Journal of Information Technology application in Education, Economy, Health and Agriculture
Vol. 01, No. 03, October 2024, pp. 50~56
ISSN: xxx-xxxx, DOI:-

Determining Potential Players For The Indonesian Senior National Team In The 2026 World Cup Qualifications Using K-Means

Slamet Risnanto[1], Fikri Alfian[2], Moh Imam Faiz[3], Moh. Nizar[4], Dinny Wahyu Widarti[5]

[1] Department of Informatics Engineering, University of Sangga Buana, Bandung, Indonesia
[2,3,4] Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[5] Department of Information Technology, STMIK Pradnya Paramita, Malang, Indonesia

Article Info

Article history:

Received August 03, 2024
Revised September 11, 2024
Accepted October 20, 2024

Keywords:

football
K-Means clustering
Indonesian National Team
2026 World Cup qualification

ABSTRACT

Football is a very popular sport, and the Indonesian National Team is the pride of the Indonesian people. In an effort to improve team performance, especially in facing the 2026 World Cup qualifiers, optimal player selection is a major challenge. This study applies data mining technology to determine potential players who can strengthen the Indonesian Senior National Team. Player data is taken from the Transfermarkt site which includes attributes such as player market value, club, and league. The methods used include data collection, data cleaning and normalization, and analysis using the K-Means clustering algorithm. The analysis process successfully grouped players into four clusters based on their potential. Players in clusters 1 and 3 have high potential to fill the main lineup, while players in cluster 0 show long-term development prospects. Visualization and manual evaluation support the interpretation of the results for strategic decision making. This study shows that the use of data mining can improve efficiency and accuracy in player selection, providing a more objective data-based approach. However, this study has limitations, such as the lack of consideration of non-technical factors. With the addition of data from other sources and the use of additional algorithms, this method can be further developed to support the performance of the Indonesian National Team optimally in the future.

Corresponding Author:

Mohammad Fikri Alfian
Department of Informatics Engineering
Faculty of Engineering, University of Widyagama Malang
Jl. Borobudur No. 35 Malang, Jawa Timur
Email fikrihalfian@gmail.com

1. INTRODUCTION

Football is a very popular sport in the world, with the Indonesian National Football Team representing the pride of the Indonesian people. In an effort to improve team performance, one of the main challenges is determining potential players who can provide maximum contribution in the match. The selection of these players often still relies on manual analysis from coaches and staff, which may be at risk of not being optimal if applied in important matches, such as the 2026 World Cup qualifiers.[1]. The higher level of competition in the 2026 World Cup qualifying phase requires quicker and more precise decisions, which can be achieved through data-driven analysis to maximize player potential and team strategies.

In the modern era, technologies such as data retrieval from online platforms and data mining offer great opportunities to improve efficiency and accuracy in player selection decision making. In this study, data will be taken from the transfermarkt.co.id site, a platform that provides comprehensive information about football players, including age, statistical data, career history, and market value.[2]. Meanwhile, data mining helps analyze the data to identify the most effective patterns or combinations based on team needs,

so that it can be used to analyze players who have the potential to be called up to strengthen the Indonesian Senior National Team in the 2026 World Cup qualifiers.[3][1].

The use of this technology, including data retrieval from Transfermarkt, can provide an objective data-based solution to help coaches determine the right players.[1]. Player samples were taken randomly, both from those who play in Liga 1 Indonesia and players who have careers abroad. The player classification process is determined based on three main parameters, namely Player Market Value, Club Market Value, and League Market Value. By utilizing comprehensive data and analyzing it using data mining techniques, the process of selecting potential players becomes more efficient, factual, and based on overall performance.[3]. This data-driven approach could help improve Indonesia's chances of qualifying for the 2026 World Cup.

This study aims to study the use of data processing and data mining techniques in determining the most potential players to be called up to the Indonesian Senior National Team in the 2026 World Cup qualifiers. Hopefully, this study can provide a significant contribution to the development of national football and ensure that player selection decisions are based on in-depth and accurate data analysis.

2. METHOD

2.1 Research Approach

This study uses a quantitative approach based on data obtained from the Transfermarkt platform. Player sample data was taken randomly, both players who compete in Liga 1 Indonesia and those who play abroad. The data was then processed using data mining methods to identify potential players who can be called up to strengthen the Indonesian Senior National Team in the 2026 World Cup qualifiers.

2.2 Tools and Technology

In the development of this research, Python is used as the main programming language supported by several important libraries, namely Pandas for data processing, Matplotlib which is used in creating data visualization, and Scikit-learn which plays a role in conducting data mining analysis. The entire coding and script execution process is carried out using the Google Colab platform as a development environment.

