Applied Computing and Informatics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

Sport analytics for cricket game results using machine learning: An experimental study

Kumash Kapadia [a], Hussein Abdel-Jaber [b,*], Fadi Thabtah [a], Wael Hadi [c]

[a] Digital Technologies, Manukau Institute of Technology, Auckland, New Zealand
[b] Information Technology and Computing, Arab Open University, Saudi Arabia
[c] Department of Computer Information Systems, Petra University, Amman, Jordan

ARTICLE INFO

Article history:
Received 1 September 2019
Revised 20 November 2019
Accepted 21 November 2019
Available online xxxx

Keywords:
Cricket
Data science
Machine learning
Prediction
Sport analytics

ABSTRACT

Indian Premier League (IPL) is one of the more popular cricket world tournaments, and its financial is increasing each season, its viewership has increased markedly and the betting market for IPL is growing significantly every year. With cricket being a very dynamic game, bettors and bookies are incentivised to bet on the match results because it is a game that changes ball-by-ball. This paper investigates machine learning technology to deal with the problem of predicting cricket match results based on historical match data of the IPL. Influential features of the dataset have been identified using filter-based methods including Correlation-based Feature Selection, Information Gain (IG), ReliefF and Wrapper. More importantly, machine learning techniques including Naïve Bayes, Random Forest, K-Nearest Neighbour (KNN) and Model Trees (classification via regression) have been adopted to generate predictive models from distinctive feature sets derived by the filter-based methods. Two featured subsets were formulated, one based on home team advantage and other based on Toss decision. Selected machine learning techniques were applied on both feature sets to determine a predictive model. Experimental tests show that tree-based models particularly Random Forest performed better in terms of accuracy, precision and recall metrics when compared to probabilistic and statistical models. However, on the Toss featured subset, none of the considered machine learning algorithms performed well in producing accurate predictive models.

1. Introduction

Cricket is a well-known sport and with its increasing popularity and viewership, change of formats and innovations in tournament played became necessary. To cater for potential future growth, global market research was commissioned by the International Cricket Council (ICC) which revealed that cricket has more than one billion fans worldwide, with the potential for significant growth. Among all formats of cricket, the popularity of Twenty20 Internationals (T20) was the highest with 92%, with 87% of the fans stating that they would like T20 to be included in the Olympic Games [14]. Indian Premier League (IPL) is a T20 tournament which involves players from all over the world [20]. IPL is held once a year, usually during April-May and it is around 2 months long. In 2017, Star India bought the five-year global media rights of IPL for $2.55 billion and the Board of Control for Cricket in India (BCCI) disclosed that IPL contributes $600 million a year to its revenue [3].

Sports analytics is a promising research field which involves deriving valuable information about the game, based on past games played, or even games in progress [24]. The prediction of the final outcome of the match proves very beneficial to team members, team coaches and also bettors. For example, games tactics can be developed by club managers based on the outcome of previous matches or statistics related to certain players [6]. IPL being a very dynamic league, bettors and bookies are incentivised to bet on the match results or during a game. The sports betting industry is growing at a fast rate. For example, in 2009 the global online gambling market was around $20 billion and increased to $40 billion in 2016, of which about 40% was sports betting [11].

* Corresponding author.
E-mail addresses: kapa74@manukaumail.com (K. Kapadia), habdeljaber@arabou.edu.sa (H. Abdel-Jaber), Fadi.fayez@manukau.ac.nz (F. Thabtah), whadi@uop.edu.jo (W. Hadi).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.aci.2019.11.006
2210-8327/

One of the primary approaches used in sport analytics research is machine learning. Machine learning techniques are utilised to predict the match result variable by developing classification models based on certain independent variables such as player's position, weather, location, etc. [23,31]. The process involves training the model based on previous matches played, then the developed model gets evaluated on an independent future match to measure its effectiveness [1]. Often, machine learning models' effectiveness is measured using metrics such as predictive accuracy and error rate among others [30]. Since cricket matches are recorded using multiple independent variables within a historical dataset and one dependant variable, (the outcome of the match) this problem can be dealt with using predictive analytics (classification methods) within machine learning. A classification algorithm will process the input dataset to construct a classification model based on the available historical matches to predict the outcome of future matches as accurately as possible.

