Predicting Match Winner Using Machine Learning

Rushikesh Bhor[1], Ajay Jagdale[2], Shubham Pisal[3], Rohit Korke[4]
[1,2,3,4]Department of Computer Engineering, Skn Sinhgad institute of technology and science kusgaon bk., Pune, India

***Abstract* - In India the game of cricket is not nearly a game but a religion which is followed and loved mustafas as there are multiple format in this game T20 format is most popular one of it is very difficult who is win the game until the last ball is bolde so we pounded why not develop a machine learning model winning a cricket match multiple factor impact the result of match the ground condition past performance records at the venue from of a particular player and team etc. this paper stresses on the key factor impact of the master and suggest regression model which gives the best prediction.**

Keywords: Naïve Bayes Classification, Euler's Strength Formula, Cricket Prediction, Supervised Learning, KNIME Tool, Cricket prediction, sports analytics, multivariate regression, neural network ,Ensemble technique.

***Index Terms* - Application, Architecture, Data, Proposed methodology, use case diagram.**

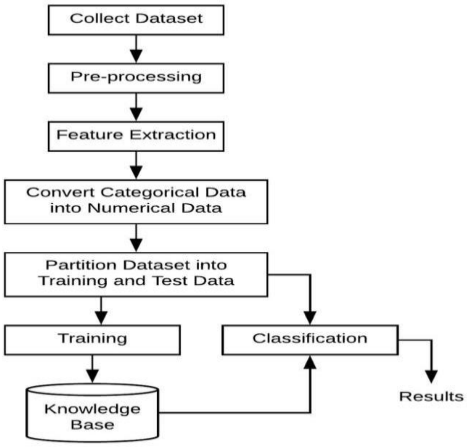
INTRODUCTION

After football, cricket is most loved and watched by many individuals in the world but in India cricket is the most loved sport. In the past few years, lots of research papers are published and lots of work is done which predicts the result of a cricket match by using the factors that affect the match outcome and they are using the supervised machine learning algorithms to predict the outcome of the match like Linear regression, support vector machines, logistic regression, decision tree, Bayes network, random forest. Cricket is one of the most well-liked sports in the world. The Twenty20 format is very popular as it is a fast-paced form of the game that attracts the spectators at the ground and the viewers at home. The Indian Premier League (IPL) is a professional Twenty20 cricket league that is governed by the Board of Control for Cricket in India (BCCI). The Indian Premier League is conducted every year and participating teams represent a city in India. Various natural factors affect the game, the hype given by the media, and a huge market like fantasy 11 and betting on sites has provided a lot of importance to the model. The rules of the game, the skill of the players, their form, and various other natural factors are very important in the prediction accuracy of the result of a cricket match. As the technology is growing and the apps like fantasy 11 and betting sites are getting popular, people are going to use the predictions given by the machine learning model. The use of machine learning makes life easier in many aspects. To predict the outcome of a cricket match we are not going to rely on a single machine learning algorithm we are going to use all the machine learning algorithms. In machine learning, there are two types of learning: supervised learning and unsupervised learning. In Unsupervised learning, the data is not properly labelled so the machine has to sort the data according to patterns, combinations without any training given. But in supervised learning, the data is labelled with the proper classification so the machine can easily analyze it and produce the correct result. For our application, the unsupervised learning models are not of any use because the data of cricket matches are properly labelled. So we are going to use the supervised learning models. In Supervised learning, there are again two types: classification and regression. Classification is used to classify among categories like red or blue and Regression is used when the output is a real number like rupees or height. In our model, we are going to use regression because the outcome will be the winning percentage and it is of type number. Our main objective is to find the key factors that affect the match outcome and select the best machine learning model that best fits this data and gives the best results. Some works have already been published in this area of predicting the outcome of a cricket match. In some papers, only a few key factors are taken for prediction so the accuracy is less. Whereas in some papers the machine learning model is not appropriate. So it is important to take all the key factors that can affect the match outcome and as well as to select the best model for training and testing the data. This will increase the prediction accuracy drastically.