2.3 Research Stages

a. Data Collection from Transfermarkt website

The first stage in this research is data collection which is done manually from the transfermarkt.co.id site. The selection of player samples is done randomly, including players who play in Liga 1 Indonesia and in foreign clubs. The data collected includes attributes such as player name, player market value, club market value, and market value of the league where the player competes. After that, the data is arranged in csv table data format to facilitate the analysis process. The resulting raw dataset contains complete information about the players and is ready to be further analyzed to determine their potential.

b. Data Cleaning and Processing

Perform data cleaning process to overcome the problem of empty data, invalid data, and data that is not relevant to the analysis objectives. Next, perform data normalization to ensure that the attribute scales that have different value ranges, such as player market value, club market value, and league market value, can be compared proportionally. This normalization aims to avoid the

dominance of attributes with a larger scale in the analysis and modeling process, so that the results obtained are more accurate and can be better interpreted.[4].

c. Data Analysis Using Data Mining Techniques

Use the K-Means clustering algorithm to group players based on the attributes of player market value, club market value, and league market value. With K-Means, players will be grouped into several clusters based on the proximity of their attribute values using the elbow method, allowing to identify players with similar characteristics in terms of player market contribution, club, and league in which they play.[4]. Once the clusters are formed, identify the patterns of each cluster to determine players with high potential. This will provide clearer insight into players with high market value, both at the individual, club, and league levels, so that it can help in selecting players with the potential to strengthen the Indonesian Senior National Team.

d. Evaluation of Analysis Results

The results will be visualized and analyzed manually.

2.4 Expected results

The structured dataset contains information on potential players for the Indonesian National Team, allowing coaches to conduct data-driven analysis in decision-making. This process supports increased efficiency in player selection for important matches, such as the 2026 World Cup qualifiers, by providing a clearer picture of player potential based on key attributes such as player market value, club market value, and league market value. This helps coaches in selecting the right players to strengthen the team, focusing on the quality and market contribution of the players.

2.5 Research Flow Diagram

1. Data Cleaning: Missing Values → Data Cleaning → Normalization
2. Clustering: Player Attributes → K-Means → Pattern Analysis
3. Results Evaluation: Data Visualization → Manual Analysis

2.6 Code Implementation in Google Colab

The entire process of data collection and analysis is done using Google Colab, which ensures ease and efficien

3. RESULTS AND DISCUSSION

3.1Research result

a. Data Collection from Transfermarkt

Data collection was carried out on the Transfermarkt.co.id site. The data that was successfully collected includes:

- Player information: Name, age, player market value, club market value, and league market value.
- Total data: A total of 22 players relevant to the research criteria, including players of Indonesian descent and active players in national and international leagues.

b. Data Cleaning and Normalization

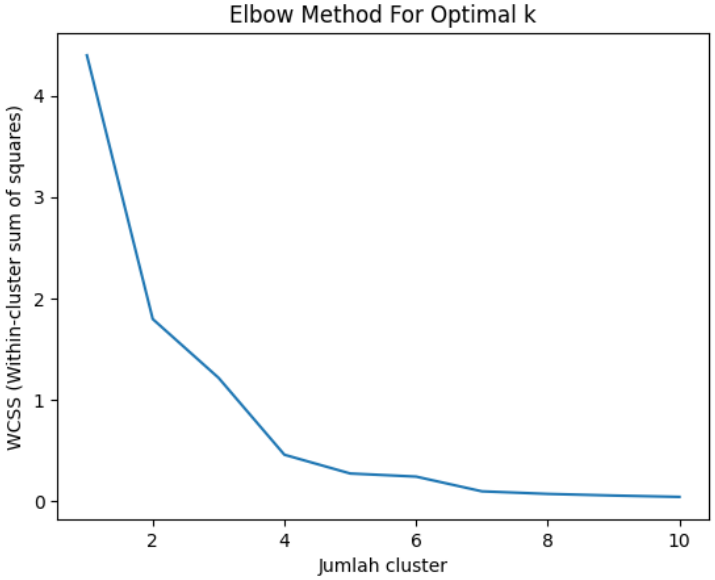
```
df = df.dropna()
scaler = MinMaxScaler()
df[["Market Value Player","Market Value Club","Market Value League"]] =
scaler.fit_transform(df[["Market Value Player","Market Value Club","Market Value League"]])
```

Data Cleaning: Duplicate and empty data are removed. Data Normalization: Player Market Value, Club Market Value and League Market Value columns are normalized so that the values are not too large compared to other attributes, using a scale of 0–1.

c. Cluster Count Testing

```
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(df[["Player Market Value","Club Market Value","League Market
Value"]])
    wcss.append(kmeans.inertia_)
```

In the Elbow method, the number of clusters (K) is varied from 2 to 11. For each value of K, the WCSS (Within-Cluster Sum of Squares) is calculated, which is the sum of the squares of the distances of each data point to its cluster center. When the WCSS is plotted against the value of K, the resulting graph resembles an elbow shape. The largest WCSS value is found at K = 1 and decreases as the number of clusters increases. On the graph, there is a point where the decrease in WCSS slows down significantly and the graph begins to move almost parallel to the X-axis. This point indicates the optimal number of clusters or the best K value. The following are the results of data testing using the Elbow method[4].



