

# Analytics and Machine Learning Approaches to Generate Insights for Different Sports

S.S. Subashka Ramesh, Nadeem Hassan, Anushka Khandelwal, Ritwiz Kaustoob, Sonal Gupta

**Abstract:** Machine Learning and Data Analytics are used in many sectors so that it can help them to improve their services and find out the future predictions as well by using the previous data. One such sector that has been increasingly using this technology is sports. Many machine learning algorithms are available for sports prediction so that one can determine the team's strength, weakness and predict the future outcome of the game. But these predictions are not always accurate. So the objective of this project is to implement common machine learning and analytics approach so that it can be used to predict the future outcomes of different games such as football, basketball and also generate insights for the same. Instead of using only one algorithm on the dataset of the scores from the previous matches, a series of an algorithm will be applied so that it can compare the final result from each algorithm and provide us with the most accurate result. Algorithms used will be the SVM model, NNR model, Random Forest Algorithm and ANN model. By generating the insights it is possible to not only determine the winner but also the position of the individual players on the field based on their respective performances. This project will thus predict the outcome of the games to a great extent which will help the teams to improve and turn their weaknesses into strength.

**Index Terms:** Machine Learning, Data Analytics, SVM model, NNR model, ANN model, Random Forest Algorithm

## I. INTRODUCTION

The "Machine Learning" is a concept based upon Artificial Intelligence that uses algorithm so that the machines such as the computer can easily learn from the fed data and predict the future result with little or no human intervention. In layman term, Machine Learning refers to an application that uses various methods and algorithms to first observe and understand the data and then look for patterns in the data so that it can be used to predict for the examples fed to make decisions accordingly.

**Revised Manuscript Received on 30 March 2019.**

\* Correspondence Author

**Dr. S. S. Subashka Ramesh\*** Assistant professor (O.G), Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Mr. Nadeem Hassan** Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Ms. Anushka Khandelwal** Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Mr. Ritwiz Kaustoob** Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Ms. Sonal Gupta** Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

These algorithms can be classified into Reinforcement machine learning algorithms, Semi-supervised machine learning algorithms, unsupervised machine learning algorithms, and supervised machine learning algorithms.

For the sports prediction, the algorithms available are Support Vector Machine (SVM) model, Hybrid Fuzzy-SVM model, Linear Regression (LR) model, Neural Network Regression (NNR) model, Artificial Neural Network (ANN) model and many more. Each algorithm has its own unique characteristic that differentiates it from one another. These models also have some drawbacks that may delay the process of generating an accurate result.

The accuracy of these algorithms is based on several factors such as the size of the dataset fed to it; patterns of data predicted by the algorithm and many more [2]. If any of this goes wrong then the entire prediction made by the algorithm goes wrong. So depending upon these conditions different algorithms are suggested to be used in different situations so that higher accuracy result can be obtained. Because in some cases it may be possible that the algorithm can provide accurate result only for small dataset but as the size of the dataset increases, the accuracy of the result decreases so we cannot rely on that algorithm to provide accurate future predictions.

In order to prevent such situations and determine the precise result, multiple algorithms are applied to the dataset of scores from previous matches so that the algorithm can predict the insight for the next upcoming games and also help the team to turn their weakness into strength.

## I. BACKGROUND

The interest of people in sports has increased rapidly in the past few decades. This has led to several measures that have been opted to keep the people interested. One such expedient is sport prediction methods. These sport prediction strategies are used not only by the team players, team managers, fans but also by gamblers.

Earlier when there were no such strategies, the team would simply play the game but were not able to identify their mistakes and weaknesses. Since the mistakes were not identified, the players were not aware of their area of improvement so there was a possibility that they would repeat the same mistakes in the next game as well. This affected the overall performance of the team. This gave rise to sports prediction. Sports prediction uses various machine learning algorithm that takes scores of the previous games as input and analyses the input to look for patterns and learns from this input to predict the output for the next game without any human intervention.



