Supervised Learning Techniques for Predicting Pro-Kabaddi League Winner

1st Agrim Agarwal
Dept. of Computer Eng (of BITS Pilani,Goa)
f20170927@goa.bits-pilani.ac.in

2nd Manhal Rahman
Dept. of EEE (of BITS Pilani,Goa)
f20180310@goa.bits-pilani.ac.in

3rd Pallavi Varshney
Dept. of Computer Eng. (of BITS Pilani,Goa)
h20190029@goa.bits-pilani.ac.in

4th Shreya Kulkarni
Dept. of Computer Eng. (of BITS Pilani,Goa)
f20180171@goa.bits-pilani.ac.in

***Abstract*—Kabaddi is a game under looked by many machine learning projects on sports analysis. The Indian sport has less popularity compared to other worldwide popular games like cricket, football or basketball. The paper focuses on prediction of winner "Pro-Kabaddi League" data by analysis and visualisation. The analysis is done with supervised learning. We are creating a discriminative model using standard models like Decision Tree ,Logistic Regression, Voting Classifier and Random Forest Tree algorithms to predict the winner of the league and compare their performance.**

I. Introduction

Kabaddi is the sport of Indian origin and is a highly strategic game that can be analysed with its data generated and rules of the sport Pro Kabaddi League (PKL) was started in 2014 with eight teams from different cities. This league is formally backed by the International Kabaddi Federation (IKF), the Asian Kabaddi Federation (AKF) and the Amateur Kabaddi Federation of India (AKFI).The Pro Kabaddi league highlighted the new, modern, international and competitive face of kabaddi game throughout the world.It consisted of 12 teams from Indian states and 7 seasons from 2014 to 2019 with around 652 matches over all the seven seasons. We have chosen the Pro Kabaddi Leauge matches as our dataset for data analysis and prediction.

II. Literature Review

With increased focus on traditional games and giving them a international exposure, several games in India saw growth in the level and also monetory value got attached to them.Sports is all about decision-making on the field and off the field, considering multiple parameters. Kabaddi, as a sport, is not different in this respect. Kabaddi can benefit from analytics as it produces a variety of data at the team level and individual player level. Analytics is changing the horizons of sports and adding new perspective to strategies. Analytics helps players in understanding their weaknesses and strengths,assists the coaches in making informed decisions rather than intuitive decisions, and helps managers optimise the costs and improve the chances of winning by analysing data and various parameters. Literature review,however, points out, in general, the lack of application of sports analytics in the game of kabaddi. The motivation of this research is to demonstrate that sports analytics can be applied to games like kabaddi and analyse the features that affect the game and implementing the model to predict the winner of the match. A research paper by Manojkumar Parmar provides a quantitative approach to profile an entire tournament to gain a general understanding of the strengths of various teams.The paper discusses and provides a quantitative perspective on traditional strategies and conceptions related to the game of kabaddi such as attack and defence strategies. The research paper by Vinod Nimbalkar focuses on Analysis and Visualization of "Pro-Kabaddi League" data.The visualization represents different aspects of the game in the graphical manner to aid understanding of data.

III. Problem Definition

Prediction of winner of Kabaddi matches. Tasks include:
1. Data collection using web scraping from the pro-kabbadi website.
2. Feature Selection
3. Model selection and Training
4. Hyperparameter tuning
5. Metric selection and error analysis

IV. Data Scraping & Preprocessing

A. Data Scraping

The data set used for the study comprises the Pro Kabaddi data obtained from the official website of pro kabaddi:
https://www.prokabaddi.com/

libraries used - Selenium, pandas

Selenium is an open-source web-based automation tool. Selenium primarily used for testing in the industry but It can also be used for web scraping. The various pages on pro kabaddi website used for making the dataset contained code

written in javascript which made it difficult to scrape data using traditional techniques like using beautiful soup library or any other similar library. Selenium opens a web page in a browser just like a human and then scrapes the data - thus, bypassing the need to deal with various complexities. pandas is used for making changes to the final csv file to make the dataset for the models used in this project.

B. Preprocessing

- Records include all matches from season 1 to season 7 which is in CSV format.
- Dropped all draw matches in all seasons
- Combined all seasons data into one file. For example we have combined all seasons raid points for team-"JAIPUR PINK PANTHERS" under one column "RAIDPOINTS"
- We have focused following features in our dataset- Team, Total Matches, Total raids, Total Points,Scores,Allouts,Tackle points, Extra points,OpExtra.
- We have also dropped winnerID and winner team name from scraped data set.
- We have divided data-set into training, validation and testing in 6:2:2 respectively.
- The data set is divided into three parts: season 1 to 5 matches data is used for training, season 6 data is used for cross validation and season 7 data is used for test data.

