Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 1, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v9ix.xxx

Clustering and Profiling Analysis for FIFA Football Player using K-Means

Salman Yuris Adila Azzami[1], Farrikh Al Zami[2], Heru Pramono Hadi[3], Candra Irawan[4], Aris Nurhindarto[5], MY Teguh Sulistyono[6]

[123456]Information Systems Study Program, Faculty of Computer Science, Dian Nuswantoro University
Jl Imam Bonjol No.207, Pendrikan Kidul, Central Semarang District, Semarang City, 50131, Indonesia

Info Artikel

Riwayat Artikel:

Received 2024-11-01
Revised 2025-02-19
Accepted 2025-02-11

Corresponding Author:

Heru Pramono Hadi

Email:heru.pramono.hadi@dsn.dinus.ac.id

Abstract *– The selection of football players is a complex process involving talent evaluation based on various performance indicators, combining objective measures with subjective assessments by coaches and scouts. This research aims to improve the football player selection process using the K-Means clustering method based on the attributes of transfer price, performance, body specifications, position, and player ability. The dataset used consists of 17.947 players taken from the FIFA 19 edition of the soFIFA.com platform, which includes complete information such as transfer price, performance, body specifications, position, and player ability. The data was processed using principal component analysis (PCA) to reduce the dimensions, followed by the Elbow Method to determine the optimal number of clusters. The clustering results show the distribution of players based on their on-field roles, such as center back, goalkeeper, striker, and left wing back. The profiling of players from each cluster is identified based on position, body type, dominant foot usage, transfer price, and rating. This research provides useful insights for coaches and scouts in selecting players that suit specific roles in the team using better analysis. The findings also highlight the importance of player clustering for data-driven decision-making, which can optimize team composition and overall performance.*

Keywords: *K-Means; Clustering; Football; Performance; Principal Component Analysis.*

Abstrak *– Pemilihan pemain sepak bola adalah proses yang kompleks yang melibatkan evaluasi kemampuan berdasarkan berbagai indikator kinerja, menggabungkan pengukuran obyektif dengan penilaian subyektif oleh pelatih dan pencari bakat. Penelitian ini bertujuan untuk meningkatkan proses seleksi pemain sepak bola menggunakan metode clustering K-Means berdasarkan atribut harga transfer, performa, spesifikasi tubuh, posisi, dan kemampuan pemain. Dataset yang digunakan terdiri dari 17.947 pemain yang diambil dari platform soFIFA.com edisi FIFA 19, yang mencakup informasi lengkap seperti harga transfer, performa, spesifikasi tubuh, posisi, dan kemampuan pemain. Data tersebut diolah menggunakan principal component analysis (PCA) untuk mereduksi dimensi, dilanjutkan dengan Elbow Method untuk menentukan jumlah cluster yang optimal. Hasil clustering menunjukkan distribusi pemain berdasarkan peran mereka di lapangan, seperti bek tengah, kiper, striker, dan bek sayap kiri. Profil pemain dari setiap cluster diidentifikasi berdasarkan posisi, tipe tubuh, penggunaan kaki yang dominan, harga transfer, dan rating. Penelitian ini memberikan wawasan yang berguna bagi para pelatih dan pemandu bakat dalam memilih pemain yang sesuai dengan peran tertentu dalam tim dengan menggunakan analisis yang lebih baik. Temuan ini juga menyoroti pentingnya clustering pemain untuk pengambilan keputusan berbasis data, yang dapat mengoptimalkan komposisi tim dan kinerja secara keseluruhan.*

Kata Kunci*: K-Means, Clustering, Football, Performance, Principal Component Analysis.*

I. INTRODUCTION

The selection of football players is a complex process that involves the evaluation of talent based on various performance indicators, combining objective measures with subjective assessments by coaches and scouts; research emphasizes that sprint speed, specifically 30 m sprint time at the age of 9 or 10 years, is a significant predictor of selection to elite football academies, as it can reliably predict player selection at the age of 12 years [1]. This selection process is often seen as an evolutionary framework, where athletes are selected based on performance excellence or traits that meet coaches' criteria [2]. Longitudinal studies suggest that while the rate of improvement in various performance metrics is critical, early metrics tend to have greater predictive power for selection [1]. However, the selection process may not always accurately reflect an athlete's potential, as it can be influenced by the biases or preferences of selectors, which may result in overlooking athletes who could develop later in life [2].

Football player selection has evolved significantly with the integration of machine learning (ML) and data science, enabling clubs to make more informed and data-driven decisions. Proper selection can significantly impact a player's development and career trajectory, making it a crucial area of study. Machine learning algorithms, such as clustering and classification, are now widely used to identify athletes with the appropriate physical and technical

attributes, improving both performance and injury prevention. For example, advanced data analysis techniques, including correlation matrices and Principal Component Analysis (PCA), are employed to uncover relationships between variables such as sprint speed, lower limb strength, and injury risk[3]. Furthermore, considering factors such as lower limb strength and balance in the selection criteria can help reduce the risk of injury; studies show that strength training plays a very important role in the prevention of football player injuries. The evaluation of players based on physical abilities, such as the ratio of hamstring muscles to quadriceps, allows for better selection practices by placing priority on long-term health [4]. Moreover, it is also financially affirming, as injuries can lead to significant costs for clubs, thus stressing the importance of effective selection strategies to minimize such risks. Therefore, while player selection is important, the broader context such as team dynamics and the role of coaching in player development should also be considered as it can affect overall success in football.[4].

