ISSN No:-2456-2165

Quantifying and Analyzing the Performance of Cricket Player using Machine Learning

Dr. Chaitanya Kishore Reddy.M[1], Sk. Arshiya Mobeen[2], P. Mounika Sridevi[3], U. Nithin Kumar[4]
[1]Professor & Dean-IT, [2,3,4]UG Scholar, Dept. Of IT,
NRI Institute of Technology, A.P-521212

Abstract:- In cricket, automation for learning, analyzing, guessing, and predicting is important. As cricket is a sport that is having high demand, no one knows who will win the game until the last over. And there are various factors inclusive of men or women, crew performances, and some diverse environmental elements that need to be taken into consideration in planning a recreation method as a result, we decided to create a machine-learning model to analyze the game using previous match data For this interest, we used a records evaluation and statistical equipment to procedure statistics and bring some suggestions. Implemented models can help selection makers throughout cricket games to test a crew's strengths in competition to the other and environmental elements. Right here we're got used sklearn, preprocessing, and label encoder, and for compilation were got used random woodland classifier set of rules to illustrate the conditions and recommendations for problem fixing We can also predict match outcomes from past experiences by using some algorithms like Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (KNN) are used for classification of match winner and Linear Regression and decision tree for the prediction of an inning's score. The dataset contains huge data on the previous performance of bowlers and batsmen in matches, many Seven features have been identified that can be used for prediction. Based on those features, models are built and evaluated using certain parameters.

Keywords:- *Random Forest Classifier, Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (KNN), NumPy, Data Mining, Analysis.*

I. INTRODUCTION

Cricket is played in a variety of countries around the world. There are lots of tournaments being held in many countries which play cricket. Cricket is a game played between two teams comprising 11 players in each team. The result is either a win, a loss, or a tie. Moreover, the game is also extremely fluctuating because at every stage of the game, the momentum shifts to one of the teams between the two. A lot of times the results get decided on the last balls of the match when the game gets really close. So, keeping all the possibilities, this report aims at studying the problem of analyzing the game results before the game has started based on the data available from the data set. These are different ways to do predictions. The analysis can be done taking into consideration the players' performances as well as the team's performances. This makes the challenge in analyzing the accurate outcome of the cricket match. Sports have gained much importance at both national and international levels. Cricket is one such game, which is marked as the most prominent sport in the world. The suggested analysis model makes use of SVM and KNN to fulfill the objective of the problem stated. Our work novelty is to analyze runs for each ball by keeping the runs scored by the batsman in the previous ball as the observed data and to verify whether our prediction fits into the desired model.

Predictive modeling using data science is increasingly important in the world of sports. One of the well-known sports in India is cricket. Any team, on a given day, has a chance to win the game with its play. This makes it difficult to make an accurate prediction. The result of the cricket game. At the national and international levels, sports have grown significantly in significance. One such game that is recognized as the most popular sport worldwide is cricket. One of the cricket formats recognized by the International Cricket Council is T20 (ICC).

SVM and KNN are used in the suggested prediction model to achieve the problem's stated goal. There haven't been many studies done in this area of cricket match prediction. In our study, we discovered that the work done so far for assessing and forecasting the results of the match is based on data mining. By using the runs the batter scored on the previous ball as the observed data, our approach is new in that we predict runs for each ball and then check to see if our prediction fits the intended model.

