Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1853 - 1863
Received 15 April 2021; Accepted 05 May 2021.

Applying Data Science for Cricket Predictions

M. Arun Manicka Raja[1], Vallabhajyosyula Vishnu Laxmi Manasa[2], D. SreeNikitha Reddy[3], K. Soma Sundari[4]

[1]Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Puduvoyal, Tiruvallur, Tamil Nadu, India. E-mail: arunmcse@rmkcet.ac.in

[2]Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Puduvoyal, Tiruvallur, Tamil Nadu, India. E-mail: vjvlakshmimanasa@gmail.com

[3]Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Puduvoyal, Tiruvallur, Tamil Nadu, India. E-mail: sreenikithareddy@gmail.com

[4]Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Puduvoyal, Tiruvallur, Tamil Nadu, India. E-mail: somasundari072000@gamil.com

ABSTRACT

In any sport selecting the best players is an important and crucial task. In the game of Cricket, the performance of the players depends on various factors such as the team they are playing against, venue of the match etc. From 15 to 20 players, cricket management team, coach and captain select 11 players for the match. They select them by analyzing and using various statistics and strategies. Every batsman contributes to the game by scoring maximum runs possible and each bowler contributes by taking maximum wickets and avoid the opponent team from scoring runs. This paper predicts the performance of players such as strike-rate of the batsman, number of wickets a bowler can take and his economy. These predictions are regression problems and various machine learning regression algorithms have been used to predict the statistics of each player. We used random forest, decision tree, linear regression, KNN, Gradient boosting regressors to generate the prediction models for the problems. For each statistic prediction certain regression algorithms are found to be accurate and effective and are discussed further in this paper.

KEYWORDS

Prediction, Cricket, Machine Learning, KNN,Random Forest, Decision Tree, Gradient Boosting, LinearRegression.

Introduction

Cricket is a world-famous sport played with two teams consisting of eleven players each. A perfect team is said to have a right number of bowlers, batsman and allrounders. A batsman has to score as many runs as possible to improve the overall match's run-rate and a bowler has to take as many wickets as possible to avoid the opponent team from scoring runs. Allrounders can do the job of both a batsman and a bowler. Each player has his contribution to achieve the success in the game. Performance of each player can be influenced by the factors like the team they are playing against, venue of the match etc. For our research we have only considered T20 matches. Twenty20 matches often last for around 3 hours, where each inning will be for about 70 to 90 minutes with an interval of 10 minutes. One inning consists of 20 overs. Eachteamwill consist of 11 players.This time span of T20 is far less when compared to the traditional form of the cricket that is Test Cricket. The main motto of reducing the time period of the game is to attract the audience and television viewers. Since the introduction of T20s the game has took a turn worldwide which resulted in many premier leagues that turned out very successful. One of the best examples is the Indian Premier League.In this paper, our aim is to predict each player's performance, overall match's run-rate and suggest suitable best playing eleven players for an upcoming T20 match. Each player's performance includes Strike rate of the bats-man, Economy of the bowler and number of wickets the bowler can take in the upcoming match. For every prediction, there are certain parameters considered accordingly. For example, prediction of economy of the bowler will be based on the parameters such as bowler's name, team they are playing against, venue of the match and past economies of that particular bowler. Since overall performance of the match is crucial, we predict best playing eleven players based on the past matches and the team they will be playing against.

Related Works

After researching online and surfing lot of websites we found that there were very few articles related to performance of player in the game of cricket. Very few researchers have observed the performance of cricket players.

Received 24 October 2020; Accepted 09 December 2020.

