Player Valuation in Indian Premier League Auction using Data Mining Technique

Prince Kansal, Pankaj Kumar, Himanshu Arya
Student, B.Tech (CSE)
Maharaja Surajmal Institute of Technology
New Delhi, India

Aditya Methaila
Student, B.Tech (COE)
Netaji Subhas Institute of Technology
New Delhi, India

***Abstract*— The Indian Premier League is a new T20 League which completed its inaugural season in 2008. Players' auctions are not new phenomena in the world of sports. However, in the game of cricket auctioning of players was first time used in Indian Premier League (IPL). No fixed method was used before to evaluate the performance of a player and determining its base price. In this study, we build several predictive models for predicting the selection of a player in the Indian Premier League, a cricket league, based on each player's past performance. Using One-Day International (ODI) variables and T-20 variables of both batting and bowling, we have found a number of interpretable variables that have explanatory power over auction values. The models that are developed can help decision makers during the auction to set salaries for the players.**

Keywords — Data Mining, Naïve Bayes, Decision Tree, MLP, Cricket, IPL.

I. INTRODUCTION

Using Mathematical Analysis for selection of players is not new to domain of sports. Forecasting future from the past data is highly subjective and thus requires extraordinarily expert decision-making. It becomes more prominent when a huge amount of money is involved. In this paper we have used several data mining tools for the prediction of base price group on the basis of their past performance in different versions of the game. Accordingly the different versions; test cricket which is played between two countries over a duration of 5 days, one day international cricket played between two international teams where each team normally gets a chance to play a maximum of 50 overs. During the last fifty years game of cricket has seen several changes and all these changes were directed to make this game more popular among the masses and to expand the reach of cricket to non-cricket playing nations. The latest form of cricket is the Twenty20 or T20 cricket. It involves two teams with each team batting maximum of 120 balls i.e. twenty overs and is completed in an about two and half hours, a much shorter duration. The T20 format gave a dynamic platform to the IPL, marked arguably the biggest business revolution in the sport of cricket in the 130-year formal history of the game.

Indian Premier League (IPL) is a Twenty20 cricket tournament where different franchise teams participate for the title. The tournament started in 2008 and from then it usually takes place every year in the months of April – June. The recent was the IPL-7 held this year (2014). IPL is the most-watched Twenty20 cricket league in the world and also known for its commercial success. During the sixth IPL season (2013) its brand value was estimated to be around US$3.03 billion. Live rights to the event are syndicated around the globe, and in 2010, the IPL became the first sporting event to be broadcast live on YouTube. IPL has hosted seven seasons till now from its year of inauguration in 2008. There are 8 teams which participate in the IPL. Each team represents a state or a part of nation, India.

II. MODELS FOR PREDICTING THE BASE PRICE

We have collected data from www.iplt20.com, www.thatscricket.com, www.espncricinfo.com. Our data consists of playing factors from all the 2 formats of the game ODI and T20 cricket. The playing factors such as average, strike rate, no of wicket, catches etc. help us to predict the selection of batsman, bowler and all-rounder in the teams. We created an individual dataset for bowler, batsman and all-rounder. The batsman dataset consists of 40 attributes, the bowler dataset consists of 32 attributes and the all-rounder dataset consists of 64 attributes.

In the IPL auction, base groups are formed according to base price. Players are allowed to choose from any of the base groups mentioned in TABLE II according to their own wish.

Data mining tools are used to predict the base price group for players which will prove to be automated and a fair way of doing it. The player's past performance has been used to predict their base price.

In this paper data containing the information of 78 batsmen, 90 bowlers and 49 all-rounders of different countries who have registered themselves for the auction was collected. These datasets contained the information of the players from their international ODI and T20 career, and also IPL records from the six seasons of the IPL from 2008 to 2013.

