Predicting the winner of Indian Premiere League match: A Comparative Study

Aniket Bhawkar
x17170885

Bhumi Patel
x18114865

Prasad Mahajan
x17163773

***Abstract*—All real-life sports contain a vast amount of statistical information regarding individual players, team, games and seasons. Cricket is one of the most popular sports played in the world. Each team has a desire for winning and choosing the best playing eleven has been the most challenging task. Indian Premier League(IPL) is being played for the past 10 years. Abundant data is being gathered and mining over this data can lead to some hidden patterns in understanding the likelihood of a team winning. In addition, years of experience, current form, the impact of the players, batsmen available and the number of runs to be chased can have an influence on the outcome; hence these factors are likewise expected to be considered. A model can be developed which can predict the outcome of the match at various stages based upon the impact factor whose formula is self-derived and explained in the following paper. State-of-the-art data mining techniques like will be used in the making of the model; Knowledge Data Discovery (KDD) approach will be considered. The performances of data mining techniques such as Naive Bayes, Support Vector Machine, Decision Trees and Multi-Linear Regression model will be implemented using various different metrics.**

I. Introduction

Cricket is an outdoor sports game played between two teams and the team that scores maximum runs wins the game. This game is played at domestic and international levels and is played across three different formats: one day international(50-over match), Test match (5-day match) and T20(20-20) format respectively. Cricket is followed and loved by more than billion people all around the world.

The recent T20 format has gained huge recognition throughout the globe. India winning the inaugural T20 World Cup in 2007, lead to a massive foundation for the introductory edition of Indian Premier League in 2008. This league was followed by the immense population due to the fast-paced fixtures. Out of all domestic leagues around the world, IPL is leading in terms of money, entertainment, footfall, popularity, number of views etc. After the enormous success of IPL, various other countries also started a similar type of franchise league. In IPL, players from various nationalities feature in a different franchise. Selection of the players depends upon the auction which is carried prior to each season. The franchise with the highest bidder for a particular player gets the ownership to feature that player. There is no upper limit for bid price but the overall budget is limited.

In this project, prediction of a match is considered before the start of the game depending upon the impact-factor of each player, playing in that respective team. The winning prediction would be done after every 5-overs of the second inning utilising various algorithms like Naive Bayes, Logistic Regression, Keras - Deep Learning, Linear Modeling. Player impact factor can be formulated by considering their previous performances. This information comprises of factors such as batting strike rate, runs scored, 30+ scores, bowling economy, wickets taken and many others. The overall impact factor of that team can then be formulated by taking a mean of all the impact factors of the featuring 11 players.

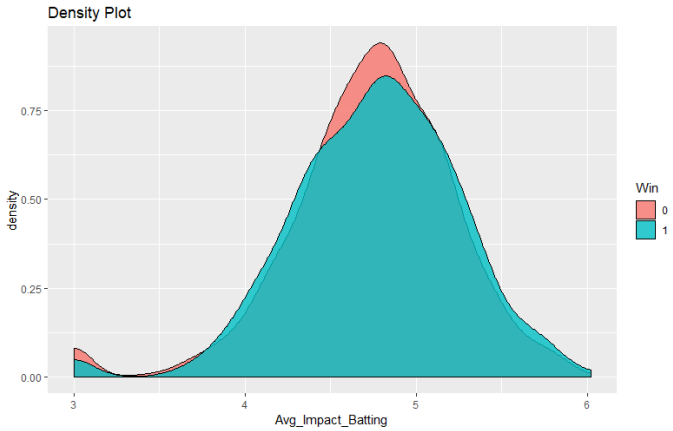

Fig. 1. Density Plot - Average Impact Batting

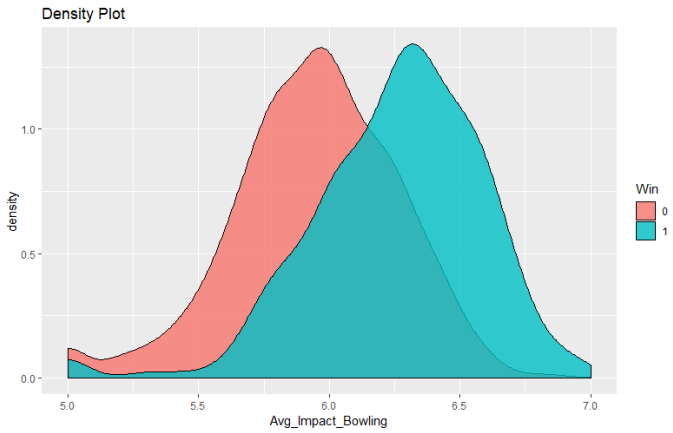

Fig. 2. Density Plot - Average Impact Bowling

The pre-processed dataset is divided into a training and testing datasets. The division of dataset is done in 75:25