II. TECHNOLOGIES USED

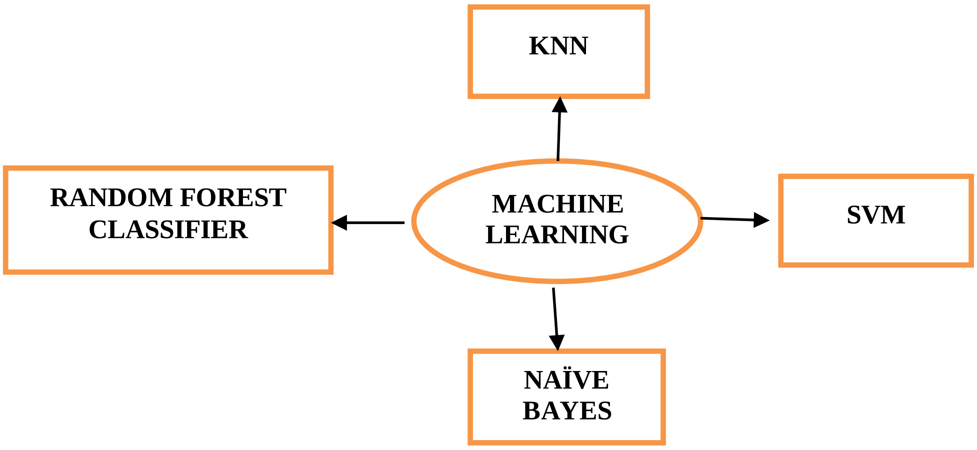

Fig. 1: Machine Learning

A. *MACHINE LEARNING*

Machine learning is the method of choice for categorizing or predicting data to aid people in making important decisions. Algorithms for machine learning are trained on instances or examples, which allows them to study historical data and learn from previous experiences. Building models alone isn't sufficient; you also need to properly tune and optimize the model so that it gives you two accurate outcomes. In order to achieve the best outcomes, improvement strategies entail tweaking the hyperparameters. As it practices using the examples repeatedly, it will spot patterns that allow it to make decisions more precisely[fig:1].

B. *SUPERVISED*

The ideal paradigm for machine learning is supervised learning. Since it is the easiest to understand, it is also the easiest to put into practice. Learning a function that converts an input into an output with the help of example input-output pairs is the challenge at hand. It infers a function from tagged training data made up of a collection of coaching instances. Each example in supervised learning may be a pair made up of an input item, such as Typically, a vector and an output value are used.

- *SVM:*

A supervised machine learning method called SVM (Support Vector Machine) can be utilized to resolve classification and regression issues. However, it is usually used to address categorization issues. The value of each feature corresponds to a certain coordinate in the SVM algorithm, and each piece of data is represented as a point in n-dimensional space (where n is the number of features you have).

A linear model called the Support Vector Machine, or SVM, can be used to address classification and regression problems. It is helpful for a variety of applications and can solve both linear and nonlinear problems. SVM is a fundamental idea:

The approach uses a line or hyperplane to partition the data into classes. It separates the words into categories during the training phase, such as happy, sad, and so on, based on the training dataset, and then predicts the mood of the input song based on the words and the correspondingly score. The top similarity songs' moods are predicted using SVM and arranged in increasing order, after which songs with the same mood are suggested. (fig:2)

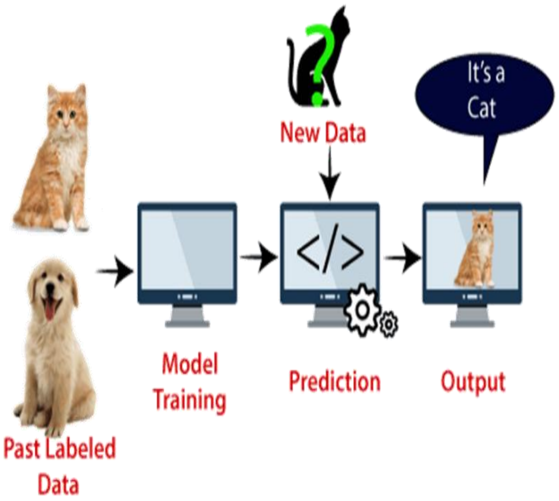

Fig: 2: SVM

- *KNN*

The K-Nearest Neighbor (K-NN) model for recommendations is an item-based strategy that searches for neighbors between objects, in contrast to user-based algorithms that look for neighbors between users. The best model for implementing item-based collaborative filtering and a great place to start when developing a recommendation system is K-Nearest Neighbor. A non-parametric learning technique is the K-NN approach. This technique uses a database with categorized data points to draw conclusions for new samples. K-NN makes no assumptions about the distribution of the underlying data and only relies on the similarity of item attributes. K-NN ranks the "distance" between each item in the database and the target item when it arrives to a decision about an item.

