Survey on machine leaning based game predictions

Sallauddin Mohmmad¹, V. Nikhitha Madishetti², Yarra Nitin³, Bonthala Prabhanjan Yadav⁴, Bommagani Sathyra Sree⁵, Beesupaka Marvel Moses⁶
¹School of Computer Science & Artificial Intelligence, S R University, Warangal, Telngana, India.
²,³,⁵,⁶Department of Computer Science and Engineering, S R Engineering College, Warangal, Telngana, India.
⁴Sumathi Reddy Institute of Technology for Women, Warangal, India.
¹sallauddin.md@gmail.com

Abstract: In the world wide millions of people interested on games and competitive matches. The stakeholders stand for one team and produce the sponsorship to the players. Huge amount of money transferred from one hand to other hand. So that stakeholder wants to select a good players into his teams. Here Machine Learning based multi variant regression algorithms used to calculate the progress of each player based on previous datasets to predict the performance at on-going match. To extract the features from on-going match characterized with learned datasets by implementing the Support Vector Machine (SVM), Gaussian Fit-chime (GAU) and KNN algorithms which perform the optimal classification on trained datasets. Feature selection and game predictions are become critical analytical process. The performance of the model effected and produces the outcome based on the feature selection. In this process some irrelevant variables removed to reduce the burden of algorithms and input datasets dimensions. This process speed up the dataset learning using various algorithms to produce the game predictions. The machine learning models mostly preferred algorithms to implement in feature selection are Linear Regression, Decision Tree Regression, Random Forest Regression and Boosting Algorithm like Adaptive Boosting (AdaBoost) Algorithm. In this paper we discussed about how to predict the game score based on trained datasets using various algorithms on Machine Learning platform.

Keywords: SVM, GAU, KNN, Linear Regression, Decision Tree Regression, Random Forest Regression and Boosting Algorithm.

1. Introduction
With technology innovation developing increasingly more progressed over the most recent couple of years, a top to bottom securing of information has gotten generally simple. Thus, Machine Learning is be-coming a significant pattern in sports examination due to the accessibility of live just as historical data. Sports analytics is the way toward gathering past match’s information and examining them to separate the basic information out of it, with an expectation that it encourages in decision making [1]. Decision making be anything including which player to purchase during a closeout, which player to set on the field for the upcoming match, or something more key assignment like, constructing the strategies for forthcoming matches dependent on players’ past performances. Machine Learning can be utilized affectively over different events in sports, both on-the-field and off-the-field. At the point when it is about on-the-field, AI applies to the investigation of a player’s wellness level, plan of strategies, or choose shot choice. It is additionally utilized in anticipating the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under licence by IOP Publishing Ltd
2. Framework of game prediction

The game prediction on the cricket is most noticeable with huge datasets in coverage of more number of years. Here the ML base predictions applied on single player and team to assess the on-going game winning strategy [8]. The objective of our proposal has twofold. First, we need to identify the features set which created major impacts on the result of games in the CRICKET. Second, when the feature set is known, we will use that data and use ML algorithm to fabricate a prediction model. For this situation, supervised learning appears to be the most fitting technique for such a goal. In supervised learning, the information will incorporate a training dataset with independent factors, for assist, steal, and free throws made[9][10]. Every one of these factors shows the group's capacities against a relevant variable (the result of past games). A short time later, the point is to foresee the result factors by applying a model from recorded cases (subordinate factors just as obvious objective variable qualities). This model will be used to gauge the objective variable incentive in a concealed game (test information). In this exploration venture, the attention was on factors identified with groups, players and rivals, for example, RUNS, WICKETS, OVERS, RUNS IN LAST FIVE OVERTS, WICKETS IN LAST FIVE OVERTS and TOTAL SCORE (LABEL).

![Framework for games results classification and prediction](image)

**Figure 1:** Framework for games results classification and prediction.

Game predictions mostly done by supervised learning such as regression or classification. Figure 1
explained the steps of complete machine learning based prediction on data model. Initially the system need to ready with primitive data sets of statistics and the features extracted. By implementing the label on data sets machine learning algorithms applies the supervised learning strategies to produce the predicted result[19]. If we want to predict a model initially system need to learned dataset by either supervised or unsupervised learning algorithms. Consider a model \( y=f(x) \) to predict this model we need to create the dataset's=\((x_1 ,y1),(x_2 ,y2),(x_3 ,y3),....(x_n ,yn))\). Here the output(y) type also a key point. Based on the output type only algorithms operate on D. Supervised learning perform the regression which produce output in continuous value based and classification which produce the discrete kind of values[21-23]. In the dataset various columns created for to predict the performance of a single player. The columns are number boundaries, number of catches, number matches played, number wickets, Number of over’s played and average strike rate. The multivariate regression model produce the out \( y \) based on above six mentioned attributes[18].

\[
y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6
\]  
(1)