In this paper, different types of classification techniques are evaluated to seek accurate models that can predict the outcome of a match. The research question we seek to answer is:

- Can machine learning technology derive accurate predictive models for cricket matches related to IPL, and if so, which machine learning models are the best with respect to accuracy, precision and recall evaluation measures?

To answer the research question different machine learning approaches are experimentally evaluated including probabilistic, Random Forest, statistical and Decision Trees [8,5,17,9]. We used 10 years' data collected from the IPL-T20 tournaments [28]. More details on the data and empirical results are discussed in Sections 3 and 4 respectively. Another important aim of this paper is to seek influential features in a cricket match that might have influence on the outcome of a match. This paper focuses on creating a simplified but effective model to predict the match outcome based on two scenarios: Home Ground & Toss Decision, respectively. Both of these are pre-conditions in a cricket match and can be known easily prior to start of any match. Most of the researchers have tried to explore One-Day Internationals (ODI) and Test match format of cricket but as T20 is new and dynamic, it will be intriguing to investigate. This study can benefit cricket club managers, sport data analysts and scholars interested in sport analytics, among others.

The paper is organised as follows: Section 2 includes a brief overview about the game of cricket, previous work related to sports analytics and the application of machine learning to predict match outcomes. Section 3 discusses the framework for the prediction model and methods applied. Section 4 is dedicated to the description of the dataset and experimental results. Finally, conclusions and future work are presented in Section 5.

2. Literature review

Nimmagadda et al. [24] applied statistical techniques to predict a T20 match result while the match is in progress. The authors have designed a model using a statistical approach to achieve the optimum outcome. Firstly, a multiple regression model is tested to develop a prediction model. Using runs scored per over in the first inning and second inning, algorithms such as Logistic Regression with multi-variable linear regression and Random Forest [4,5] are used to predict the final result. The software used for modelling is Anaconda and Python libraries like pandas, NumPy and IPython to work with the data structure and applying algorithms [2,21]. The main result obtained was based on the impact of toss winner and resultant match winner. The predictive model considered the innings score at regular intervals and the final scores to predict the match result. The model predicted score and run rate projected score were quite near to the final score, in particular the score predicted by the model was more accurate to the actual score. When no feature selection was applied to the dataset the model's accuracy was not satisfactory, i.e. slightly above 50%.

Pathak & Wadhwa [25] investigated the prediction of the result for cricket matches using data mining techniques. They experimented on predicting the outcome for ODI (One Day International) match format based on various factors such as home ground, toss decision, innings, fitness of team players and other dynamic strategies. In addition to the techniques implemented by [16], a Support Vector Machine (SVM) method was used to predict the result [12]. Evaluating the accuracy of these techniques, they developed a tool COP (Cricket Outcome Predictor), which gives the probability for winning an ODI match. The data under study was the international cricket match data from 2001 to 2015 for ODI format and scraped from [7]. Results obtained clearly showed that the classifiers derived by the SVM method outperformed those of Naïve Bayes and Random Forests methods [8,5]. SVM produced 62% accuracy, whereas the accuracy rates of the other methods were around 60%. The COP tool developed in R software [27]enabled a user to select the features to predict the match outcome, and the user could change between the classifiers to make multiple predictions. A notable result was observed when COP system was applied on the India vs. Australia series in which Naïve Bayes derived more competitive classifiers in terms of predicting the match outcome.

Jhanwar & Pudi [15] conducted an experimental study to predict the outcome for ODI cricket matches using data mining techniques. The authors investigated the match result using team players' performance individually in batting and bowling aspects. Initially the potential of 22 players was studied using their career statistics and KNN, Support Vector Machine (SVM), Random Forests, Logistic Regression and Decision Trees techniques were applied [18,12,5,4,9]. To predict the outcome of the match, the relative strength of each team is studied, along with the venue of the match and toss result. The data considered under the study was cricket matches from 2010 to 2014 for 9 country teams in international One-Day format. The accuracy of the KNN model was higher than the other models in predicting the relative strength of the team players giving almost 71% accuracy for the ODI match. There was no feature selection involved in this study.