The top K items are then suggested as the most comparable items. The K-Nearest Neighbors method's algorithm is as follows: 2012 (Han et al.)

- Establish the parameter k (number of nearest neighbors).
- Determine the separation between all training data and the data that will be examined.
- Sort the distances created (in ascending order) and find the one that is the closest.
- Include the proper class (c).
- Determine how many classes are closest neighbors, and then identify the class as the data evaluation.

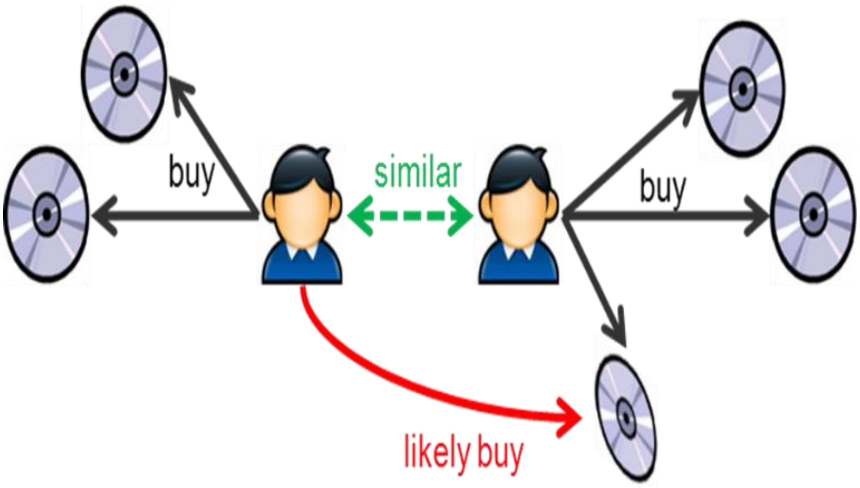

Fig. 3: KNN

➢ *Naive Bayes*

The Naive Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. One of the most straightforward and efficient classification algorithms is the Naive Bayes Classifier, which aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur[fig:4].

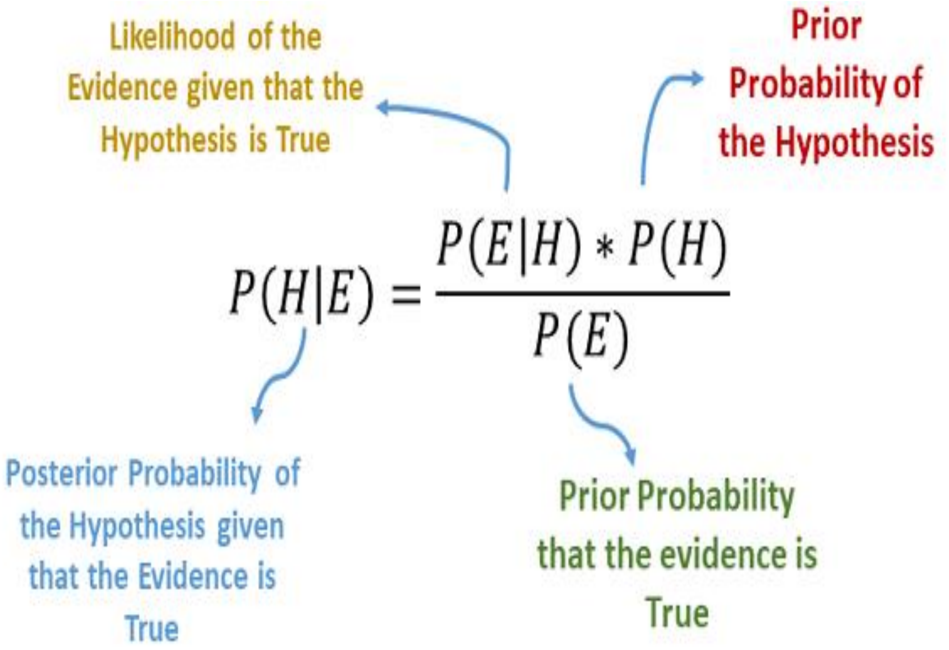
Likelihood of the Evidence given that the Hypothesis is True
Prior Probability of the Hypothesis
$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
Posterior Probability of the Hypothesis given that the Evidence is True
Prior Probability that the evidence is True