	Team	Op Team	score	Opscore	RaidPoints	OpRaid	TacklePoints	OpTackle	Allout	OpAllOut	ExtraPoints	OpExtra	win	season
0	JAIPUR PINK PANTHERS	U MUMBA	28	44	25	28	1	12	2	4	0	0	1	1
1	BENGALURU BULLS	DABANG DELHI K.C.	42	28	23	15	13	13	6	0	0	0	0	1
2	BENGALURU BULLS	PUNERI PALTAN	39	34	20	20	14	11	4	2	1	1	0	1
3	BENGAL WARRIORS	U MUMBA	24	35	17	24	7	9	0	2	0	0	1	1
4	DABANG DELHI K.C.	PUNERI PALTAN	35	31	23	22	10	7	2	2	0	0	0	1
5	U MUMBA	TELUGU TITANS	34	35	23	24	9	7	2	4	0	0	1	1
6	U MUMBA	PATNA PIRATES	35	32	22	23	9	7	4	2	0	0	0	1
7	BENGALURU BULLS	BENGAL WARRIORS	46	30	25	19	15	11	6	0	0	0	0	1

Fig. 1. dataset preperation from web-scraping

V. Methodology

- We have calculated the sum of all previous season scores for "Team" and "OpTeam"for all matches between "Team" and "OpTeam" so that we can get combined history as new feature during training the model.
- Eg for a match in season 6 between Team "U Mumba" and OptTeam "Puneri Paltan" we include the columns prev-scores which include sum of scores of Team "U Mumba" in all matches against OpTeam "Puneri Paltan" in past 5 seasons. Similarly prev-op-scores column will include sum of scores of Team "Puneri Paltan" in all matches against "U Mumba" in past 5 seasons. This gives us an idea of their head on matches history and is used as a feature for model training.
- Summarized the Data Set: Break down the data to group appropriately.Calculate the features required on the basis of given data and include it into the dataframes.
- Features selected from the given model are raid-points,tacklepoints, allout points,extra points, scores and the previous scores (of head-on matches in between the two teams in previous seasons).
- Evaluating Algorithms: We have run following algorithms- Logistic Regression, SVM(Support Vector Machines),Decision Tree,Voting classifier algorithm on our data set.

VI. Data visualization

We have drawn correlation matrix and found which features are more correlated to each other. Also, plotted bar graph which represents number of times won by each team over seven seasons.