APPLICATION

The main objective of sports prediction is to improve team performance and enhance the chances of winning the game. The value of a win takes on different forms like trickles down to the fans filling the stadium seats, television contracts, fan store merchandise, parking, concessions, sponsorships, enrollment and retention. Any betting app,or online cricket show platform

MODEL ARCHITECTURE



we use one hot encoding for data preprocessing and when an organization has a huge amount of data then we use ensemble technique.

DATA

Real world data is dirty. We can’t expect nicely formatted and clean data as provided by Kaggle. Therefore, data pre-processing is so crucial that I can’t stress enough how important it is. It is the most important stage as it could occupy 40%-70% of the whole workflow, just to clean the data to be fed to your models. I scraped three scripts from Crickbuzz website comprising of rankings of teams as of May 2008, details of the fixtures of 2008 ipl twenty-twenty and details of each team’s history in previous ipl cups match. I stored the above piece of data in two separate csv files. This was done as the results of the last few years should only matter for our predictions. Since I didn’t get the data for 2010 and 2011. Then I did manual cleaning of the data as per my needs to make a machine learning model out of it.

PROPOSED METHODOLOGY

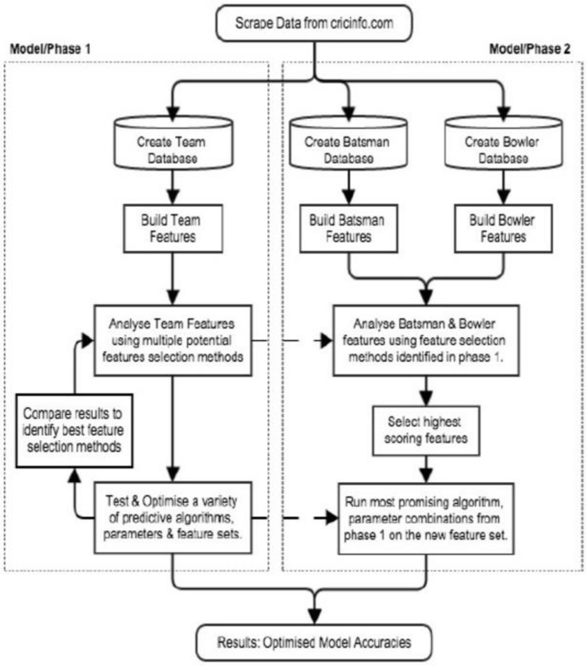

Scrape Data from cricinfo.com
Model/Phase 1
Model/Phase 2
Create Team Database
Create Batsman Database
Create Bowler Database
Build Team Features
Build Batsman Features
Build Bowler Features
Analyse Team Features using multiple potential features selection methods
Analyse Batsman & Bowler features using feature selection methods identified in phase 1.
Compare results to identify best feature selection methods
Select highest scoring features
Test & Optimise a variety of predictive algorithms, parameters & feature sets.
Run most promising algorithm, parameter combinations from phase 1 on the new feature set.
Results: Optimised Model Accuracies

Figure 1. Methodology flowchart

FEATURE ENGINEERING & SELECTION

This is probably the most important part in the machine learning workflow. Since the algorithm is totally dependent on how we feed data into it, feature engineering should be given topmost priority for every machine learning project.

