Data Analysis and Visualization on Pro-Kabaddi League

[1]Vinod Nimbalkar, [2] Durvesh Mundhe, [3]Prof. Dr. Suhasini VijayKumar
[1]Student, [2]Student, [3]Professor
[1]Master Of Computer Application
[1]Bharti Vidyapeeth's Institute of Management & Information Technology, Belapur (CBD) India

***Abstract*:** Data Visualization makes the complex information easy to understand. It helps to reveal patterns and trends; it uses color to easily distinguish trends and it allows the viewer to get a sense of the data. Statistical numbers are hard to consume but effective data visualization has ability to summarize a ton of information and in the process, save time and effort. Data analysis is the method or methods that can be used to analyze data. Data analysis comprises of inspecting data followed by cleaning, transforming and modeling it to emphasize on crucial pieces of information to help in decision making. The paper focuses on Analysis and Visualization of "Pro-Kabaddi League" data. The analysis is done with Supervised Learning. We are creating a Deterministic Model using KNN and Logistic Regression algorithms to find the accuracy. The visualization represents different aspects of the game in the graphical manner to aid understanding of data**.**

***Index Terms*- K Nearest Neighbour Algorithm, Logistic Regression, Pro Kabaddi, Sports, Visualization.**

I. INTRODUCTION

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.

This paper will provide analysis for first four league seasons using visualizations to show different aspects of the game. The paper aims to make efficient use of the available data set to draw conclusions and to develop a model using different algorithms like K-means, KNN and Logistic Regression to calculate correctness of analyzed data.

II. LITERATURE REVIEW

Kuan-Ta Chen et.al proposed a learning approach for recognizing game bots. The trajectories of human players and those of bots are altogether different can be found. In addition, although game bots can able to mimic player's decisions, but certain human behavior cannot be detected by bots because they are AI-hard. Kuan-Ta Chen did a case study on Quake 2 using kNN, SVM, Isomap and the accuracy was 98% or higher on a trace of 700 seconds[1]. The project by Sujit Nair is to find out if they can predict the result of football matches using classification algorithms viz. Naïve Bayes and K Nearest Neighbour (KNN). For the dataset we have acquired the results of league games from across eleven different domestic leagues across Europe. The matches range from the 2008-09 season to the 2015-16 season. Out of these the training dataset consists of the results of the first six seasons and the last two seasons can be used as testing dataset. Using kNN it is found that team will win if the home player rating is 8. The first six seasons are taken as training set with the remaining two seasons as test set. The accuracy found 67% to 70% accurately predicted result [2]. Ashutosh Tripathi et.al system predicts cricket match outcome by analyzing pre-stored match data using simple but effective K-means clustering algorithm. They used the algorithm in the system and finally represented the analyzed result, showing the performance of the algorithm in visualizing number of runs scored which is the apex determinant of match outcome. The data from both past and expected is selected from the cluster which contains maximum number of entries. The major problem in detection of player strategy is whether the selected strategy correctly determines action of the players[4]. In this paper, Fernando Palero et.al

studied a Real Time Strategy (RTS) game. In Real Time Strategy games the participants make use of units and resources to secure areas of a map and/or destroy the opponents' resources. They extracted real time strategies of the player strategies from various game plays through a web platform. Once the information is gathered , the model will be evaluated as unsupervised learning[3]. Kabaddi is a high power content game that has started from antiquated time till date in India. It has headed out to different nations in the area. There are a wide range of styles of kabaddi; out of these, four renowned styles in India are Sanjeevani, Gaminee, Amar, and Punjabi. Kabaddi is likewise the national session of Bangladesh and it is additionally prevalent in many conditions of India like Maharashtra, Tamil Nadu, Bihar, Andhra Pradesh, Telangana and Punjab. India is the most effective group on world stage; eight universal groups took part in the 2014 Asian Games at Incheon, South Korea. The amusement is genuinely mainstream in the rustic territories of India. The Indian national groups have been ruling the occasion and have won as far back as the diversion was incorporated in the Asian Games in 1990. KNN, K-Means, Logistics Regression can be connected to sports in a scope of routes, with information now available about nearly anything. Most famous games like Cricket (ODI, IPL), Basketball (NBA), Soccer and so forth are utilizing this calculation to examination. These amusements comes about exhibited that there is a roof at around 78% percent precision that can't be outperformed utilizing the accessible information. Obviously, it is misty how much a calculation would enhance with more itemized details.