Fig. 4: Naive Bayes

➢ *RANDOM FOREST*

The popular machine learning algorithm Random Forest is a part of the supervised literacy methodology. It can be applied to ML issues involving both bracket and retrogression. It's erected on the idea of ensemble literacy, which is a system of integrating colorful classifiers to address delicate issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the handed dataset and pars them to increase the dataset's prophetic delicacy(fig5).

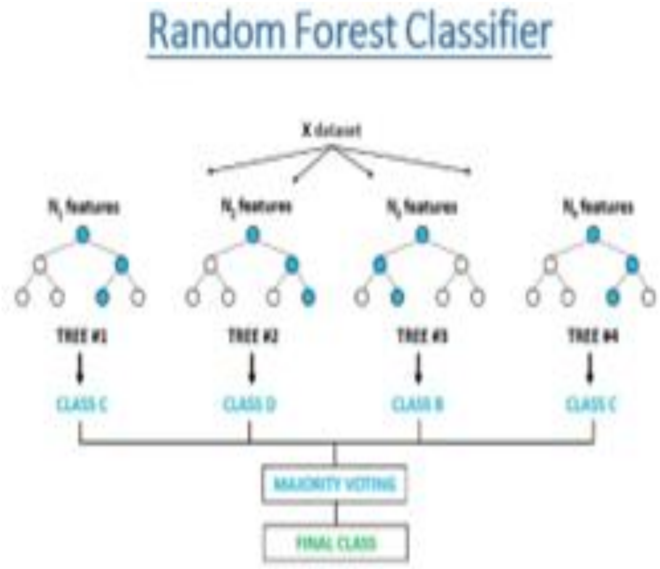

Fig. 5: Random Forest Classifier

C. UNSUPERVISED:

Unsupervised learning is a machine learning technique where supervision of the model is not necessary. Instead, you should allow the model to carry out its own data collection calculations.

It primarily works with unlabeled data and searches for previously undiscovered patterns in a batch of data that has no pre-existing classifications and little to no human oversight.

Unsupervised learning, also known as self-organization, allows for the modeling of probability densities across inputs in contrast to supervised learning, which occasionally uses human-labeled data.

III. SOFTWARE REQUIREMENTS SPECIFICATION

- Python 3.0 or later
- Windows XP, 7, 8, or 10; NumPy,
- pandas, sklearn, and matplot; and an operating system.
- Internet browser (google chrome or Firefox)

A. Introduction to Python

Python is a general-purpose, interpreted programming language. Python's straightforward syntax places a strong emphasis on readability, which lowers system maintenance costs. Modules and packages are supported by Python, which encourages system organization and code reuse. Although it takes a little longer to compile its code, it saves space. While coding, indentation needs to be considered. The following is what Python does:

- Python can connect to database systems and can be used on a server to build web applications.
- Files can also be read and modified by it. Big data management and advanced mathematical operations are both possible with it.
- It can be used to create software that is ready for production.