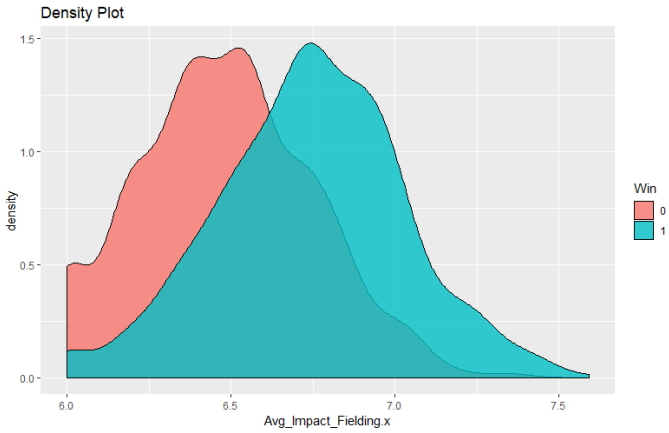

Fig. 3. Density Plot - Average Impact Fielding

ratio. Its often observed that more the data gets trained, more accurate will be the results. In identifying the best variables for training the model, feature extraction on the training dataset is supposed to be carried out. Figure 1, figure 2 and figure 3 elaborates the winning and losing density plot depending upon the Average Impact Factor based upon all three aspects i.e. batting, bowling and fielding respectively. It's clearly visible that a team outperforming in bowling and fielding, and equally compatible in batting have a better probability of winning the fixture.

The model is trained with all the featured variables from the training dataset and then it is tested for accuracy using testing dataset. The following report gives a brief overview of related works, methodology and usage of an appropriate algorithm. The further section contains the evaluation of the outcomes of the models and comparison for identifying the better model. The last section includes the conclusion and future work.

II. Literature Review

In the traditional sport predictive modelling, team performance was taken into consideration in analysing the odds of a team winning. However, there is a need to implement a predictive model for the IPL cricket league, which could predict the winning side after regular intervals of the ongoing game. Hence, to fulfil this gap, the proposed system would predict the winning after every 5 overs and take into consideration critical factors like average team impact in every department of the game, toss, score at that instance, wickets lost and runs required to be scored.

Every franchise dreams of having a strong bench strength and a few core players. In order to achieve this, a predictive model was implemented using various data mining techniques [4]. These techniques are used for selecting players in the IPL through the auction. Several attributes such as batting, bowling skills in ODI and T20 formats are been considered for evaluating the performance of a player. This model was also capable to set the salaries of individual players which are set during the auction. As IPL, is a very competitive tournament, every player gets limited opportunities to feature for the respective franchise. It is important for every player to perform well to make sure that he plays for next season.

As seen earlier, the better bowling and fielding sides have the advantage of winning the game. To tackle the challenges faced by bowlers, a predictive model was developed which was capable to predict the problems and performance evaluation of newbie bowlers. In order to achieve this, the bowling performance of each bowler for first three seasons of IPL was studied and then a predictive model to judge the performance of new bowlers who were going to make their debut in IPL season 4 was developed [5]. Thus by evaluating their performances, team franchises can set their limit of the budget for a particular bowler according to his performance.

Similar to bowlers, all-rounders play a critical role while making a perfect team combination. All-rounders often act as an amalgam between the batsman and bowlers; which boost up the overall team performance. A dataset of 35 all-rounders was collected from IPL season 4 and a prediction was made about their performances at the end of the season [6]. Another literature was published towards the performance of bowlers by the end of the season. Thus it can be concluded that a franchise with quality bowlers and fielding can have a larger impact on the overall outcome of the game.

Often it is observed that the odds of a team winning depends upon the performance they showcase on a particular fixture. A predictive model was developed using techniques like Linear Regression and Naive Bayes Algorithm to investigate the outcome of both the innings of the One-Day International(ODI) match [3]. The approach was to consider and predict the score of every 5^{th} over in a 50 overs match. The prediction of scores is based on various factors such as the number of wickets in hand, venue of the match, etc. An incremental curve was observed and thus the theory was statistically proven. However, the model comprised of numerous flaws, like players yet to bat, death-over bowling of the opponent, etc were not considered. Moreover, the model was not tested for the T20 format of cricket which is furthermore unpredictable.