The clustering process in football player selection aims to group players based on similar attributes to improve team performance and devise more effective strategies. By utilizing data-driven insights, clustering helps identify player profiles that fit specific roles within the team. K-Means clustering, a widely used unsupervised learning method, is particularly effective in grouping players based on performance metrics such as goals, assists, and possession statistics [5]. However, K-Means has limitations, such as its sensitivity to initial centroid placement and its assumption of spherical clusters, which may not always align with the complex distribution of player data. In contrast, Gaussian Mixture Models (GMM) offer a more flexible approach by accommodating data with non-spherical distributions, making them suitable for analyzing multifaceted player attributes[6], [7], [8].

The use of machine learning algorithms in football has revolutionized data analysis, predictive modeling, and performance optimization. These algorithms can efficiently handle large amounts of data generated from matches, player statistics, and fan interactions, leading to deeper insights and better decision-making. For instance, correlation matrices are often applied before clustering to identify key relationships between variables, such as the correlation between ball possession and goals scored. Additionally, ML models can analyze transfer market data to predict player performance based on historical trends, helping clubs identify high-value players at lower costs[9], [10]. Moreover, the ability to analyze different types of data, similar to microbial community profiling, enables the identification of patterns and trends in player performance and game outcomes[11]. In terms of predictive analysis, machine learning models can predict match outcomes and player performance, similar to educational data mining techniques that forecast student success based on historical data [12]. These algorithms can adapt to changes in data distribution, ensuring that predictions remain relevant over time, as demonstrated in healthcare settings during the COVID-19 pandemic [13]. Although machine learning provides a powerful tool for football analysis, challenges such as data efficiency and model interpretability remain important considerations for effective implementation [14].

The correlation matrix helps in understanding the relationships between various variables, such as match progress and performance achievements, providing insights into how these factors influence each other. Meanwhile, Principal Component Analysis (PCA) is utilized to reduce data dimensionality and highlight the most critical variables in determining outcomes. This approach enables teams to make data-driven decisions that enhance player development and optimize overall performance. For example, one study showed that performance variables had a significant influence on the team compared to match data[15]. As another example, PCA can detect a strong relationship between ball speed and the game space matrix that can show the interaction of these factors during a match [16]. [15]. PCA can generally make complex datasets simple by reducing variables to principal components and revealing underlying patterns [16]. When such data has been used for the analysis process using spatiotemporal methods, it can show the position of the player against the course of the game [17]. While correlation matrices and PCA provide valuable insights, they may overlook contextual factors such as player psychology and team dynamics, which also significantly affect performance outcomes in football.

The use of FIFA data has become a crucial tool in modern football analysis, enabling the assessment of player performance, team strategy, and match outcome predictions. The integration of this data into machine learning models, such as K-Means clustering, further enhances the ability to uncover hidden patterns and trends within player statistics, team performance, and physical attributes. Several studies from the last five years explore how data science and machine learning techniques are applied to improve understanding of football dynamics. Research by Sweeney (2022) in Nature explains how the use of big data, including data from FIFA[18], has changed the way football analysis is conducted. By utilizing player and team statistics, this research shows how machine learning algorithms can be used to predict match outcomes and aid in managerial decision-making. This data-driven modeling can improve prediction accuracy as well as help in determining team strategies[19].

K-Means clustering, in particular, has been widely applied to FIFA data to group players based on performance metrics, such as goals, assists, and defensive actions. This approach helps clubs identify undervalued players who may excel in specific roles, thereby optimizing recruitment strategies. Research by Brown and Smith (2023) in the Journal of Sports Analytics shows how the K-Means method can be used to cluster players based on their performance statistics taken from FIFA data. By analyzing various attributes such as goals, assists, and possession, the study was able to identify groups of players with similar characteristics, which can help teams with scouting and game strategy development [19], [20].

The urgency of adopting data science in football is evident from an industry perspective, where clubs are increasingly leveraging machine learning (ML) techniques to recruit high-potential players at lower costs. This not only ensures financial sustainability but also helps maintain competitive performance. The advancements in data science highlight its transformative potential in modern football, providing clubs with both competitive and financial advantages through data-driven decision-making[21]

II. METHOD

The methodology of this study centers on clustering analysis, a robust approach for grouping unlabeled data based on inherent similarities. This section details our systematic process, including data collection, preparation, preprocessing steps, and the implementation of clustering algorithms as seen in figure 1. The analysis framework consists of three main stages: data preprocessing, model training, and results evaluation.

Clustering is considered as one of the effective methods in grouping unlabeled data. Basically, the clustering process aims to group data based on similar characteristics. This method also serves to separate data into groups that have significant differences in characteristics.

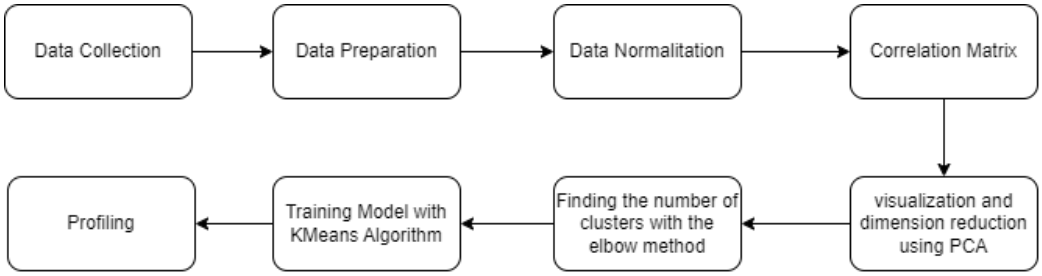

Figure 1. The main stages in the clustering process include preprocessing, model training, and final results

A. Data Collecting

The dataset used is a public dataset obtained from the soFIFA.com website. This website is a third-party website that provides data from a football game called FIFA. This site is often used as a reference for football fans, regarding detailed information about one of which is the character of the player. The dataset used was taken from the football game from the FIFA 19 edition, which was accessed on August 12, 2024. The amount of data used was 17,947 data, with 63 columns. The collection of the dataset was done by downloading the data directly from soFIFA.com, which provides the data in a structured format ready for further processing for analysis purposes. This dataset was used in the research to perform player profiling, clustering, and performance analysis based on the attributes.

