates a good balance between the compactness and separation of clusters. This index reflects that variance between clusters is considerably greater than variance within clusters, affirming the effectiveness of feature selection and clustering approach.

4.3.2 Random Forest Regressor Performance

RFR model validation involved splitting dataset into training and testing sets in an 80-20 ratio. This split ensured that model was trained on a substantial portion of data while still being evaluated on a separate, unseen dataset. The model performance was evaluated using two primary metrics, which are MAE and the R² score.

The MAE for the Random Forest model was approximately 480,858.25 euros. This error measure represents the average deviation of the predicted player market values from their actual market values. Considering the wide range of player valuations, with stars like Kylian Mbappé valued over 180 million euros and lesser-known players valued much lower, this MAE is reasonable. It suggests that while the model is quite accurate, there is variability, particularly at the extremes of the valuation spectrum where market influences and individual player characteristics might not be fully captured by available data.

The R^2 score achieved was approximately 0.93. This high score indicates that model explains 93.4% of the variance in player market values based on selected features. Such a high explanatory power underscores the importance and relevance of the features used, such as ‘age,’ ‘wage,’ ‘contract length’.

4.4 Conclusion

The detailed analysis and interpretation of these results highlight the critical aspects of feature selection, model tuning, and evaluation. The primary goal of this research was to create a comprehensive model that enables recruitment teams to accurately identify players who exhibit similar characteristics and performance levels to any given player and estimate their value. This was achieved by developing a robust player similarity model using K-means clustering and a RFR model.

CHAPTER 5: CONCLUSION AND DISCUSSION

5.1 Interpretation of Outcomes in Relation to Research Questions

RQ1

How can advanced data analysis be utilised to enhance traditional scouting methods and improve player scouting and recruitment in football?

The research demonstrated that advanced data analysis techniques, such as K-Means clustering and RFR, can streamline player scouting and recruitment. This enhancement is achieved by leveraging these models to first narrow down a vast pool of players based on specific performance attributes. K-Means clustering model groups similar players, thus enabling scouts to focus on a smaller, more relevant group that matches club's specific needs. For instance, if a club is looking to find a player similar to Kylian Mbappé, the model can identify a list of players who exhibit similar attributes and keep the same level of performance.

Subsequently, the RFR model estimates the market value of these identified players, allowing clubs to prioritise potential signings based on their budget constraints. This approach saves considerable time and resources that would otherwise be spent on extensive scouting trips and subjective assessments. Instead, clubs can focus on a select few players who not only fit their tactical requirements but also fall within their financial means. The final step involves traditional methods of scouting, such as live game observations, which are now more targeted and efficient. This integration of data-driven models into traditional scouting practices enhances the accuracy and efficiency of the recruitment process, enabling clubs to make more informed decisions with fewer resources.

Subsequently, RFR model estimates market value of these identified players, allowing clubs to prioritise potential signings based on their budget constraints. This approach saves considerable time and resources that would otherwise be spent on extensive scouting trips and subjective assessments. Instead, clubs can focus on selecting a few players who not only fit their tactical requirements but also fall within their financial means. The final step involves traditional methods of scouting, such as live game observations, which are now more targeted and efficient. This integration of data-driven models into traditional scouting practices enhances the accuracy and efficiency of recruitment processes, enabling clubs to make more informed decisions with fewer resources.

RQ2

How can a data analysis model enhance success in football recruitment?

In football, spending large sums on player recruitment does not necessarily guarantee success. A striking example is Chelsea FC, which invested over £1 billion in new players since May 2022 to finish in 12th and 10th place in the English Premier League. This outcome underscores the notion that spending big money is no substitute for spending wisely. A data-driven approach to player recruitment, utilising advanced analytics can significantly increase the likelihood of success by providing a more accurate assessment of players' potential impact.

K-Means model utilised key performance and position-specific metrics to comprehensively evaluate a player's capabilities when compared to any given player. These metrics were chosen based on their strong correlation with success metrics such as overall and potential ratings, ensuring they are relevant and distinct enough to avoid multicollinearity. This helps in raising recruitment success bar in football.

RQ3

How can player similarity and valuation model help clubs with limited financial resources or those constrained by FFP?