III. DATA SET

The data set used for the study comprises the Pro Kabaddi data obtained from the official website of pro kabaddi. The records were obtained for season 1, season 2, season 3 and season 4, which is in csv format.
For this study, we focused on the Team, Player, Player Type, Total Matches, Total raids, Total Points etc.

Team	Player	PlayerType	TotalMatches	TotalPoints	TotalRaidPoints
BengaluruBulls	Sunil Jaipal	Raider	25	94	71
Umumba	Nitin Madane	Raider	25	160	152
BengalWarriors	Shrikant Tewthia	All Rounder	40	104	71
TeluguTitans	Rakesh Kumar	All Rounder	45	237	179
DabangDelhi	Bajirao Hodage	Defender, right cover	43	83	0
DabangDelhi	Meraj Sheykh	All Rounder	43	161	119
JaipurPinkPanthers	Jasvir Singh	Raider	46	300	277
PatnaPirates	Pardeep Narwal	Raider	38	263	256
PuneriPaltan	Girish Maruti Ernak	Defender	48	111	3

Table 1: Some data in Data Set

Furthermore, we selected Match Result.csv, Player Leader Board.csv, Player Stats.csv, Team Stats etc and did Cleaning, transform process on it for achieving desired result.

IV. PRELIMINARY

As our intention is to choose the best algorithms for Pro Kabaddi datasets which can be integrated in our Model, we have to search among those that can support data analysis with numeric data, handle incomplete data and find accurate player type with given input. The Visualizations is created using python libraries such as Numpy, Pandas, Matplotlib. Pandas aims to be the fundamental high level building block for doing practical with real world data analysis in Python. Numpy is powerful n dimensional array object,

sophisticated functions, useful linear algebra and random number capabilities. It will help to represent data in graphical manner for easy understanding.

To predict the accuracy of data, algorithms such as K-Means, KNN, and Logistic Regression are used. K-Means is iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. The KNN technique assumes that the entire training set includes not only the data in the set but also desired classification for each item. In effect, training data become the model. Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

V. METHODOLOGY

The methodology used is as follows:
1) Load the Data Set: In this step, the data set is loaded from the csv file URL
2) Summarize the Data Set: Break down the data to group appropriately.
3) Visualization: We create bar graph from the input data to get the idea of the distribution.
4) Evaluating Algorithms: The KNN, K-Means and Logistic Regression algorithm are used. The data set is divided into 2 parts; 70% training data and 30% testing data. We try to find which algorithm is good on this problem.