B. Data Preparation

Data preparation is a process that aims to clean the data. The data to be modeled must be clean, and consistent so that it can be analyzed further.

1) *Missing Value:* To ensure that the analyzed data is accurate, we must first examine the presence of missing data in each feature. After further investigation, we found several features that have missing values. In practice, we divide into 2 processes in handling missing features. If there is a missing feature below 15%, it will be tolerated by filling the missing value with 0. However, if the missing value is more than 90%, the data will be deleted.

   Features that will be filled with 0 can be seen in table 1:

TABLE 1
FEATURES THAT WILL BE FILLED WITH 0

Feature Name	Presentation missing value
release_clause_euro	10,24%
value_euro	1,42%
wage_euro	1,37%

In table 2 are some of the features that will be removed:

TABLE 2
LIST OF FEATURES TO BE REMOVED

Feature Name	Presentation missing value
national_team	95,26%
national_rating	95,26%
national_team_position	95,26%
national_jersey_number	95,26%

2) *Data Encoding:* Before processing the data using the modeling algorithm, we must convert the categorical data into numeric. The conversion is adjusted according to the example in table 3 with the body_type feature :

TABLE 3
EXAMPLES OF FEATURES TO BE TRANSFORMED

Categorical Data	Numeric Data
Lean	0
Normal	1
Stocky	2

3) *Duplicate Data:* Ensuring that there is no duplicate data is also very important. Duplicate data causes information to be biased. This will result in inaccuracies in the analysis process. After the existing data was processed, no bias was found in the dataset used. So this data does not need to be handled again regarding data duplication.

C. Correlation Matrix

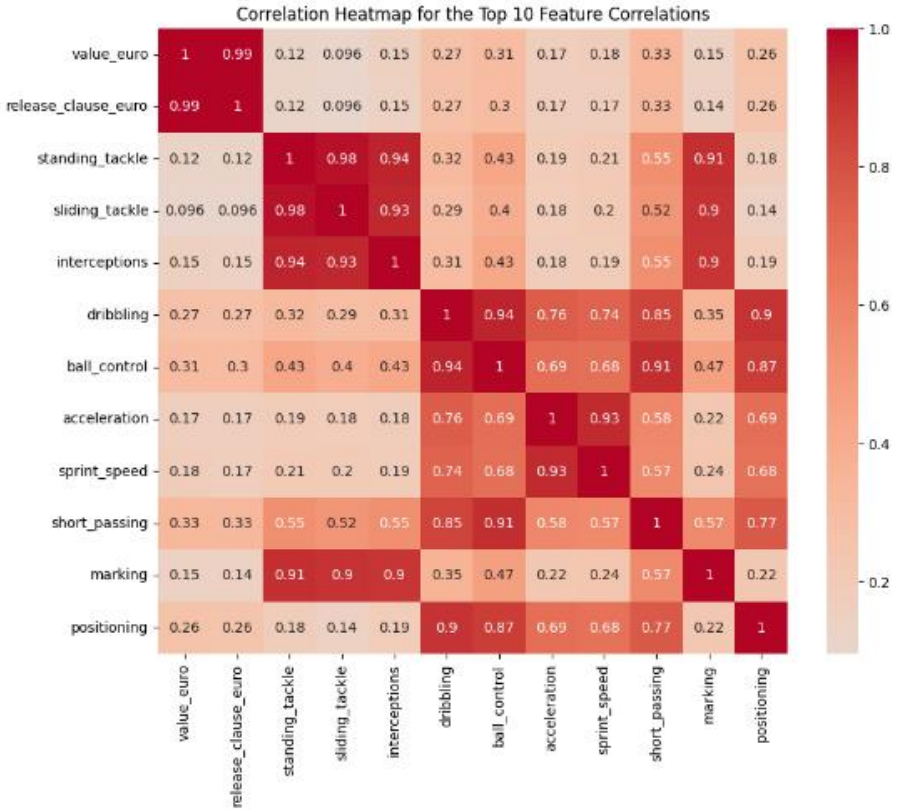

Figure 2. 10 features with the highest correlation.

Correlation Matrix is a table that shows the relationship between one feature and another. Each cell in this table shows how strongly the variables are related to each other. Due to the dataset contains more than 63 features, we simplified the correlation matrix using highest score in figure 2. Here, figure 2 shows the 10 features that have the highest correlation with each other. The following is the formula for the correlation matrix,

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y}) \qquad (1)$$

Information:
$Cov(X,Y)$ = Covariance between variable X and Y
n = Number observation (data points)
Xi = Individual value from variable X
Yi = Individual value from variable Y
$\bar{X}$ = Average of variable X
$\bar{Y}$= Average of variable Y

D. PCA

PCA (Principal Component Analysis) is a linear reduction technique that aims to reduce high-dimensional features. PCA is used to reduce the number of variables in a dataset while retaining as much information as possible. PCA can help simplify machine learning models so that the training process can be faster, and reduce overfitting. Here are the basic steps and formulas in PCA:

The first step that needs to be done when using PCA is to standardize the data which is known as Z-score normalizationso that each feature has a mean of 0 and a variance of 1. This process is done by subtracting the mean and dividing by the standard deviation for each feature:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \qquad (2)$$

Information :
X_{ij} = value feature j from row i
μ_j = average of feature j
σ_j = standard deviation from feature j
After the data is normalized, the next step is to measure the relationship of each feature in the dataset.