The data mining tool Weka 3.6.6 is used for experiment. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

978-1-4799-6629-5/14/$31.00 ©2014 IEEE

TABLE I.
ATTRIBUTES OF BATSMAN, BOWLER AND ALL-ROUNDER

Batsman	Bowler	All-rounder
Indian/Overseas(1/0)	Indian/Overseas(1/0)	Indian/Overseas(1/0)
Age	Age	Age
Capped/Uncapped (1/0)	Specialism	Capped/Uncapped (1/0)
Base_Group	Capped/Uncapped (1/0)	Base_Group
ODI_Matches	Base_Group	ODI_Matches
ODI_Innings	ODI_Matches	ODI_Innings
ODI_Runs	ODI_Balls	ODI_N#O#
ODI_Avg	ODI_RunsC	ODI_Runs
ODI_S#R#	ODI_Wickets	ODI_HS
ODI_100s	ODI_BAvg	ODI_Avg
ODI_50s	ODI_Eco	ODI_S#R#
T20s_Matches	ODI_S/R	ODI_100s
T20s_Innings	ODI_5s	ODI_50s
T20s_Runs	ODI_10s	ODI_4s
T20s_Avg	T20s_Balls	ODI_6s
T20s_S#R#	T20s_RunsC	ODI_Catches
T20s_100s	T20s_Matches	T20s_Matches
T20s_50s	T20s_Wickets	T20s_Innings
T20s_Catches	T20s_Bavg	T20s_N#O#
IPL_Matches	T20s_Eco	T20s_Runs
IPL_Innings	T20s_S/R	T20s_HS
IPL_Runs	T20s_5s	T20s_Avg
IPL_Avg	T20s_10s	T20s_S#R#
IPL_S#R#	IPL_Matches	T20s_100s
IPL_100s	IPL_Balls	T20s_50s
IPL_50s	IPL_RunsC	T20s_4s
IPL_Catches	IPL_Wickets	T20s_6s
ODI_4s	IPL_Bavg	T20s_Catches
ODI_6s	IPL_Eco	IPL_Matches
ODI_N#O#	IPL_S/R	IPL_Innings
ODI_HS	IPL_5s	IPL_N#O#
T20s_N#O#	IPL_10s	IPL_Runs
T20s_HS		IPL_HS
T20s_4s		IPL_Avg
T20s_6s		IPL_100s
IPL_N#O#		IPL_50s
IPL_HS		IPL_4s
IPL_4s		IPL_6s
IPL_6s		IPL_Catches
		ODI_Balls
		ODI_RunsC
		ODI_Wickets
		ODI_Bavg
		ODI_Eco
		ODI_S/R
		ODI_5s
		ODI_10s
		T20s_Balls
		T20s_RunsC
		T20s_Wickets
		T20s_Bavg
		T20s_Eco
		T20s_S/R
		T20s_5s
		T20s_10s
		IPL_Balls
		IPL_RunsC
		IPL_Wickets
		IPL_Bavg
		IPL_Eco
		IPL_S/R
		IPL_5s
		IPL_10s

TABLE II.

Base Price (`)	Base Group
20,000,000	A
15,000,000	B
10,000,000	C
5,000,000	D
3,000,000	E

III. DATA MINIG ALGORITHMS

Three algorithms are applied namely decision tree, naïve Bayes and MLP on the data set to predict the base price of players. The algorithms are applied in order on batsmen, bowlers and all-rounder data set respectively.

A. J48 Algorithm (decision tree)

Tree is a popular classifier which is simple and easy to implement. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are easier to read and interpret. The drill through feature to access detailed patients profiles is only available in Decision Trees.

B. Naïve Bayes

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The advantage of using Naive Bayes is that one can work with the Naïve Bayes model without using any Bayesian methods.

C. Multilayer perceptron (MLP)

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable. The multilayer perceptron consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes and is thus considered a deep neural network.

IV. RESULTS

The functions used for quantifying the accuracy of a class is given as follows:

- *TP Rate*: True positive rate (the number of items correctly labeled as belonging to the positive class).
- *FP Rate*: False positive rate (which are items incorrectly labeled as belonging to the class).
- *Precision*: Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant.

- *Recall*: Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved.
- *F-measure*: The f-score (or f-measure) is calculated based on the precision and recall.
- *ROC area*: A ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Each prediction result or instance of a confusion matrix represents one point in the ROC space.

A. Batsmen Data Set

This paper used 78 instances for the data classification.

Fig. 1 shows the decision tree output of the data set. It can be seen that IPL_S#R# is the most important attribute for prediction.

TABLE III shows the accuracy by class when the J48 algorithm was applied on the data set. TABLE IV shows the accuracy by class when the naïve bayes algorithm was applied on the data set. TABLE V shows the accuracy by class when the MLP algorithm was applied on the data set.