MODEL BUILDING

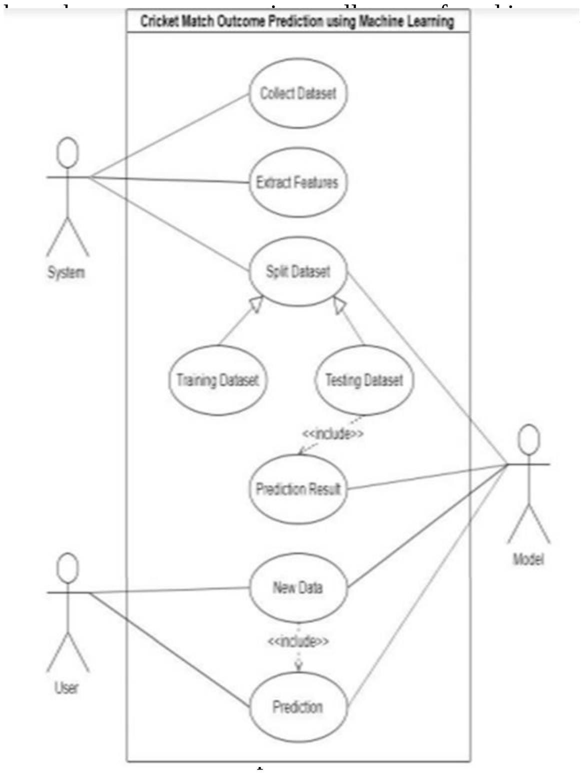
Now that we have processed and cleaned our data, we have to build the machine learning models on this cleaned data. We have implemented SGDRegressor, KNN-Regressor, Linear Regression using LeastSquare Estimates, Weighted KNN-Regressor and compared our models with the Scikit learn’s model and we have achieved almost similar results to that of the Scikit Implementation.

LITERATURE SURVEY

An extensive online search produced very few articles related to players' performance prediction in the game of cricket. A very small number of researchers have studied the performance of cricket players. Muthuswamy and Lam [1] predicted the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They used back propagation network and radial basis network function to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match. Wikramasinghe A. Rabindra Lamsal and Ayesha Choudhary [1] In this paper, they have taken the data of matches from the official website of the Indian Premier League. That data had many features so they analyzed the data and selected some key features. They had used the scikit-learn machine learning library to pre-process the data and applied some selection models. They removed the low variance, univariate and recursive features. By Using these feature selection models they found 5 key features from 15. The features are home team, away team, venue, toss winner, toss decision and winner. They had trained the Random Forests and Multiple Linear Regression model by 10 seasons of IPL data and trained by the 11th season's data. 41 out of 60 matches their model are able to classify correctly. So their accuracy is 68.33% which is not so good. Limitations of this model are it uses only 5 features and only 2 machine learning models. B. Abhishek Naik, Shivanee Pawar, Minakshee Naik, Sahil Mulani [2] This paper processes the data dynamically and gives a prediction as the match progresses. Before the match starts their prediction depends on the factors like batting, bowling, batting order, captain of both the teams and batting-bowling stats on that ground against that opponent and after the match starts their prediction depends on batsman-bowler performance and batting bowling order of particular player. They are predicting only the one day international (ODI) matches by using the logistic regression and K-means clustering. In this paper, they had only tested this model on one match which is India vs Australia which happened on 26th March 2015 at Sydney cricket ground and their prediction was correct. Their predictions can go wrong sometimes because they are fluctuating on every ball. C. Singhvi, Arjun, Ashish Shenoy, Shruthi Racha and Srinivas Tunuguntla [3] In this paper, they have taken 16 features to train the model. Features are like average runs scored by a player, the average number of 4s and 6s hit by a player, average strike rate of a player, number of times the player is not out, numbers of the 50s and 100s scored by the player, total number of matches played by a player, current and average batting position, average number of wickets taken in a match by bowler, average economy and average runs conceded, average number of wide and no-balls bowled and last is average number of maiden overs bowled. They had taken the data of all T20 matches domestic, league matches and international. Many machine learning algorithms are used like Randomizes Forest, Naive Bayes, Decision Trees, Linear SVM, Non-Linear SVM and they are trained by data of 5390 T20 matches. After testing the model the Support Vector Machine given the best accuracy of prediction which is 63.89%. D. Swetha, Saravanan.KN [4] In This paper only briefs about the key factors that cricket match depends on. No machine learning model is trained to predict the match result. The factors discussed in this paper are pitch, toss, and team strength, past records, home ground advantage, current performance, and weather. Pitch plays a very important role in the match because how the ball will behave is totally dependent on it. Toss is also important the teams chasing first wins more matches as the target is known and dew comes in play after evening. By calculating the average of all players and the current form of players we can easily find out the team strength. Past performances play a vital role in prediction, what is the performance of a team on the ground against a particular opponent is very important. If a team is playing on the home ground then crowd support becomes the 12th man of the team also the players are familiar with the playing conditions. The current form of the team and players is also important to predict the winner. The weather condition also affects the swing of the ball and the match outcome. These features can be used to train the model to get better prediction accuracy. E. Geddam Jaishankar Harshit, Rajkumar S [5] This paper compares various supervised machine learning algorithms that can be used to predict the match result. A dataset of 5000 one day international matches is taken from Cricinfo and 70% is used to train the model and 30% is used to test the model. They are using Support Vector Machine, Logistic Regression, Decision Tree and Bayes Classifier as machine learning algorithms. They got 60%, 65%, 67% and 72% respectively. So as we can see the Bayes classifier has the best accuracy among all.