$$C = \frac{1}{n-1}Z^T Z \qquad (3)$$

Information :
Z = Data matrix after normalization
n = number of existing rows
We need to calculate the eigenvalue and eigenvector of the covariance matrix C, to get the principal components. The eigenvalue (λ) indicates the amount of variance explained by each principal component, and the eigenvector (v) is the direction of the principal component.

$$Cv = \lambda v \qquad (4)$$

Information :
λ = the magnitude of the variance in the eigenfactor (eigenvalue)
v = calculation of the direction of the main components
After getting the eigenvectors (principal components), we can use them to transform the original data into the principal component space. If we want to reduce the data dimension to k dimensions, we select the k largest eigenvectors (with the highest eigenvalue), then we multiply it with the original data Z to get the data in the principal component space:

$$Z' = ZW_k \qquad (5)$$

Information :
W_k = matrix p×k consisting of from k eigenvector the biggest.
Z'= result transformation from data to in room new dimension k.

Sorting eigenvalues from largest to smallest eigenvector associated with the largest eigenvalue will become the main component, followed by the second largest eigenvalue to the smallest. These principal components will be used to reduce the dimensionality of the data.

E. Elbow Method

Elbow method is an algorithm used to determine the optimal number of clusters in a dataset. One of the calculation schemes in the Elbow Method is the Distortion score. Distortion score is the process of calculating the total squared distance error between the data points in the cluster and their centroids. The smaller the distortion score, the better the data in the cluster is centered around the centroid.

$$Distortion\ (Inertia) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \qquad (6)$$

Information :
k = Number of clusters
C_i= Cluster i
$x \in C_i$= Data points that are within the clusterC_i
μ_i = Centroid from clusterC_i
$\|x - \mu_i\|^2$= Euclidean distance squared between data $x$points and centroids μ_i

F. K-Means

K-Means is an algorithm used to divide groups of data into different clusters based on the similarity between data points. The goal of this algorithm is to form clusters that are as similar as possible while ensuring that different clusters have the lowest possible similarity. The number of clusters is determined beforehand using the Elbow Method. In implementing K-Means, several libraries are commonly used: Pandas for data preprocessing, Scikit-learn for PCA and K-Means clustering, and Matplotlib/Seaborn for data visualization. The following is the formula of the K-Means algorithm:

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - c_{kj})^2} \qquad (7)$$

Information :
x_i= vector feature from i-th data point
c_k= centroid of the kth cluster
n = number feature

III. RESULT AND DISCUSSION

To understand the grouping of data into groups based on their similar characteristics, we must use a technique commonly called clustering. One of the most frequently used clustering methods is K-Means. However, before running the K-Means algorithm, we must first determine the number of clusters to be used and using standardize the data with Z-score normalization. Before doing that, we need to first look at the correlation between the data using a correlation matrix. The purpose of this correlation matrix is to see the relationship between features. If you look at the previous figure, it can be seen that we have a lot of features. Therefore, before modeling, we will reduce the data using Principal Component Analysis (PCA). This algorithm is used to reduce the dimensionality of the data.

To evaluate the impact of PCA on clustering performance, we compare the clustering results before and after applying PCA. The Hopkins Score before PCA was 0.8488, while after applying PCA, the score improved to 0.9581. This indicates that PCA helped in forming more well-separated and compact clusters, enhancing the overall clustering performance. As seen in Figure 3. After we reduce the data using PCA, the next step is to use the data to calculate the number of clusters that will be used during the modeling process. In this case, the author uses the Elbow Method with distortion score as the algorithm. The range of the number of clusters determined starts from K = 1 to K = 10. This calculation shows that K=4 with a score of 155388.302 is the most optimal number of clusters. Therefore, we will create 4 clusters using K-Means.

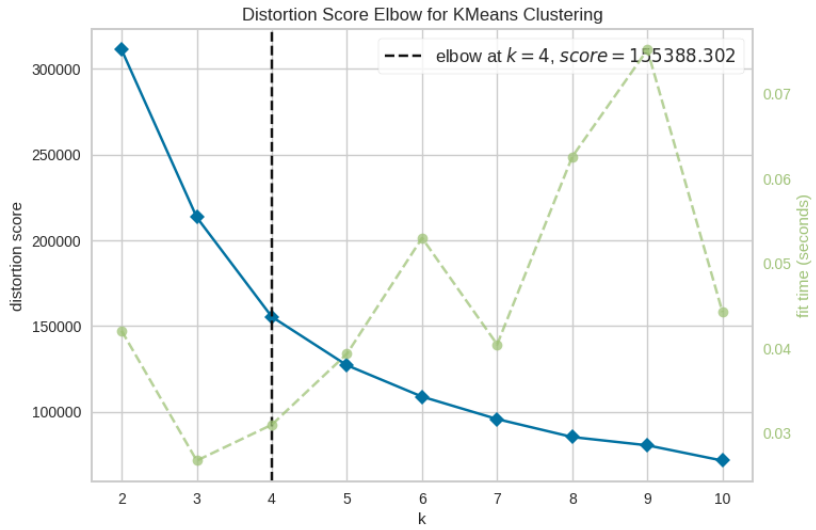

Figure 3 Metode Elbow Score

After we determine the number of clusters, then we can use it in the K-Means algorithm to do modeling. Figure 4 is the distribution of data from the modeling results,