The bar graph in *Fig. 2* shows the comparison between all 3 algorithms. It can be concluded from the Bar graph that MLP outperforms the other two algorithms. The accuracy for MLP is 94.18% which is way better than naïve bayes and j48 algorithm.

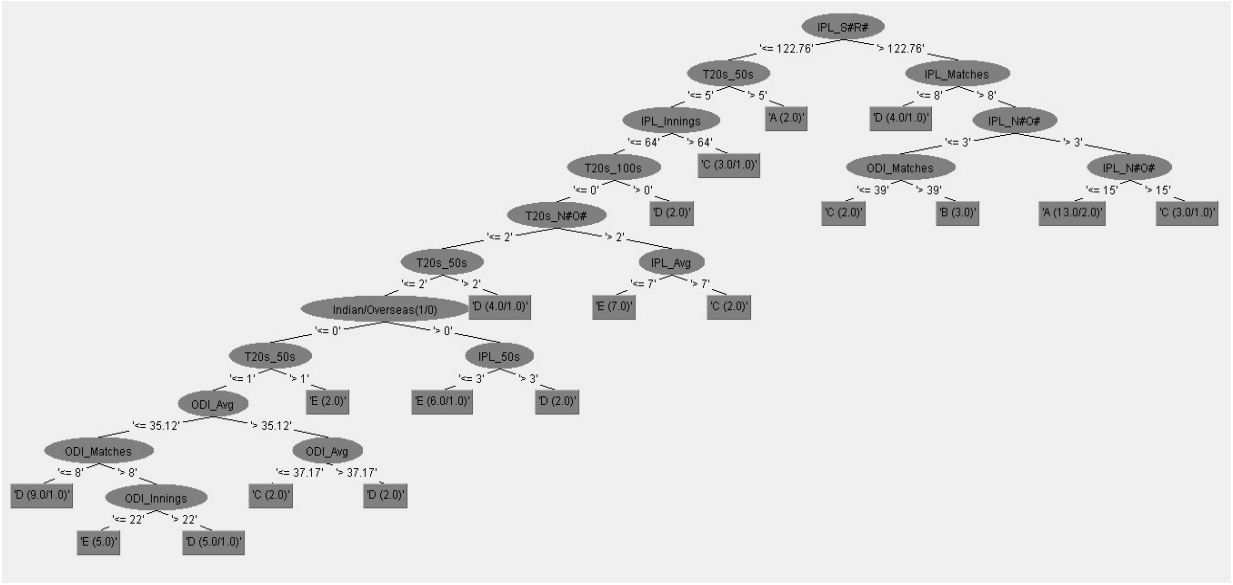

Fig. 1. Decision tree output of the batsmen data set

TABLE III. ACCURACY BY CLASS WITH J48 ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.813	0.032	0.867	0.813	0.839	0.973	A
0.429	0	1	0.429	0.6	0.932	B
0.909	0.03	0.833	0.909	0.87	0.986	C
1	0.074	0.857	1	0.923	0.978	D
0.95	0.017	0.95	0.95	0.95	0.995	E
0.885	0.038	0.892	0.885	0.876	0.979	←Weighted Avg.

TABLE IV. ACCURACY BY CLASS WITH NAÏVE BAYES ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.938	0	1	0.938	0.968	0.941	A
0.857	0	1	0.857	0.923	0.893	B
0.818	0	1	0.818	0.9	0.925	C
1	0.056	0.889	1	0.941	1	D
1	0.017	0.952	1	0.976	1	E
0.949	0.022	0.954	0.949	0.948	0.968	←Weighted Avg.

TABLE V. ACCURACY BY CLASS WITH MLP ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.938	0	1	0.938	0.968	0.941	A
0.857	0	1	0.857	0.923	0.893	B
0.818	0	1	0.818	0.9	0.925	C
1	0.056	0.889	1	0.941	1	D
1	0.017	0.952	1	0.976	1	E
0.949	0.022	0.954	0.949	0.948	0.968	←Weighted Avg.