The player similarity and valuation models offer significant advantages for clubs with limited financial resources or those constrained by FFP regulations. RFR, with high R^2 score of 0.93, provided accurate player valuations based on variables like 'age', 'wage', 'contract length', and 'release clause'. This high accuracy helps clubs make informed decisions about player acquisitions, ensuring they do not overspend and remain compliant with FFP rules. The models also enable clubs to identify undervalued or affordable talent that fits their budget and tactical needs, making them valuable tools for strategic planning and financial management in the transfer market.

5.2 Limitations and Challenges

The effectiveness of any data-driven model is highly dependent on the quality and comprehensiveness of the input data. In the context of this study, two primary models were employed, K-means clustering model for identifying player similarities and RFR model for estimating player values. Both models face distinct limitations and challenges related to data accuracy and availability.

5.2.1 Data Accuracy and Descriptiveness in Player Similarity Model

K-means clustering model groups players based on performance metrics to identify similar players. However, accuracy of these groupings heavily relies on descriptiveness of available data. Traditional metrics like ‘shooting’ are often insufficiently descriptive, as they fail to capture nuances of a player's style and positional attributes. This lack of specificity can lead to inaccurate player groupings and misguided recruitment decisions. To address this, more granular and position-specific metrics are necessary. Advanced metrics like Expected Goals (xG) provide nuanced insights into specific aspects of player performance, such as a striker's ability to convert chances. xG measures the quality of scoring opportunities a player encounters, offering a deeper understanding of a player's finishing skills beyond mere goal counts.

5.2.2 Challenges in Player Valuation and Data Availability

RFR used for estimating player values also encounters significant challenges, particularly regarding availability and reliability of financial data. Accurate player valuation requires comprehensive data on contractual details, market trends, and individual performance. However, such data especially contractual information, is often private and inaccessible to the public, posing a significant limitation. This secrecy means that the model may have to rely on estimations or outdated data, which can compromise its accuracy.

5.3 Future Work

The current study has laid a solid foundation for improving football recruitment through data-driven models. However, integration of AI for future work can further refine and expand these capabilities.

5.3.1 AI-Generated Scouting Reports

One promising area for future work involves developing an AI-generated scouting report. This feature would provide detailed insights into player performance, strengths, weaknesses, and overall impact on the team. Leveraging technologies such as the Google Generative AI Library, these reports could offer nuanced analyses that go beyond traditional metrics, offering scouts and decision-makers a comprehensive understanding of a player's capabilities and potential fit within a team. This tool could be especially useful for uncovering hidden gems or identifying players who excel in specific tactical roles.

5.4 Conclusion

The integration of advanced data analysis techniques into football scouting and recruitment processes has demonstrated considerable potential to revolutionise traditional practices. This thesis explored the use of K-Means clustering and RFR models to enhance player identification, valuation, scouting, and recruitment, particularly useful for clubs operating under financial constraints such as those imposed by UEFA FFP regulations.

K-Means clustering model successfully grouped players based on specific performance metrics, allowing recruitment teams to identify players with similar attributes and potential. This approach not only streamlines the scouting process but also provides a more objective and comprehensive assessment of players' capabilities, aiding in the identification of suitable replacements or additions to a team. The use of positional metrics further refined these groupings, ensuring that players were categorised accurately according to their on-field roles and contributions.

RFR, applied to Transfermarkt data, provided a robust mechanism for estimating player market values. The model's high R^2 score indicated its strong predictive accuracy, making it an invaluable tool for clubs looking to make financially prudent decisions in the transfer market and adhering to FFP regulations.

This research also highlighted limitations, the need for more granular data and the challenges of accounting for market dynamics and contractual complexities. In conclusion, this thesis contributes to the growing field of sports analytics by demonstrating how data-driven models can be leveraged to optimise player recruitment strategies, ultimately increasing their competitive edge in the ever-evolving world of football.