USE CASE DIAGRAM

Use-case diagrams *describe the high-level functions and scope of a system.* These diagrams also identify the interactions between the system and its actors. The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not [illegible]



trained and tested using a dataset. Match dataset contains all the features and labels. Match dataset is split into training and testing datasets.

Predicting IPL-2020 Winner Classification and Regression are the two branches of Supervised Learning in the field of Machine Learning. These are the basic topics that one should learn when starting their journey with Machine Learning. Doing projects is the only way through which one can learn and master these topics

CONCLUSION AND FUTURE WORK

In this paper, we selected 17 key features and 6 machine learning models that give the best possible prediction accuracy. As we can see in the below table all the papers are using a different number of features and different machine learning algorithms. Also, they are targeting different cricket formats. Some papers have only discussed features whereas some papers have discussed which machine learning algorithm will be best. The lowest accuracy is of [3] which is 63.05% and the highest accuracy is of [8] which is 85%. So we analyzed every paper and found all the key factors that increased prediction accuracy and algorithms that predicted with the best accuracy. The highest prediction accuracy is 85% and in our paper, we are getting an accuracy of nearly 90%. By using this model we are going to predict the outcome of twenty 20 matches, one-day international matches, and test matches also. This model can be used for predicting the outcome of other sports also like football, hockey, tennis, baseball, rugby, etc. From the study there are numerous elements which impact the result of any IPL match. Main factors that fundamentally impact any IPL match could be their host group, non-home group, arena, winner of toss and many more. This relatively helped in the calculation of strength. Different ML techniques were handed down for the IPL data set which contributed to this study. The data set consists of all the IPL matches that were held from the past 6 years, that is from 2014 to 2019. The prepared models were utilized to foresee the result of IPL matches. T20 cricket has a scope for changeability, because even a few balls can totally change the game. IPL was started 12 years back, there were very less number of games played compared to 50-50 and test games. Thus, structuring ML for anticipating game results with a precession of 75% is exceptionally good at this stage.

REFERENCE

[1] [1] A. L. Samuel, “Some studies in machine learning using the game of checkers. iirecent progress,” in Computer Games I, pp. 366–400, Springer, 1988.

[2] A. Bandulasiri, “Predicting the winner in one day international cricket,” Journal of Mathematical Sciences & Mathematics Education, vol. 3, no. 1, pp. 6–17, 2008.

[3] Indian Premier League Official Website

[4] P. Langley, W. Iba, K. Thompson, et al., “An analysis of bayesian classifiers,” in Aaai, vol. 90, pp. 223–228, 1992.

[5] S. Kampakis and W. Thomas, “Using machine learning to predict the outcome of English county

twenty over cricket matches," arXiv preprint arXiv:1511.05837, 2015.

[6] L. Passfield and J. G. Hopker, "A mine of information: can sports analytics provide wisdom from your data?," International journal of sports physiology and performance, vol. 12, no. 7, pp. 851–855, 2017.

[7] T. H. Davenport, "What businesses can learn from sports analytics," MIT Sloan Management Review, vol. 55, no. 4, p. 10, 2014.

[8] Muhammad Yasir, LI CHEN, Sabir Ali Shah, Khalid Akbar, M.Umer Sarwar, "Ongoing Match Prediction in T20 International", International Journal of Computer Science and Network Security, Volume: 17 Number: 11 (November 2017)