B. Python libraries

➢ *NumPy:*

A general-purpose package for handling arrays is called NumPy. It offers a multidimensional array object with outstanding speed as well as capabilities for interacting with these arrays. It is the cornerstone Python module for scientific computing. It has a number of characteristics, including the following crucial ones:

- Tools for integrating C/C++ and Fortran code;
- Tools for integrating C/C++ and Fortran code;
- Beyond its apparent applications in science, NumPy also functions well as a multi-dimensional storage container for general data.

➢ *Pandas:*

Built on top of the NumPy library is the open-source Pandas library. It is a Python package that provides a number of data structures and actions for working with time series and numerical data. It is quick and offers consumers exceptional performance & productivity. Python programming language offers high-performance and simple-to-use data structures and data analysis tools. Pandas are utilized in a variety of academic and professional subjects, including economics, statistics, analytics, and other areas.

➢ *Sklearn:*

The most effective and reliable Python machine-learning library is called Skearn (Skit-Learn). It is an open-source Python library that uses a uniform interface to implement various machine learning, pre-processing, cross-validation, and visualization methods. Through a consistent Python interface, Sklearn offers a variety of effective methods for statistical modeling and machine learning, including classification,regression, clustering, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib and was written primarily in Python.

IV. EXISTING SYSTEM

Predictive models are used in historical analysis to identify potential match results. This requires extensive number crunching, data science expertise, visualization tools, and the capacity to include more recent findings. Numerous Python packages that offer higher-level functions related to data analytics are built ona solid foundation called NumPy. These tools are frequently used to get real-time insights that aid in making decisions for outcomes that can change the course of a game, both on the field and in business operations centered around cricket.

A. A viewpoint on utilizing machine learning to analyze match results.

In this study, a novel model is created for forecasting runs using the batsman's prior runs scored as the observable data. A sizable dataset with information on 577 IPL matches was taken into account within dependent and dependent factors in order to complete the objective. RNN and HMM provide the best prediction accuracy for forecasting runs in IPL, according to research presented in this publication. This model is unique in its own right because IPL match results were rarely predicted using machine learning techniques. The created model aids in the analysis and forecasting of IPL match outcomes. Similar work can be arranged for other game types like test cricket, ODI games, and T20 games. The model can be improved further to incorporate crucial elements of numerous other factors that affect the final result of the game, such as weather conditions, injured players, etc.

B. IPL match winner prediction using machine learning techniques.

In sports, and cricket in particular, it can be difficult and confusing to predict the winner. But this may be significantly simplified and made easier by using machine learning. The numerous elements that affect a match’s result in the Indian Premier League were found in this study. The participating teams, the match location, the city, the toss winner, and the toss decision were among the factors that had a substantial impact on an IPL match&#39; s outcome. The points won by each team were calculated using a generic function for the classifier model that took into account team 1, and team 2, the location of the game, the winner of the coin flip, the city, and the result of the toss. On the IPL dataset created for this work, various machine learning algorithms with a classification focus were trained. We employed the approaches of logistic regression, decision trees, random forest, and K-nearest neighbors to arrive at the final evaluation. The Random Forest classifier and Decision Tree offered the highest accuracy of 89.151% among these methods.

C. Utilizing Machine Learning to Predict Cricket Match Results.

The numerous elements that affect a match&#39; s result in the Indian Premier League were found in this study.The home team, the away team, the toss winner, the toss decision, the venue, and the weight of therespective teams are the seven variables that have a substantial impact on the outcome of each IPL match.The points earned by each player are determined by a multivariate regression-based model that takes intoaccount their past performances, which include

- The number of wickets taken
- The number of dot balls given
- The number of ours hit
- The number of sixes hit
- The number of catches
- The number of stumpings.

D. Disadvantages:

Since the accuracy level in predicting or evaluating the data of prior match results has been so low, it has been challenging to anticipate the precise outcome of the match using previous prediction models.

Predicting the average performance of bowlers and batsmen as well as their collective team performance is likewise a difficult issue. To solve the overfitting issue that had previously arisen, we devised a model that would result in greater accuracy.

