Multimedia Tools and Applications (2024) 83:29897–29914
https://doi.org/10.1007/s11042-023-16819-3

K Means clustering and descriptive analytics based performance recommending system for Kabaddi team and player

Vikas Khullar[1]

Received: 4 August 2022 / Revised: 26 April 2023 / Accepted: 31 August 2023 /
Published online: 14 September 2023

Abstract
In the contemporary era, Kabaddi is considered as a commercial game. Manual analytics were used in old times in which the best team or player was selected based on the background. However, in the present era, statistical analysis and machine learning are used in place of conventional methods of sports analytics in commercial sports such as Cricket, Football, etc. This study intends to analyze correlations for features of accumulated online datasets and to make predictions for the team and player performance using correlated features by using statistical machine learning approaches. The suggested methodology includes feature extraction, correlation identification, and implementing appropriate machine learning approaches. Initially, the parameters of the Kabaddi game concerns such as the impact of tosses, cards, and home ground on results were analysed. Subsequently, based on team and player data correlation and cluster analysis, important characteristics were identified and appropriate rank-based scoring was established. Finally, the regression-based prediction was suggested with an r2 score and a cross-validation score greater than 0.91 with the least errors. Finally, a trained machine-learning model with greater outcomes was suggested by verifying the parameters that were analysed. After the completion of analysis, the proposed techniques would be utilized in real-time scenarios using visual dashboards, deep learning, the Internet of Things, etc.

Keywords Sports analytics · Machine learning · Regression · Clustering · Correlation analysis · Forecasting

1 Introduction

The Kabaddi was known as a common sport of Indian rural youth and was famous in several Indian states like Punjab, Haryana, Maharashtra, Kerala, Tamil Nadu, etc. After Indian Independence for bringing interest Kabaddi in Indian youth, the Kabaddi Federation of India was established in 1950 for standardizing the rules and regulations of the game [21].

✉ Vikas Khullar
vikas.khullar@gmail.com

[1] Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

However, it was in 1978 that the Kabaddi became familiar to Asia and the world which resulted in the establishment of the Asian Kabaddi Federation in 1978 and the International Kabaddi Federation in 2004 [7]. Unfortunately, until 2010, Kabaddi was a very less perceived game community as compared to other sports such as Cricket, Football, Hockey, etc., both national and international. Due to the efforts made by famous players and entrepreneurs, to improve and financially support the game of Kabaddi, the current status and standard of Kabaddi has changed. Kabaddi game has gained fame in the world of professional sports and generating higher revenues in the form of commercial leagues such as the Pro Kabaddi League (PKL), etc. [7]. In 2021, PKL media rights sold in the amount of 900 crore Indian Rupees, and the player with the highest auction price in PKL was 93 lakhs Indian Rupees [22]. KP sponsorship rights were sold in 2017 to VIVO in the amount of 300 crores Indian Rupees [24]. According to data obtained from the ESPN Sports network, the total prize money in PKL raised 300% from 2 to 8 crores Indian Rupees [2]. According to the Broadcast Audience Research Council, the last PKL was successful in attracting 1.2 billion viewers [23] Kabaddi is gaining recognition and popularity in the collaborated ecosystem as a professional game for earning name-and-fame. For understanding, the basic anatomy of the Kabaddi game, it is a contact-based team game, which includes two teams. Both groups operate at their ends one by one around the group area, division line indicated at a rectangle. Turn by turn, a single player raids across the centreline to tag rival players and arrive back at their place within a specified time [3].

In further sections, "related work" is concluded with the focused objectives of this study. In the section "Materials and methods" the details about implementing techniques, data collection, pre-processing, feature selection, and implementation procedure have been discussed. Further, in the results section, a detailed analysis is conducted including, "team-wise exploratory analysis", "analysis of home and away teams to win", "analysis of defending and raid points on wins", "analysis of toss results on wins", "analysis of team scored points on cards issues", "team data cluster analysis", "player data /cluster analysis", "prominent features identification", and "team-player combined recommendation system".

2 Related work

Analogous to the other local games, manual or conventional analytics were conducted in local-level games to enhance the performance of the team or players. While doing manual or traditional procedures for analysis, there are many drawbacks. For example, more time required, the probability of encountering the human error, getting partial results, having less mathematical accuracy, etc. From the current literature, it is identified that computer-based sports analytics can conquer the lacunas of traditional sports analytics. In sports analytics, descriptive analytics with fundamental statistical theories and current trending machine learning (ML) approaches are used [20]. Ofoghi et al., implemented a decision-making technique to analyse triathlon players' performance by implementing Bayesian networks and propose to find better split times to improve the performance of players [13]. Kabaddi player performance as an independent variable was analysed by applying regression analysis over dependent variables, i.e., cardiovascular endurance, flexibility, agility speed, and explosive strength [18]. Descriptive analytics refers to data visualizations, statistical analysis, and many more, whereas ML algorithms have been able to supply game predictions or forecasting regarding future happenings by available data [12, 14]. Knobbe, et al., statistically explored historical data of professional speed skating players and used

hidden facts to improve training patterns. Aggregation functions such as addition, average, maximum, minimum, etc., were implemented to identify the duration, load, and intensity of applied training exercises, and accordingly, training sessions were allocated to players. Further regression models were implemented to identify the relation between the identified parameters and outcome for future predictions [9, 10]. Data is key for any prediction or forecasting. Constantinou and Fenton focused on smart data collected through knowledge discovery approaches and then filtered for required features. Further, a Bayesian neural network was implemented for time series forecasting using featured data and resulted in a better comparison of raw data sets [4]. Autoregressive integrated moving average (ARIMA) and recurrent neural networks (RNNs) based on an approach were proposed for the English Premier Football League. The proposed approach was capable of predicting the score for individual players or teams.