Muthuswamy[1] and Lam contributed their work in predicting the performance of Indian bowlers against international teams. Back International Journal of Data Mining & Knowledge Management Process, propagation network and radial basis network function was used by them to predict number of runs a bowler is likely to concede and maximum number of wickets a bowler is likely to take in an ODI match. Wikramasinghe[2] used linear model to predict the performance of batsmen in test cricket. Barr and of getting out and used a 2D representation with Strike Rate on X axis and P(out) on Y axis. Then they defined criterion based on strike-rate, P(out) and batting average of the batsmen. Iyer[3] and Sharda classified batsman and bowler into three different categories that is – performer, average and failure. Based on this information, they suggested if a particular player has to be included in world cup 2007 or not. Jhanwar and Pudi[4] predicted the result based on the strength of two teams. He calculated performances of all the players in each team and used it for prediction. Lemmer introduced a new parameter called Combined Bowling Rate. He defined it as the combination of strike-rate, bowling average and economy. He used this for the further predictions. Bhattacharjee and Pahinkar[7] used the previously proposed combined bowling rate for predictions in Indian Premier League (IPL). Similarly, in a paper that was published in 2010 named ‘CricAI’, a classification-based tool to predict the outcome in ODI cricket many unique parameters were considered such as home game advantage,day/night match, winning toss and batting first etc. Algorithms used by them are Bayes theorem (independent variables), decision tree, bagging and boosting. Data set was organized in Attribute relation file format (ARFF). Also, in a research paper named Data Mining and Machine Learning in Cricket Match Outcome Prediction they used following parameters: Cricket player performance analysis, Cricket match simulation, Cricket team selection (relative team strength). They also mentioned about social media, fan sentiments on twitter and score prediction by fans.

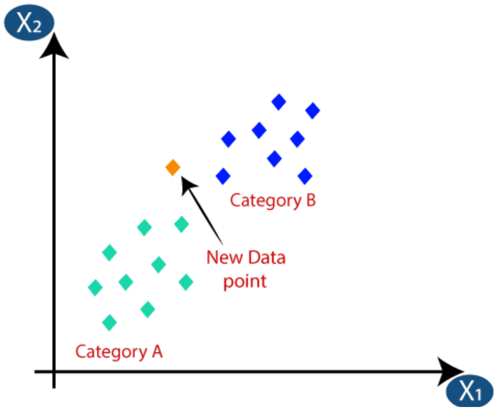
Our work is probably one of the very few generalized approaches to predict number of runs a batsman will score and number of wickets a player gain on a particular match day. We predict entire run-rate of the team, batsman strike-rate, bowler’s economy and also predict best playing eleven for the team. We used supervised machine learning algorithms to predict the performance and statistics of any player in future matches.

Architecture

Learning Algorithms

K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest yet effective Machine Learning algorithms. It is a Supervised Learning technique has intense application especially in pattern recognition, data-mining and intrusion detection. The KNN algorithm works with the concept that similar things exist in close proximity. In other words, similar components are near to each other.



Received 24 October 2020; Accepted 09 December 2020.

Working of KNN Algorithm

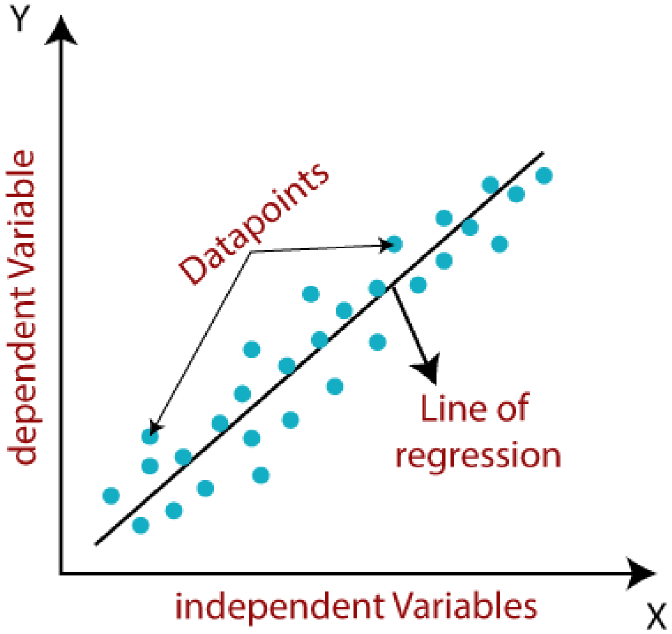
KNN finds the distances between all the existing examples of the data and new data. This done by selecting the examples (K) closest to the new data point, then the most frequent label (in the case of classification) or averages the labels (in the case of regression) are considered.

