# Improvised Linear Regression model to scout the best Manager for Manchester United

Manikandan Rajagopal[1], Mansurali A[2], Raja P[3] and Harish V[4]

{ manikandan.rajagopal@christuniversity.in[1], mansurali@cut.ac.in[2], rajaperumal@cut.ac.in[3], harish@psgim.ac.in[4] }

CHRIST(Deemed to be University), Bangalore[1], Central University of Tamilnadu, India[2,3], PSG Institute of Mangement, Coimbatore[4]

**Abstract.** The paper presents the efficacy of employing Artificial intelligence in sports analytics. The study envisages on the possibility of employing machine learning methodologies to aid in the decision-making procedure of identifying an appropriate manager for a football club through the utilization of machine learning models. This study is centered on the collection and analysis of data pertaining to managers within the top five leagues in Europe. The data pertaining to the pool of football managers that were available was subjected to pre-processing techniques and afterwards analysed in order to extract valuable insights on their historical performance. Subsequently, a variety of machine learning algorithms, including as clustering, random forest, logistic regression, and improved multiple linear regression, were employed to forecast the optimal manager for a club. The models underwent training using a dataset of historical performance data. Subsequently, their performance was assessed by comparing their predictions against actual outcomes, and metrics such as accuracy and F1 score. This study offers significant contributions to the utilization of machine learning models in addressing practical challenges within the sports business. The results of this study provide valuable insights for organizations encountering comparable difficulties and may serve as a point of reference for future investigations in this domain.

**Keywords:** Football Manager, Machine Learning, Clustering, Random Forest, Logistic Regression, Multiple Linear Regression.

## 1 Introduction

Manchester United Football Club is a renowned professional football organization situated at Old Trafford, Greater Manchester, England. It actively participates in the Premier League, which represents the highest tier of English football. The football club, commonly referred to as "the Red Devils," was established under the name Newton Heath LYR Football Club in 1878. However, in 1902, it underwent a name change and became known as Manchester United. Subsequently, the team relocated to its present stadium, Old Trafford, in 1910. According to the study conducted in[1]. . The club is widely recognized as being among the most renowned and

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343259

accomplished globally, boasting a substantial and devoted following. As detailed in [2],. The previously lucrative club is currently seeing a decline in attendance, resulting in a notable decrease in stadium occupancy. The manager's job is crucial in determining the success of a football club. This paper aims to explore various strategies that can aid in the decision-making process of selecting an appropriate manager for a football club. This study offers significant contributions in understanding the utilization of machine learning models and proposes an improvised linear regression model for addressing real-world challenges in the sports sector. The implications of the study's findings may prove valuable for other organizations encountering comparable obstacles, and could serve as a point of reference for other investigations in this domain.

## 2 Literature Review

The decision-making processes within the football industry were predominantly based on subjective judgements and intuitive judgements, particularly in matters pertaining to player and coach selection. In[3], The utilization of data analysis is facilitated that has helped teams in acquiring a more comprehensive comprehension of player and team performance, According to [4], Machine learning has had a substantial influence on the domain of coach and management selection, representing a critical area of impact. This approach involves assessing objective criteria, including the individual's win-loss record, tactical acumen, and ability to effectively manage players as discussed in [5]. The authors in [6] discussed that the Premier League, La Liga, Bundesliga, Serie A, and Ligue 1 are widely acknowledged as the most competitive and renowned football leagues globally. In [7-8], it is discussed that the clubs participating in these competitions exhibit a perpetual pursuit of enhancing their performance and maintaining a competitive edge. The identification of proficient managers who possess the ability to instill a mindset of triumph inside the team, hence fostering achievements in both domestic leagues and UEFA tournaments [9]. The utilization of data analysis and machine learning in the football sector is expected to expand due to ongoing technological advancements and improved data accessibility, rendering it a captivating and evolving domain[10]. According to [11], the Premier League is home to renowned and highly accomplished football clubs, like Manchester United, Chelsea, Liverpool, and Arsenal. These clubs have established themselves as global powerhouses, consistently attracting elite players and coaches from other countries.

With a huge global following and significant revenue generated from broadcast rights and sponsorship deals, the Premier League has become one of the most important and lucrative football leagues in the world[13]. There has been an increase in football- specific research during the past 20 years. Even though association football (soccer) is the subject of most of this research, publications about other football codes have been steadily rising[14]. There is proof that players who compete at the highest level have anthropometric profiles that are affected by more methodical training and selection. In order to meet match demands and the necessity for position- specific requirements, fitness is being optimized [15].