The algorithms used earlier were simple and would only predict the winner of the game with very less accuracy but when more advanced and complex algorithms were implemented it was found that it is possible to predict other features of the game as well such the position of the player on the field in order to improve his performance and a large amount of input can be given to improve the accuracy of the result generated.

## II. RELATED WORK

The existing systems have used different machine learning algorithms for sports prediction which are elaborated below,

Harmandeep Kaur [3], proposed a system where the prediction is done by using Hybrid Fuzzy-SVM model. The Support Vector Machine model is effective and efficient but lacks rule generation capabilities. Hybrid Fuzzy-SVM model gives more accurate result compared to the SVM model result.

Zifan Shi [4], states that attributes are more important than the model and are to be used as an upper limit to predict the quality of the result. Trusting the capabilities of ML techniques, particularly classification learners, to uncover the importance of features and learn their relationships, evaluated a number of different paradigms on this task.

Jaak Uudmae [5], proposed a system that used the Linear Regression (LR) model and Neural Network Regression (NNR) model. The LR approach achieved 64% accuracy in predicting the winning team while the NNR approach achieved 65% accuracy to determine the winner.

Grant Avalon [6], proposed a system that used the Linear Regression (LR) model, Gaussian Discriminant Analysis, Principal Component Analysis coupled with support vector machines, random forest and adaptive boosting.

Rory P. Bunker [7], provided a critical analysis in Machine Learning, focusing on the application of Artificial Neural Network (ANN) to predict sports results. Identified various learning methodologies utilized, data sources, appropriate means of model evaluation, and specific challenges of predicting sports result.

Renato Amorim Torres [8], proposed a system that will predict not only the winner but also tries to classify the position of one player based on his features.

Each of these existing systems has its own advantages and disadvantages. The main idea of this project is to utilize the advantages of these systems to develop a more potentially stable and reliable system that gives a more accurate outcome.

## III. METHODOLOGY

The Game prediction is normally treated as an order issue, with one class (win, lose, or draw) to be anticipated. Even though a few researchers have likewise taken an interest in the numeric prediction issue, where they foresee the triumphant edge – numeric esteem. In the game forecast, vast quantities of highlights can be gathered including the authentic execution of the groups, consequences of matches, and information on players, to enable diverse partners to comprehend the chances of winning or losing approaching matches. The choice of which group is probably going to win is vital on account of the monetary resources engaged with the

wagering procedure; hence bookmakers, fans, and potential bidders are altogether keen on predicting the chances of winning or losing ahead of time. When an anticipated outcome for the match is acquired, an extra issue is whether to place a bet on the match or not, given the bookmaker's chances. Also, sports managers are endeavoring to show fitting methodologies that can function admirably to evaluate the potential rival in a match. In this manner, the test of foreseeing sports results is something that has for quite some time been important to various partners, including the media. The expanding measure of information identified with games that are presently electronically accessible has implied that there has been an expanding enthusiasm for creating clever models and expectation frameworks to estimate the aftereffects of matches.

Traditionally all the ML models use one of the algorithms to predict the result for the match. In the proposed system a set of data is first given as an input to the model and all the algorithms are run to get the result. The algorithm through which we get the most accurate result is then displayed on the system. It will even show the co-relation between different parameters in order to know which parameter is affected the most by which parameter. We can even use parallel computing in order to reduce the time taken by the system that is each system is run with an algorithm simultaneously. These results can further be used to improve the performance of the team in the weak areas.

## IV. PROPOSED SYSTEM

The Fig.1 is a flow diagram that explains the basic components of the system. In the first step we collect the data for the sport we want to predict the outcome to form different resources present on the internet such as NBA or IPL websites or kaggle. Now the data preprocessing comes in which is our second step, as we cannot directly feed the data to the classifier at once as there are several things we need to take care of in order to get the proper accuracy.