Prediction of the winning team by considering each aspect of cricket is supposed to be done [7]. A custom formula was developed, which could rank the player based upon their batting and bowling skills. These skills take into scrutiny factors like runs scored, average, economy, wickets taken, dot balls, catches taken, strike rate, 50s, 100s, 4s, 6s scored, etc. Based on the analysis, a similar formula can be constructed in order to imply to analyse the impact factor of each player. This impact factor can then be utilized in calculating the overall team impact factor in various departments i.e. batting, bowling and fielding.

After calculating the necessary impact factors, data mining technique could provide an explorable view while making sports prediction. Data mining in sports gives deep insights on that game and could be used in signifying the hidden information [8]. This eventually helps sports professionals to plan out their strategies which will contribute them to success in various sport events.

Not only outdoor games but a predictive model was built on

online games as well. This model used logistic regression in predicting the odds of winning [9]. Replay data of online game DOTA2 was taken into consideration and by using logistic regression a noteworthy accuracy was achieved. Similarly, a model was built for the prediction of NBA games using various classification techniques [10]. The model was developed to predict the outcome of each match and further, they are compared with each other. Football match winner predictive model was demonstrated to predict the match results of English premier league [11]. Promising results were obtained by using various machine learning techniques, thus by exploring a predictive model of any sport can be developed.

This proposed system comprises of IPL league data collected from the year 2008-2017. The report will be emphasized on predicting the outcome at each distinctive stages of a match.

III. Research Questions

To what extent a machine learning model can predict the chances of a team winning before the game is played and after every 5 overs are bowled in the second inning of a IPL T20 match?

IV. Methodology

SEEMA, CRISP-DM and Knowledge Discovery Database (KDD) are some renowned methodologies available for data mining. For the development of the model, the KDD process would be used. KDD is also defined as the process of extraction of knowledge from the data through databases.

KDD can be elaborated as follows:

1) To create a deeper understanding of the basis of the application domain, respective prior knowledge and the output goals required.
2) Sports data from figure 4 signifies the selection of appropriate variables and columns which will be further used for implementation from which the above-specified queries can be solved.
3) Next step is of data cleaning and data pre-processing. It includes the removal of noise and outliers. Various tactics would be considered for handling missing data fields.
4) Data reduction will be used in finding useful features to present the data depending upon the queries. Transformation methods will be used to minimize the effective number of variables.
5) This step evolves for choosing an appropriate data mining task such as classification, regression, clustering etc.
6) This step will disclose an appropriate method thereby searching the various patterns in the data. After that, models and the parameters required will be decided.
7) Interpretation of data mining patterns.
8) Integrating the discovered knowledge.