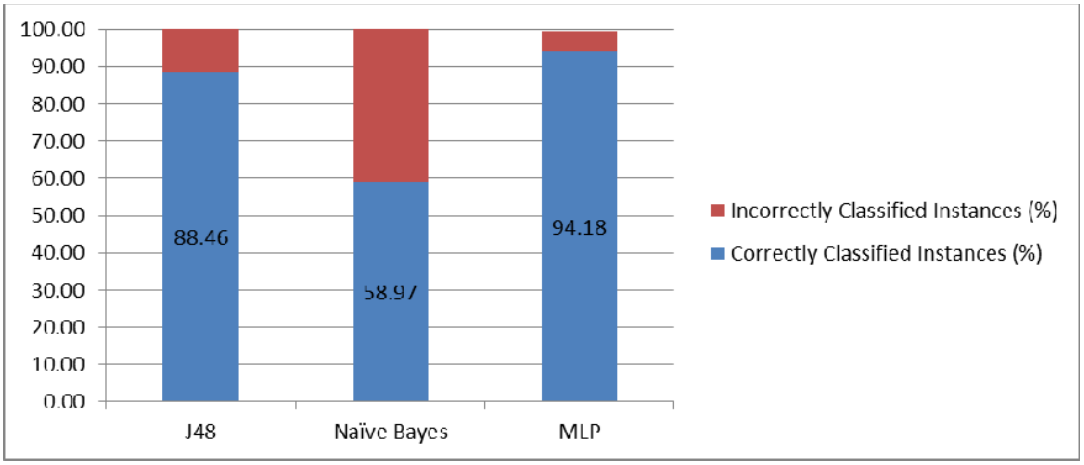

Fig. 2. Comparison between Accuracy of J48, Naïve Bayes and MLP algorithm applied on Batsmen Data Set

B. *Bowler Data Set*

This paper used 90 instances for the data classification *Fig. 3* shows the decision tree output of the data set. It can be seen that IPL_Wickets is the most important attribute for prediction.

TABLE VI shows the accuracy by class when the J48 algorithm was applied on the data set. TABLE VII shows the accuracy by class when the naïve bayes algorithm was applied on the data set. TABLE VIII shows the accuracy by class when the MLP algorithm was applied on the data set.

The bar graph in *Fig. 4* shows the comparison between all 3 algorithms. It can be concluded from the Bar graph that MLP outperforms the other two algorithms. The accuracy for MLP is 94.44% which is way better than naïve bayes and j48 algorithm.