First, we need to fill or drop the empty rows, if there are less number of rows which are empty in a larger dataset, we can drop those rows as they will cause a problem in our classification. If there is some column which is having a significant number of empty rows then we need to check the variation of data and fill the mean or median if the entire column into the particular rows as dropping a large number of columns will affect our accuracy score. After dealing with the lost values we encode the data into numeric forms as simple classification model won't understand textual data unless NLP is applied which will increase the time taken for preprocessing the data.

Now the third step is of feature selection which needs to be done very intelligent manner as taking in too many features won't be useful and will unnecessarily increase the time taken to process the data and taking too less feature will decrease the accuracy. The best method employed for this is taking a smaller dataset and checking which rows are actually

affecting the target value which we want to predict and recursively removing the unnecessary columns. The fourth step is feeding the processed data to our models.

Here we are using different classification models and choosing the best accuracy out of it as it may happen that some models are giving better results than others on a specific dataset. Here we have chosen KNN, Decision-tree, Lgboost, random forest and SVM. Since we are trying to make a generalized system here for different sports we tend to create a web application where we will pass the independent labels and the different algorithms as arguments to the model and the model will return the best algorithm suited with the accuracy score. The last part of the application will give the factors affecting the target variable so as the team is able to work on their strength and weaknesses.

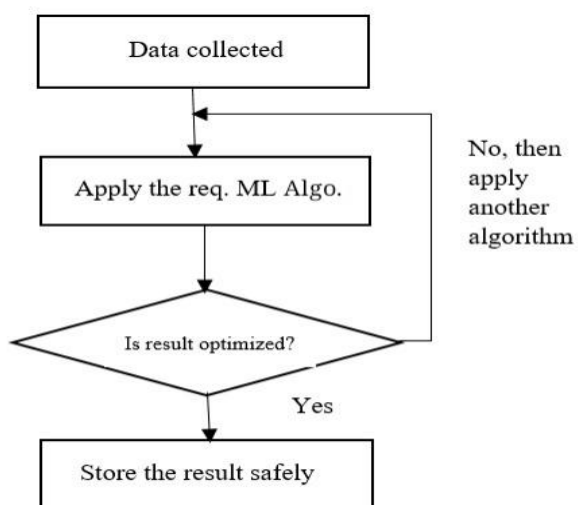


Fig.1: Components of the system

**A. Benefits and Impacts**

The system provides a model which will be more accurate as it won't depend on one algorithm. The process of creating insights will be automated. The system will also be showing the correlations between the independent factors to give the team better insights so that they can create better strategies to win the matches. This data collected by the device will be stored onto the cloud so that the data can be retrieved whenever required. The usage of AES and BRA ensures the data is stored and retrieved from the cloud securely.

**V. SYSTEM ARCHITECTURE**

The system architecture is divided into five major modules. They are:

- Collection of dataset
- Training the machine
- Testing the decision tree
- Linear Gradient Boosting
- Storage of results

The Fig.2 depicts the architecture of the system. The working of the system can be explained as, the system is connected to a high-speed Wi-Fi connection, and the system contains some inbuilt algorithm to process the data fed to it. The Central system uses Linear Gradient processing that converts the data into a simplified dataset and gives it to the

decision tree. The decision tree then trains the data for the individual players and generates an outcome. Once the outcome has been generated, it is tested for efficiency and accuracy. If the outcome is not efficient enough it is again passed to the decision tree. This loop occurs until an efficient outcome has been generated and is then passed as the result. This, in turn, gives the game outcome based on precision.