V. PROPOSED SYSTEM

The entire body of work has been neatly arranged in this architecture. Dataset preparation and loading into the backend are the first steps. The five steps follow are first Data Acquisition, second Data Pre-processing, third Feature Selection, and fourth Training Classification Methods, and finally, Testing Data are the steps followed to get the accurate result. The user interface is then given access to a number of functions for use with players or matches. Predictions may also be made using it.Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest. This model gives more accurate results when compared to existed system. Extra-order categories are also included in this model in predicting the output.

The following modules for analysis, prediction, ranking, and visualizations are implementable.

- Overall group efficiency
- Batman evaluation and ranking
- Batman evaluation and ranking
- Bowler evaluation and ranking
- match evaluation
- team evaluation
- head-to-head evaluation

VI. ADVANTAGES OF THE PROPOSED SYSTEM

To ensure that your predictions are accurate, take into account the following:

- types of players
- The pitch's condition
- injured athletes
- Comparison statistics

Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

VII. SYSTEM ARCHITECTURE

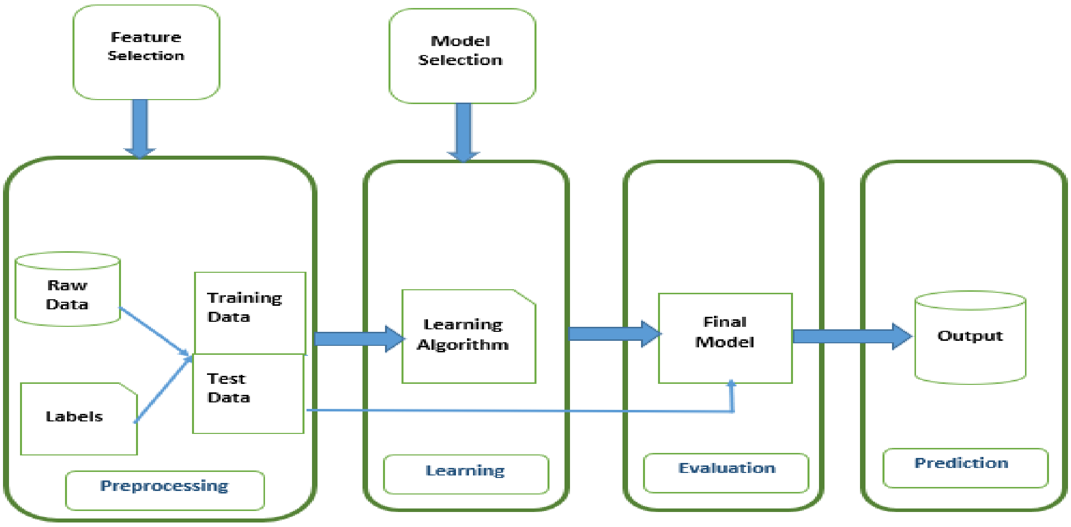

Fig. 6: System architecture

VIII. FUTURE SCOPE

This project can be improved in a number of different ways with future work.

- The data set may contain certain external variables, such as player weariness, player injury, winning streaks with particular teams, overall winning streaks, average runs scored by a team in prior matches against a particular team, etc. We can try to forecast outcomes based on these variables and track how accurate our predictions turn out to be.
- The prediction can take into account the performance of the players in the team, such as the total number of runs scored by a player in the tournament, the player's form guide, the number of men of the match awards earned, etc. in addition to high-level information about the different matches, such as the winner of the toss, the outcome of the toss, the home team, etc.
- There are no web/mobile applications or user interfaces in my project. Therefore, it is conceivable to develop a web application that would receive the entire set of data as input and output the predictions for each occurrence as a pdf or text file.