APPENDIX

Appendix A

Additional Modern Scouting Practices in Football (Improvements in Performance Metrics)

The field of football analytics has led to significant advancements in performance metrics, revolutionising talent scouting and recruitment. One pivotal development is expected goals (xG), which calculates the likelihood of a shot resulting in a goal based on various shot characteristics (Caley, 2015). Unlike actual goals, which can be influenced by randomness and luck, xG offers a more reliable assessment of team and player performance (Brechot & Flepp, 2020). For example, Liverpool FC used an xG model to evaluate Jurgen Klopp's performance at Borussia Dortmund, highlighting his team's underlying strength despite poor league results (Schoenfeld, 2019). In addition to xG, modern football utilises various advanced data collection methods and performance metrics. Technologies like Catapult One and STAT Sports provide detailed player tracking data, including distance covered, speed, acceleration, and heart rate, providing a comprehensive view of a player's physical performance. This data assists scouts and coaches in evaluating a player's athleticism and endurance, which are crucial factors in modern football.

In 2021, the Premier League entered a partnership with Oracle Cloud to harness the extensive data collected from every match since its inception. This data provides valuable insights into team strategies, player performances, and match outcomes (Ariel Kelman, 2021). Opta Sports, is another prominent data provider, which compiles comprehensive statistics on passes, tackles, and interceptions, enabling in-depth analysis of player contributions and team dynamics. These advanced metrics and technologies underpin modern scouting and recruitment practices, surpassing traditional methods of observation, subjective judgment, and intuition. By leveraging data-driven insights, clubs can pinpoint players with specific attributes that align with their tactical requirements and reduce reliance on potentially biased human observations. For example, metrics such as passes completed under pressure, progressive carries, and expected assists (xA) offer nuanced views of a player's technical and creative abilities. These metrics are invaluable for scouting midfielders and attackers who need to excel in high-pressure situations and create goal-scoring opportunities. Similarly, defensive metrics like blocks, clearances, and defensive actions per 90 minutes assist scouts in identifying solid defenders who can effectively protect their team's goal.

The integration of these advanced metrics into the scouting process represents a substantial improvement over the traditional method. By utilising objective data, clubs can make more informed decisions, minimising the risks associated with player recruitment. This data-driven approach ensures a more consistent and reliable talent assessment, aligning with modern football clubs' strategic goals.

Appendix B

Further Research Directions and Future Works

This section outlines additional research directions in areas of player scouting, team building, and creation of a user-friendly web application**.**

Team Builder and Player Recommendations

An extension of this work could involve developing a "Team Builder" feature. This system would recommend players to replace or complement current squad members based on various features like play style, league experience, skills, and age. By utilising similarity analysis methods, such as cosine similarity, the system can identify players with comparable attributes and suggest potential replacements or additions. This would enable clubs to make informed decisions when looking to strengthen specific positions, ensuring that new signings align with the team's tactical needs and long-term strategy.

Interactive Web Application

To facilitate accessibility and usability, creation of interactive web application using tools like Streamlit could be highly beneficial. This platform would present data and insights in an intuitive and interactive format, allowing users to easily search for players, view scouting reports, and explore player similarities. The web application could be built using Requests for handling HTTP requests and more libraries, while dotenv will be used for managing environment variables securely.

This application could serve as a comprehensive tool for scouts, coaches, and analysts, offering real-time data updates and customised reports.

The outlined future work has the potential to significantly enhance capabilities and utility of data-driven models in football recruitment. By integrating advanced analytics, AI-generated insights, and user-friendly interfaces, these developments can provide invaluable support to football clubs in making strategic, informed scouting and recruitment decisions.

REFERENCES

Aalbers, B., & Van Haaren, J. (2018). Distinguishing Between Roles of Football Players in Play-by-Play Match Event Data. *Paper presented at the International Workshop on Machine Learning and Data Mining for Sports Analytics.*

AlShboul, R., Syed, T. Q., Memon, J. and Khan, F. M. (2017). Automated player selection for a sports team using competitive neural networks.

Apatean, A., Szakacs, E. and Tilca, M. (2017). Machine-learning based application for staff recruiting.

Baroncelli, A., & Lago, U. (2016). Italian Football. *Journal of Sports Economics, 7*(1), 13-28. doi:10.1177/1527002505282863.

Bekkers, J., & Dabadghao, S. (2019). Flow Motifs in Soccer: What can passing behavior tell us? *Journal of Sports Analytics,* 5(4), 299-311.