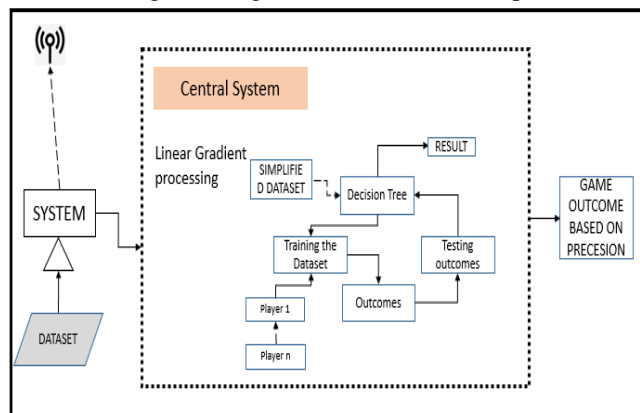


Fig.2: A schematic representation of the system

**A. Collection of Dataset**

This is the very basic module which deals with the collection of a right dataset. The dataset that is to be used has to be filtered on the basis of various aspects. The parameters that might affect the outcome of a game are

- Past records of players in the game and/or in the bench.
- The height, weight and medical test results.
- The weaknesses/health issues of players.
- Running speed and dodging capabilities of players.

The collection of the dataset is based on the acceptability of players or teams and the integrity of data that is being shared. Once the right quality of data is collected, it then can be used to train a machine and to generate predictive results [9].

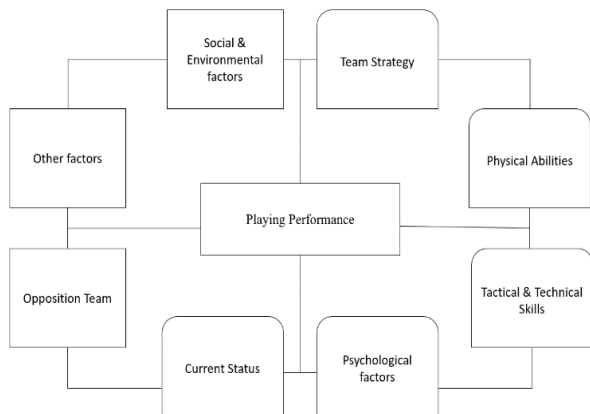
**B. Training the machine**

This module explains how the machine is trained to carry out an operation based on algorithms that are being used. The right amount and quality of data is the foundation of this module. It basically follows the concept of machine learning where the data is fed to the machine and it tries to learn and understand the data and look for patterns which will help it in future predictions [10]. Once the machine has understood the data it can also apply the same to other possibilities. It is used to simplify the dataset so that the data can be collected at an individual level. For this system, the individual level represents the data collected from the past performance of the individual players in the previous games. The machine is fed with relevant data with an algorithm to work upon. The results are thus jotted down for analysis. The generic data is then filtered out to increase the precision of the result.



### C. Testing Decision Tree

This module describes about how we take Decision tree algorithm into consideration. The decision tree takes one data at a time and compares the various parameters based on the calculative measures. Once the data is produced it carries out a comparison between the data available for various levels of tree. [11] Since the accuracy of a Decision tree is poor, we use algorithms like LG Boosting to increase the accuracy of our prediction, impacting the final outcome. The existence of this module is to give a base to our research and produce much accurate result.



**Fig.3: Factors affecting the performance of the player**

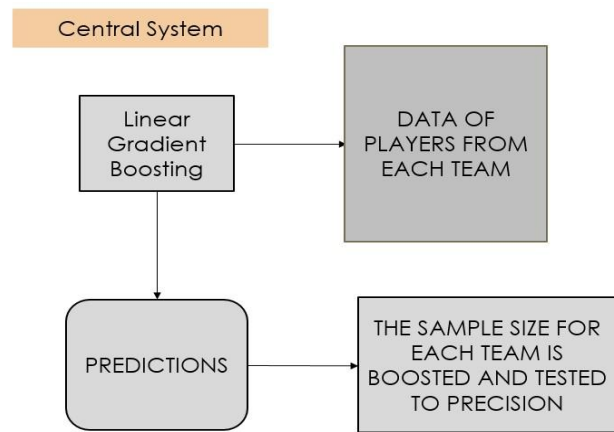
Fig.3. shows the factor affecting the performance of the player. The decision tree tries to consider these factors based on the performance of the individual player in the previous game. The Social & Environmental factors include climate on the day of the match, jetlag faced by the player. The Team Strategy includes tactics, playing system, knowledge about the opposition team. Physical Abilities include speed, strength and flexibility of the player. The tactical and technical skills explain creativity, tackling, anticipation. Psychological factors consist of motivation; will to win and concentration by the player. Current status describes about nutritional, training level of each individual player. Other factors include gender, luck, referee and match type information.