Linear Regression

Linear Regression is a supervised machine learning algorithm. It predicts the continuous-output and has a constant slope. It has a linear relationship between a dependent, and multiple-independent variables. Linear regression has a linear relationship, which means it finds value of the dependent variable with respect to the value of the independent variable.

Mathematically, linear regression can be given as:

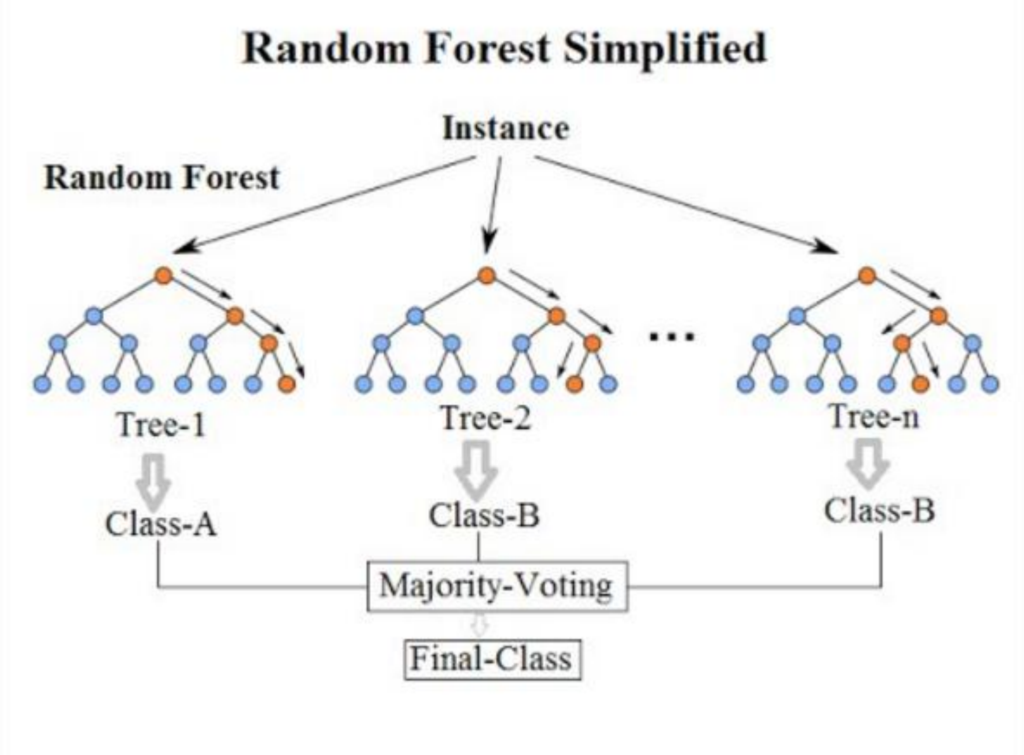
$y= a_0+a_1x+ \varepsilon$
Here,
Yis Dependent Variable
X is Independent Variable
a0 is intercept of the line
a1 = Linear regression coefficient
ε = random error



Random Forest

Random forest algorithm is a very popular algorithm that belongs to supervised learning. The term '*Random*' is because of the ***'Randomly created Decision Trees'***.It can be used for Classification as well as Regression problems in ML. It works on a concept of combining multiple classifiers and solves a complex problem easily.It solves the problems of overfitting and underfitting.

Received 24 October 2020; Accepted 09 December 2020.