Kampakis & Thomas [16] conducted a study to predict the outcome of cricket matches in twenty over format. The competition under study was the English Cricket Cup and the model was tested on seasons 2009 to 2014, based on the data from previous matches. A model was developed on simple prediction and then further investigation was carried out on complex features for in-depth analysis. Initially the team data was used and then player data was analysed. Feature selection methods utilised were Chi-square testing, mutual information and Pearson correlation. The authors utilized Naive Bayes, Logistic Regression, Random Forests and Gradient Decision Trees on the selected features from the data [8,4,5,10]. By applying these methods to predict the match outcome, it was found that the model derived by Naïve Bayes offered around 64% prediction accuracy on the dataset used. At the same time comparing the accuracy of different techniques, Naïve Bayes produced the highest level of accuracy, the lowest was Gradient Decision Trees.

Munir et al. [23] experimented with twenty over format cricket matches to predict the outcome using various data mining techniques. The main aim of the study was to combine pre-game and in-game data to predict the outcome. They considered the T20 International match data along with IPL data till 2015 as the training data set. In depth analysis was conducted by segmenting the data on the basis of venue, one team against all other teams, bat-

Please cite this article as: K. Kapadia, H. Abdel-Jaber, F. Thabtah et al., Sport analytics for cricket game results using machine learning: An experimental study, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.11.006

ting first and so on. Decision Tree was applied to predict the match outcome, and produced models with around 78% accuracy for the team that bats first and 75% when it bats second. IG technique was used for feature selection.

3. Methodology

This research attempts to evaluate different machine learning techniques to the problem of predicting the outcome of IPL cricket matches. We design intelligent models to predict a match outcome based on the impact of home ground and toss winner respectively. The team that wins the toss contemplates factors such as weather, pitch and outfield to decide whether to bat or field first, with the intention of securing a strategic advantage. Two models are formulated in this paper, one depicting the impact of home ground and the other considering the effect of toss decision. The former considers 6 variables and the latter, 7 variables.

Fig. 1 depicts the framework of cricket match intelligent models. Initially, the input dataset is pre-processed by eliminating any incomplete records so that there are no missing values in the dataset. Data with no match result were excluded from the classification (5 instances). More importantly, we eliminate features that have no direct impact on the performance of the training phase by applying feature selection. Features including Match ID, Match date and Venue among others have been discarded prior to the training phase of the machine learning techniques (more details are given in Section 4). Once the input dataset has been pre-processed, it is split into two features sets; one feature set that concerns features related to the home ground and another one for the toss decision features.

Once the dataset is split, a number of different learning algorithms are applied on the two feature sets to derive predictive models for the match result. The testing methods used to derive the classifiers is ten-fold cross validation with stratification [32]. These models are then compared to seek the one(s) that can be utilized for forecasting upcoming matches results. The machine learning algorithms that have been implemented to derive the predictive models are Naïve Bayes, Random Forest, K-nearest neighbour and Model Decision Tree. The choice of these methods is based on the diverse learning they adopt to develop the models (Section 4.1 gives more details).

4. Data & result analysis

4.1. Data and features description

Historical data of Indian Premiere League (IPL-T20) tournaments is captured to perform prediction analysis. We consider IPL Cricket matches for 10 years (2008 to 2017) and store them in a dataset. The dataset used consists of 17 variables and 637 instances and was downloaded from [28]. MatchSK and MatchID are unique to matches played. Matches are played on home ground (matches played in the home city of team 1) and a few matches were played on international grounds (South Africa and United Arab Emirates (UAE)). Only two seasons were played out of India, IPL-2009 and IPL-2014 in South Africa and UAE respectively. The dataset variables are depicted in Table 1.

According to the rules of IPL only 8 individual teams participate in each season. However, a few teams have been created and dissolved during the period covered by the data, so the dataset has 13 individual teams. Most seasons are played by Royal Challengers Bangalore, Kings XI Punjab, Delhi Daredevils, Mumbai Indians and Kolkata Knight Riders. Kochi Tuskers Kerala played only one season whereas Gujarat Lions have been part of two seasons. Some teams represented a city but have been dissolved and created again with a new name. For instance, Pune Warriors was dissolved in 2014 and Rising Pune Super giants came into existence in 2016. Sunrisers Hyderabad, created in 2013 was formerly known as Deccan chargers.