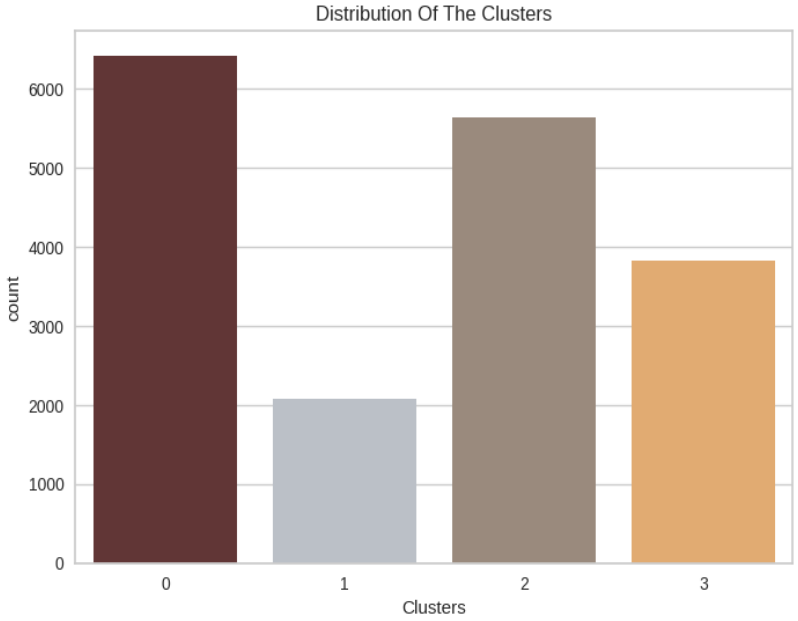

Figure 4. Cluster Distribution After Modeling Using K-Means

In addition, since PCA was used previously, there are also visualization results of the clusterized PCA. Figure 5 shows the visualization results of the clusterized PCA,

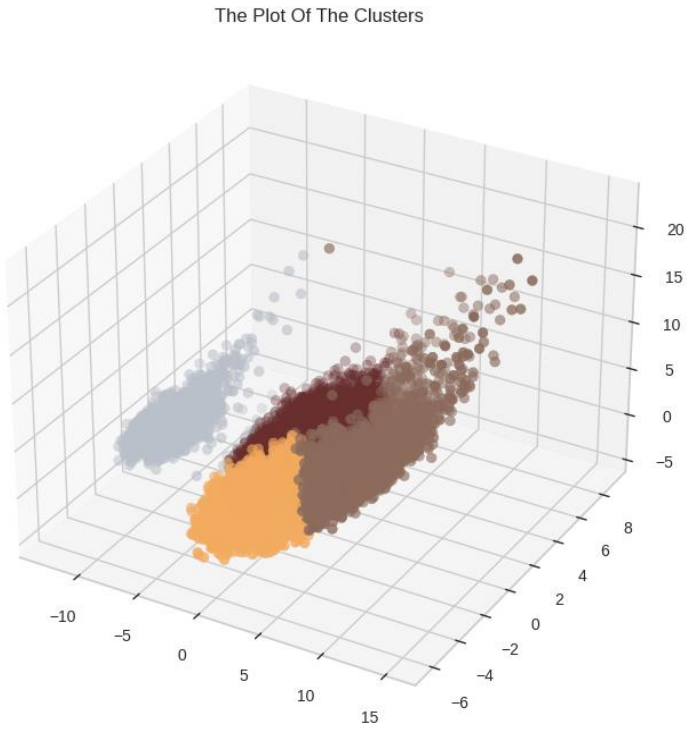

Figure 5. PCA Visualization Results That have been Modeled Using K-Means

Following This is results profiling obtained from results clustering using K-Means which has been done previously,

A. Positioning

The clustering results indicate that height and strength play a crucial role in classification, with taller and stronger players often grouped as defenders (CB), while shorter, agile players tend to be classified as strikers (ST) or midfielders (LWB). Additionally, agility and speed significantly impact forward players (ST) and wingers (LWB), as these attributes are essential for quick movements and dribbling. On the other hand, diving and reflexes are the primary distinguishing factors for goalkeepers (GK), forming a distinct cluster specific to their role.

The following are the clustering results of each cluster in the Positioning feature:

1) Cluster 0:

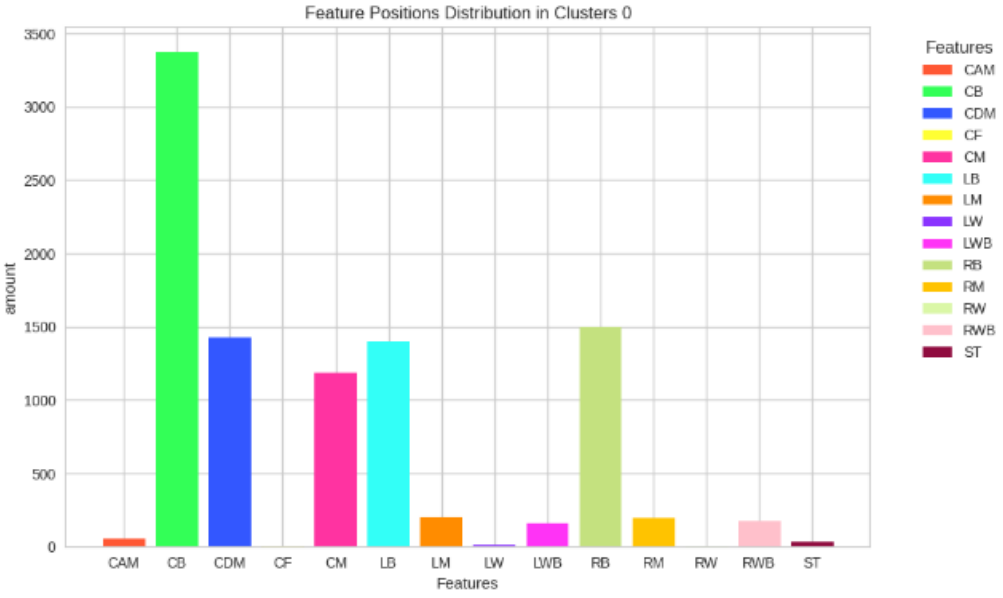

Figure 6. Data distribution of feature positions from cluster 0

In Figure 6, it can be seen that Cluster 0 has the most positions as a Central Back(CB), characterized by high strength and defensive attributes. which functions to block the opponent's attack and secure the central defense area in front of the goalkeeper.