Here \( \beta \) weight value of each attributes. The total team weight prediction evaluated based on each individual payer strength which gained by outcome \( y \). The total team weight \( T_w \) given as:

\[
T_w = \frac{\sum_{i=1}^{n} y_i}{\text{total appearance}}
\]  
(2)

| Table 1: Sample dataset variables and their descriptions. |
|----------------------------------------------------------|
| Mid | 1 |
| Date | 18-04-2008 |
| Venue | M Chinnaswamy Stadium |
| Bat_team | Kolkata Knight Riders |
| Ball_team | Royal Challengers Bangalore |
| Batsman | BB McCullum |
| Bowler | P Kumar |
| Runs | 1 |
| Wickets | 0 |
| Overs | 0.2 |
| runs_last_5 | 1 |
| wickets_last_5 | 0 |
| striker | 0 |
| non-striker | 0 |
| total | 222 |

[17]The Sample dataset variables and their descriptions are shown in the Table 1. In the game predictions initially perform the data pre-processing before the learning of datasets to filter the all kinds of noise. The empty columns, missing values, normalized certain variables and etc[11][12]. Mostly the below mentioned pre-processing techniques implement on datasets given as:

- Removing unwanted columns.
- Keeping only consistent teams.
- Removing the first 5 over’s data set in every match.
- Converting the column 'date' from string into date time object.
- Handling categorical features.
- Splitting dataset into train and test set on the basis of date.

3. Feature selection and learning models

Feature selection and game predictions are become critical analytical process. The performance of the model effected and produces the outcome based on the feature selection. In this process some irrelevant variables removed to reduce the burden of algorithms and input datasets dimensions. This process speed up the dataset learning using various algorithms to produce the game predictions [13][14]. The machine learning models mostly preferred to implement in feature selection given as:
• Linear Regression.
• Decision Tree Regression.
• Random Forest Regression.
• Boosting Algorithm like Adaptive Boosting (AdaBoost) Algorithm.

Mean while some errors also rectified with respect to algorithms and the value of error vary based on performance of the algorithm[15]. The errors are probably Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)[16].

3.1. Methodology

In our project after preprocessing the data we have 9 out of 15 and rows of 76,014. Then we have removed few batting teams and bowling teams so has to consider only consistent teams which were presently playing the match in data pre-processing itself.

consistent_teams = ['Kolkata Knight Riders', 'Chennai Super Kings', 'Rajasthan Royals', 'Mumbai Indians', 'Kings XI Punjab', 'Royal Challengers Bangalore', 'Delhi Daredevils', 'Sunrisers Hyderabad']

The above teams were only considered out of some other teams then dataset was reduced to 53811 rows and 9 columns. Then we were removing the first five over’s such that at least 5 over’s data is required for good prediction. So after removing these rows we get the 40108 rows and 9 columns. Then we applied one hot encoding replacing with numerical data since before the data is categorical then data is replaced with 0’s and 1’s which is easy to predict the final score. Then based on this numerical data we have spitted the data into test train splitting. this splitting is done based on time since data set is a time series kind. Then test train splitting is done >2017 is taken as test and remaining from 2008 -2016 taken as training.

Training set: (37330, 21) and Test set: (2778, 21).

Using logistic regression we were getting very low prediction like 4.8956083513318935.

Table 2: Model of evaluation with different algorithms and their error values

| Model Evaluation       | Error Type | Error Value         |
|------------------------|------------|---------------------|
| Linear Regression      | MAE        | 12.118617546193295  |
|                        | MSE        | 251.00792310417455  |
|                        | RMSE       | 15.843229566732111  |
| Decision Tree Regression| MAE    | 16.904967602591793  |
|                        | MSE        | 530.4694024478042   |
|                        | RMSE       | 23.031921379854616  |
| Random Forest Regression| MAE    | 13.611577573794097  |
|                        | MSE        | 322.42698682030436  |
|                        | RMSE       | 17.95625202597425   |
| AdaBoost Regression    | MAE        | 12.137835661931923  |
|                        | MSE        | 247.04286032001912  |
|                        | RMSE       | 17.95625202597425   |

In the processing of the result system model evaluation and error type with error rates presented in Table 2.Finally we used simple regression model for the better model prediction. Some of the predictions were:

Prediction 1:
- Date: 14th April 2019
- IPL : Season 12
- Match number: 30
- Teams: Sunrisers Hyderabad vs. Delhi Daredevils
- First Innings final score: 155/7

The above mentioned data is actual data after using simple linear regression we get output prediction score near to actual score that given as:
- The final predicted score (range): 157 to 172.
Similarly we tried to see same prediction by taking different inputs were this model works better which is given as:

Prediction 2:
- **Date**: 10th May 2019
- **IPL**: Season 12
- **Match number**: 59 (Eliminator)
- **Teams**: Delhi Daredevils vs. Chennai Super Kings
- **First Innings final score**: 147/9
- **The final predicted score (range)**: 137 to 152.