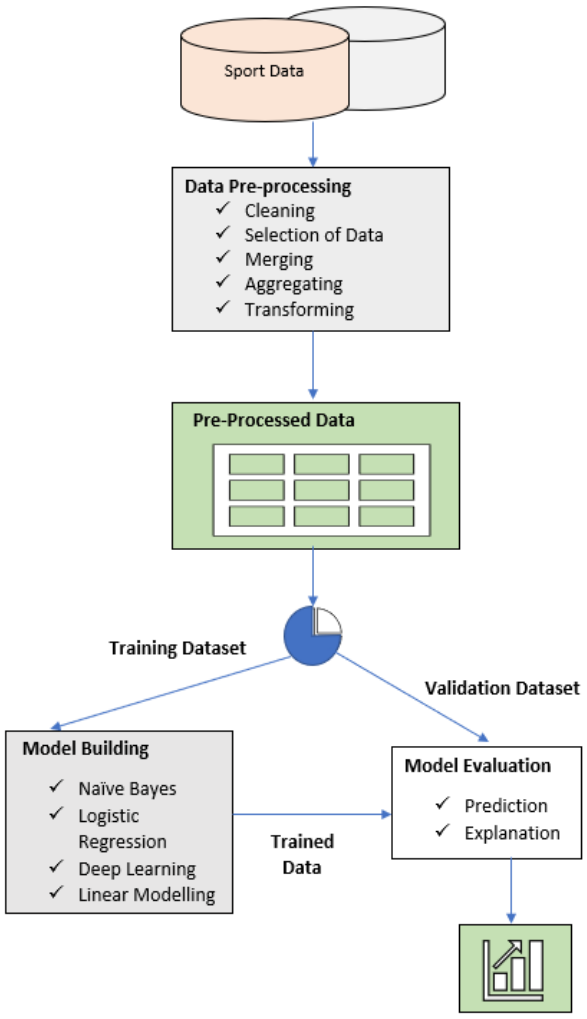

Fig. 4. Knowledge Discovery Database and System Architecture

A. Data Sources

The data set which is used for this project is downloaded from Data World[1]. Multiple raw sub-datasets are available over the mentioned source. It comprises of raw details of all the deliveries being bowled in IPL seasons from 2008-2017. It consists of around 150k rows and 47 columns. This data is enabled with various classes however, the imbalance is observed, thereby increasing the complexity. The second sub-dataset provides Match information. It contains comprehensive details of each match played. The data incorporates around 577 rows and 19 columns. Players data provides the attributes about the player who featured over the years. Team dataset gives an overall insight about all the franchises in IPL 2008-2017. Lastly, Player_match comprises of thorough information about the team members in each match. Therefore by using all these datasets, the aim of the project could be satisfied.

[1]https://data.world/raghu543/ipl-data-till-2017

B. Data Pre-Processing

Data cleaning and Pre-processing embark a significant importance in data mining. Data cleaning is defined as the process of identifying inaccurate data and then replacing it with appropriate values or by zeros, by using various data wrangling tools. The Blank spaces, NA values present in this dataset are replaced by 0 or by the mean of that column with respect to the specificity of that column. Data Pre-processing deals with converting the cleaned data into a proper, usable and understandable format. Therefore in this dataset, as an example, name of the teams are assigned to numerical values where it is then again converted into factors. This data can be served as a pilot for Data Transformation.

C. Data Transformation

The data is supposed to be transformed into appropriate structure as per the required format in order to get the final dataset. The dataset consists of ball by ball data. From that data, the statistics of all the player can be calculated. This data comprises of total runs scored, wickets taken, 4s, 6s, number of catches, strike rate etc. Aggregate of team score at the end of every 5th over can also be calculated.

Thereafter, for analysing the summary of every player, a custom formula of impact factor was formulated by taking various factors under consideration. Based upon the performance of the player, a calculate value was assigned as impact factor. This impact factor was categorized into batting impact, bowling impact, fielding impact. The overall team impact factor can be evaluated by taking into consideration the impact factors of featuring eleven players. Formula for impact factor is given by

$$BIF = SR + No.of4s + Noof6s + RS$$

$$BwIF = W + E + DB$$

$$FIF = C + RO$$

BIF = Batting Impact Factor of Player, BwIF = Bowling Impact Factor of Player, FIF = Fielding Impact Factor of Player, SR = Strike rate, E = Economy, RS = Runs scored, C = Catches, RO = Run Outs, W = Wickets Taken, DB = Number of Dot Balls

As this report deals with predicting match score at four different instances of the match. Appropriate data reduction and useful data columns were mutated to find the answer of each query.

D. Data Mining and Algorithms

The main aim of this project is to predict the match winner

1) Before the match starts
2) After 5 overs
3) After 10 overs
4) After 15 overs

using the team impact factors and toss winner. The predictions would be done by using Naive Bayes classifier, Logistic Regression classifier, Deep Learning - Keras modeling. Alternatively, Linear modeling method for time series prediction is being implemented. By using all these various type of machine learning techniques, results will be compared based upon the accuracy of the model.

Supervised Machine learning is a method that helps machine to classify the data based on the past information which are fed to the machine. The model is trained using the training data and then the accuracy of the model is evaluated by its capability to predict the results of testing data. In supervised machine learning algorithm, the dependent variable is predicted using independent variables. Thus by using different set of independent featured variables, a function will be generated which will automatically map input to the output. This process is continued till the model achieves adequate accuracy. The supervised machine learning algorithms such as Naive Bayes, Logistic Regression and Linear Modelling are used.