4.2. Feature selection

Different feature selection methods were tested to get the influential attributes of the considered dataset. Methods considered are

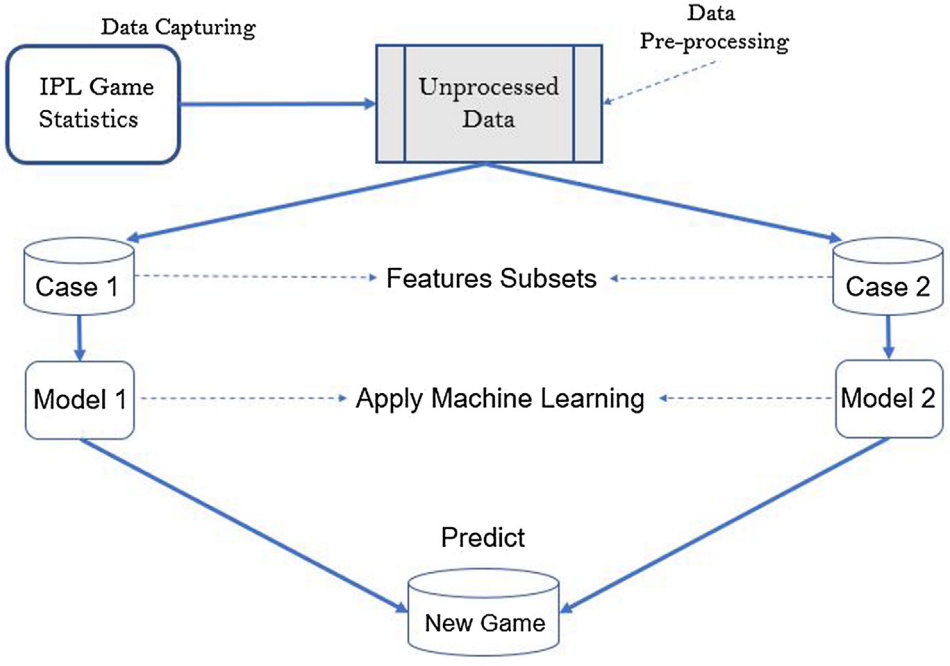

Fig. 1. Proposed Methodology.

Please cite this article as: K. Kapadia, H. Abdel-Jaber, F. Thabtah et al., Sport analytics for cricket game results using machine learning: An experimental study, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.11.006

Table 1
The dataset variables' description.

Attribute	Description
Match_SK	Serial Number for matches played
match_id	Unique ID for match
Team1	Home team for the match
Team2	Away Team for the match
match_date	Date when the match was played
Season_Year	IPL seasons over the years
Venue_Name	The cricket stadium where match was conducted
City_Name	Home Team city for the match
Country_Name	Country for the tournament
Toss_Winner	The team which won the toss for the match
match_winner	Winner of the match
Toss_Name	Decision to bat or field decided by the Toss winner
Win_Type	Winning the match by runs or by wickets
Outcome_Type	Match result that is win, lose or tie
ManOfMatch	The team player that was important to win the match
Win_Margin	Margin by which a team wins the match
Country_id	ID of the country for the tournament

filter based in nature and include Correlation-Based Feature Selection (CFS) [13], IG [26], ReliefF [19] and Wrapper [18]. For the Wrapper we used the following classification algorithms: Decision Tree (C4.5), Naïve Bayes and KNN. For ReliefF and Wrapper, K-Nearest Neighbours (K = 10) was used for evaluating the attributes. Table 2 depicts the results produced against the considered dataset. The results of filter methods have shown that Team1 is the top feature followed by Toss_Winner, Season_Year and City. Whereas, ReliefF selected the Toss related features at top rank and surprisingly Team1, City and Season_Year bottom rank. Considering the results from Wrapper method, the results show significant discrepancy due to the different base classification algorithms used yet Team1 was a consistent choice in the considered methods. After evaluating these results, the features selected for the experiment were Team1, Team2, City, Season_year, Win_Type and Outcome_type. Toss related features were excluded to avoid the risk of overfitting the models derived by the classifiers.

4.3. Experimental settings and evaluation measures

For this research, all experiments have been conducted on WEKA (The Waikato Environment for Knowledge Analysis), a machine learning tool which has automated intelligent techniques [32]. To visually explore the IPL data insights, visualisation tool Tableau has been used [29]. The processing machine used to conduct the experiments is an Intel i5 6th generation processor with 8 GB RAM on Windows 10, 64-bit operating system.