2) Cluster 1:

As seen in Figure 7, cluster 1 only has players with the goalkeeper position (GK), This cluster is strongly influenced by diving, reflexes, and positioning attributes. so we can assume that this cluster only contains players with the goalkeeper position.

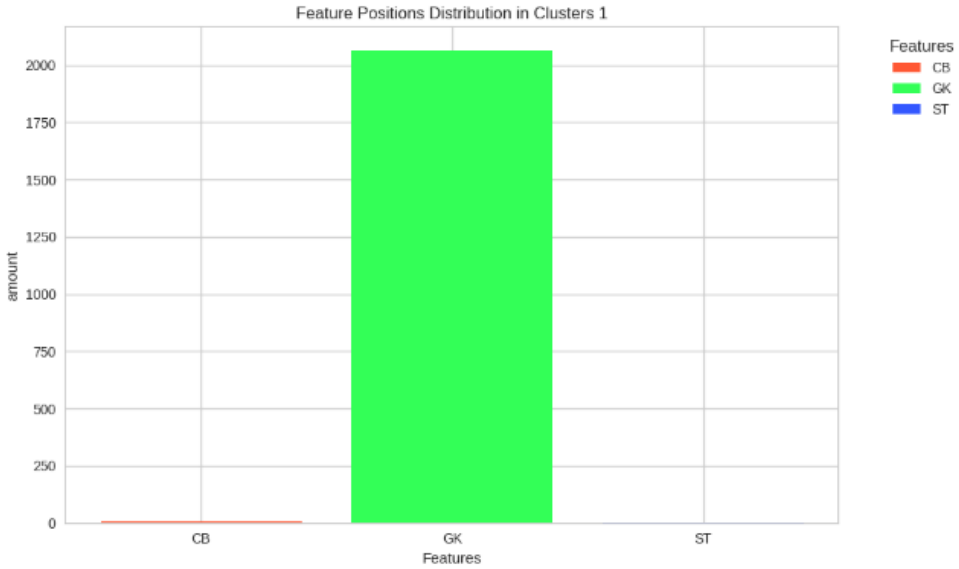

Figure 7. Data distribution of feature positions from cluster 1

3) Cluster 2:

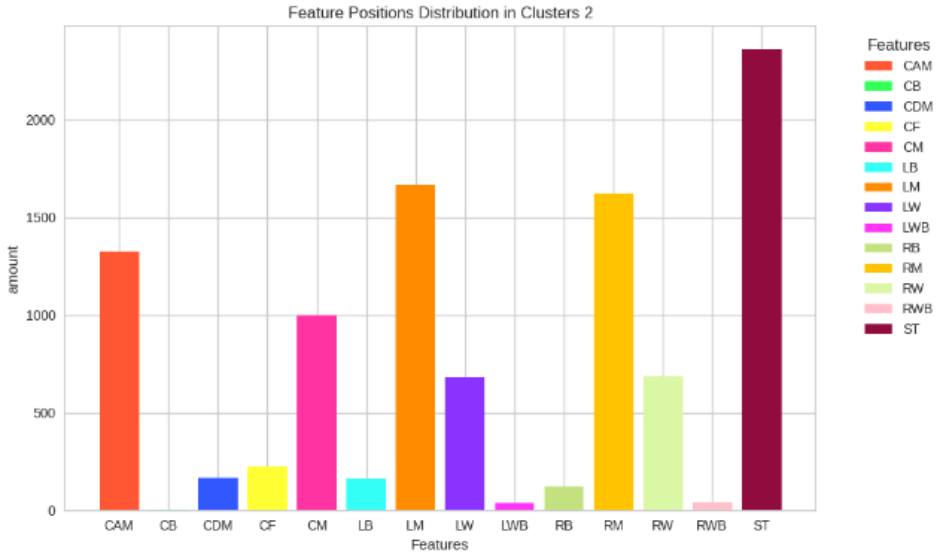

Figure 8. Data distribution of feature positions from cluster 2

Cluster 2 has the most players in the striker position (ST). This position has a role to score as many goals as possible and is usually at the forefront of the attack. This can be seen in Figure 8.

4) Cluster 3:

In figure 9, it can be seen that cluster 3 has the most players from the left wing back (LWB) position which has a defensive and offensive role on the left side as well as supporting attacks and helping defend.