IX. CONCLUSION

As can be seen in the Results, we used the Random Forest Classifier Algorithm to develop our prediction model and achieved a high accuracy rating. As a proof of concept, we have also included two additional algorithms. The first one is Multinomial Logistic Regression, which was one of the most frequently used algorithms in earlier prediction models, and the second one is AdaBoost, which has not previously been used in cricket match prediction models but is less prone to overfitting and is regarded as a good fit in this model. As can be seen from the data, AdaBoost is a good fit but not the best for our model that attempts to predict the outcome of IPL and T20 matches, whereas Multinomial Logistic Regression is completely unsuited. As a result, every percentage point gain in the model's accuracy would be regarded as extremely crucial. Additionally, we intended to create a model that outperforms earlier iterations of such a model because cricket is a game whose outcome depends on a number of variables. We proceeded with the Random Forest Classifier Prediction Model as a result. On the input data, the Support Vector Machine (SVM), Naive Bayes, and k-Nearest Neighbor (KNN) algorithms are applied to determine the optimum performance. Performance indicators are used to compare these strategies. Support Vector Machine (SVM) provides a higher accuracy score on test data than the other two algorithms, according to the study of the metric.

REFERENCES

[1.] College students’ prevalence and perceptions of text messaging while driving Author links Open overlay PanelMarissa A.Harrison.R.

[2.] Rabindra Lamsal and Ayesha Choudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”, arXiv:1809.09813v5 [stat.AP] 21 Sep 2020.

[3.] Pallavi Tekade, Kunal Markad, Aniket Amage, Bhagwat Natekar, “Cricket Match Outcome Prediction Using Machine Learning”, International Journal Of Advance Scientific Research And Engineering Trends, Volume 5, Issue 7, July 2020, ISSN (Online) 2456-0774.

[4.] Ch Sai Abhishek, Ketaki V Patil, P Yuktha, Meghana K S, MV Sudhamani, “Predictive Analysisof IPL Match Winner using Machine Learning Techniques”, International Journal of InnovativeTechnology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S,December 2019.

[5.] Shubhra Singh, Parmeet Kaur, “IPL Visualization and Prediction Using HBase”, Procedia Computer Science 122 (2017) 910–915.

[6.] Deepak Saraswat, Vijai Dev, Preetvanti Singh" Analyzing the performance of the Indian Cricket Teamusing Weighted Association Rule Mining"2018International Conference on Computing, Power andCommunication Technologies (GUCON) GalgotiasUniversity, Greater Noida, UP, India. Sep 28-29, 2018

[7.] Manuka Maduranga Hatharasinghe, Guhanathan Poravi "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links"2019 5th International Conference for Convergence in Technology (I2CT) Pune, India. Mar 29-31, 2019.

[8.] Indian Premier League - https://www.iplt20.com/, 2020.

[9.] Gagana S, K Paramesha, "A Perspective on Analyzing IPL Match Results using Machine Learning", IJSRD - International Journal for Scientific Research & Development| Vol. 7, Issue 03, 2019 | ISSN (online): 2321-0613.

[10.] https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020.

BIOGRAPHIES:

Dr. Chaitanya Kishore Reddy. M is currently working as a Professor and Dean in the Department of Information Technology at NRI Institute Of Technology, Pothavarappadu, Agiripalli, Krishna(dist.), India. He received Ph.D. in Computer Science and Engineering and M. Tech in Computer Science and Engineering at Jawaharlal Nehru Technological University, Kakinada. He has Published 40 research papers in various National and International Journals and International Conferences. He is a member of ISTE, CSI, and IAENG. His research areas are Mobile Ad-hoc Networks, IoT, and Cloud Computing.

SK. Arshiya Mobeen is currently studying B.Tech with a specification in Information Technology atthe NRI Institute of Technology. She has donea summer internship on Cricket match analysis.

P. Mounika Sridevi is currently studying B.Tech with the specification of Information Technology at NRI Institute of Technology. She has donea summer internship on Cricket match analysis.

U. Nithin Kumar is currently studying B.Tech with a specification in Information Technology at NRI Institute of Technology. He has done a summer internship on Cricket match analysis