Biermann, C. (2016). Moneyball im Niemandsland-Midjyllands Revolution. *11FREUNDE*. Recovered From: Midtjyllands Revolution - 11FREUNDE

Bleaney, R. (2014). Football Manager computer game to help Premier League clubs buy players. *In*.

Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: how to rethink performance evaluation in european club football using expected goals. *Journal of Sports Economics,* 21(4), 335-362.

Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction, *Applied Computing and Informatics* 15(1): 27 – 33. Retrieved From: http://www.sciencedirect.com/science/article/pii/S2210832717301485

Caley, M. (2015). Premier League Projections and New Expected Goals. Retrieved from

https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals

Christensen, M. K. (2009) An eye for talent: Talent Identification and the 'practical sense' of top-level soccer coaches. *Sociology of Sport Journal,* 26, 3, 365-382.

Dellal, A., Chamari, K., Wong, d. P., Ahmaidi, S., Keller, D., Barros, R., . . . Carling, C. (2011). Comparison of physical and technical performance in European soccer matchplay: FA Premier League and La Liga. *European Journal of Sport Science,* 11(1), 51– 59.

Elberse, A. (2013). Ferguson’s formula. *Harvard Business Review,* 91(10), 116-125.

Franck, E. (2014). *Financial Fair Play in European Club Football What is it all about?* Retrieved from https://EconPapers.repec.org/RePEc:zrh:wpaper:328

Franks, I.M. and Miller, G. (1986) Eyewitness testimony in sport. *Journal of Sport Behavior,* 9, 1, 38-45.

Geey, D. (2020). Done Deal: An Insider's Guide to Football Contracts, Multi-Million Pound Transfers and Premier League Big Business. London. *Bloomsbury Publishing PLC.*

Gerrard, B. (2016). Analytics, technology and high performance sport. Critical issues in global sport management, 205–218.

Gerrard, B. (2017). The Role of Analytics in Assessing Playing Talent. In J. Baker et al. (Hrsg.), Routledge Handbook of Talent Identification and Development in Sport (422-431). *Routledge Handbooks Online.* Recovered from: https://doi.org/10.4324/9781315668017.ch30

Hakes, J. K., and Sauer, R. D. (2006). An Economic Evaluation on the Moneyball Hypothesis. *Journal of Economic Perspectives.* 20(3), 173-186. Recovered From: An Economic Evaluation of the Moneyball Hypothesis - American Economic Association (aeaweb.org)

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players’ market values: Accuracy and evaluation attributes of an online community. *Sport Management Review*, 17(4), 484–492.

Hughes, S. (2009) *Secret Diary of a Liverpool Scout.* Trinity Mirror Sport Media Liverpool.

Jannat, M., Chowdhury, S. S. and Akther, M. (2016). A probabilistic machine learning approach for eligible candidate selection.

Jantan, H., Hamidah, Hamdan, A. and Othman, Z. (2010). Human talent prediction in hrm using c4.5 classification algorithm, *International Journal on Computer Science and Engineering 2.*

Johansson, A. and Fahlén, J. (2017) Simply the best, better than all the rest? Validity issues in selections in elite sport. *International Journal of Sports Science and Coaching,* 12, 4, 470-480.

Kerr, A., Barraclough, S., Till. K., and Emmonds, S. (2022). Methodological Approaches to Talent Identification in Team Sports: A Narrative Review. *Sports*, 10(6), 1-16. Recovered From: Sports | Free Full-Text | Methodological Approaches to Talent Identification in Team Sports: A Narrative Review (mdpi.com)

Lewis, M. (2004). Moneyball: The Art of Winning an Unfair Game. *W. W. Norton & Company.*

Li, Y.-m., Lai, C.-y. and Kao, C.-p. (2011). Building a qualitative recruitment system via svm with mcdm approach, Applied Intelligence 35(1): 75–88. *Copyright - Springer Science+Business Media, LLC 2011*. Retrieved From: https://ezproxy.ncirl.ie/login?url=https://search.proquest.com/docview/873356714?accountid

Lund, S. and Söderström, T. (2017) To see or not to see: Talent Identification in the Swedish Football Association. *Sociology of Sport Journal,* 34, 248 -258.