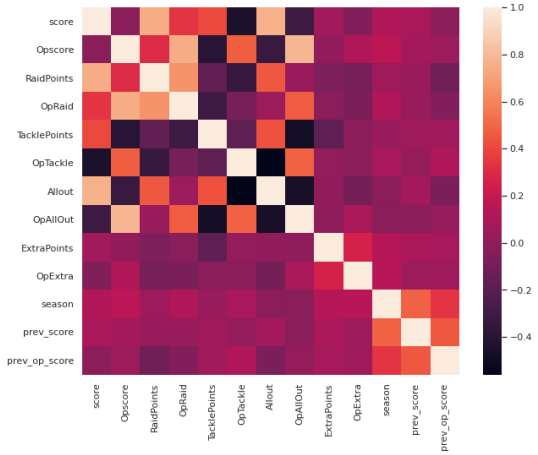

fig 2: correlation matrix

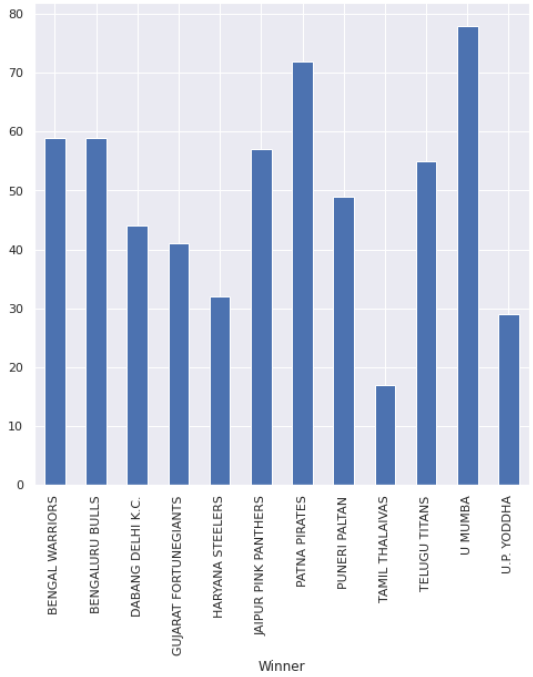

fig 3: distribution of matches won by teams

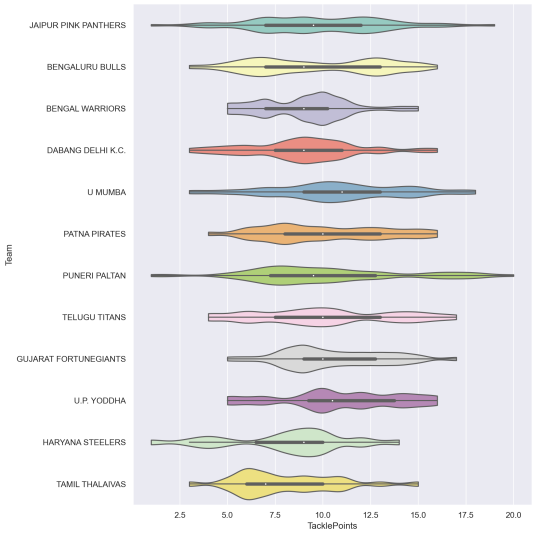

fig 4: Violin plots of winning team tackle points

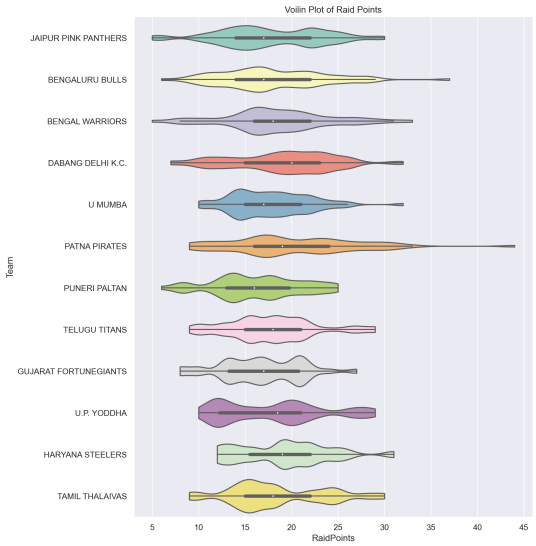

fig 5: Violin plots of winning team raid points

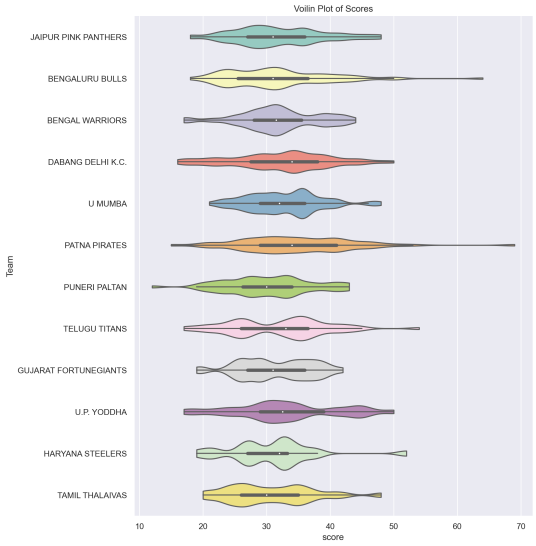

fig 5: Violin plots of total score points

VII. Results

The results of visualisation, hypothesis testing, and predictive model building are quite encouraging. The results can suffer from extreme biases induced by small sample size and is important point to be considered.The violin plot discussed as part of descriptive analytics is a useful tool to visualise the behaviour of teams comprehensively. Granular information provided in the violin plot can be used to characterise a team which is particularly helpful in devising winning strategies.The code also provides the name of the winning team of the input match

Model	Valid Accuracy	Test Accuracylex¿lex
Linear SVM	1.0	1.0
Logistic Regression	1.0	1.0
Voting Classifier	0.93	0.97
Decision Tree	0.93	0.87

Model	Valid dataset F1 score	Test dataset F1 Score
Linear SVM	1.0	0.99
Logistic Regression	1.0	1.0
Voting Classifier	0.93	0.97
Decision Tree	0.93	0.88

Model	Validation AUC Score	Test AUC Score
Linear SVM	1.0	0.992
Logistic Regression	1.0	1.0
Voting Classifier	0.999	0.968
Decision Tree	0.935	0.871

VIII. Conclusion

We found that logistic regression is giving best accuracy i.e. 1 on both validation as well as test data.Thus the winner of the match given can be predicted using this model. Further improvements in the predictions can be done with more data, which will be generated as the years pass by and more and more leagues happen. Also, including different styles of Kabaddi, will have more rules and in-game performance metrics, and hence can make it possible to include more features for the classification problem. Hence, the performance is highly biased on the style of game the model is trained on. Further scope of the study may include multi label classification which can label between win, lose or draw.

References

[1] Parmar, Manojkumar. (2017). Sports Analytics: Kabaddi. 10.13140/RG.2.2.23850.52164.

[2] Vinod Nimbalkar,Durvesh Mundhe,Prof. Dr. Suhasini VijayKumar. (2018) Data Analysis and Visualization on Pro-Kabaddi League. http://www.ijcrt.org/papers/IJCRT1813461.pdf

[3] Bagchi, Amritashish Raizada, Shiny Mhatre, Aniket Augustine, Anthony. (2019). Forecasting the winner of pro kabaddi league matches.

[4] https://www.prokabaddi.com/