Picture1K-Cluster Determination Graph (Elbow)

Number of Clusters (K): Determined as many as 4 clusters based on evaluation using the elbow method.

d. K-Means Model

Model building is done using the KMeans class from the scikit-learn library.

```
kmeans = KMeans(n_clusters=4)
df["Cluster"] = kmeans.fit_predict(df[["Market Value Player","Market Value Club","Market Value League"]])
```

Table1Data cluster results

	Name	Player Market Value	Market Value Club	Market Value League	Cluster
0	Paes	0.199136	0.612038	0.244687	3
1	Walsh	0.110115	0.394248	0.186154	0
2	Rizky Ridho	0.054501	0.026536	0.002696	2
3	Idzes	0.33257	0.858858	1	3
4	The Verdonk	0.276957	0.441516	0.257932	3
5	Hey	0.165729	0.133916	0.257932	0
6	Ferdinand	0.03223	0.242298	0.304086	0
7	Oratmangoen	0.065637	0.193002	0.186154	0
8	Struick	0.007207	0.032278	0.006347	2
9	Arhan	0.015591	0.087316	0.018949	2
10	Hilgers	1	1	0.257932	1
11	Yuda Edith Pratama	0.001669	0	0.002696	2
12	Diks	0.499476	0.95138	0.064585	1
13	Very	0.08234	0.099469	0	2
14	Adhitya Harlan	0	0	0.002696	2
15	Jacob Sayuri	0.03223	0.016675	0.002696	2
16	Muhammad Ferrari	0.023911	0.026536	0.002696	2
17	Shayne Patty's name	0.03223	0.088326	0.013071	2
18	Nathan Tjoe-A-On	0.037798	0.557668	0.304086	3
19	Witan	0.029478	0.026536	0.002696	2
20	Lucky Caraka	0.021159	0.008641	0.002696	2
21	Eliano Reinders	0.060069	0.290925	0.257932	0

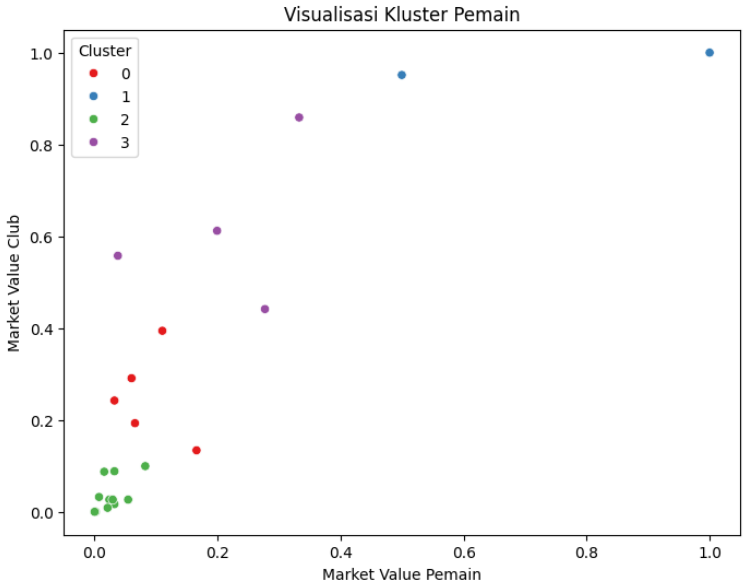


Figure 1. Distribution form between clusters

2.2 Discussion

1. Player Pattern Analysis in Each Cluster
   - Cluster 0: Players in this cluster show good potential, especially young players who are expected to develop rapidly in the future. Although their market value is not as high as players in other clusters, their long-term potential is quite promising.
   - Cluster 1: Players in this cluster have very good potential, with many playing for big clubs. An example is a player like Hilgers, who has the highest market value for both the player and his club, showing very competitive quality and performance.
   - Cluster 2: Players in this cluster show lower potential due to their relatively small market value, reflecting the lower level of competitiveness in their league. This suggests that they may struggle to compete at the 2026 World Cup qualifying level.
   - Cluster 3: Players in this cluster have quite good potential, with most playing in clubs with high market value. An example is a player like Idzes, who plays in the league with the highest market value, shows very good quality and competes in a more competitive league compared to players in other clusters.
2. Benefits of Clustering Results for the Indonesian National Team
   - Most Potential Players in the Main Lineup: Players included in clusters 1 and 3 are very suitable to fill the main lineup of the Indonesian National Team. On average, players in these two clusters have solid experience, both individually and in the league where they compete. Their presence in the main position can provide significant contributions to the team.
   - Players as Backups to the Main Lineup: Players in cluster 0 are good backups in case a key player is injured. While they may not have the individual skill or league experience of players in other clusters, they show great potential for growth, making them reliable backup options.
   - Players Who Still Need More Experience: Players in cluster 2 have dual characteristics. For young players, they still have time and opportunities to hone their skills and develop better in the future. Meanwhile, senior players in this cluster may not need to be called up to the Indonesian national team, because their declining stamina can affect their performance in the 2026 World Cup qualifiers.
3. Limitations and Recommendations