Working of Random Forest

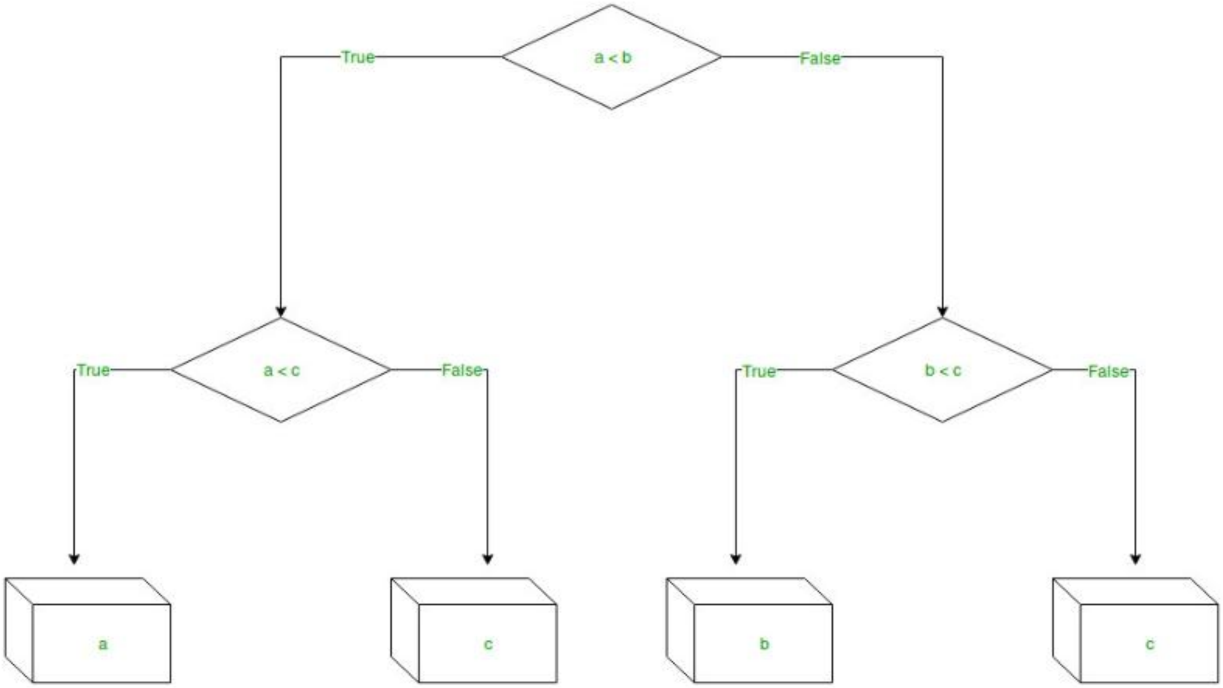
- Pick the value of "k" from the total features,‘m’ by chance, wherek<=m.
- Choosethebestsplitpointandmeasurethe"d"node among the functions "k."
- Perform node splitting into daughter nodes again using the strongest break.
- Repeat the three steps above before hitting "1" number ofnodes.
- Repeatabovefourstepstodevelopforesttoestablish 'n'trees

Decision Tree

Decision tree builds classification or regression models as a tree structure, with datasets broken up into ever-several subsets with developing the decision tree, literally in a tree-like structure way with branches and nodes. It is used to explain the sequence of actions that must be performed to get the desired output. Decision-tree algorithm falls under the category of supervised learning algorithms. It can handle both categorical and numerical data.
The branches represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

In the example below, which shows a decision tree that weighs the smallest of three numbers, the branches reflect the truth/falsity of the argument and make a decision based on that:



Decision tree regression examines an object's characteristics and trains a model in the shape of a tree to forecast future data and generate measurable continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Gradient Boosting

"Boosting" in machine learning is a way of combining multiple simple models into a single composite model. This is also why boosting is known as an additive model. Boosting is a machine learning ensemble meta-algorithm for reducing bias primarily and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. When a weak learner can be implemented efficiently, boosting provides a tool for aggregating such weak hypothesis es to approximate gradually good predictors for larger, and harder to learn, classes. As we combine more and more simple models, the complete final model becomes a stronger predictor. The term "gradient" in "gradient boosting" comes from the fact that the algorithm uses gradient descent to minimize the loss.

Gradient-based learning drawson the fact that it is generally much easier to minimize a reasonably smooth, continuous function than a discrete(combinatorial) function.

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual. This residual predicted by a weak model is added to the existing model input and thus this process nudges the model towards the correct target. Repeating this step again and again improves the overall model prediction.