### 4. Conclusions

In this paper we discussed about game prediction based on the trained datasets of a player. By implementing of Machine Learning algorithms the stakeholder can select a good team based on the players previous performance result which analyzed by algorithms. The technology usage on game prediction succeeded in some cases but not in all cases. In the real time scenario this could helps to predict game winners no obvious. In the entire game time our feature extraction will not balance the concept game prediction. In the future I will continue my research on game predictions and analyze the better way of algorithms implantation for game predictions.

### 5. References

[1] N Abdelhamid, F Thabtah and H Abdel-jaber 2017 Phishing detection: A recent intelligent machine learning comparison based on models content and features *IEEE International Conference on Intelligence and Security Informatics (ISI)* Beijing 72-77.

[2] Kain, K J and Logan T D (2014) Are sports betting markets prediction markets? Evidence from a new test *Journal of Sports Economics* 15 45–63.

[3] Weissbock J Viktor H and Inkpen D 2013 Use of Performance Metrics to Forecast Success in the National Hockey League *Workshop on Sports Data Mining at ECML/PKDD*.

[4] Agha, N and Tyler B D 2017 An investigation of highly identified fans who bet against their favorite teams *Sport Management Review* 20 296–308.

[5] An J Y 2016 Improving protein–protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model *Protein Sci* 25 10 1825–1833.

[6] Dadi Ramesh, Syed Nawaz Pasha and Mohammad Sallauddin Nov 2018 Cognitive-Based Adaptive Path Planning for Mobile Robot in Dynamic Environment *Advances in Intelligent Systems and Computing Springer* 117 117-123.

[7] Yang, Y et al 2016 Args-oap: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured arg-database. *Bioinformatics* 32 2346–2351.

[8] Mohammed Ali Shaik, P Praveen and R Vijaya Prakash June 2019 Novel Classification Scheme for Multi Agents *Asian Journal of Computer Science and Technology* 8 S3 54-58.

[9] Arango-Argoty and G et al 2018 Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data *Microbiome* 6 23.

[10] J Bhavana and Komuravelly Sudheer Kumar 2018 A Study on the Enhanced Approach of Data Mining Towards Providing Security for Cloud Computing *Indian Journal of Public Health Research & Development* 9 11 1176-1179.

[11] Junsomboon N and Phienthrakul T 2017 Combining over-sampling and under-sampling techniques for imbalance dataset *Proceedings of the 9th International Conference on Machine Learning and Computing* 243–247.

[12] R Guns and R Rousseau 2014 Recommending research collaborations using link prediction and random forest classifiers *Scientometrics* 101 2 1461–1473.
[13] Sandeep CH, Thirupathi V, Pramod kumar P and Naresh kumar S 2019 Goals and Model of Network Security *International Journal of Advanced Science and Technology* **28(20)** 593-599.

[14] Coleman B J 2017 Team Travel Effects and the College Football Betting Market *Journal of Sports Economics* **18** 4 388–425.

[15] Dr M Sheshikala, Sallauddin Mohmmad and Shabana 2018 Survey on Multi Level Security for IoT Network in cloud and Data Centers *Jour of Adv Research in Dynamical & Control Systems* **10** 10 134-146.

[16] Harshavardhan A, Suresh Babu Dr and Venugopal T Dr 2017 “Brain tumor segmentation methods – A Survey ” *Jour of Adv Research in Dynamical & Control Systems* **11** 240-245

[17] Harshavardhan A, Suresh Babu Dr and Venugopal T Dr 2016 “3D Surface Measurement through Easy-snap Phase Shift Fringe Projection”, *Springer conference International Conference on Advanced Computing and Intelligent Engineering Proceedings of ICACIE* 1 179-186

[18] Harshavardhan A, Suresh Babu and Dr, Venugopal T Dr 2017 “An Improved Brain Tumor Segmentation Method from MRI Brain Images” 2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT) IEEE 1–7. DOI.org (Crossref) doi:10.1109/ICECIT.2017.8453435.

[19] A.Harshavardhan Syed Nawaz Pasha Sallauddin MD, D.Ramesh 2019 “Techniques used for clustering data and integrating cluster analysis within mathematical programming” *journal of mechanics of continua and mathematical sciences* **14**(6) 546-57, DOI.org (Crossref)https://doi.org/10.26782/jmcms.2019.12.00038

[20] Harshavardhan A and K. Shruthi. 2017 “An effective implementation of faulty node detection in mobile wireless network.” *International Journal of Advanced Research in Computer Science* **8**(8) 705–08. DOI.org (Crossref) doi:10.26483/ijarcs.v8i8.4877.

[21] Rajasri I, Guptha AVSSKS and Rao YVD 2016 Generation of Egts: Hamming Number Approach Procedia Engineering 144 537-542 10.1016/j.proeng.2016.05.039

[22] Mahender K, Kumar TA and Ramesh KS 2017 Performance study of OFDM over multipath fading channels for next wireless communications International Journal of Applied Engineering Research 12(20) 10205-10210

[23] Seena Naik K and Sudarshan E 2019 Smart healthcare monitoring system using raspberry Pi on IoT platform ARPN Journal of Engineering and Applied Sciences 14(4) 872-876.