### D. Linear Gradient Boosting

The Linear Gradient Boosting algorithm is being used in this module. [12] The results from decision tree are erroneous and lack precision. Thus, LG Boosting is chosen algorithm which will work on a simplified data such as,

- i. The data is filtered with relevant observations.
- ii. The parameters which are more effective are taken into consideration for filtering.
- iii. The amount of data is decreased based on relevancy.
- iv. The time taken to generate results is lesser since the data is minimal.

Once the simplified data is obtained, it is used to train the machine to carry out predictions.



**Fig.4: Working of Linear Gradient Boosting**

The Fig.4 shows the working of Linear Gradient Boosting. It takes the data of the individual players in each team such as the position of the player on the field, number of goals or runs scored by that player and many more characteristic feature of that player from his previous games. It then applies predictions on this data which is obtained from the Decision tree. The sample size for each team is boosted and tested to precision in order to obtain efficient result.

### E. Storage of Result

The last module describes how the result of the model can help to predict the winning probability of a team based on certain parameters. The result also shows the weak areas of particular team/players. Thus, can help to train the team members better and to collectively decide the area of improvement. This can help save time. The task of this module is to store the information safely, we use a cloud management system and Encryption algorithms to upload and download the data safely. The user authentication system control to make sure only authorized entities is using the data.

## VI. PROCESS FRAMEWORK

Here a web application is built which will take the input of the data and produce the insights. The frontend part of the application will be built through HTML5 and CSS3 which will be used by user to interact with the application. The user can upload their data here in a CSV or TSV format which will go on the flask backend. The users need to select the target column which he wants to predict the backend will consider the other variables as the labels for the target column. The application needs to run on a flask server. [13] The backend will return the response in JSON format which will be processed and shown to the user by the front end part. After making the decisions the application will also show the correlation between all the factors to the user which are affecting the target or in our case the winning or losing of the team which will help the team to work on their strengths and weaknesses. Here the application will also update the CSV files after each match and auto run the analysis to give real time predictions [14].

The project has been tested all the different listed algorithms on a dataset of IPL matches and the different accuracy scores achieved by different algorithms and the best was achieved by Random Forest algorithm which was around 88% on the test dataset.

The Fig.5 shows the framework of the system where in the first level the individual player is observed for which their score for that game is stored onto the dataset. Then in the second level the machine learning algorithms are applied onto these data such that it can be used to predict the outcome for the future games. These predictions are thus stored onto the database so that it can be referred from anywhere at any point. The data stored uses security mechanisms so that only the authorized person can view the data.

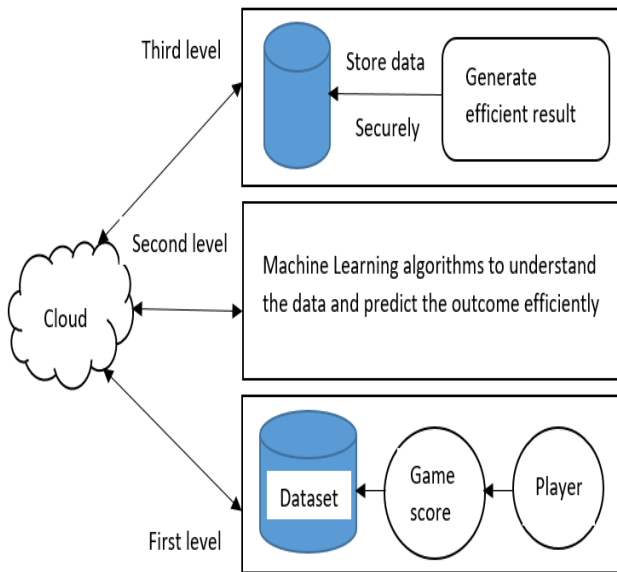


Fig. 5: Framework of the system

VII. MATHEMATICAL ANALYSIS

A step to reduce error and to achieve precision can be explained with the help equation below:

$$Y = M(x) + \text{error}$$

The error is treated as white noise which can have some effect on the outcome(Y).