Data Collection

The main aim is to predict the performance of each player based on their performance in the past matches. Not only the performance of the player is to be predicted but also depending on the previous records and other parameters a decisionhastobepassedwhethertheplayerisideal to be included in the team. In order to achieve a reliable accuracy, we analyzed large amount of data. For this, the initial step was to collect data for all possible matches. Data from the year 2010 to 2020 has been taken from www.cricinfo.com using various scraping tools.Two data sets have been considered for this project. First dataset consists of 132 T20 International cricket matches that India played against various teams. Second data set consists of details about each ball played in these 132 matches that is, there are details of 31040 balls played.

Features used in these data sets include innings, over, venue, batting, bowler, striker, non-striker, runs, out, extras, overtype, match ID, playing against. Further, cleaned data is split into training(80%) and testing data(20%). Training data is given to all the machine learning models and the accuracy of each model is carefully observed. Thesystem makesuseof 'RandomForest', Gradient Boosting, KNN, Linear regression, Decision treealgorithmswhichwerefound to be the best and most adaptable for each parameter used in this paper. Out of these algorithms ultimately one best algorithm for each prediction has been selected that produced most optimal result.

Data Pre-Processing

Statistics such as strike rate and economy are not availabledirectly for each game. Hence such attributes have beencalculated using functions and mathematical formula.These attributes that are generally used tomeasure theperformance of the players are as follows:

Attributes of Batting

Inning**s:**Total number of innings the batsman has battedtill the day of the match. This attribute signifiesthe experience of the batsman. The experience of thebatsman Is determined by the innings of the match.Batting Average: The average number of runs scored perinnings is called batting average. This attributeindicates how reliable the player is.

Strike Rate: The average number of runs scored per 100balls faced is called as strike rate. In T20 cricket, it isimportant to score runs as fast as possible. Runs scored in aslow manner is rather harmful to the team and increases the probability of loosing of the match as there are lesser number of overs. Strike-rate is an attribute that depicts how many runs the batsman can score.

The number of innings in which the batsman scored morethan 100 runs are called centuries. This attribute is nothing but the performance of the player to play long innings and score maximum runs. When the batsman scored more than 50 runs in innings,the score is then called as fifties. The number of innings in which the batsman was dismissed without scoring a single run is called as zeroes. This attribute depicts the number of times a batsman failed to score the runs. Highest Score is the maximum number of runs that a batsman can score in any single innings among all his performances.

Attributes of Bowling

Innings**:** The number of innings during which the bowlerbowled a minimum of one ball. It represents the bowling.

Experience of a player**:** The more innings the player hasplayed, the experienced the player is.

Overs: The number of overs bowled by a bowler.Thisattribute also indicates the experience of the bowler. Themore overs the bowler has bowled, the experienced thebowler is.