*1) **Naive Bayes**:* This technique is based on the Bayes Theorem. Naive Bayes model is easy to implement and is used for larger datasets. It is used to predict the probability of the dependent variable based upon different attributes and mainly used in the problems which are having multiple classes. Naive Bayes is formulated as

$$P(A/B) = [P(B/A)XP(A)]/P(B)$$

Where P(A/B) = Posterior probability of class given predictor; P(A) = Probability of prior class; P(B/A) = Likelihood of probability of predictor given class; P(B) = Prior probability of predictor

In Naive Bayes , number of libraries such as Naive Bayes, dplyr and psych have been used. All the continuous numerical values are converted into required categorical format. Three-fourth of the data is used as a training data and remaining data is used as testing data. The winning of team is predicted based upon the toss and impact factor of a team (batting, bowling and fielding). This algorithm is used to predict the team winning based on four conditions a) before the match and b) after 5, 10, 15 overs respectively.

*2) **Logistic Regression**:* Logistic regression method is a classification method that it is used to predict dichotomous values based on the independent variables. It is used for prediction of probability for which output lies between 0 and 1. The formula for logistic regression is given by

$$Logit(a) = ln(A/(1-A))$$
$$Logit(a) = B0 + B1(X1) + B2(X2)...Bn(Xn)$$

Where, A = Probability of the characteristic of interest; X1, X2, Xn = Independent variables; B0, B1, B2 = Regression coefficients

This model can be used to predict the outcome of match for every 5 overs and before the match as well. Libraries such as *dplyr* and *e1071* have been used. The data is splited into training and testing in a 75:25 ratio. For this project Win has

been predicted by using help of independent variables such as toss winner, average impact batting factor, average impact bowling factor and average impact fielding factor.

3) ***Linear Modelling****:* Linear modelling can be used to predict the next possible value in a given sequence of numbers. For the given query, the sequence of scores per over could be considered. This sequence can be used to predict the score in the upcoming overs. The model could be treated in a Time-Series format.

$$Y = M(X) + B$$

Where, B = constant; Y = Value of Dependent variable; X = Value of Independent variable; M = Regression coefficient

The score is predicted by using training data and then it is compared with testing data to understand the accuracy given by the model. Libraries such as *plyr*, *dplyr* and *tidyverse* have been used.

4) ***Deep learning using Keras****:* Keras is a Python based Deep Learning framework which is the high-level API of TensorFlow. It can configured to execute over Theano, TenserFlow or CNTK. Being an open-source Deep Learning framework, Keras models are developed by stacking layers and connecting graphs. It can be classified as semisupervised learning which is an amalgum of both supervised and unsupervised learning. Supervised machine learning is used for labelled data whereas unsupervised machine learning is used on unlabelled data therefore semisupervised learning algorithms are trained on a combination of both labelled and unlabelled data. Keras can be used for building neural networks. Multi-layer Preceptron will be built for multiclass classification.

In this method, *keras*, *TensorFlow* and *readr* libraries would be used. Prediction would be done which will be based on test data. Evaluation of model, interpretation of the result will eventually help the user to add and hide layers which will later be used in adjusting the optimization parameters for achieving better results.

E. ***Software and Libraries***

In the development of the models, various softwares and libraries were utilised. R studio was used for statistical analysis and Github facilitated with a provision of generating backup of the code at regular intervals. Libraries like dplyr, plyr, readr, tidyverse, psych, e1071, Keras and TensorFlow were employed in retrieving the accuracy.

V. Performance Metrics

For every model, accuracy is the only performance metrics which is considered to compare every model for each query. Accuracy is defined as total number of correct predictions divided by total number of predictions in a dataset. Range of accuracy varies from 0 to 1. Accuracy is formulated as

$$Accuracy = 1 - ((TP + TN)/(TP + TN + FP + FN))$$

Where TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative.

VI. Evaluation

In the implementation of the models, various algorithms were applied. Coin toss model can defines a benchmark accuracy on the respective dataset. Prediction of match winner with coin toss model provided an accuracy of 54.06%. The evaluation could be carried out in two different stages.

1) Time-Series Analysis
2) Classification Analysis

A. Time-Series Analysis

In this approach, Linear Modeling technique with its ability to predict the next possible number given a sequence of previous numbers could be used. In this case, the sequence of numbers would be the score after every over till the 15th over and predictions would be made to identify the score at the end of the 20th over. The number received would be then compared with the target score and with the help of confusion matrix the accuracy could be calculated.