$$\text{error} = G(x) + \text{error}_a$$

Now each step is to regress the error rate.

$$\text{error}_a = H(x) + \text{error}_b$$

This step can make the precision up to 84.00%.

$$Y = M(x) + G(x) + H(x) + \text{error}_b$$

Combining it together will give us an accuracy of more than 84.00%.

Now we can find the optimal weights for each of the three learners,

$$Y = \alpha * M(x) + \beta * G(x) + \gamma * H(x) + \text{error}_c$$

If the system is able to generate good weights, the probability of a better model with greater accuracy is higher. The principle on which the linear boosting algorithm works is hence explained.

VIII. RESULTS

The Fig.6 shows the probability graph for a team to win the game based on the statistics of winning the toss. It compares two graph where the possibility of a team winning a game by winning the toss is compared to the team winning the game graph.

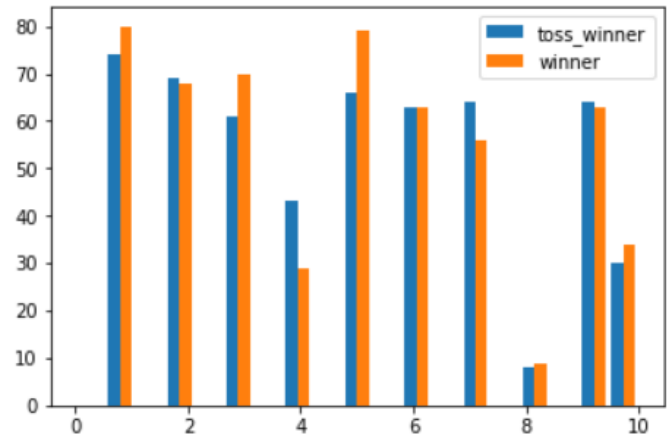


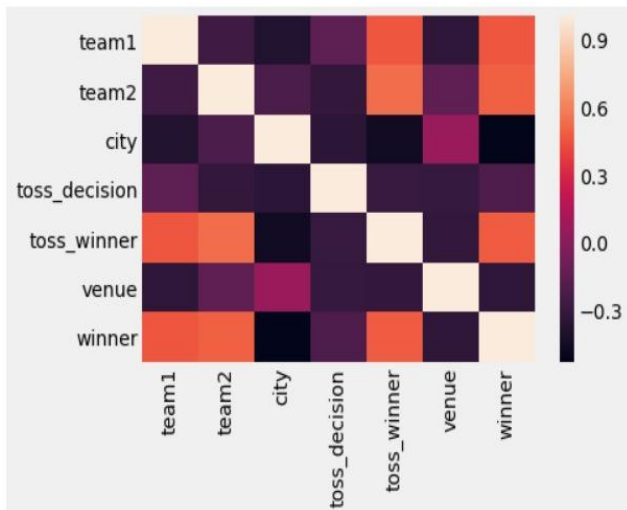
Fig. 6: Probability of match winning by winning toss

The Fig.7 shows a tabular representation of the possibilities for the team to win the game that depends on various factors like city, toss decision, toss winner, venue, winner and many more.

	team1	team2	city	toss_decision	toss_winner	venue	winner
team1	1.000000	-0.084381	-0.046003	0.022750	0.425401	-0.012728	0.412322
team2	-0.084381	1.000000	0.038132	-0.030446	0.480245	0.079119	0.449836
city	-0.046003	0.038132	1.000000	-0.100636	-0.048419	0.095562	-0.084256
toss_decision	0.022750	-0.030446	-0.100636	1.000000	0.008358	-0.052083	0.047102
toss_winner	0.425401	0.480245	-0.048419	0.008358	1.000000	0.037618	0.348551
venue	-0.012728	0.079119	0.095562	-0.052083	0.037618	1.000000	0.029805
winner	0.412322	0.449836	-0.084256	0.047102	0.348551	0.029805	1.000000

Fig.7: Factors affecting the possibility to win the game

The Fig. 8 shows a graphical representation generated that shows the future outcome to predict the winner of the game considering the various factors.