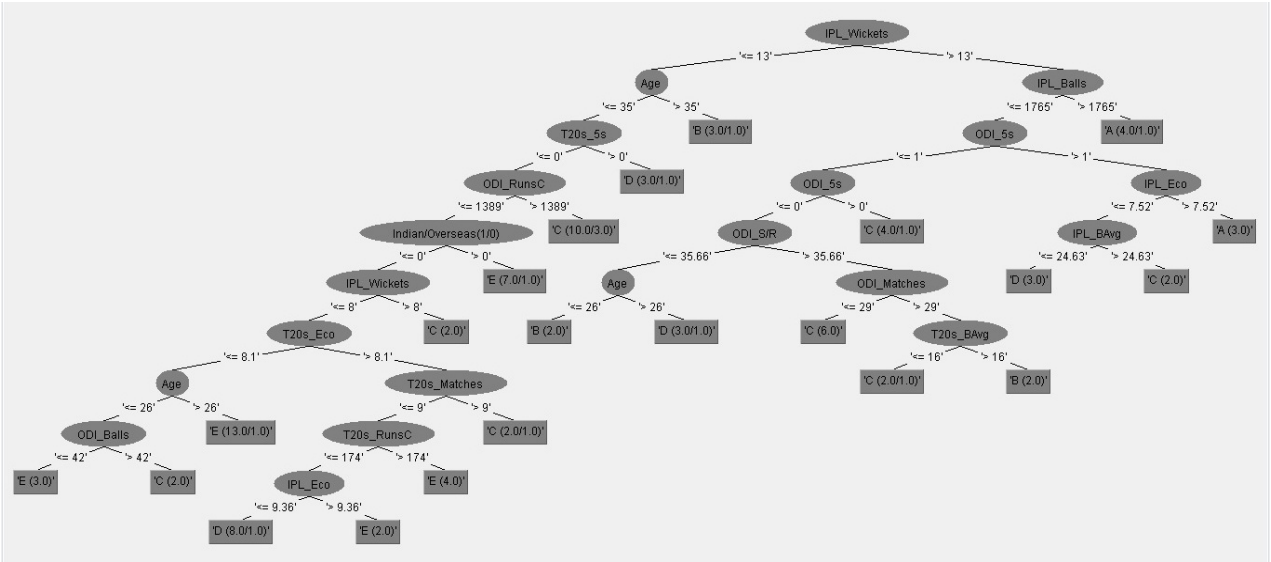

Fig. 3. Decision tree output of the bowler data set

TABLE VI. ACCURACY BY CLASS WITH J48 ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.923	0.094	0.8	0.923	0.857	0.973	C
1	0.012	0.857	1	0.923	0.997	A
0.857	0.012	0.857	0.857	0.857	0.994	B
0.667	0.043	0.824	0.667	0.737	0.932	D
0.9	0.033	0.931	0.9	0.915	0.978	E
0.856	0.05	0.857	0.856	0.853	0.968	←Weighted Avg.

TABLE VII. ACCURACY BY CLASS WITH NAÏVE BAYES ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.385	0.047	0.769	0.385	0.513	0.793	C
1	0.06	0.545	1	0.706	0.968	A
0.857	0.084	0.462	0.857	0.6	0.933	B
0.238	0.087	0.455	0.238	0.312	0.572	D
0.9	0.25	0.643	0.9	0.75	0.857	E
0.6	0.128	0.615	0.6	0.565	0.785	←Weighted Avg.

TABLE VIII. ACCURACY BY CLASS WITH MLP ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.031	0.929	1	0.963	0.998	C
1	0	1	1	1	1	A
0.714	0	1	0.714	0.833	0.835	B
0.952	0.043	0.87	0.952	0.909	0.954	D
0.933	0	1	0.933	0.966	0.998	E
0.944	0.019	0.949	0.944	0.944	0.975	←Weighted Avg.

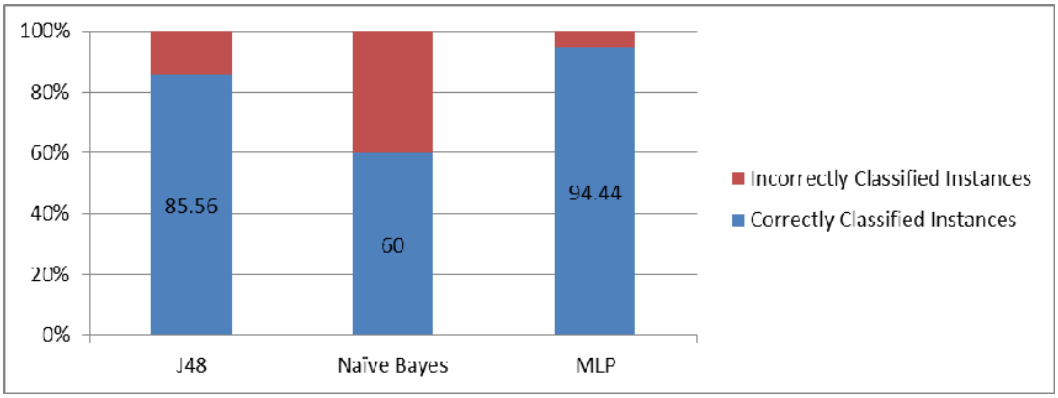

Fig. 4. Comparison between Accuracy of J48, Naïve Bayes and MLP algorithm applied on Bowler Data Set

C. *All Rounder Data Set*

49 instances of all-rounders for the data classification were used. *Fig. 5* shows the decision tree output of the data set. It can be seen that IPL_RunsC is the most important attribute for prediction.

As the All Rounder dataset would contain instances sharing the attributes of both as a bowler and a batsman. So, prediction would require the maximum number of attributes as, an all-rounder category comprises the attributes of other two categories, i.e a bowler and a batsmen.

TABLE IX shows the accuracy by class when the J48 algorithm was applied on the data set. TABLE X shows the accuracy by class when the naïve bayes algorithm was applied on the data set. TABLE XI shows the accuracy by class when the MLP algorithm was applied on the data set.