Maier worked to predict the Biathlon Shooting performance of players by identifying factors affecting performance and then applying ML models based on identified factors. In this, the author achieved 60, 61, and 62 percent area-under-curve by applying a logistic regression model, artificial neural network model, and decision tree boosting model [11]. Apostolou & Tjortjis worked for football season datasets to identify three aims, (a) classification of footballer position, (b) prediction number of goals by renowned players, and (c) the number of shoots by a player in a particular match. Hybrid Random Forest and Sequential Minimal Optimization were implemented to achieve mentioned aims with 85 percent accuracy [1]. Recent trends in sports analytics for identifying the performance of players and teams playing basketball under the National Basketball Association (NBA) were discussed [19] and highlighted the capabilities of statistics and ML in sports analytics. By using ML approaches, Dieu, et. al, determined player performance by considering player parameters such as structural, functional, technical, contextual, and expertise. The features of collected data of considered parameters were identified by applying the Principal Component Analysis (PCA) algorithm and further, these parameters were used in Random Forest for prediction analysis [5]. Here, the author compared the subjective (manual) data analysis with objective (ML) based data prediction and found ML-based objective methods were better than subjective ones. With the enhancement in data volume, big data platforms and deep learning algorithms have also established their role in sports analytics. Team behavior is one of the requirements for improved performance. In Fujii, et. al, data-driven machine learning approaches have been utilized for quantitative behavioral analysis of football & basketball teams, and further, it was compared with quantitative rule-based approaches [6]. Kaur, et. al., had worked to identify the popularity of the player using deep learning-based text sentiment analysis over the big data environment [8].

The aim of the current study was established based on the literature reviewed. In searching the literature, a very less number of papers related to Kabaddi were identified. However, in other sports such as Cricket, Football, etc., sports analytics was one of the dominant fields of study which were used to improve team or individual performance. The main objective of this study is to create a robust and flexible system to provide a descriptive, predictive, and perspective level of analytics for Kabaddi to understand and enhance the level of the game. The aimed objectives for the study were as follows:

1. *To find the number of wins correlated to home teams and away teams.*
2. *To find the number of wins correlated to defending points and raid points.*
3. *To find the number of matching results correlated to the number of toss results.*
4. *To find the number of team points correlated to the number of cards issued.*

5. *To apply Teams and Players data cluster analysis to find the collective features of teams and players*
6. *To create a regression-based predictive model to forecast team and player combined ranking.*

3 Materials and methods

This section provides details for data collection, data pre-processing, feature selection, and implementation procedure. The further details are divided into subsections as mentioned below:

3.1 Implemented techniques

In this paper, correlation analysis and cluster analysis were utilized to verify the mentioned hypothesis and also to find out prominent player and team performance features. Correlation helps to identify prominently related features of a group of features. It allows us to focus on the features of concern and to identify features that are not of use. Cluster analysis is the sub-technique of discriminant analysis. It is implemented with a group of similar observations, including several variations. Cluster analysis is grouping up the individual objects of the same group based on their statistical similarity. The resultant goal of cluster analysis is to find out similar properties and fix them into the same group. There are various clustering algorithms are available that operate based on different properties such as distance, centroid, density, distribution, etc. By using cluster and correlation analysis, the best-fitted features were identified in this study. Regression is an approach to identifying the relationship between dependent and independent variables that further could be used for making predictions. In this case, regression was used to create predictive potent models for team and player-selected data.

3.2 Data collection, pre-processing, and feature selection

The data of Kabaddi was collected from web scrapping, open data source websites, the dataset from published resources, etc. Then, the utilized resources and data-related other information were mentioned in Table 1. Along with the data collection, data pre-processing was employed for clean and relevant data collection in the form of comma-separated format files. The final table was created and then the details of the other characteristics were mentioned. Table 2 was formed including details regarding team statistics and covering diverse team features including points scored, win-lose detail, and cards issued during

Table 1 Data Sources Detail

Method	Source Link	References
Pro Kabaddi League 2019 Data Online Source	https://www.kaggle.com/sujaypandit/prokabbadi-league-2019	[15]
Pro Kabaddi League 2019 Data Online Source	https://github.com/kirtiraj23/ProKabaddiHackaThon	[16]
Pro Kabaddi Leagues Data Online Source	https://www.prokabaddi.com/	[17]

Table 2 Pro Kabaddi Team Details Master Data

Column Name	Data Type
Team Name	String
Team ID	Integer
Defending Points	Integer
Raid Points	Integer
Number of Win Toss and Win Game	Integer
Number of Win Toss and Loose Game	Integer
Total Played Matches	Integer
Total Wins	Integer
Total Looses	Integer
Player Received Red Cards	Integer
Player Received Yellow Cards	Integer
Player Received Green Cards	Integer

matches. The collected data was scaled in percentage concerning the total played matches. Table 3 was made which included team-wise details for different scored points at the granule level in terms of raid points, tackle points, super points, do-die points, etc. Detail for the Pro Kabaddi player was given in Table 4. Data features mentioned in the tables were collected from different resources in partition and combined to draw meaningful insights.