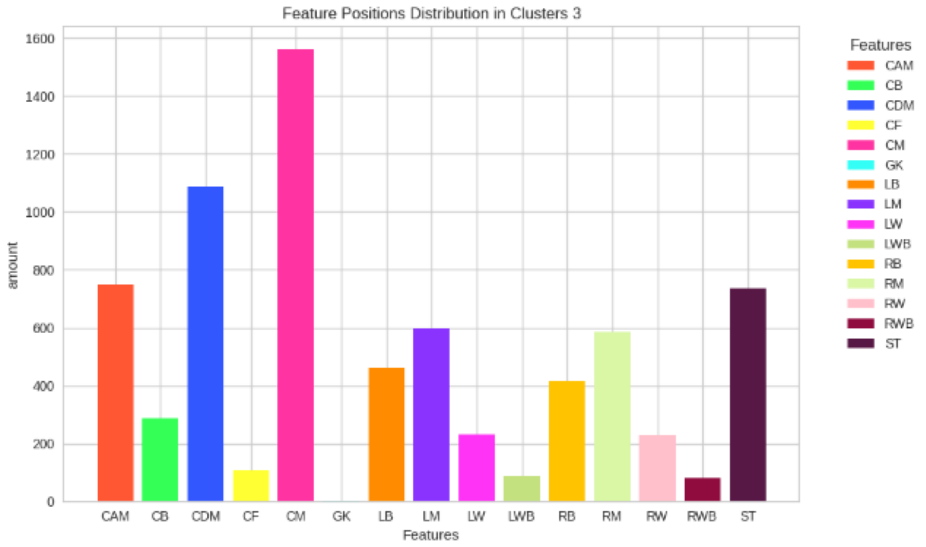

Figure 9. Data distribution of feature positions from cluster 3

B. Use of both feet for players

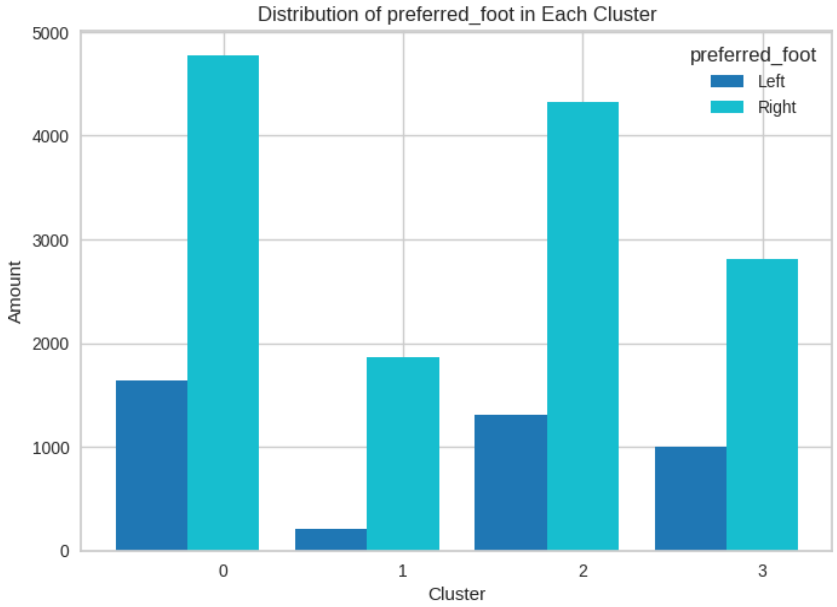

Figure 10. Data visualization of preferred_foot features

Figure 10 shows that the four clusters above are similar in the dominant use of the player's foot. It can be seen that all of them predominantly use the right foot. Skillful use of both feet (including the left foot) will also have a positive impact in terms of agility. Because players who are dominant in both feet will certainly be more points when in difficult conditions that require two-foot skills.

C. Body Type

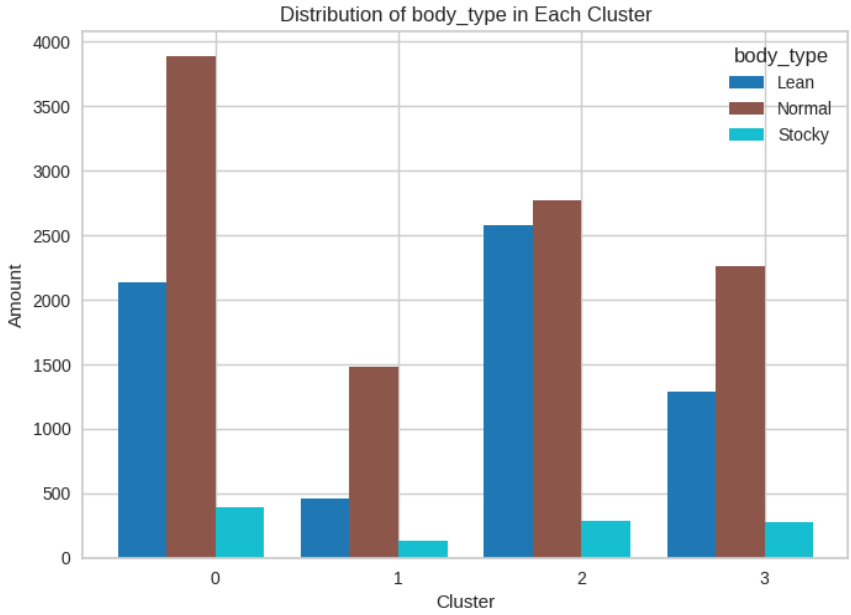

Figure 11. Data visualization of body_type features

As seen in Figure 11, clusters 0 and 3 show that very much players who have normal body. However, it looks that more from more from half player with normal body that has thin body. In Cluster 2 it is seen that there is players who have normal body, and quite thin balanced. In cluster 1 it has players who have the most normal body. This suggests that physical attributes could be an influencing factor in player categorization.

D. Ratings and Transfer Prices

Cluster 0 contains players with fairly high ratings and low fees. Cluster 2 contains players with low transfer fees but also low ratings. Cluster 3 has players with high fees and the highest rating. While as explained earlier, rating 1 is a collection of players with a position as a goalkeeper. The ratings and transfer fees in cluster 1 are quite varied ranging from the lowest to the highest. This is quite clearly seen in Figure 12.