**Fig.8: Graphical Representation generated to predict the winner**

The project provided the prediction for the winner team with an accuracy of 84%. However, this depends on various factors because of which it can be challenging to predict the outcome with such accuracy.

### IX. CONCLUSION AND FUTURE WORKS

Regardless of the expanding utilization of Machine Learning models for the game forecast, progressively precise models are required. This is because of the high amount of betting on the game, and for game directors looking for helpful information for demonstrating future coordinating systems. In this way, Machine Learning appears as a fitting approach for game forecast since it produces prescient models that can anticipate the outcomes of different matches utilizing predefined dataset which is available online.

However, it is difficult to determine the actual outcome of an upcoming game but it is possible to predict the outcome up to a certain extent. There are several factors that affect the outcome of the game and if there is any change in these factors then it will certainly affect the outcome.

With the increasing technology and new methods being developed there is a possibility that a system will be developed that can predict the result with greater accuracy. It can also be possible to predict the characteristics of an individual player that can help their team to win. These characteristics may include the position of the player on the field, minimum score to be scored by an individual player and more.

### REFERENCES

1. Machine Learning Definition URL: <https://www.expertsystem.com/machine-learning-definition/>
2. Ethem Alpaydm. *Introduction to Machine Learning*. Massachusetts Institute of Technology, 2010.
3. Harmandeep Kaur & Sushma Jain, "Machine Learning Approaches to Predict Basketball Game Outcome".
4. Zifan Shi, Sruthi Moorthy & Albrecht Zimmermann, "Predicting NCAAAB match outcomes using ML techniques – some results and lessons learned".
5. Jaak Uudmae, "PredictingNBAGame Outcomes".
6. Grant Avalon, Batuhan Balci, and Jesus Guzman, "Various Machine Learning Approaches to Predicting NBA Score Margins" published in CS 229 Final Project - Autumn 2016.

7. Rory P. Bunker & Fadi Thabtah, "A machine learning framework for sport result prediction".
8. Renato Amorim Torres, "Prediction of NBA games based on machine learning methods".
9. Peter Norvig Fernano Pereira and Alon Halevy. „The Unreasonable Effectiveness of Data“. In: *Journal IEEE Intelligent Systems* 24 Issue 2 (2009), pp. 8–12.
10. Jason Brownlee. How to Use Ensemble Machine Learning Algorithms in Weka. 2016. URL: <http://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka>
11. Niels Landwehr, Mark Hall, and Eibe Frank. „Logistic Model Trees“. In: 95.1-2 (2005), pp. 161–205.
12. Jason Brownlee. Boosting and AdaBoost for Machine Learning. 2016. URL: <http://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning>.
13. Jason Brownlee. Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning. 2016. URL: <http://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning>.
14. Arjun Krishnan. What is the best way to explain the bias-variance trade-off in laymens terms? 2014. URL: <https://www.quora.com/What-is-the-best-way-to-explain-the-bias-variance-trade-off>.

### AUTHORS PROFILE

**Dr. S. S. Subashka Ramesh** Assistant professor (O.G) in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Mr. Nadeem Hassan** IV year, B. Tech student in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Ms. Anushka Khandelwal** IV year, B. Tech student in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Mr. Ritwiz Kaustob** IV year, B. Tech student in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India

**Ms. Sonal Gupta** IV year, B. Tech student in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai -89 (T.N), India