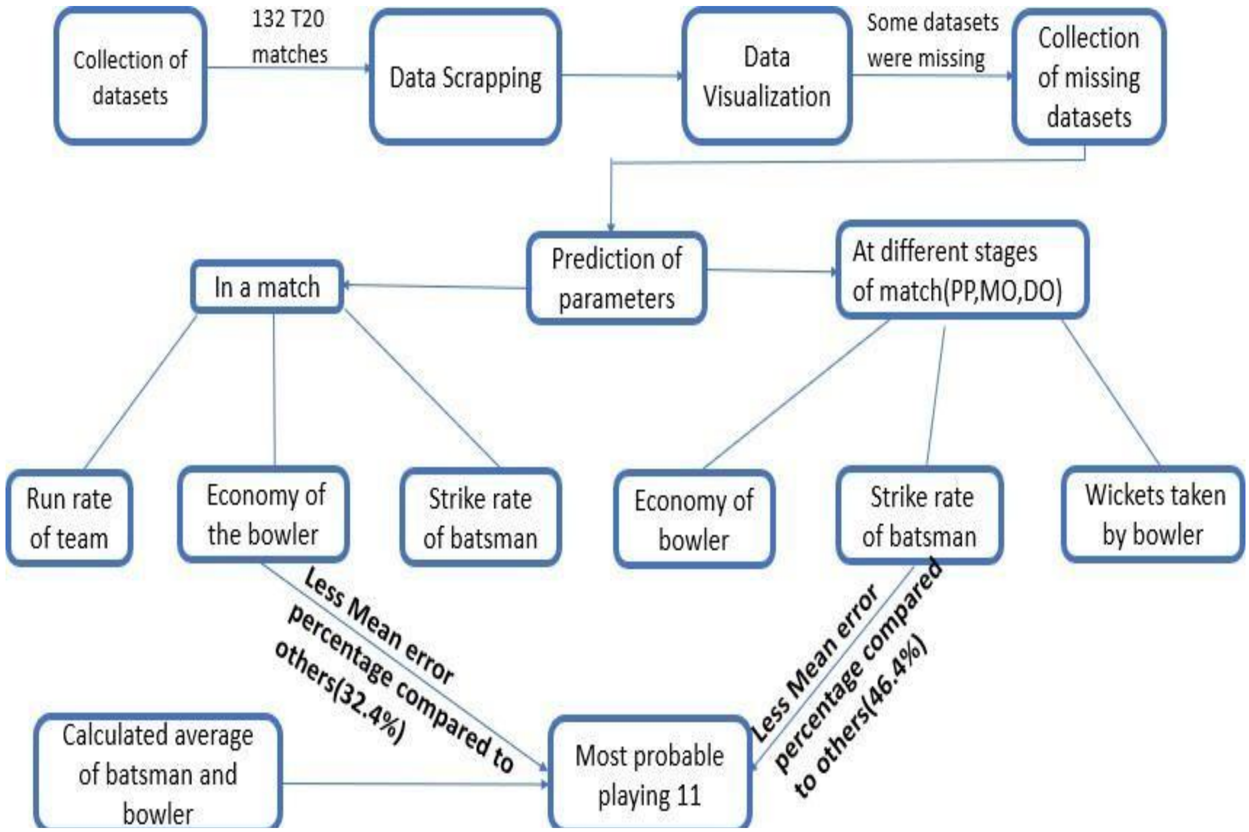
Bowling Average: Bowling average is defined as thenumber of runs stopped by a bowler per wicket. Thisattribute indicates the efficiency of a bowler to concede thebatsmen from scoring runs. Lower values of bowlingaverage indicate more capabilities.

Bowling economy**:** The number of balls bowled per wickettaken is called bowling economy. This attributeindicates the wicket taking capability of the bowler. Lowerthevalues, higher the efficiency of the bowler.

Cleaning

A large number of data that's related to opposition teams were null values. The reason behind this is mainly because a player has not played any match in that particular venue against a particular team. We considered them as values that are missing and these values have been replaced with average of the other corresponding attributes. We also applied a technique to solve the imbalance caused since many of the records fall under class 1. This imbalance effects the performance in greater level. Thus, we used oversampling technique to solve unbalanced data.