The bar graph in *Fig. 6* shows the comparison between all 3 algorithms. It can be concluded from the Bar graph that MLP outperforms the other two algorithms. The accuracy for MLP is 97.95% which is way better than naïve bayes and j48 algorithm.

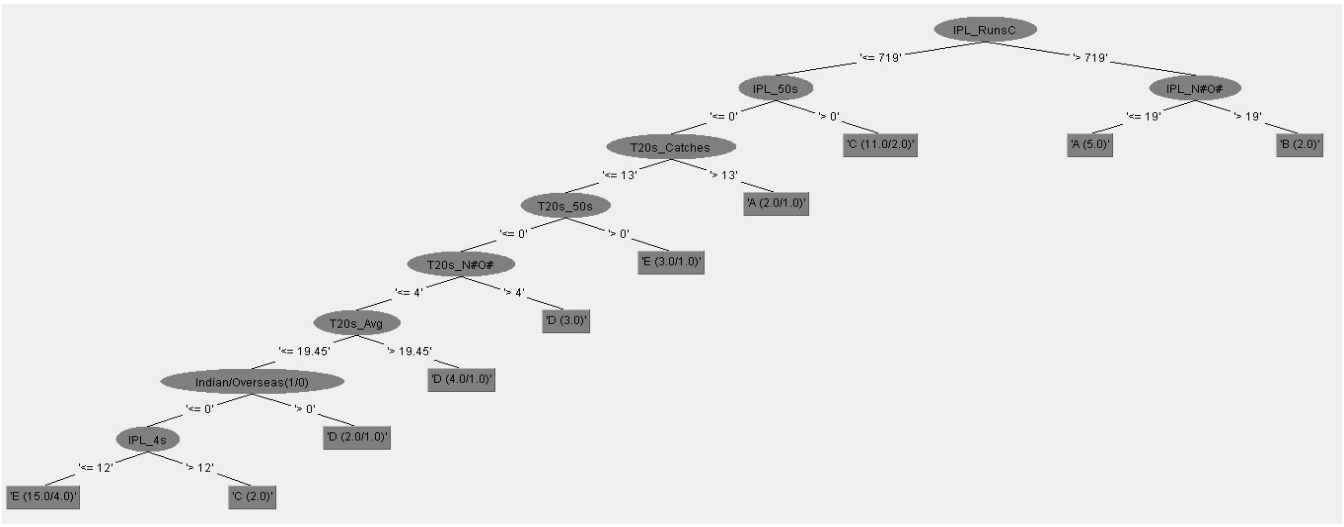

Fig. 5. Decision tree output of the all-rounder data set

TABLE IX. ACCURACY BY CLASS WITH J48 ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.75	0.024	0.857	0.75	0.8	0.968	A
0.786	0.057	0.846	0.786	0.815	0.933	C
0.5	0	1	0.5	0.667	0.953	B
0.875	0.049	0.778	0.875	0.824	0.97	D
0.867	0.147	0.722	0.867	0.788	0.914	E
0.796	0.073	0.811	0.796	0.793	0.94	←Weighted Avg.

TABLE X. ACCURACY BY CLASS WITH NAÏVE BAYES ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.5	0.024	0.8	0.5	0.615	0.933	A
0.5	0.171	0.538	0.5	0.519	0.833	C
0.75	0.022	0.75	0.75	0.75	0.933	B
0.625	0.195	0.385	0.625	0.476	0.851	D
0.667	0.118	0.714	0.667	0.69	0.886	E
0.592	0.123	0.627	0.592	0.599	0.877	←Weighted Avg.

TABLE XI. ACCURACY BY CLASS WITH MLP ALGORITHM

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	A
1	0	1	1	1	1	C
0.75	0	1	0.75	0.857	0.967	B
1	0	1	1	1	1	D
1	0.029	0.938	1	0.968	1	E
0.98	0.009	0.981	0.98	0.978	0.997	←Weighted Avg.