3.3 Implementation procedure

This paper mainly utilized correlation analysis and cluster analysis to achieve the defined hypothesis. Subsequently, based on team and player data correlation and cluster analysis, important characteristics were identified and appropriate rank-based scoring was established. The stepwise procedure is explained as follows:

Table 3 Pro Kabaddi Team Wise Point Details

Column Name	Data Type
Team Id	String
Allouts Conceded	Integer
Allouts Inflicted	Integer
Average Points Scored	Integer
Average Raid Points	Integer
Average Tackle Points	Integer
Conceded Points	Integer
Do or Die Points	Integer
Total Points Scored	Integer
Raid Points	Integer
Number of Successful Raids	Integer
Number of Successful Tackles	Integer
Super Raids	Integer
Super Tackles	Integer
Tackle Points	Integer

Table 4 Pro Kabaddi Players Detail

Column Name	Data Type
Player Name	String
Match Played	Integer
Points	Integer
Career Best Points	Integer
Not Out Percentage	Integer
Raids	Integer
Integer Successful Raids	Integer
Unsuccessful Raids	Integer
Empty Raid	Integer
Successful Raid Percentage	Integer
Raid Touch Points	Integer
Raid Bonus Points	Integer
Total Raid Points	Integer
Super Raids	Integer
Super 10 s	Integer
Tackles	Integer
Successful Tackles	Integer
Unsuccessful Tackles	Integer
Successful Tackles Per Match	Integer
Tackle Bonus Points	Integer
Tackle Success Rate	Integer
Super Tackles	Integer
High 5 s	Integer

1. Correlation analyses were conducted to identify relational dependencies as follows:
   a. The number of wins versus home teams and away teams.
      The exploratory analysis explained the behavior associated with the feature in terms of central tendencies, deviations, skewness, kurtosis, interquartile ranges, and so on. The impact of home or away teams to win percentage was determined in this study by evaluating the association between total victories and home/away teams. A number of wins versus defending points and raid points. The correlation between defending points and raid points was categorized against match wins and losses.
   b. The number of match results versus the number of toss results.
      A correlation study was performed to determine the impact of toss winning or losing on the number of match victories.
   c. The number of team points versus the number of cards issued to the team.
      The impact of the number of cards issued during their ongoing game was also analyzed to identify the impact of the game results.

2. Cluster Analysis was conducted to identify features dependencies as follows:
   a. Teams Data Cluster Analysis.
      A cluster analysis was implemented in team data to segregate or group the individual features based on their statistical similarity.

b. Players Data Cluster Analysis.
A cluster analysis was implemented in the player data for grouping individual features based on their statistical similarity. In this section, three different clustering algorithms were implemented for cluster analysis i.e., k-means, algometric, and mean-shift.
c. Selecting Top Features from team and player data.
Prominent feature identification from the team and players' data is implemented for predicting the best combination of team and player.

4 Results and discussions

The aimed outcomes as mentioned in the objectives were achieved and verified in this section. Various areas of the Pro Kabbadi league dataset were explored in terms of correlation and cluster analysis to achieve decided aims. Result-wise details were discussed in further sections.

4.1 Team wise exploratory analysis

Table 5 explained the team-wise statistical descriptions of considered database features. The conducted exploratory analysis explained the feature's behavior in terms of central tendencies, deviations, skewness, kurtosis, interquartile ranges, etc. All 12 teams were at the same experience level as all teams had played games in the average range of 18 to 20. Teams had a high deviation in points-related features like Total points scored, Total points conceded, Raid points, and Tackle points resulting in a standard deviation of 55.4, 45.53, 41.98, and 16.84. These high deviation features made a winning or losing impact on the

Table 5 Team wise Exploratory Analysis

Variable	N	Min	Max	Mean	Median	Standard Deviation	Skew	Kurtosis	IQR
Games	12	18	20	19.33	19.5	0.75	-0.63	-0.96	1.00
Total points scored	12	551	742	631.67	633	55.44	0.47	-0.55	60.00
Total points conceded	12	561	723	631.67	633.5	45.53	0.32	-0.50	47.75
Average points scored	12	25.8	31.2	28.63	28.985	1.87	-0.14	-1.45	3.53
Successful raids	12	223	361	286.33	282.5	35.63	0.51	-0.05	31.50
Raid points	12	289	437	358.83	351	41.98	0.40	-0.66	50.75
Average raid points	12	16	21.9	18.55	18.17	1.94	0.30	-1.26	3.00
Successful tackles	12	160	207	177.50	179	12.76	0.66	0.01	15.25
Tackle points	12	171	232	194.83	191.5	15.84	1.01	0.48	11.00
Average tackle points	12	8.55	11.6	10.08	10.08	0.75	0.04	0.12	0.74
Super raid	12	4	15	8.42	7.5	3.52	0.71	-0.84	4.25
Super tackles	12	8	32	18.67	17.5	6.49	0.35	-0.56	9.75
Do-or-die raid points	12	42	73	59.42	59.5	10.32	-0.25	-1.25	17.25
All-outs inflicted	12	13	34	23.00	21	6.62	0.27	-0.93	8.50
All-outs conceded	12	17	33	23.00	22	4.30	1.03	0.50	2.25

game. High-skewed features could influence the training of machine learning algorithms. However, low-skewed features are better regression models. Here, the kurtosis values ranged between + 1 to -0.7 which reflects an appropriate possibility to support prediction modeling, as the accepted range was between + 3 to -3. The possibility of outliers in features identified with a higher interquartile range is also high. Therefore, it would help us to identify and remove outliers from the available database. As shown in Fig. 1, the box plot explained the available outliers in data with high interquartile ranges such as Total Points Conceded, Successful Raids, Successful Tackles, Tackles Points, etc.