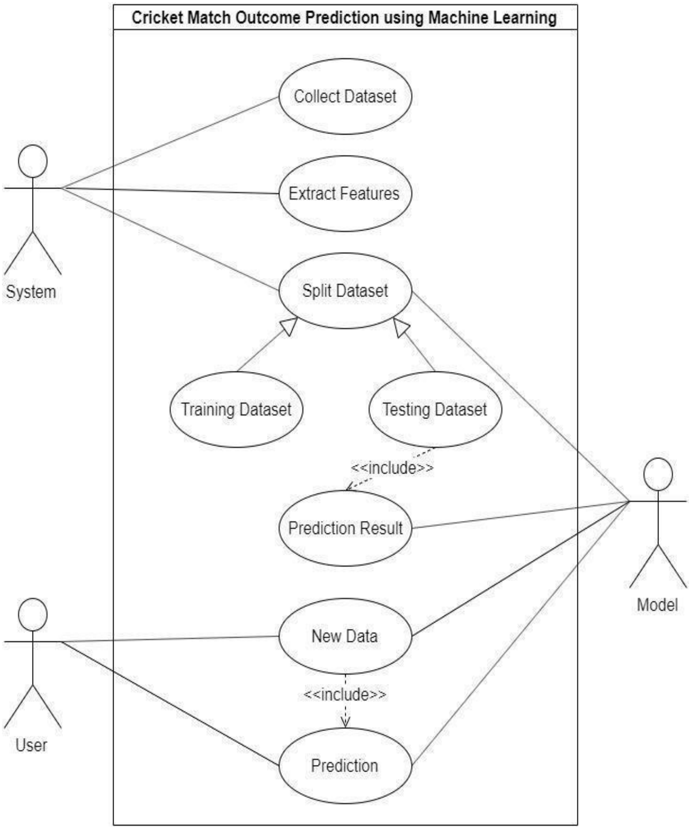
DataFlow Diagram



Use Case Diagram

REGRESSOR	ACCURACY %
KNN regression	42.4%
Linear regression	29.29%
Random Forest	55.5% (Selected algorithm)

Received 24 October 2020; Accepted 09 December 2020.



Results and Discussion

We used different machine learning algorithms to find the best combination that gives the most accuracy. We used five machine learning algorithms: Decision Trees, Random Forest, linear regression, Gradient boosting and KNN in our experiments. The results are tabulated as follows. Table 1 depicts the accuracies of the algorithms for predicting run rate,table 2 depicts the error percentages of the algorithms for predicting strike rate of the batsman, table 3 depicts the error percentage of algorithms for predicting economy of the bowler and table 4 depicts accuracy of the algorithms for predicting how many wickets a bowler can take.

Team Run-rate Prediction

Variables considered for Prediction of run rate

1. Ground Name
2. Playing Against
3. Batting Innings

Here Ground name and Playing against are string values. But for a machine its hard or impossible to learn the string values, So, we assigned an integer value for each string value present in the dataset

Random forest regressor has been used for predicting run rate of the team since it gives highest accuracy.

Batsman’s Strike-rate Prediction

Variables considered for Prediction of strike rate

1. Ground Name
2. Playing Against

3. Match ID
4. Over type
5. Innings
6. Bowler

For a machine its hard or impossible to learn the string values, So, we assigned an integer value for each string value present in the dataset.

REGRESSOR	ERROR-PERCENTAGE
KNN regression	26.46%
Linear regression	23.44%
Random Forest	23.53%
Gradient boosting	23.20% (Selected algorithm)
Decision tree	30.67%

Gradient boosting regressor has been used to predict strike rate of the batsman.This is an ensemble learning method for regression. The percentage error of this algorithm is less when compared to other algorithms because this is a type of decision tree algorithm which does required number of decision tree iterations and tries to decrease the error percentage for each iteration.Whereas in random forest and decision tree algorithm, they do not try to minimize the error percentage for next iteration.

Bowler’s Economy Prediction

Variables Considered for prediction of economy of the bowler.

1. First, we will short list the bowlers with minimum experience required.
2. Bowler Name
3. Ground Name
4. Playing Against
5. Batting Innings

For a machine its hard or impossible to learn the string values, So, we assigned an integer value for each string value present in the dataset.

REGRESSOR	ERROR-PERCENTAGE
KNN regression	24.41%(Selected Algorithm)
Linear regression	31.01%
Random Forest	28.50%
Gradient boosting	29.24%
Decision tree	34.06%

The most suitable algorithm for finding the economy in the respective type of over is KNN regression algorithm with an error percentage of 24.41 which is least when compared to other algorithms.

We can build a regression function that lies within the interval of the training data.

Wickets Prediction

Variables considered for Prediction of wickets:

1. Ground Name
2. Playing Against

3. Striker
4. Over type
5. Innings
6. Bowler
7. Match ID

For a machine its hard or impossible to learn the string values, So, we assigned an integer value for each string value present in the dataset.

Classifier	**Accuracy-Percentage**
KNN classifier	74.96%
Random Forest classifier	71.11%(Selected Algorithm)
Gradient boosting classifier	74.81%
Decision tree classifier	62.66%

In classification problems if values of y are imbalanced that is if there is a single value repeatedly more than 50% of the list length then it is called imbalanced data. For such kind of data some of the classification algorithms will not learn or get trained from the data and gives the value which is repeated the greatest number of times in y as output. This is called Imbalanced Class Behavior. For this the accuracy will be more but always give the same output for any kind of input.