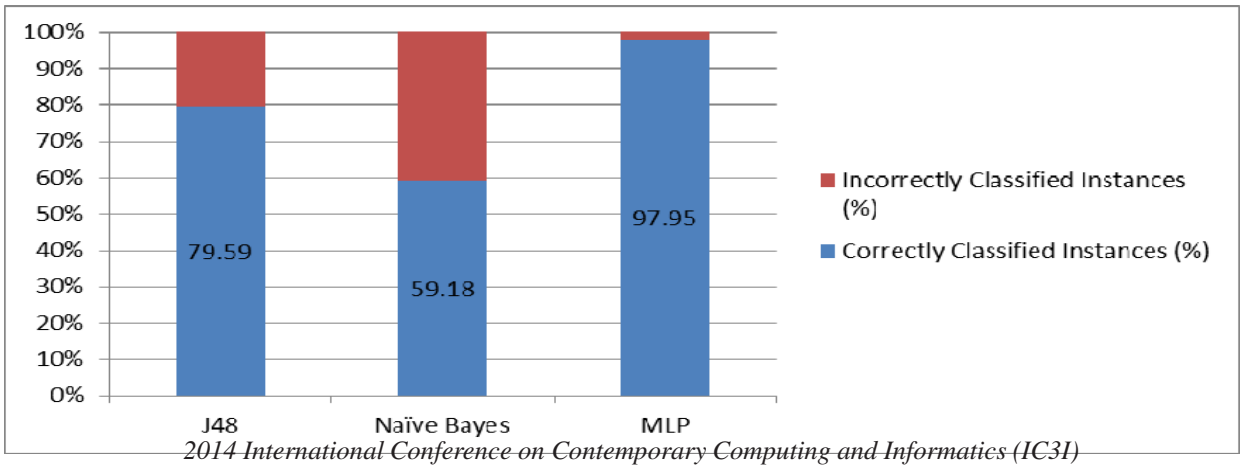

Fig. 6. Comparison between Accuracy of J48, Naïve Bayes and MLP algorithm applied on All-Rounder Data Set

CONCLUSION

By using various classifications based data mining techniques such as J48 algorithm, naïve Bayes and MLP (multilayer Perceptron), a new model is developed for predicting the base price group of a player on the basis of his past performance. Research in this paper concluded from the result that MLP (Multi-layer Perceptron) gives the best accuracy among all other algorithms for each category, a batsman, a bowler and an all-rounder. J48 gave a suitable accuracy whereas Naïve Bayes didn't provide the satisfactory result for the classification. Two algorithms gave a suitable result, hence opening the dimensions for the usage of data mining algorithms in IPL auctioning. This newly developed model could help the IPL franchises in auction to classify a player on his past performance and set his base price accordingly. It is a simple and an accurate mathematical approach used for classification of players using effective data mining algorithms. Whereas in present, traditional methods like coach evaluation or player's brand is used to decide a player's value during IPL auction. This model will provide a mathematical fact for the fixation of certain value for a particular player. This model is itself unique in its own ways, as data mining was never used before for fixing a player's base price during an IPL auction giving a player an amount, what he deserves and neglecting all the false manipulations. So, the use of data mining techniques for selection of players as a valuation factor in the biding can save franchise a lot of spending on players who are poor performers. These models have the ability to build a talented team with minimum cost.

REFERENCES

[1] Singh, S, Gupta, S., & Gupta, V. (2011). "Dynamic bidding strategy for players auction in IPL," International Journal of Sports Science and Engineering, 05(01), 03-16.

[2] Singh, S. (2011), "Measuring the performance of teams in the Indian Premier League,"American Journal of Operations Research, 01, 180-184.

[3] Pankush Kalgotra, Ramesh Sharda, Goutam Chakraborty, "Predictive Modeling in sports leagues: an application in Indian Premier League," in SAS Global Forum 2013, pp. 019-2013.

[4] L. J. A. Lenten, W. Geerling and L. Kónya, "A Hedonic Model of Player Wage Determination from the Indian Premier League Auction: Further Evidence", School of Economics and Finance, La Trobe University, Australia.

[5] S.R. Clarke and J.M. Norman." Dynamic programming in cricket: protecting the weaker batsman,". Asia-Pacific Journal of Operational Research. 1998, 15: 93-108.

[6] Sharda, R. & Delen, D. (2006)," Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, 30, 243-254.

[7] Parker, D, Burns, P, & Natarajan, H. (2008, October)"Player valuations in the Indian Premier League," Frontier Economics.

[8] http://www.iplt20.com

[9] http://www.thatscricket.com