4.2 Analysis of home and away teams to wins

As per human thoughts, there may be an impactful relation between home and away teams on the number of wins. In this section, analysis of the impact of home-playing or away-playing teams on win percentage was calculated by finding the correlation between total wins and home/away teams as shown in Fig. 2. Experimented results reflected 0.32 as the correlation value in terms of the home team and 0.49 in terms of the Away team. Hence, the away team was expected to have more chances to win in comparison to the home team as mentioned.

4.3 Analysis of defending and raid points on wins

As per Kabaddi rules, the number of wins depends on the defending points or raid points. However, out of both defending and raiding for points, the one that required more team focus was identified in this section. Here, the correlation between defending points and raid points was categorized against match wins and losses. The applied method resulted

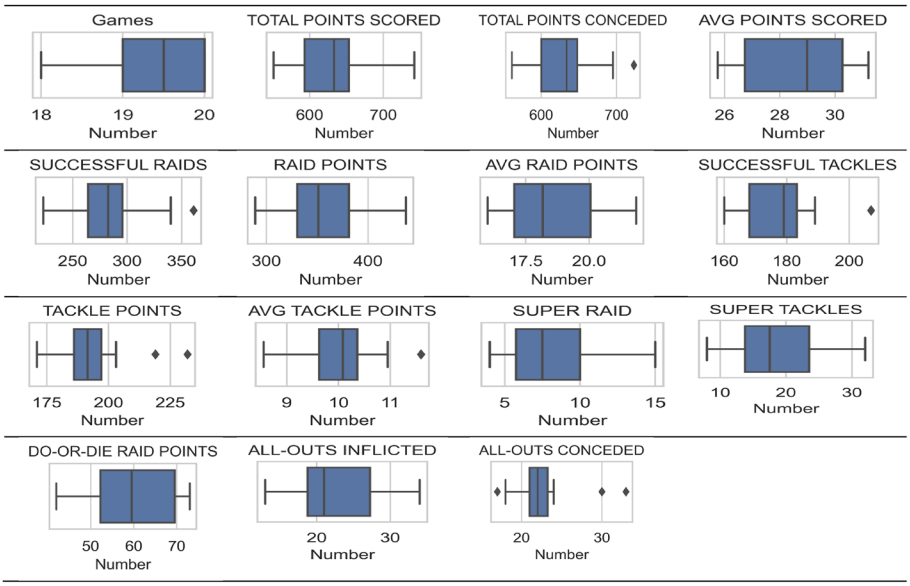

Fig. 1 Box Plots for Team Features

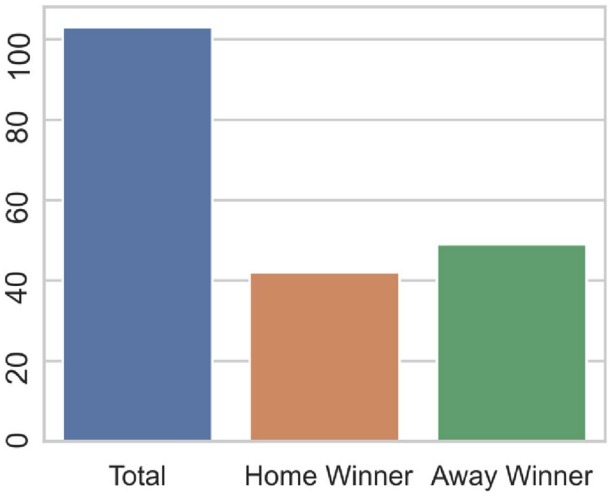

Fig. 2 Analysis of Home and Away Teams on Wins

in a correlation value of 0.42 for the raid–win scenario on a scale of 0 to 1, whereas the defending-win scenario resulted in 0.51. As shown in Fig. 3, there is a minor difference in both correlations; however, defending points resulted from more help in game-winning in comparison to raiding points.

4.4 Analysis of toss results on wins

In sports, "win tosses win games" is a very popular phase. In this section, a similar kind of analysis was conducted to identify the impact of toss wins or losses over the number of match wins by implementing correlation analysis. From Fig. 4 it was observed that the winning toss impacted on win game as it resulted in a positive correlation (0.70) and a loose game with the inverse correlation (-0.51). A positive correlation could be considered sufficient proof to explain the impact of winning tosses on winning games.