B. Classification Analysis

Classification Analysis could be carried out in making predictions based on the independent variables. As justified earlier, the impact factors calculated would be utilised in this approach. Other independent variables like, Team, Opponent, Toss Winner, Target Score, Actual Score after particular number of overs and the number of wickets lost could be considered. Models like Naive Bayes, Logistic Regression and Deep Learning with Keras would be used.

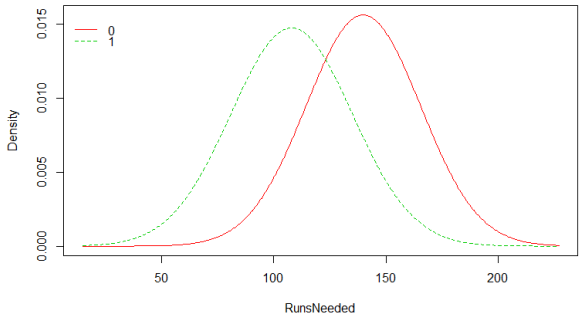

Fig. 5. Runs Needed 5th Over

As seen in figure 5, 6 and 7, it is observed that there is a significant difference between the runs required to classify the probability of a side winning or losing. Hence, the target score and the actual score after particular overs do have a noteworthy importance in the evaluation. This difference can create an influence on the prediction of the model. The initial 6-overs of a T20 game are known as the powerplay overs. The team which takes the complete advantage of this initial period can have lesser runs to score in the remaining 15-overs, thus decreasing the pressure and increasing their odds of winning. It is also visible that the normal distribution of winning and losing gets more significant as the match progresses.

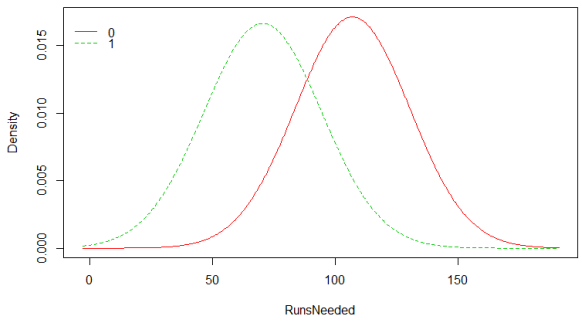

Fig. 6. Runs Needed 10th Over

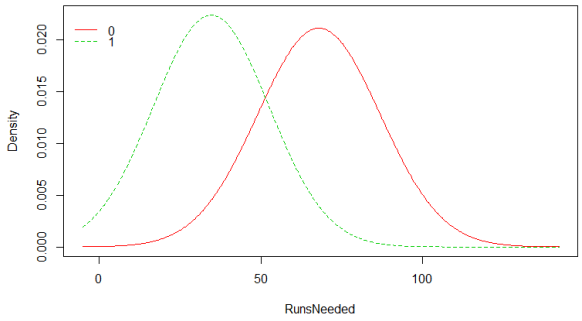

Fig. 7. Runs Needed 15th Over

VII. Results

Accuracy received after implementing every models at different instances of the game are important in making a comparison. The below table highlights the accuracy of predictions made using Naive Bayes classifier, Logistic Regression, Deep Learning with Keras and by Linear Modelling at different stages of the game. In this project, predictions are done before the match and after every 5 overs of the second inning.

Models	PreMatch	5th	10th	15th
Naive Bayes	53.45%	66.02%	72.3%	79.1%
Logistic	53.7%	67.9%	71.2%	78.47%
Keras	67.18%	76.2%	77.05%	84.41%
L.M.	-	-	-	51.2%

It can be concluded that Naive Bayes, Logistic Regression and Keras - Deep Learning can provide better predictions as compared to Linear Modeling. Linear Modeling provided an accuracy which was below-par the benchmark. Hence it can be justified that time-series analysis is not a suitable technique in predicting the winner of the game without considering the independent variables. For all the test-cases such as Pre-match, after 5th, 10th and 15th overs, Keras was able to achieve the maximum accuracy as compared to the remaining models. Figure 8 compares the accuracy and an incremental curve can be observed as the game proceeds.