Different algorithms are adopted to deal with the research problem including Naïve Bayes, Random Forest, K-nearest neighbour and Model Decision Tree [8,5,17,9]. These algorithms are selected as they adopt different learning approaches. The hyperparameter settings are the defaults used in WEKA. The predictive models derived by the machine learning algorithms have been evaluated using various metrics including classification Accuracy, Precision and Recall on the basis of confusion metrics analysis. We used ten-fold cross validation with stratification as a testing method to derive the models [32]. Using this method, the training dataset will be divided into ten folds arbitrary with stratification. Then, the learning algorithm will be trained on nine-fold and then tested on the hold-out fold. The process gets repeated ten times to produce an average predictive accuracy of the model.

4.4. Results analysis

4.4.1. Case 1: Home team features set

The aim of predicting the match result in the first model is to evaluate the impact of home ground advantage. In this experiment, the variable "Result" is derived based on the Home Team (Team 1) winning the match, when the match is played on home ground. For example, it is the frequency of Chennai Super Kings winning the match when it plays in its home ground Chennai. The attribute "Result" will be the target class for predicting the outcome by classification. The format of the tournament is that each team is designated one city as its home ground, and two matches are played by combination of two teams, playing once at the first team's home ground and once at the other team's home. In the considered dataset, two seasons were played in a foreign country and for these matches the venue is changed to the home ground of their respective teams. A few teams, such as Kings XI Punjab have different home grounds, but their original home ground is Mohali. For this experiment, only those features that have an impact on the home ground are considered.

The classification results derived by the considered machine learning techniques against the Home Team features set are shown in Figs. 2A & 2B. Based on Fig. 2A, it is apparent that Naive Bayes is the most accurate model to predict the winner. The accuracy of Naïve Bayes is 57% which is relatively low; Random Forest and Model Trees algorithms also produce equally low results with 54% and 56% accuracy respectively. Accuracy produced by KNN is the lowest with only 52%. This means that using the considered features machine learning techniques were unable to improve predictive accuracy as all models showed an unacceptable level of accuracy.

The Precision results shown in Fig. 2B is consistent with the accuracy results as Naive Bayes algorithm outperformed the rest of algorithms with 60.5% precision, due to lower false positives (FPs). According to the confusion matrix results Naïve Bayes algorithm had misclassified 136 instances whereas Model Tree, KNN and Random Forest misclassified 171, 146 and 160 instances respectively. These misclassifications have increased the FPs and subsequently lowered the Precision results, especially for KNN and Random Forest algorithms.

Fig. 2B shows that the Recall rate for Model Trees is higher than the other algorithms. The Recall rate obtained by Model Trees was 68.6% as this algorithm achieved low false negatives (FNs), just 106 instances. Precision and Recall rates for KNN is around 55%, which

Table 2
Attributes chosen by the different filters and wrappers.

CFS	IG	ReliefF	Wrapper with C4.5	Wrapper with Naïve Bayes	Wrapper with KNN
Team1	Team1	Toss_Winner	Team1	Team1	Team1
Season_year	City	Toss_Name	Season_year	Toss_Name	Team2
City	Toss_Winner	Win_Type	Win_Margin		Toss_Winner
Toss_Winner	Season_year	Team2			Toss_Name
Toss_Name	Team2	Win_Margin			Win_Type
	Toss_Name	Outcome_Type			
	Win_Type	Team1			
	Outcome_Type	City			
	Win_Margin	Season_year			

Please cite this article as: K. Kapadia, H. Abdel-Jaber, F. Thabtah et al., Sport analytics for cricket game results using machine learning: An experimental study, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.11.006

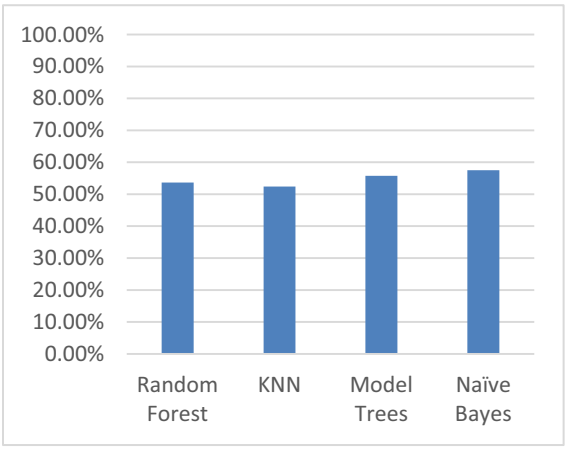

Fig. 2A. Accuracy figures derived by the machine learning techniques from Home Team features set.