VI. IMPLEMENTATION

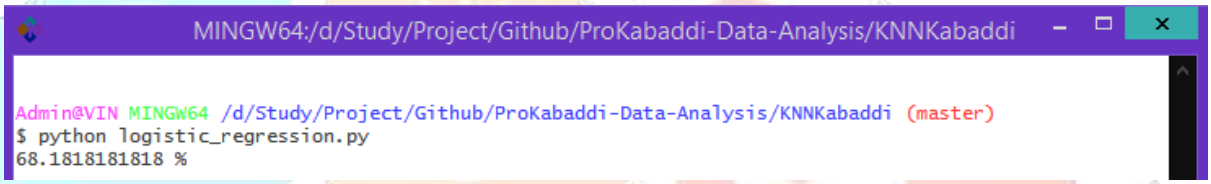

Figure 1: Logistic Regression

The accuracy of data by Logistic Regression algorithm is 68%

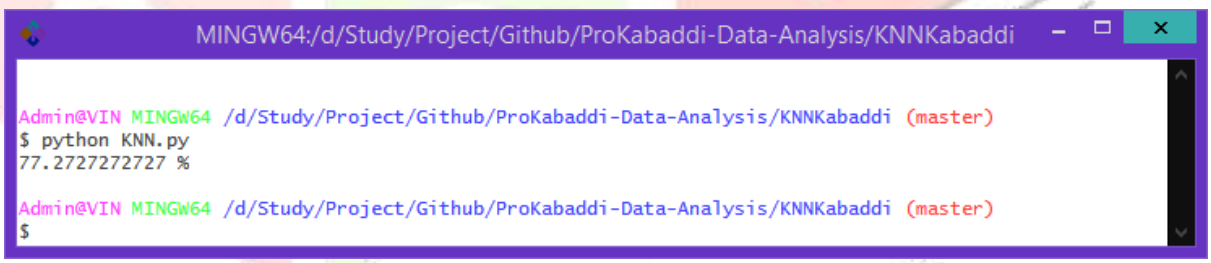

Figure 2: KNN

The accuracy of data by KNN algorithm is 77.27%

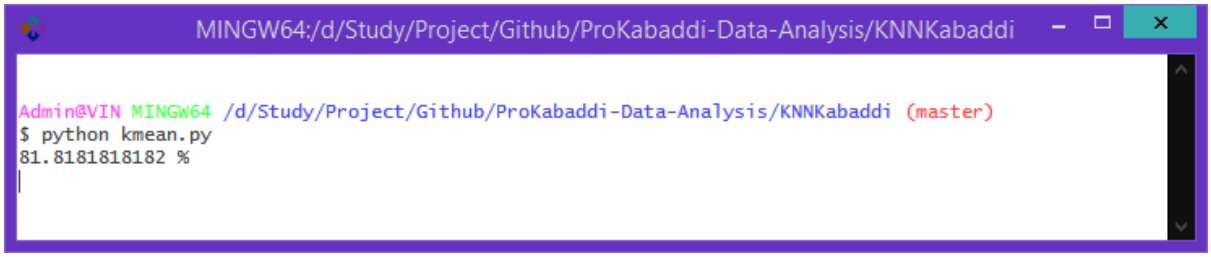

Figure 3: K-Means

The accuracy of data by K-Means algorithm is 81.81%

VII. DATA VISUALIZATION

The data set is analyzed to show different patterns and stats of the game in the form of graphs. Data visualization provides ability to find and interpret rich data sources helping in understanding the data and communicate the data insights/findings. The visualizations results are as follows:

VISUALIZATIONS

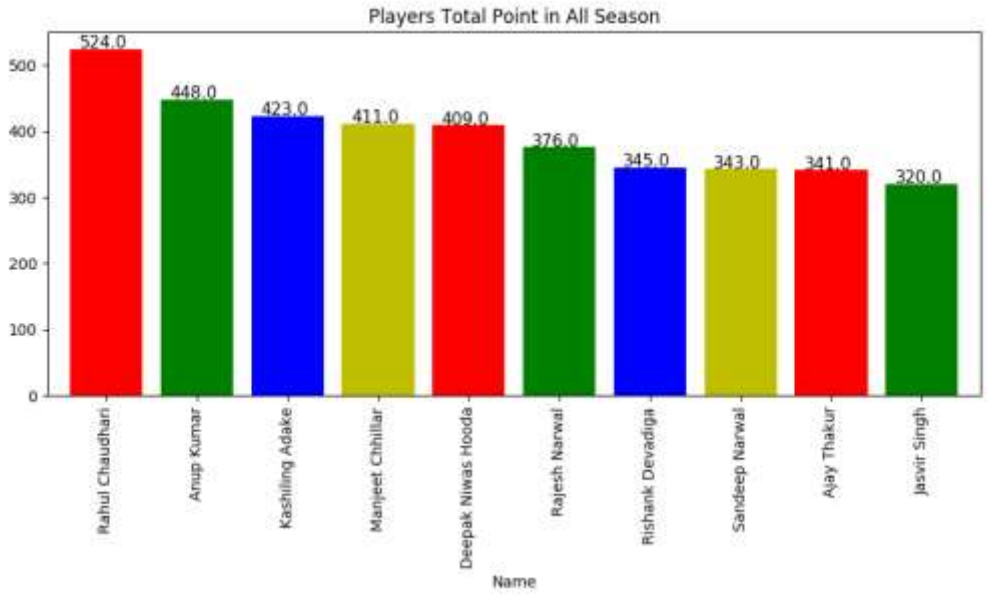

Figure 4: Player Raid Points

Figure 4 shows the visualization of the players who earned maximum raid points in the league. It can be viewed Rahul Chaudhari has taken 524 raid points.