So even though remaining algorithms has greater accuracy, Random forest is chosen as the better algorithm since this does not produce Imbalance Class Behavior.

Best Playing Eleven Prediction

Based on all the above parameters and individual player’s predictions, best eleven players are selected for the upcoming match. This is the ultimate and the most important prediction of the project.

Conclusion and Future Work

Selecting right players in every match is a crucial task for team’s victory. Accurate predictions on number of runs a batsman can score and number of wickets a bowler is can take in a particular match will contribute to the team management in selecting best players for every match. In this paper, we considered match, batting and bowling datasets based on players’ statistics, previous performances and characteristics. Not just these but there are some other features that affect the performance of players like the type of wicket, climate conditions etc. These weren’t considered in this paper due to lack of data in any of the websites. Five regression algorithms were used for predictions and the results were compared. Random Forest turned out to be the most accurate for both the datasets for predicting run-rate of the team. Similarly, gradient boosting turned out best for the prediction of strike-rate, KNN for economy of the bowler and random forest for wickets. Since our paper mainly targets T20 matches, this work can be further extended to all forms of cricket such as test cricket and ODIs. The models other forms of cricket can be further improved to reflect the characteristics of batsman and bowler, for example batsmen will need patience and stamina to play for longer innings in test matches whereas score maximum runs in less overs in T20 matches. Similarly, bowlers will need to have much stronger wicket taking abilities in test cricket and maximum economy rate that is conceding less runs in T20 matches. Moreover, we can attempt to improve accuracies of the regressors for T20 matches.

References

[1] S. Muthuswamy and S. S. Lam, "Bowler Performance Prediction for One-day International Cricket Using Neural Networks," *In Industrial Engineering Research Conference*, 2008.

[2] I. P. Wickramasinghe, "Predicting the performance of batsmen in test cricket," *Journal of Human Sport &Exercise*, vol. 9, no. 4, pp. 744-751, May 2014.

Received 24 October 2020; Accepted 09 December 2020.

[3] G.D.I. Barr and B. S. Kantor, "A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket," *Operational Research Society,* vol. 55, no. 12, pp. 1266-1274, December 2004.

[4] S. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," *Expert Systems with Applications,* vol. 36, pp. 5510-5522, April 2009.

[5] M. G. Jhanwar and V. Pudi, "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),* 2016.

[6] H. H. Lemmer, "The combined bowling rate as a measure of bowling performance in cricket," *South African Journal for Research in Sport, Physical Education and Recreation,* vol. 24, no. 2, pp. 37-44, January 2002.

[7] D. Bhattacharjee and D. G. Pahinkar, "Analysis of Performance of Bowlers using Combined Bowling Rate," *International Journal of Sports Science and Engineering,* vol. 6, no. 3, pp. 1750-9823, 2012.

[8] S. Mukherjee, "Quantifying individual performance in Cricket - A network analysis of batsmen and bowlers," *Physical A: Statistical Mechanics and its Applications,* vol. 393, pp. 624-637, 2014.

[9] P. Shah, "New performance measure in Cricket," *ISOR Journal of Sports and Physical Education,* vol. 4, no. 3, pp. 28-30, 2017.

[10] D. Parker, P. Burns and H. Natarajan, "Player valuations in the Indian Premier League," *Frontier Economics,* vol. 116, October 2008.

[11] I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, Dhaka, Bangladesh, 2018, pp. 500-505, doi: 10.1109/CEEICT.2018.8628118.

[12] A. Kaluarachchi and S. V. Aparna, "CricAI: A classification based tool to predict the outcome in ODI cricket," 2010 *Fifth International Conference on Information and Automation for Sustainability, Colombo,* 2010, pp. 250-255, doi: 10.1109/ICIAFS.2010.5715668.

[13] C.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.

[14] D.M. M. Hatharasinghe and G. Poravi, "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links," 2019 *IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India*, 2019, pp. 1-4. doi: 10.1109/I2CT45611.2019.9033698.