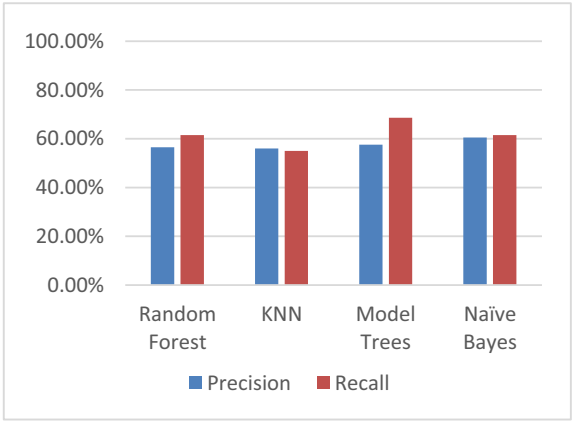

Fig. 2B. Precision and Recall figures derived by the machine learning techniques from Home Team features set.

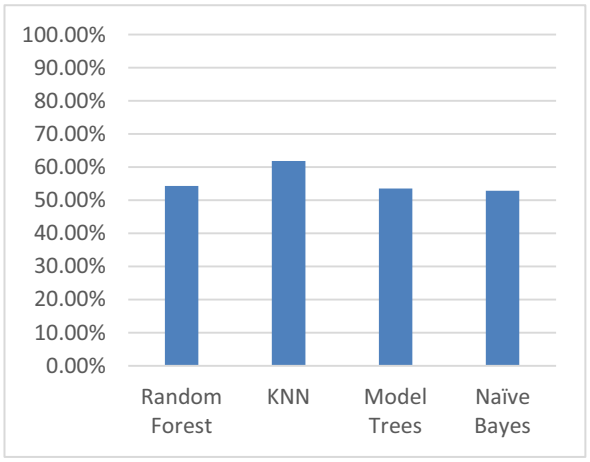

Fig. 3A. Accuracy figures derived by the machine learning techniques from Toss Winner features set.

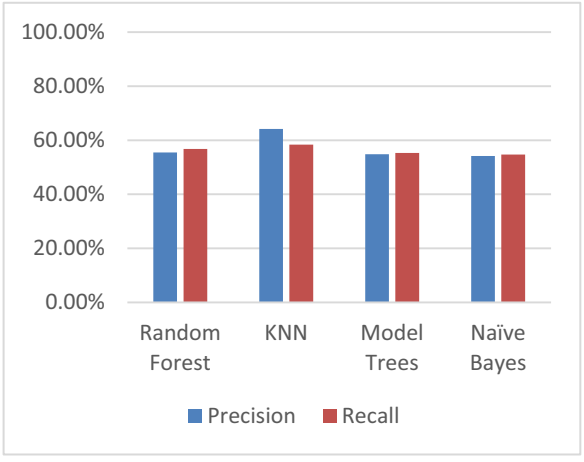

Fig. 3B. Precision and Recall figures derived by the machine learning techniques from Toss Winner features set.

is the worst of the figures obtained by the other machine learning algorithms. Recall rates for Random Forest and Naïve Bayes are the same with 61.5%. Overall, it seems that Random Forest, Model Tree and Naïve Bayes achieved acceptable performance with respect to the Recall metric for this feature set.

Looking closely at the confusion matrix results, 146 instances which should belong to the "Lose" class have been misclassified by KNN to the "Win" class. This has contributed to a low Precision rate and the reason for low Recall rate by KNN is that 152 instances that should be "Win" are classified as "Lose". By contrast, for Naïve Bayes, only 136 were wrongly predicted as "Win" instead of "Lose", that is 47% FP, and 130 instances were misclassified as "Lose" in place of "Win" resulting in 38% FN rate. The results achieved from Random Forest and Naïve Bayes algorithm are quite similar, in terms of Recall rate, as the instances incorrectly classified as "Lose" were 130. The Recall rate for Model Trees was 69% due to 106 instances misclassified as "Lose" instead of "Win" (FNs). When we compare the ratio of misclassified "Win" (FPs) and misclassified "Lose" (FNs) for KNN and Naïve Bayes, the results derived from the latter are more accurate.