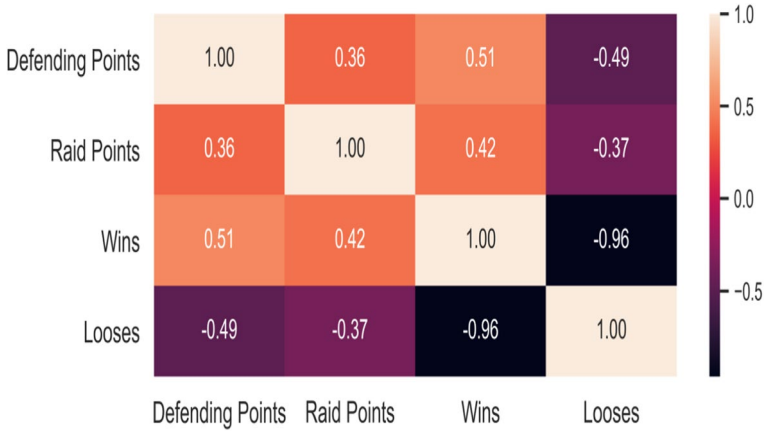

Fig. 3 Analysis of Defending and Raid Points on Wins

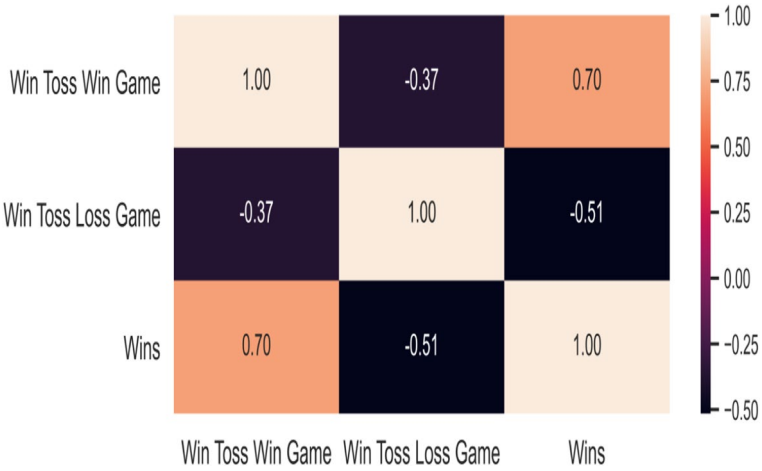

Fig. 4 Analysis of Toss Results on Wins

4.5 Analysis of team scored points on cards issues

During games, various levels of cards such as Green, Yellow, and Red are issued to control the players for maintaining game-appropriate behavior amongst players. The impact of number of cards issued during their ongoing game could also impact the game results. In this section impact in terms of the correlation of issued cards in the game, results were identified. As shown in Fig. 5, a very low correlation value between issued cards and wins /losses had identified, so from the results the considered hypothesis was not proved.

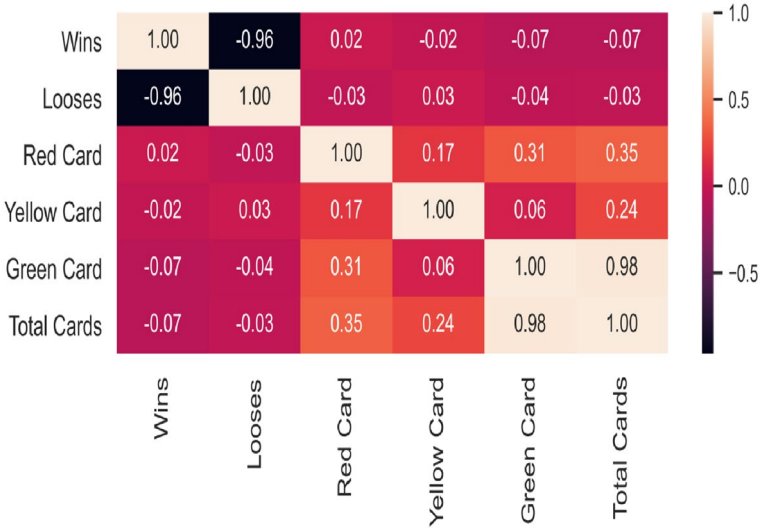

Fig. 5 Analysis of Team Scored Points on Cards Issues

Table 6 Silhouette Score for an implemented Clustering algorithm

Clustering Algorithms	Silhouette Score
K Means	0.69
Algometric	0.68
Mean Shift	0.59

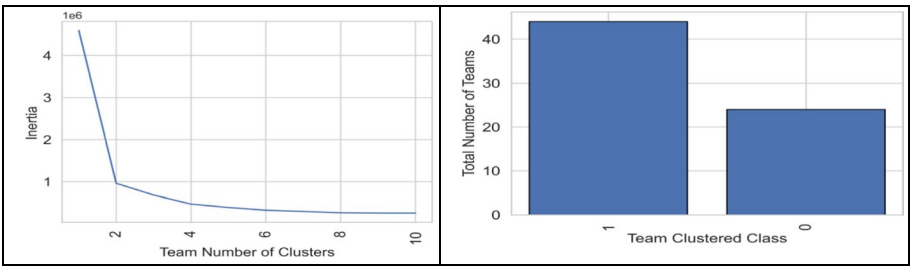

Fig. 6 Clustering Results for Teams Data

4.6 Team data cluster analysis

In this section, cluster analysis was implemented in team data to segregate or group the individual features based on their statistical similarity. Three different clustering algorithms were implemented for cluster analysis i.e. k-means, algometric, and mean-shift. Initially, by applying the knee method, several clusters were identified that resulted in two maximum clusters in team data. So, by working with two clusters silhouette scores were fitted to identify the best algorithm out of all three, in which k-means resulted in the best values. K Means resulted in the highest parametric score of 0.69 in comparison to other mentioned algorithms as presented in Table 6. Further cluster analysis was conducted using k-means. The results of the clustering process were shown in Fig. 6. Through k-means, team data was added with cluster names or labels, and now the team data can be used as classification data. In Fig. 7, team data correlation with identified labels through k-means was depicted.

4.7 Player data /cluster analysis

In this, cluster analysis was implemented in the player data for grouping individual features based on their statistical similarity. In this section, three different clustering algorithms were implemented for cluster analysis i.e. k-means, algometric, and mean-shift. In the first step of applying the knee method, the number of clusters was identified that resulting in two maximum clusters in player data. So, by working with two clusters silhouette scores were fitted to identify the best algorithm out of all three, in which k-means resulted best. K Means resulted in the highest parametric score of 0.41 in comparison to other mentioned algorithms as presented in Table 7. Further cluster analysis was conducted using k-means and the results of the clustering process are shown in Fig. 8. Through k-means, player data was added with cluster names or labels, and now our team data can be used as classification data. In Fig. 9, team data correlation with identified labels through k-means was depicted.