## 3 Research Methodology

The following process flow diagram in context to the research methodology.

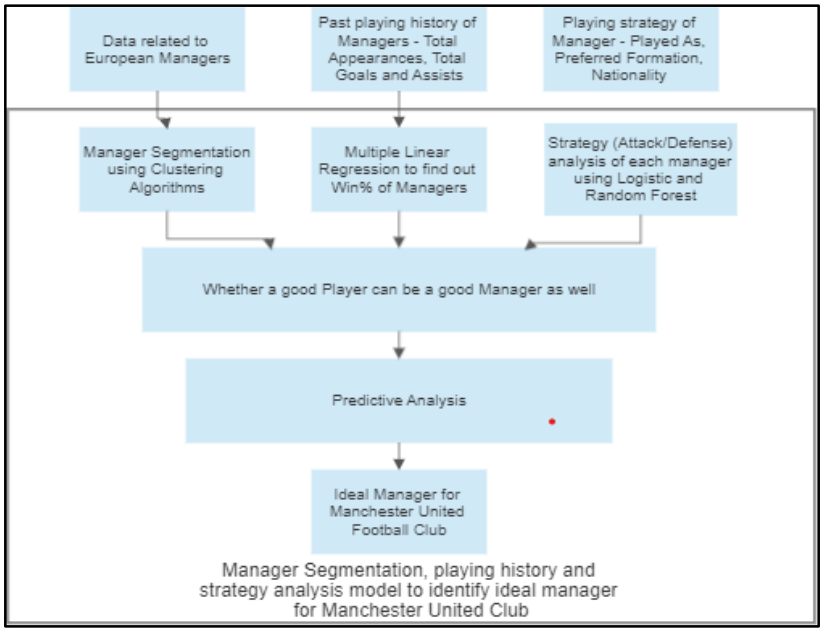


Figure 1 – Proposed framework

First, data related to European managers, which will be helpful in clustering the managers based on the common characteristics they carry in their strategy and approach for playing the match. This will help in grouping those kinds of managers in a one group to consider them as a total for approach understanding. Second; past playing history of managers such as appearances in total match, goals and assists which will directed to the application of the machine learning algorithm in multi linear regression, to find out the win % of the percentage out of the total match he had played throughout his career. Third, leads to understanding the manager's strategy and approach, like whether he believes or is stronger in defense or in attack by applying the improvised logistic regression, to help in preparing a plan for strategy in game against that particular manager team. Data was manually collected from 5 different websites totaling about 31 variables.

## 3 Result and Findings

### 3.1 Clustering: Best Manager for Manchester United

From the cluster profile, it can be examined the statistical summary of the variables for each cluster. For example, here we have taken the mean of each variable for each cluster. This helps us in understanding the overall patterns in the data and identify any notable differences between the clusters. It is also useful to examine the frequency of observations in each cluster. This can give us a sense of how well the data has been partitioned into the different clusters. Like here we can see in the first cluster the frequency is 26, in the second its 41 and in the third cluster its 37. Now for Manchester United to succeed again, they need to start winning the domestic league, which is the premier league again so that they can look at cluster 1 of frequency 26

because cluster 1 has a major number of "Major league winners' managers". Again, regaining their image as Europe's top giants in football can be a starting phase. Also, from a long-term perspective, they can target a manager who has already won some UEFA leagues so that the managers already have some mindset to be successful not only domestically but also across Europe, for that too cluster 1 seems to be the most logical cluster. Now having a manager from England is always better because he has played most of his life across those pitches, both in clubs and internationally. Now for this, Cluster 3 is much better. “Erik TEN HAG” can be the best possible manager for Manchester United. Erik Ten Hag is considered the best manager for Manchester United based on clustering results of managers from top 5 leagues. It may depend on various factors such as the manager's past records, team performance, Nationality, Preferred Formation, etc. that were used for clustering the managers. The advantage of using clustering in this scenario is that it allows for grouping similar managers based on these parameters and can provide useful insights to the club management for making a decision. The outcome of this analysis would aid the club in finding a manager who can help the team perform well in both domestic and UEFA leagues, thereby solving a crucial business problem of improving the team's performance and increasing revenue generation. The CLUSPLOT obtained from the result is presented in Figure 2.