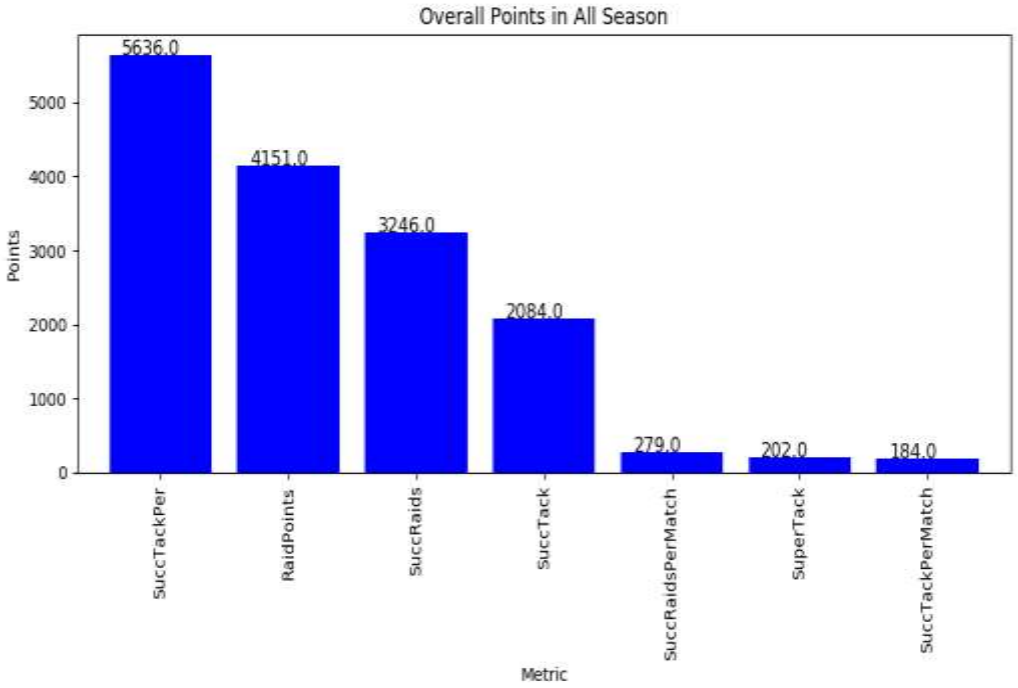

Figure 5: Overall Points

Figure 5 shows the visualization consisting of the aspect which earned maximum points in the league. It can be viewed that maximum points are earned through Super Tackle.

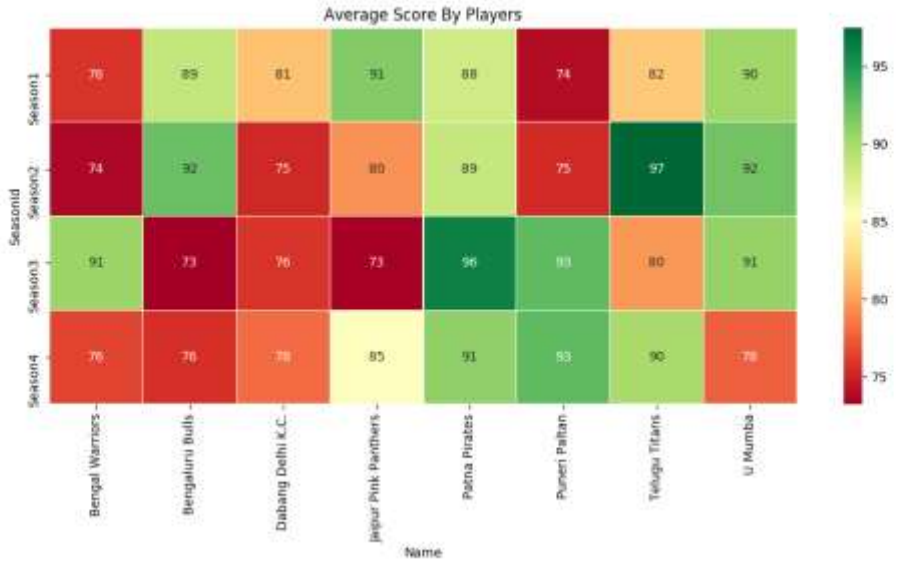

Figure 6: Average Score Heat Map

Figure 6 visualization consists of heat map which shows the points scored by players in the team per season. It can be viewed that the team Patna Pirates players have performed better throughout all the seasons.