Figure 12. Data visualization of Income and Spending features

IV. CONCLUSION

Based on the profiling results above, Cluster 0 consists of players who possess strong abilities at a relatively low price. In terms of positioning, this cluster primarily comprises center-backs. Most players in this cluster have a normal body type and predominantly use their right foot. Cluster 1 mainly consists of goalkeepers, with a wide range of transfer prices and skill levels. Cluster 2 includes players with low transfer prices and abilities. This cluster may require further training to enhance performance. Cluster 3 represents players with the highest performance levels, accompanied by high transfer prices. This cluster is optimized for players with exceptional abilities, ready for high-level competitions. Most players in this cluster are positioned as left wing-backs. Compared to traditional selection methods, clustering provides a data-driven approach that enhances objectivity and consistency in player categorization. While conventional selection heavily relies on subjective judgments from coaches and scouts, clustering systematically groups players based on quantifiable performance metrics, thereby minimizing human biases

.

ACKNOWLEDGEMENT

This research has been conducted in collaboration with IDSS Research Center Faculty of Computer Science Universitas Dian Nuswantoro.

BIBLIOGRAPHY

[1] D. Fortin-Guichard, I. Huberts, J. Sanders, R. Van Elk, D. L. Mann, and G. J. P. Savelsbergh, "Predictors of selection into an elite level youth football academy: A longitudinal study," *J. Sports Sci.*, vol. 40, no. 9, Art. no. 9, May 2022, doi: 10.1080/02640414.2022.2044128.

[2] J. Baker, K. Johnston, and N. Wattie, "Survival Versus Attraction Advantages and Talent Selection in Sport," *Sports Med. - Open*, vol. 8, no. 1, Art. no. 1, Dec. 2022, doi: 10.1186/s40798-022-00409-y.

[3] D. Berrar, P. Lopes, J. Davis, and W. Dubitzky, "Guest editorial: special issue on machine learning for soccer," *Mach. Learn.*, vol. 108, no. 1, Art. no. 1, Jan. 2019, doi: 10.1007/s10994-018-5763-8.

[4] M. Beato, S. Maroto-Izquierdo, A. N. Turner, and C. Bishop, "Implementing Strength Training Strategies for Injury Prevention in Soccer: Scientific Rationale and Methodological Recommendations," *Int. J. Sports Physiol. Perform.*, vol. 16, no. 3, Art. no. 3, Mar. 2021, doi: 10.1123/ijspp.2020-0862.

[5] S. Mahallati, J. C. Bezdek, M. R. Popovic, and T. A. Valiante, "Cluster tendency assessment in neuronal spike data," *PLOS ONE*, vol. 14, no. 11, Art. no. 11, Nov. 2019, doi: 10.1371/journal.pone.0224547.

[6] W. Lu, D. Ding, F. Wu, and G. Yuan, "An Efficient Gaussian Mixture Model and Its Application to Neural Network," Aug. 13, 2024, *Computer Science and Mathematics*. doi: 10.20944/preprints202302.0275.v3.

[7] F. Chamroukhi and H. D. Nguyen, "Model-based clustering and classification of functional data," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 4, Art. no. 4, Jul. 2019, doi: 10.1002/widm.1298.

[8] T. Li, G. Kou, Y. Peng, and P. S. Yu, "An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data," *IEEE Trans. Cybern.*, vol. 52, no. 12, Art. no. 12, Dec. 2022, doi: 10.1109/TCYB.2021.3109066.

[9] D. Tuia *et al.*, "Perspectives in machine learning for wildlife conservation," *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Feb. 2022, doi: 10.1038/s41467-022-27980-y.

[10] R. M, R. B, L. R, and S. K, "Player Performance Prediction in Sports Using Machine Learning," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, Gurugram, India: IEEE, May 2024, pp. 1–6. doi: 10.1109/ISCS61804.2024.10581086.

[11] R. B. Ghannam and S. M. Techtmann, "Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1092–1107, 2021, doi: 10.1016/j.csbj.2021.01.028.

[12] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, Art. no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[13] C. Duckworth *et al.*, "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Nov. 2021, doi: 10.1038/s41598-021-02481-y.

[14] A. Adadi, "A survey on data-efficient algorithms in big data era," *J. Big Data*, vol. 8, no. 1, Art. no. 1, Jan. 2021, doi: 10.1186/s40537-021-00419-9.

[15] J. M. Oliva-Lozano, H. Martínez-Puertas, V. Fortes, R. López- Del Campo, R. Resta, and J. M. Muyor, "Is there any relationship between match running, technical-tactical performance, and team success in professional soccer? A longitudinal study in the first and second divisions of LaLiga," *Biol. Sport*, vol. 40, no. 2, Art. no. 2, 2023, doi: 10.5114/biolsport.2023.118021.

[16] B. Gonçalves, D. Coutinho, J. Exel, B. Travassos, C. Lago, and J. Sampaio, "Extracting spatial-temporal features that describe a team match demands when considering the effects of the quality of opposition in elite football," *PLOS ONE*, vol. 14, no. 8, Art. no. 8, Aug. 2019, doi: 10.1371/journal.pone.0221368.

[17] L. Pappalardo *et al.*, "A public data set of spatio-temporal match events in soccer competitions," *Sci. Data*, vol. 6, no. 1, Art. no. 1, Oct. 2019, doi: 10.1038/s41597-019-0247-7.

[18] D. Sun, "An Overview of Machine Learning Applications in the Football Field," *Appl. Comput. Eng.*, vol. 8, no. 1, pp. 318–322, Aug. 2023, doi: 10.54254/2755-2721/8/20230178.

[19] "Maths predicts World Cup winner — and more of this week's best science graphics," *Nature*, pp. d41586-022-03809-y, Nov. 2022, doi: 10.1038/d41586-022-03809-y.

[20] D. Adam, "Science and the World Cup: how big data is transforming football," *Nature*, vol. 611, no. 7936, pp. 444–446, Nov. 2022, doi: 10.1038/d41586-022-03698-1.

[21] V. Bhatia and A. More, "Implementing Football Prediction System Using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 11, pp. 392–396, Nov. 2024, doi: 10.22214/ijraset.2024.65056.