- Limitations:
  - The analysis does not take into account non-technical factors, such as player mentality and adaptation.
- Recommendation:
  - It is necessary to add data from other sources, such as coach reports or player GPS data during training.[5].
  - Use of other algorithms such as Decision Tree or Random Forest to evaluate more variables that affect player performance[6][7]

4. Conclusion

This study successfully utilized data mining technology to support the process of selecting potential players for the Indonesian Senior National Team, especially in preparation for the 2026 World Cup qualifiers. The following are the main conclusions:

1. Data Collection Efficiency

   By utilizing the power of the transfermarkt.co.id site, this study is able to collect comprehensive player data, including player name attributes, player market value, club market value, and market value of the league where they play.

2. Patterns Identified Through Clustering

   The K-Means Clustering technique groups players into four clusters based on the attributes of player market value, club market value, and league market value in which they play:
   - Cluster 0 shows young players with great potential to develop, although their market value is not as high as other players. These players have promising long-term prospects.
   - Cluster 1 includes players with exceptional potential, many of whom play at big clubs with high market value, such as Hilgers, who demonstrate high competitive qualities.
   - Cluster 2 consists of players with lower potential, reflected by their smaller market value, indicating that they may face challenges competing at international level, such as the 2026 World Cup qualifiers.
   - Cluster 3 contains players with quite good potential, playing in clubs with large market value, such as Idzes, who compete in competitive leagues and show better quality compared to players in other clusters.

3. Data-Driven Decision Enhancement

   Data-based analysis of this study provides an objective solution for coaches in determining the most suitable players for team strategy. This is expected to increase the efficiency of player selection and increase the chances of the Indonesian National Team in facing important matches such as the 2026 World Cup qualifiers.

4. Limitations and Suggestions for Further Research
   - This research is still limited to the data available on Transfermarkt, so there is an opportunity to integrate other data sources.
   - Non-technical factors such as player mentality, team adaptation and physical fitness need to be considered in future analysis.

With an approach that combines data mining technology, this research contributes to the development of national football through more objective data-based player selection.

REFERENCES

[1] SY Pradita and PHP Rosa, "Decision Support System for Determining Player Positions in a Football Team Using the Modified Simple Multi Attribute Rating Technique (SMART) Method,"*Nas. Seminar. Comput. Science.*, vol. 4, no. 2, pp. 365–370, 2016.

[2] NNQ Lutfillah and H. Purnomo, "Determinants of Market Value of Professional Football Players in League 1 Indonesia and Thailand,"*Equity*, vol. 25, no. 2, pp. 34–49, 2022, doi: 10.34209/equ.v25i2.4554.

[3] E. Irfiani and F. Indriyani, "Data Mining for Web-Based Class Promotion Decision Making System,"*INFORMATICS Educ. Prof. J. Informatics*, vol. 2, no. 1, pp. 19–28, 2017, [Online]. Available: http://ejournal-binainsani.ac.id/index.php/ITBI/article/view/582

[4] Ahmad Harmain, P. Paiman, H. Kurniawan, K. Kusrini, and Dina Maulina, "Data Normalization for K-Means

Efficiency in Grouping Areas with Potential for Forest and Land Fires Based on Hotspot Distribution,"*Tech. Technol. Inf. and Multimed.*, vol. 2, no. 2, pp. 83–89, 2022, doi: 10.46764/teknimedia.v2i2.49.

[5] H. Pamungkas, KK Aji, R. Prasetiyo, H. Yusuf, and M. Nidomuddin, "Analysis of Professional Football Player Performance Using GPS,"*J. Educator. Sports*, vol. 12, no. 2, pp. 220–230, 2023.

[6] U. Kalsum, "Use of Decision Trees for Decision Making in Employee Hiring,"*Education*, vol. 1, no. February, pp. 22–23, 2009, [Online]. Available: http://repository.uin-suska.ac.id/10813/%0Ahttp://repository.uin-suska.ac.id/10813/1/2010_201005TIF.pdf

[7] Suci Amaliah, M. Nusrang, and A. Aswi, "Application of Random Forest Method for Classification of Coffee Drink Variants at Konijiwa Bantaeng Coffee Shop,"*VARIANCE J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.