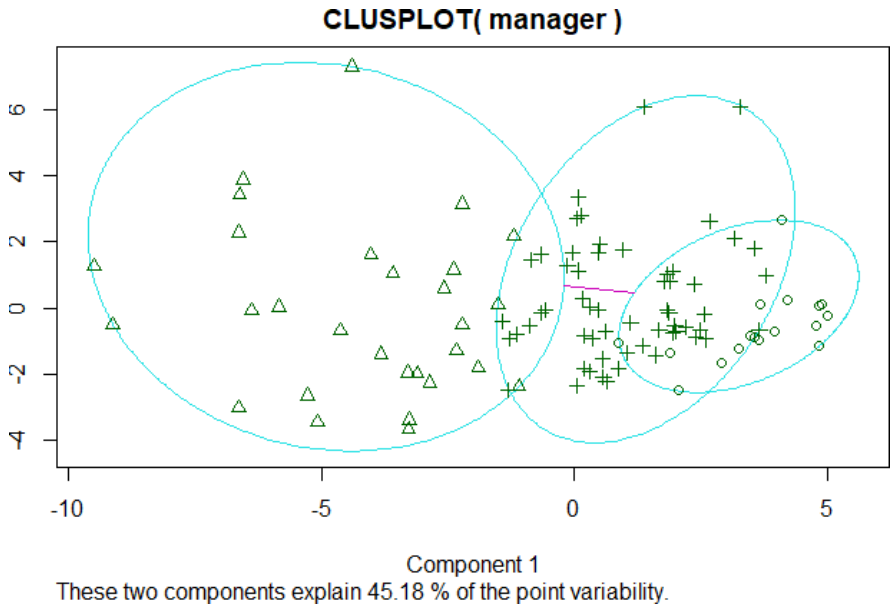


Figure 2 – CLUSPOT

### 3.2 Logistic Regression: Strategy Analysis of a Manager (Attacking/Defending)

Logistic Regression model was used to classify whether a manager is capable of winning the UEFA Championship depending on the specific position played by them. As per the dataset, seven positions were taken into consideration: Forward, Forward/Defender, Forward/Midfielder, Full-Back, Goalkeeper, Midfielder, Not Played. Model summary suggests that the chances of a manager bringing home a UEFA Champions Trophy is not dependent on the position played in the past as none of the variables is significant for the model. Accuracy

achieved on 30% test data is 9%, which is pretty low. Hence, the hypothesis is confirmed. For example, Pep Guardiola, one of the most successful managers, won almost every trophy and developed players like Lionel Messi. He was a midfielder and was having a defensive mindset but as a manager always used the '4-3-3' combination, which is an attacking strategy and focuses more on goals scoring.

# 4 Improved Linear Regression model for calculating Win% of a Manager

On using all the variables for MLR, the R- squared value and the Adjusted R-squared value had a significant difference of 56.31, which is too high and suggests that some variables are not related to the Win%. Therefore, after StepAIC and more domain knowledge, final variables which were selected were total appearances, total goals, Fifa club world cup winner, Nationality, Total Matches, Other Trophies, Total Players used and assists to predict Win%. The new model had a difference of 11.29, which indicates the model is more significant. ANOVA table that compares the two MLR models. High p-value suggests that we can reject the null hypothesis i.e., the two models are not related to each other. Hence, the variables selected after StepAIC are directly related to the Win% of a manager and can help predict how good the manager is.

## 4.1 Improved Linear Regression for determining whether a good player be a good manager

As per domain knowledge (Football), variables like "Total Appearances", "Goals Scored as a Player", "Assists Provided". An R- squared value of 0.1201 suggests that the predictors in the model explain approximately 12% of the variance in the target. It implies that a manager need not be a good player to be a good manager. For example, Jose Mourinho is one of the finest managers of the Premier League. He has not played any single match as a player but has won all major European trophies as a manager.

## 4.2 Comparative analysis of the models

Figure 3 and 4 displays the comparative analysis of the models investigated and the proposed improved linear regression models