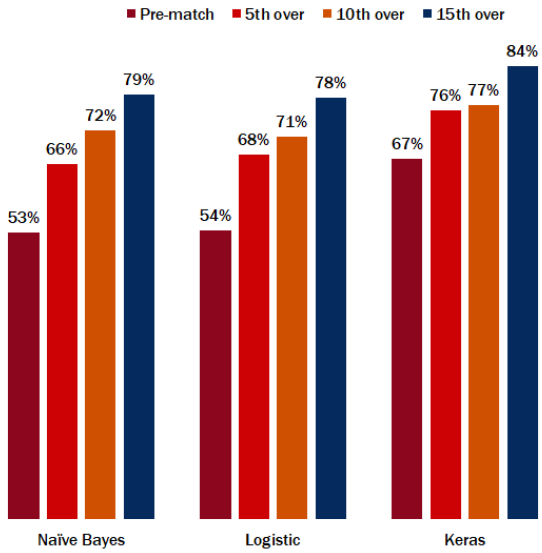

Fig. 8. Accuracy Comparison

VIII. Discussion

A predictive model, that has the capability to identify the winning team prior the start of the game makes this study crucial. The accuracy provided by Keras is significant even before the game is played and can be utilised in the bidding industry. The prediction capability improves with time which indeed makes sport analytics a wider field of interest. Depending upon the data, franchise can identify their weaker links can replace them with players who can perform better. Odds of a team winning against a particular opponent can also be identified and the team management can take step in improving the outcomes. Recognising the impact players in the opponent franchise can improve the game plan and restrict them at an inferior total.

IX. Conclusion and Future Work

The IPL cricket history and player performance data are used for predicting the outcome of a match by considering the average impact factor of the team based on featuring players. Prediction is undertaken in four cases: Before the fixture and after every 5 overs of the second inning i.e. after 5th, 10th and 15th over. Various algorithms such as Naive Bayes, Logistic Regression, Linear Modelling, Keras Deep Learning were implemented and the accuracy of each model has been compared. From the results, it can be found out that prediction of match-winner using time series gives accuracy much lesser as compared to other classification algorithms. Therefore it can be concluded that for predicting a winner during various stages of IPL game, time-series model should be refrain and should consider implementing Naive Bayes, Logistic Regression of Deep Learning classification algorithm. From the above experimentation, Keras Deep Learning algorithm provided the maximum accuracy. In future, the analysis can further be drilled down in examining the impact players

and in-form players who can provide better results to the franchise. Moreover, the team performance with respect to a particular opponent can be classified.

REFERENCES

[1] Hand, David J., 2007. Principles of data mining. Drug safety, 30(7), pp.621-622. https://doi.org/10.2165/00002018-200730070-00010

[2] Subasingha, S.A.D.P., 2019. ODI cricket match winning prediction using data mining techniques.

[3] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, 2015, pp. 60-66. doi: 10.1109/ICSCTI.2015.7489605

[4] P. Kansal, P. Kumar, H. Arya and A. Methaila, "Player valuation in Indian premier league auction using data mining technique," 2014 International Conference on Contemporary Computing and Informatics (IC3I), Mysore, 2014, pp. 197-203. doi: 10.1109/IC3I.2014.7019707

[5] Hemanta Saikia, Dibyojyoti Bhattacharjee and H. Hermanus Lemmer (2012) Predicting the Performance of Bowlers in IPL: An Application of Artificial Neural Network, International Journal of Performance Analysis in Sport, 12:1, 75-89, DOI: 10.1080/24748668.2012.11868584

[6] Saikia, H. and Bhattacharjee, D. (2011) On Classification of All-rounders of the Indian Premier League (IPL): A Bayesian Approach, Vikalpa, 36(4), pp. 5166. doi: 10.1177/0256090920110404.

[7] Jayanth, S. B., Anthony, A., Abhilasha, G., Shaik, N., & Srinivasa, G. (2018). A team recommendation system and outcome prediction for the game of Cricket. Journal of Sports Analytics, 111. doi:10.3233/jsa-170196

[8] Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data Mining in Elite Sports: A Review and a Framework. Measurement in Physical Education and Exercise Science, 17(3), 171186. doi:10.1080/1091367x.2013.805137

[9] Yang, Y., Qin, T. and Lei, Y.H., 2016. Real-time esports match result prediction. arXiv preprint arXiv:1701.03162. url: https://arxiv.org/abs/1701.03162

[10] Thabtah, F., Zhang, L. & Abdelhamid, N. Ann. Data. Sci. (2019) 6: 103. https://doi.org/10.1007/s40745-018-00189-x

[11] Rahul Baboota, Harleen Kaur, Predictive analysis and modelling football results using machine learning approach for English Premier League, International Journal of Forecasting, Volume 35, Issue 2, 2019, Pages 741-755, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2018.01.003.