Maguire, K. (2019). The price of football. *Agenda Publishing.*

Mathew, V., Chacko, A. M. and Udhayakumar, A. (2018). Prediction of suitable human resource for replacement in skilled job positions using supervised machine learning, *2018 8th International Symposium on Embedded Computing and System Design (ISED),* pp. 37–41.

rMcHale, I. (2018). Sports business analytics: The past, the present and the future. In S. Chadwick et al. (Hrsg.), *Routledge Handbook of Football Business and Management* (246-257). Routledge.

Müller, O., Simons, A., & Weirunann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, 263(2), 611-624. doi:10.1016/j.ejor.2017.05.005

Nash, C. and Collins, D. (2006) Tacit knowledge in expert coaching. *Quest,* 58, 465-477.

Oracle, 2021. Premier League selects Oracle Cloud Infrastructure to power new advanced football analytics. *Oracle Press Release*.

Retrieved From https://www.oracle.com/uk/news/announcement/premier-league-selects-oracle-cloud-infrastructure-2021-05-06/ [Accessed 28 May 2024].

Pariath, R., Shah, S., Surve, A. and Mittal, J. (2018). Player performance prediction in football game, pp. 1148–1153.

Peeters, T., & Szymanski, S. (2014). Financial fair play in European football. *Economic Policy, 29*(78), 343–390.

Peña, J. L., & Navarro, R. S. (2015). Who can replace Xavi? A passing motif analysis of football players. *arXiv preprint arXiv:1506.07768.*

Plumley, D., Wilson, R., & Ramchandani, G. (2017). Towards a model for measuring holistic performance of professional Football clubs. *Soccer & Society, 18*(1), 16–29.

Radicchi, E., & Mozzachiodi, M. (2016). Social Talent Scouting: A New Opportunity for the Identification of Football Players?. Physical Culture and Sport. *Studies and Research*, 70(1), 28–43. Recovered From: https://doi.org/10.1515/pcssr-2016-0012

Reeves, M., McRobert, A., Lewis, C. and Roberts, S. (2019) A case study of verbal reports for Talent Identification purposes in soccer: A Messi affair! *PLoS ONE,* 14, 11, 1-17.

Rivoire, X. (2011). Arsene Wenger: *The Biography*: Aurum Press Limited.

Rookwood, J. and Buckley, C. (2007) The Olympic soccer tournament. *Journal of Olympic History,* 15, 3, 6-15.

Rossi, G., Semens, A., & Brocard, J. F. (2016). Sports agents and labour markets: evidence from world football. *Routledge*.

Schoenfeld, B. (2019). How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory. Retrieved from https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Schokkaert, J. (2016) Football clubs' recruitment strategies and international player migration: evidence from Senegal and South Africa. *Soccer and Society,* 17, 1, 120-139.

Schumaker, R.P., Solieman, O.K., and Chen, H. (2010) Sports knowledge management and data mining. *Annual Review of Information Science and Technology,* 44, 115-157.

Skandalis, A., Byrom, J., & Banister, E. (2015). Brand scouting: co-creation of value in the Football Manager community. *ACR North American Advances*, 409-414.

Stats Perform. (2024). Retrieved from https://www.statsperform.com/

Szymanski, S. (2015). Money and Soccer: A Soccernomics Guide: Why Chievo Verona, Unterhaching, and Scunthorpe United Will Never Win the Champions League, Why Manchester City, Roma, and Paris St. Germain Can, and Why Real Madrid, Bayern Munich, and Manchester United Cannot Be Stopped: Nation Books.

Taylor, M. (2006) Global players? football, migration and globalisation, c. 1930-2000. *Historical Social Research,* 31, 1, 7-30.

Tranckle, P. and Cushion, C. (2006) Rethinking giftedness and talent in sport. *Quest.* 58, 265-282.

UEFA (2011). Financial Fair Play. *UEFA Direct*. pp. 12–13. Retrieved From UEFA"direct #105 (02.2011)

Williams, A. M. and Franks, A. (1998) Talent Identification in soccer. *Sports, Exercise and Injury,* 4, 159-165.

Wyscout. (2024). Retrieved from https://wyscout.com/