4.4.2. Case 2: Toss winner features set

The method of predicting the match result in the second model is to evaluate the impact of winning the toss prior to the game and making a decision whether to bat or field first. In this model, the dependant variable "Result" is derived based on the Toss winning Team, winning the match, i.e., the "Win" is determined when Toss winning Team wins the Match. The variable 'Result' will be the target class for predicting the outcome by classification. Fig. 3A and Fig. 3B depict the performance results of the machine learning algorithms against the Toss Winner features set. The figures show that the classification accuracy of KNN is 62%, making it a more appropriate model than the other models. Naïve Bayes produced a low accuracy result of around 52%, not a good fit for this type of predictive task.

Analysing the results of the models produced for the Toss Winner features set, KNN surprisingly produced the best results. For instance, Precision rate for KNN is high at 64.2% and KNN's Recall is 58.4%, both reflecting correct classification for the "Win" class label, with a TP rate of 58%. Compared to the Home Team features set, the results from Random Forest are lower, as this algorithm's FP rate is 48.35% for class 'Win'. The Decision Tree model's results are slightly lower for the Toss decision dataset compared to the Home Team features set. The FP rate for the Decision Tree algorithm is around 48% due to 147 instances misclassified as "Win" instead of "Lose", while 144 instances were incorrectly predicted as "Lose" that were supposed to be "Win" by this algorithm. Similarly, for Naïve Bayes a lower rate was observed in terms of Preci-

Please cite this article as: K. Kapadia, H. Abdel-Jaber, F. Thabtah et al., Sport analytics for cricket game results using machine learning: An experimental study, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.11.006

sion and Recall rates, 49% of "Win" were misclassified leading to the Precision rate being 6% lower compared to that achieved on the Home team subset. KNN performs extremely well when processing the Toss Winner features set.

5. Conclusions and future work

Applying machine learning for analysing cricket sports by considering historical game data, players performance, natural parameters, pre-game conditions and other features is beneficial for multiple stakeholders. In a dynamic format like T20, where the situation in a game changes on every ball, it becomes challenging to predict the match outcome. For predicting the final outcome of a T20 cricket match, we have investigated machine learning technology for the possibility of improving the prediction rate of the results of matches. We have formulated the problem in two scenarios, named for the most influential features, firstly the Home Team features set and secondly Toss Winner decision features set.

By analysing the results achieved using four different machine learning techniques on 10 years' T20 matches, the model built on Toss related features generates slightly better results than Home Advantage in terms of the evaluation measures used (Accuracy, Precision, Recall, FPs, FNs, etc). Particularly, KNN outperformed the other algorithms when processing the Toss Winner feature set by deriving higher accuracy predictive models than Decision Trees, Probabilistic and Statistical models. Furthermore, incorrectly classified instances by KNN, both FPs and FNs, are low, resulting in imporved Precision and Recall rates. KNN achieved 134 misclassified instances to "Lose" class that were supposed to belong to "Win" class and 105 instances were wrongly classified as "Win" which is around 35%. On the other hand, the results derived from Naïve Bayes on Toss Decision subset are not promising due to the class independence assumption of the algorithm. But Naïve Bayes produced better results for Home Team subset. This study is beneficial to team managers and scholars interested in cricket data analytics.