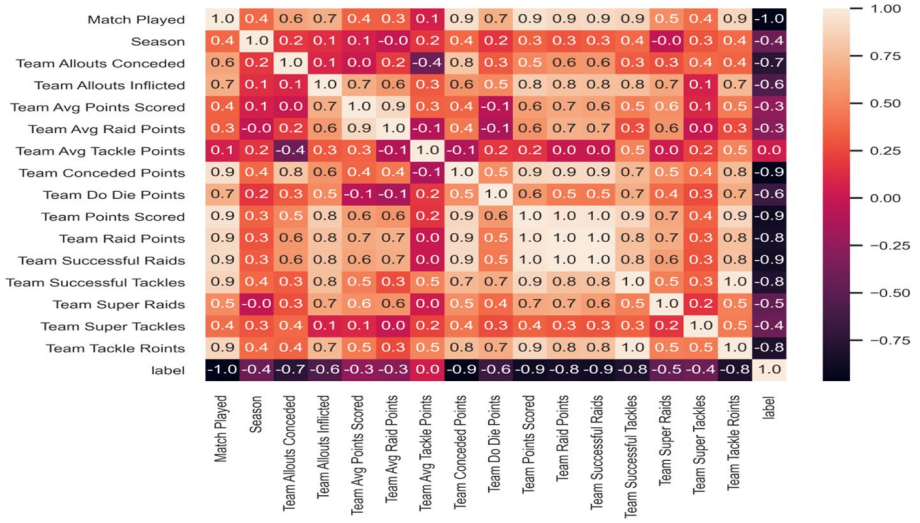

Fig. 7 Analysis of Team Data on K Means defined Labels

Table 7 Silhouette Score for an implemented Clustering algorithm

Clustering Algorithms	Silhouette Score
K Means	0.41
Algometric	0.39
Mean Shift	0.36

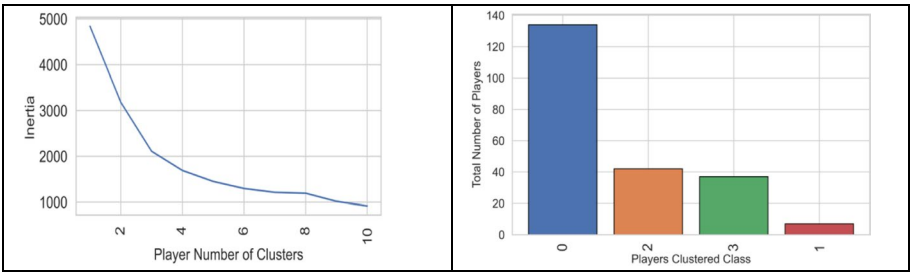

Fig. 8 Clustering Results for Players Data

4.8 Prominent features identification

Prominent feature identification from the team and players' data was required to implement models for predicting the best combination of team and player. Features highlighted with a correlation above 0.5 (either positive or negative) for the team and players in Figs. 7 and 9 were utilized in a combined manner to shortlist as prominent features. For more precise results, the range could get updated. The details of selected prominent features along with correlation parameters were shown in Table 8.

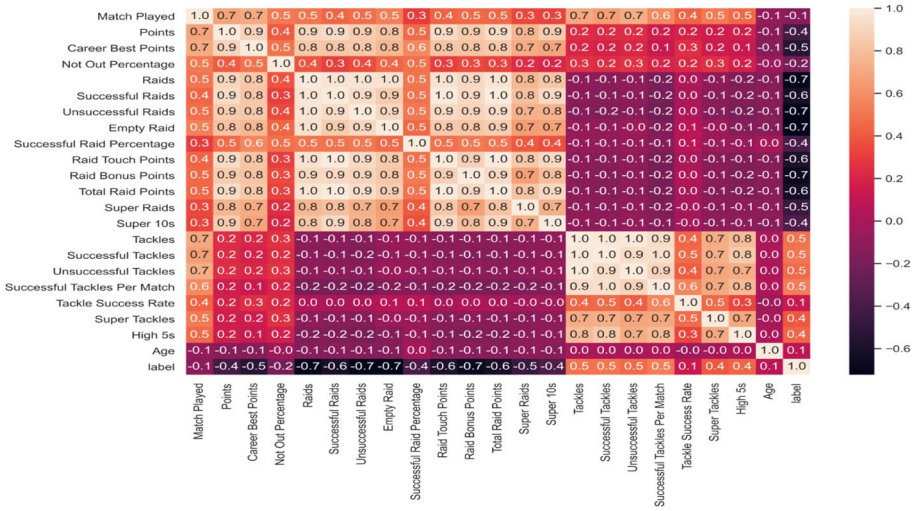

Fig. 9 Analysis of Players Data on K Means defined Labels

Table 8 Prominent Selected Features of Team and Players Data

Features	Identified Correlation
Player Match Played	0.800416097
Player Total Points	0.728359505
Player Career Best Points	0.646728381
Player Tackles	0.701177825
Player Successful Tackles	0.692634714
Player Unsuccessful Tackles	0.688584346
Player Successful Tackles Per Match	0.609096497
Player Super Tackles	0.538694142
Player High 5 s	0.548127709
Team Match Played	0.965638795
Team Allouts Conceded	0.735781198
Team Allouts Inflicted	0.551144676
Team Conceded Points	0.949361398
Team Do-Die Points	0.646469576
Team Points Scored	0.875216534
Team Raid Points	0.848100206
Team Successful Raids	0.867349794
Team Successful Tackles	0.796462629
Team Tackle Points	0.807249284