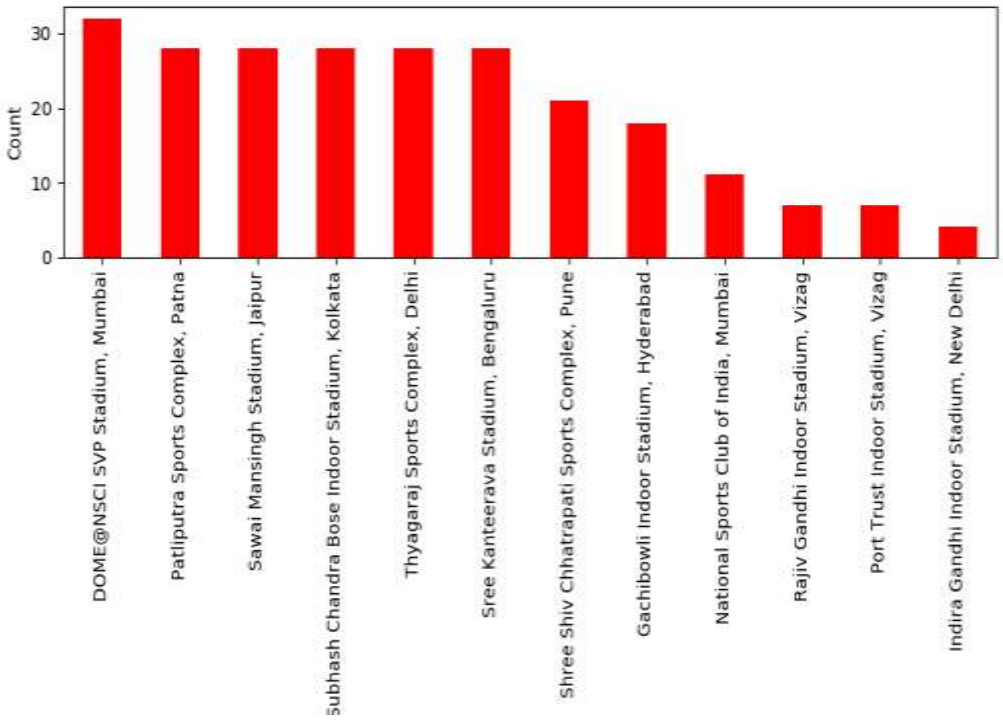

Figure 7: Matches Hosted By Stadium

Figure 7 visualization consists of the number of matches hosted by the stadiums in the league. It can be viewed that Mumbai stadium has hosted maximum number of matches.

VIII. CONCLUSION

In this paper, we compare three algorithms on Pro Kabaddi data set with some parameter. The implementation of Logistic Regression on pro kabaddi data set is less efficient than others like KNN and K-Means. In all algorithms, K-Means is more accurate and efficient in all parameter with 81.81% accuracy. Looking at the visualizations it can viewed which player got maximum points

in all seasons, Super Tackle earned maximum points in all seasons, team performance in all seasons and the stadiums which hosted the maximum games. The current system only works with static data; the future scope of the system can be to visualize the real-time data. This model can also be applied to other sports such as Cricket, Football, Basketball etc to analyze and visualize the data.

VIII. REFERENCES

1] Kuan-Ta Chen, Hsing-Kuo Kenneth Pao, Hong-Chung Chang "Game bot identification based on manifold learning" Worcester, Massachusetts — October 21 - 22, 2008 ACM New York, NY, USA ©2008 ISBN: 978-1-60558-132-3

[2] A.Joseph, N.E.Fenton, M.Neil "Predicting football results using Bayesian nets and other machine learning techniques" Volume 19, Issue 7, November 2006, Pages 544-553

[3] Sujit Nair, Varun Salian, Milind Dave, Piyush Kadve "PREDICTION OF RESULTS OF FOOTBALL MATCHES" International Journal of Research In Science & Engineering Special Issue 7-ICEMTE e-ISSN : 2394-8299 p-ISSN: 2394-8280

[4] Prof. Preeti Satao , Ashutosh ripathi , Jayesh Vankar , Bhavesh Vaje, Vinay Varekar "CRICKET SCORE PREDICTION SYSTEM (CSPS) USING CLUSTERING ALGORITHM" ISSN (PRINT): 2393-8374, (ONLIN E): 2394-0697, VOLUME-3, ISSUE-4, 2016 43

[4] Raghu Ramakrishnan, Johannes Gerhke, “Database Management System” McGraw Hill

[5] M. Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms,” John Wiley & Sons Publishing

[6] http://www.prokabaddi.com/about-prokabaddi

[7] "Python for Data Analysis by Wes McKinney", O’REILLY