Machine learning may slightly improve predicting the results based on pre-game conditions but at this stage it cannot be an acceptable solution due to lack of variables in the dataset, which can be considered as one of this research's limitations. In order for machine learning techniques to be productive, more data including live data streaming and statistics of players are needed. Considering the dynamics of the tournament, team players' data and statistics are required. It would be advantageous to predict the final score of the innings by analysing the run rate per over and also checking the probability of winning for each team depending on the actual run rate and the required run rate in the second innings. Similar models can be built for other cricket formats, i.e. test cricket and ODI series. Finally, in the near future we intend to build a classification system based on deep learning to capture more useful features that can potentially improve the accuracy of prediction while the game is in progress.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Abdelhamid, N., Ayesh, A., Thabtah, F., 2012. An experimental study of three different rule ranking formulas in associative classification mining, in: Proceedings of the 7th IEEE International Conference for Internet Technology and Secured Transactions (ICITST-2012), pp. 795–800, UK.
[2] Analytics, C., 2017. Anaconda software distribution. Computer software Vers, 2-2.
[3] Bhatia, G., 2017. The richest sport in India just keeps getting richer. Retrieved December 25, 2018 from https://www.cnbc.com/2017/09/27/indian-premier-league-cricket-a-rich-sport-is-getting-a-lot-richer.html.
[4] D. Böhning, Multinomial logistic regression algorithm, Ann. Inst. Stat. Math. 44 (1) (1992) 197–200.
[5] L. Breiman, Random forests, Machine Learn. 45 (1) (2001) 5–32.
[6] R. Bunker, F. Thabtah, A machine learning framework for sport result prediction, J. Appl. Comput. Informatic (2017), https://doi.org/10.1016/j.aci.2017.09.005. Elsevier.
[7] Cricinfo, n.d. Retrieved from http://www.espncricinfo.com.
[8] R.O. Duda, P.E. Hart, D.G. Stork, Bayesian decision theory, Pattern Classification 11 (4) (2001) 99–102.
[9] E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten, Using model trees for classification, Machine Learn. 32 (1) (1998) 63–76.
[10] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[11] Fuller, S., n.d. Topic: Sports Betting. Retrieved December 25, 2018 from https://www.statista.com/topics/1740/sports-betting/.
[12] Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS technical report, 14(1), 5–16.
[13] M.A. Hall, Correlation-based feature selection for machine, learning, 1999.
[14] ICC-Cricket, 2018. First global market research project unveils more than one billion cricket fans [Press release]. Retrieved December 20, 2018, from https://www.icc-cricket.com/media-releases/759733.
[15] Jhanwar, M.G., Pudi, V., 2016. Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016).
[16] Kampakis, S., Thomas, W., 2015. Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. arXiv preprint arXiv:1511.05837.
[17] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy k-nearest neighbor algorithm, IEEE Trans. Systems Man Cybernetics 4 (1985) 580–585.
[18] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[19] Kononenko, I., 1994. Estimating attributes: analysis and extensions of RELIEF, in: European conference on machine learning, pp. 171–182. Springer, Berlin, Heidelberg.
[20] Marwaha, D.Y., 2013. The Indian Premier League: What are the factors that determine player value?.
[21] W. McKinney, Python for data analysis: Data wrangling with Pandas, NumPy, and IPython, O'Reilly Media, Inc., 2012.
[22] R. Mohammad, F. Thabtah, L. McCluskey, Tutorial and critical analysis of phishing websites methods, Comput. Sci. Rev. J. (2015) 1–36, Elsevier.
[23] Munir, F., Hasan, M.K., Ahmed, S., Md Quraish, S., 2015. Predicting a T20 cricket match result while the match is in progress (Doctoral dissertation, BRAC University).
[24] A. Nimmagadda, N.V. Kalyan, M. Venkatesh, N.N.S. Teja, C.G. Raju, Cricket score and winning prediction using data mining, Int. J. Adv. Res. Development 3 (3) (2018) 299–302.
[25] N. Pathak, H. Wadhwa, Applications of modern classification techniques to predict the outcome of ODI cricket, Procedia Comput. Sci. 87 (2016) 55–60.
[26] Quinlan, J.R., 2014. C4. 5: programs for machine learning. Elsevier.
[27] R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved January 04, 2019 from http://www.R-project.org/.
[28] Reddy, S., 2018. IPL 2017. Retrieved December 05, 2018 from https://www.kaggle.com/somashekharareddy/ipl-2017#DIM_MATCH.xlsx.
[29] Stolte, C., Hanrahan, P., Chabot, C., n.d. Tableau: Business Intelligence and Analytics Software. Retrieved December 15, 2018 from https://www.tableau.com/.
[30] Thabtah F., 2006. Pruning techniques in associative classification: survey and comparison. J. Digital Information Manage., Volume 4:202-205. Digital Information Research Foundation.
[31] Thabtah, F., Cowling, P., Peng, Y., 2005. A study of Predictive Accuracy for Four Associative Classifiers. J. Digital Information Manage., Volume 3:202–205. Digital Information Research Foundation.
[32] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

Please cite this article as: K. Kapadia, H. Abdel-Jaber, F. Thabtah et al., Sport analytics for cricket game results using machine learning: An experimental study, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2019.11.006