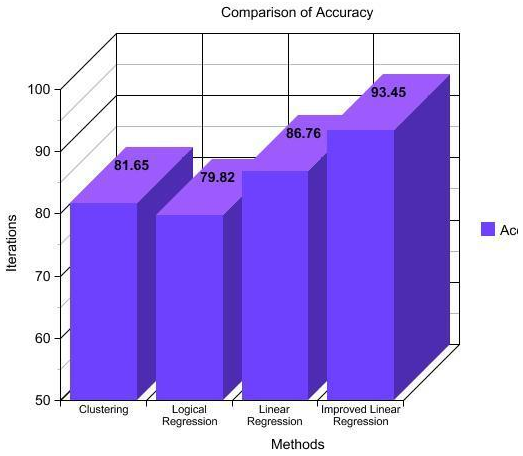


Figure 3- Comparison of Accuracy

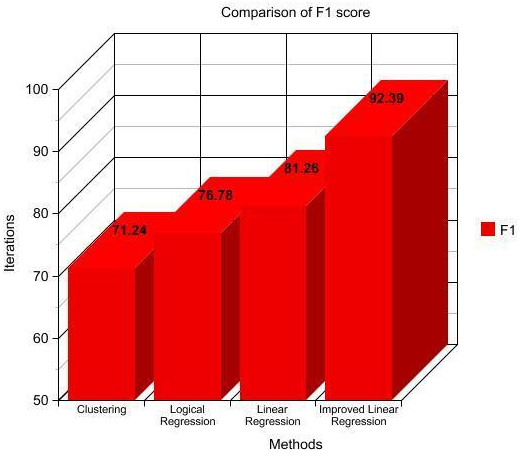


Figure 4-Comparison of F1 score

It is much evident that the novelty in applying the improved linear regression model has higher rate of F1 score of 92.39 and accuracy of 93.45% respectively.

## 5 Conclusion

This research work aimed to deploy basic artificial intelligence in sports domain. The paper implemented improvised linear regression that can help Manchester United identify the best football manager for their team. Data was collected on the managers from the top five European leagues and analyzed using machine learning models like Clustering, Random Forest, Logistic Regression, and Improvised Linear Regression. The results of the analysis revealed that Erik Ten Hag was the best manager for Manchester United based on the results. This paper's outcome could also serve as a guide for other researchers working on AI models in sports analytics and decision making. However, the research also had its limitations such as the availability of data.

## References

[1] Ahtiainen,S. (2018).Top 5 European football leagues – The association between financial performance and sporting success. https://aaltodoc.aalto.fi:443/handle/1234567 89/32207.

[2] Aria, M. a. (2022). Football analytics: a bibliometric study about the last decade contributions. 15(1). doi:10.1285/i20705948v15n1p232.

[3] Arni Arnason, M. P. (2017, aug 30). Risk Factors for Injuries in Football. 32(1). doi:https://doi.org/10.1177/0363546503258912.

[4] Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.0 03.

[5] Barros, C. P., & Leach, S. (2006). Performance evaluation of the English Premier Football League with data envelopment analysis. Applied Economics, 38(12), 1449-1458. https://doi.org/10.1080/00036840500396574.
[6] Carter, N. (2007, sep 10). Managing the Media': The Changing Relationship Between Football Managers and the Media. 217-240. doi:https://doi.org/10.1080/1746026070143 7045.
[7] Christos Tjortjis, V. C. (2020, Dec 11). Sports Analytics for Football League Table and Player Performance Prediction. doi:10.1109/IISA50023.2020.928435.
[8] Daniel Memmert, D. R. (2018, jun 11). Data Analytics in Football. Positional Data Collection, Modelling and Analysis, 186. Retrieved from https://doi.org/10.4324/ 9781351210164.
[9] 9. Danny M. Pincivero, T. O. (2012, Oct 23). A Physiological Review of American Football. 247-260. Retrieved from https://link.springer.com/ article/ 10.2165/ 00007256 - 199723040-00004.
[10] File, K. (2018). You're Manchester United manager, you can't say things like that: Impression management and identity performance by professional football managers in the media Journal of Pragmatics, Volume 127, pg 56-70, ISSN0378-2166.
[11] Gilbourne, T. R. (2008, June 13). Science and football: a review of applied research in the football codes.693-705.doi:https://doi.org/10.1080/0264041031000 102105.
[12] Haas, D. J. (2003). Productive efficiency of english football teams - A data envelopment analysis approach. In Managerial and Decision Economics (Vol. 24, Issue 5, pp. 403–410). Emerald Group Publishing Limited. https://doi.org/10.1002/mde.1105.
[13] Herold et al. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. International Journal of Sports Science & Coaching. 14. 10.1177/1747954119879350.
[14] Rosch, D., Hodgson, R., Peterson, L., Graf-Baumann, T., Junge, A., Chomiak, J., & Dvorak, J. (2000). Assessment and evaluation of football performance. The American journal of sports medicine, 28(5_suppl), 29-39.
[15] 15. Sarmento, H., Marcelino, R., Anguera, M. T., CampaniÇo, J., Matos, N., & LeitÃo, J. C. (2014). Match analysis in football: a systematic review. Journal of sports sciences, 32(20), 1831-1843.