4.9 Team-player combined recommendation system

The possible dependency of both team's statistics and the player's statistics are dependent on each other. So, the final implementation concern in this study was focused on proposing a Team-Player combined recommendation system to predict the rank of combinations in

Table 9 Example Features Quartile Conversion Results with Score

Team	Player Name	Player Match Played	Points	Career Best Points	Tackles	Unsuc-cessful Tackles	Team Match Played	Team Allouts Conceded	Team Allouts Inflicted	Team Conceded Points	Team Do Die points	Team Points Scored	Team Raid Points	Team Successful Raids	Team Successful Tackles	Team Tackle Points	Score
Bengal Warriors	Maninder Singh	3	3	3	2	2	1	3	1	2	1	1	2	2	1	1	28
Bengal Warriors	Viraj Vishnu	2	2	2	3	3	2	1	3	2	1	2	2	2	2	2	31
Bengaluru Bulls	Amit Sheoran	3	3	2	3	3	1	2	1	2	2	1	1	1	1	1	27
Dabang Delhi K.C	Sumit	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	17
Haryana Steelers	Naveen	2	3	3	2	2	2	3	1	3	1	3	3	3	3	3	37
Jaipur Pink Panthers	Karamvir	1	1	1	1	1	2	1	1	1	3	2	2	1	1	1	20

Table 10 Example Features Actual Value with Score

Team	Player Name	Player Match Played	Points	Career Best Points	Tackles	Unsuccessful Tackles	Team Match Played	Team All Outs Conceded	Team All Outs Inflicted	Team Conceded Points	Team Do Die points	Team Points Scored	Team Raid Points	Team Successful Raids	Team Successful Tackles	Team Tackle Points	Score
Bengal Warriors	Maninder Singh	20	205	19	7	7	14	25	11	515	30	430	283	213	106	122	28
Bengaluru Bulls	Mahender Singh	16	46	7	90	51	24	24	37	766	73	887	547	431	219	230	44
Dabang Delhi K.C	Aman Kadian	3	4	2	2	1	14	16	15	412	84	419	226	162	116	134	18
Gujarat Fortune-giants	Lalit Chaudhary	4	5	3	2	0	25	20	36	769	74	876	480	392	258	274	35
Jaipur Pink Panthers	Nilesh Salunke	17	55	15	7	5	22	32	22	780	75	711	415	334	202	225	40
Tamil Thalaivas	Yashwant Bishnoi	3	2	2	2	2	15	19	10	488	43	432	253	203	124	132	17

Table 11 Linear Regression Models Metrics Testing Data Comparison on Team-Player Combined Dataset

Algorithm	Mean Squared Error	Mean Absolute Error	R2 Score	Cross Validation Score
Linear Regression (LR)	3.726	1.525	0.938	0.912
Random Forest Regression (RFR)	9.518	2.544	0.843	0.835
Decision Tree Regression (DTR)	1.649	0.789	0.973	0.914
Support Vector Regression (SVR)	6.786	2.167	0.888	0.880
Multi Layered Perceptron Regression (MLPR)	6.359	2.036	0.895	0.870
Bayesian Ridge Regression (BRR)	3.777	1.530	0.938	0.912
Ridge Regression (RR)	3.727	1.525	0.938	0.912
Lasso Regression (LR)	3.975	1.568	0.934	0.913
Kernel Ridge Regression (KRR)	3.961	1.580	0.935	0.911
Passive Aggressive Regression (PAR)	7.383	2.096	0.878	0.718

terms of regression statistics. Data related to prominent features mentioned in Table 9 were utilized to create a ranking system. The concept of the ranking system was included, and each feature of the collected data was categorized into three numerical quartiles i.e. 1, 2, 3. Similarly, all features were converted into quartile categories as shown in Table 10. Other than these parameters, a new column "Score" was added in Table 9. Here, a novel method was used to identify score columns. The score column was included in the row sum of identified quartile integers for all individual features. The calculated score column was now treated as a dependent parameter for regression models along with actual feature values as independent parameters as shown in Table 10.

According to Table 10, team-player combined features and their corresponding score values were obtained. Further regression modeling was conducted using algorithms as mentioned in Table 11 and also analyzed through metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R2-Score (R2S), and Cross-Validation Score (CVS). The comparative analyses were conducted by testing data random fractions of 20 percent and algorithms trained on 80 percent data for reducing the partiality in model creation. Some of the models resulted in R2S and CVS above 90 percent, now these were focused training models. More precise results could be considered by identifying models with the least MSE and MAE along with high R2S and CVS. Such identified models were Linear Regression (LR), Decision Tree Regression (DTR), Bayesian Ridge Regression (BRR), Ridge Regression (RR), Lasso Regression (LR), and Kernel Ridge Regression (KRR). However, the least MAE and MSE along with high R2S and CVS resulted in decision tree regression.

5 Conclusion

The present era is of professional Kabaddi and required support with tools for statistical analysis and machine learning for making predictions. In this paper, initially, the correlations between Kabaddi parameters were identified. The results of correlation analysis showed, (a) the away team had more chances to win in comparison to the home team, (b)

defending points identified were more helpful in wins in comparison to raiding points, (c) toss win was identified impacted on wins, and (d) no relation was identified between issue cards during the match and wins or losses. After this, based on team and player data correlation and cluster analysis were conducted and important features were identified and appropriate rank-based scoring was identified. Finally, the regression-based prediction was proposed with an r2 score and cross-validation score of more than 0.91 with the least errors. A detailed analysis was conducted to develop a recommendation system to make a performance analysis of Kabaddi players and its team. This paper covers most of the aspects to defend the results of the proposed Kabaddi recommendation system. In the future, the created models could be used for creating applications such as Dream11, MPL, Analysing Dashboards, web embedding API's etc.

Data availability The open source dataset used in this paper [15–17]. The analyzed data will make available on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Apostolou K, Tjortjis C (2019) Sports analytics algorithms for performance prediction. 10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019, pp 1–4. https://doi.org/10.1109/IISA.2019.8900754
2. Bhagavatula M (2021) The improbable success of the Pro Kabaddi League. ESPN. https://www.espn.in/kabaddi/story/_/id/20170469/the-improbable-success-pro-kabaddi-league. Accessed 10 Jul 2022
3. Parmar MK (2017) KABADDI: from an intuitive to an quantitative approach for analysis, predictions and strategy. In: 5th International conference on Business Analytics & Intelligence (link) held at IIM Bangalore, India
4. Constantinou A, Fenton N (2017) Towards smart-data: improving predictive accuracy in long-term football team performance. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2017.01.015
5. Dieu O, Schnitzler C, Llena C, Potdevin F (2020) Complementing subjective with objective data in analysing expertise: a machine-learning approach applied to badminton. J Sports Sci 38(17):1943–1952. https://doi.org/10.1080/02640414.2020.1764812
6. Fujii K (2021) Data-driven analysis for understanding team sports behaviors. Machine learning based analysis for team sports behaviors paper, pp 1–9. http://arxiv.org/abs/2102.07545. Accessed 10 Jul 2022
7. Ghosh SS, Sarma AS (2018) The Evolution of Pro Kabaddi League in India. 4(4), 23–28
8. Kaur A, Kaur R, Jagdev G (2021) Analyzing and exploring the impact of big data analytics in sports sector. SN Comput Sci 2(3):1–19. https://doi.org/10.1007/s42979-021-00575-y
9. Kaur A, Vaid H, Mukhija L (2023) K-Means clustering for prophesy of freshmen's attainment with euclidean execution. 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), January, pp 1209–1215. https://doi.org/10.1109/IITCEE57236.2023.10090874
10. Knobbe A, Orie J, Hofman N, Van Der Burgh B, Cachucho R (2017) Sports analytics for professional speed skating. Data Min Knowl Disc 31(6):1872–1902. https://doi.org/10.1007/s10618-017-0512-3
11. Maier T, Meister D, Trösch S, Wehrlin JP (2018) Predicting biathlon shooting performance using machine learning. J Sports Sci 36(20):2333–2339. https://doi.org/10.1080/02640414.2018.1455261
12. Malik V, Mittal R, Mittal A, Singh J, Singla S, Kukkar A (2022) Applying data mining for clustering shoppers based on store loyalty. Proceedings - 2022 5th International Conference on Computational Intelligence and Communication Technologies, CCICT 2022, pp 370–373. https://doi.org/10.1109/CCiCT56684.2022.00073
13. Ofoghi B, Zeleznikow J, Macmahon C, Rehula J, Dwyer DB (2016) Performance analysis and prediction in triathlon. J Sports Sci 34(7):607–612. https://doi.org/10.1080/02640414.2015.1065341

14. Passfield L, Hopker JG (2017) A mine of information: can sports analytics provide wisdom from your data? Int J Sports Physiol Perform 12(7):851–855. https://doi.org/10.1123/ijspp.2016-0644
15. Pro-Kabaddi League (2019) https://www.prokabaddi.com/. Accessed 10 Jul 2022
16. Pro Kabaddi Hackathon (2021) https://github.com/kirtiraj23/ProKabaddiHackaThon. Accessed 10 Jul 2022
17. Pro Kabaddi League (2020) https://www.prokabaddi.com/. Accessed 10 Jul 2022
18. Sanjit S, Pandey AK (2016) An estimation of Kabaddi performance on the basis of selected phyasical fitness components. Indian J Phys Educ Sports Appl Sci 6(4):27–35
19. Sarlis V, Tjortjis C (2020) Sports analytics — evaluation of basketball players and team performance. Inf Syst 93:101562. https://doi.org/10.1016/j.is.2020.101562
20. Singh P, Parashar B, Agrawal S, Mudgal K, Singh P (2023) Kabaddi: a quantitative approach to machine learning model in Pro Kabaddi. Lect Notes Netw Syst 554:243–260. https://doi.org/10.1007/978-981-19-6661-3_22
21 Singh S, Srivastava DP, Patvardhan C (2023) Game theoretic analysis of Kabaddi. J Stat Appl Prob 12(1):313–319. https://doi.org/10.18576/jsap/120127
22. Vasudevan S (2021) Pro Kabaddi League team owners unhappy with media rights auction process. The Hindu. https://sportstar.thehindu.com/kabaddi/pro-kabaddi-league-media-rights-auction-star-india-pkl-2021-conflict-of-interest-nepotism-u-mumba-ronnie-screwvala-patna-pirates-telugu-titans/article34356692.ece. Accessed 10 Jul 2022
23. VIVO Pro Kabaddi League : A HIT amongst the masses (2021) Pro Kabaddi. https://www.prokabaddi.com/news/second-most-followed-sports-league-india. Accessed 10 Jul 2022
24. Vivo signs five-year sponsorship deal with Pro Kabaddi worth Rs 300 crore (2021) Scroll.In. https://scroll.in/field/837056/vivo-signs-five-year-sponsorship-deal-with-pro-kabaddi-worth-rs-300-crore. Accessed 10 Jul 2022