Available Online at www.ijcsmc.com/journal

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 5, May 2022, pg.31 – 39

Evaluation and Analysis of Data Mining Techniques to Predict the Winner of Indian Premier League

[1]Ramanjot Kaur; [2]Prof.(Dr.) Gurpreet Singh; [3]Er. Kajal
[1]M.Tech CSE, St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar
[2]Professor, Department of CSE, St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar
[3]Assistant Professor, Department of Computer Sc. & Engg., St. Soldier Institute of Engg. & Technology, Near NIT, Jalandhar

DOI: https://doi.org/10.47760/ijcsmc.2022.v11i05.003

Abstract: Cricket is a popular sport in India as well as the rest of the world. In recent years, the T-20 variation of this game has grown in popularity. The Indian Premier League (IPL), a tournament based on this structure, has risen dramatically in recent years. On the other hand, Cricket is known as the game of chance. Fans and followers are also concerned with predicting the victor of a tournament or match. Technology, on the other hand, is rapidly changing. After training a model, researchers always turn to data mining algorithms to predict something. So, in this research, we use various supervised learning approaches to predict the winners of Indian Premier League matches. We have the following attributes in this system: team names, match venue, toss winner, toss decision, match winner, won by how many runs, and umpires present for the match. Decision trees, Random Tree, Naive Bayes, KNN and ENHALGO are examples of supervised techniques utilized in this research.

Keywords: Decision trees, Random Tree, Naive Bayes, KNN, ENHALGO

I. Introduction

The major goal is to identify the essential parameters that influence the match outcome and to choose the optimal machine learning model that fits the data and produces the best outcomes. Some work has already been published in the domain of forecasting the outcome of a cricket match. Because only a few essential factors are used to forecast in certain papers, the accuracy is lower. However, the machine learning model is incorrect in other studies. As a result, it's crucial

to think about all the important factors that could affect the match's outcome and the best model for training and analyzing the data. This will significantly improve prediction accuracy.

II. PROPOSED SYSTEM

The IPL is one of the most-attended cricket league in the world and it ranks sixth position on all sports leagues. Therefore the proposed system focuses on analyzing the IPL matches results by applying classification algorithm in data mining. Results must be greater than the existing system and accuracy will be compared to the above algorithms. The output will be of calculating the accuracy for the IPL match, after oversampling the imbalanced dataset.

III. DECISION TREE

It is used as supervised tool for field of data mining. It consists of two processes namely training and testing. Interpretation phase of decision tree is composed of deficiencies in neural networks. Therefore tree bagging has been implemented in decision tree process. Bootstraps aggregation is represented as ensemble method of decision trees. Each tree is grown as independent drawn bootstrap of input data. Errors are referred to as “out of bag” which is not included in this replica. Tree bagging takes as leverage of predictions for predicting the unseen data.

IV. NAIVE BAYES

Naive bayes classifier is used where one class is independent of another class. Thus normal and kernel distribution were implemented in our paper.

V. RANDOM TREE

Random tree is uniquely the most prominent machine learning algorithms which consists of multiple decision trees together. Here every single individual tree will explore more deeply about its class predictions and the class with most of the votes, becomes our model’s actual prediction.
Classification type of problem always have a discrete value as the output which are completely different to each other. The main strategy behind random forest is that it divides the whole strategy into multiple trees resulting in various solutions resulting in the most prominent tree path as the final accuracy. This helps in many classification algorithm, to classify various object

depending their behavior. Here the expected prediction error is calculated for every time, this error is also known as test error.

VI. K NEAREST NEIGHBORS

This algorithm is possible for classification as well as regression type of problems, this algorithm is one of the prominent in machine learning since it is a non – parametric way where there won't be any expectations about the distribution of data. In supervised learning KNN is used in powerful application like pattern identification, data mining and intrusion detection. KNN is completely robust, it calculates the distance between the test data and the input and gives the prediction accordingly. One of the equations used for finding the distance between the input and test data.

VII. PROBLEM FORMULATIONS

Cricket is one of the famous outdoor sports that contain a large set of statistical data in real world. As IPL games rise in popularity, it is necessary to examine the possible predictors that affect the outcome of the matches. In this paper, the several years' data of IPL containing the players details, match venue details, teams, ball to ball details, is taken and analyzed to draw various conclusions which help in the improvement of a player's performance. It focuses on measuring the outcome of Indian Premier League (IPL) matches by applying the existing data mining algorithms to the balanced as well as imbalanced dataset.

VIII. OBJECTIVES OF THE STUDY

In this work, I propose a technology based on data mining algorithms for the induction of decision trees. It is well suited in our context for various reasons.

1. To study the literature of data mining algorithms including Naıve Bayes, Random Forest, K-Nearest Neighbour (KNN) .
2. To improve the accuracy of predicting the winner team of IPL matches using enhanced algorithm.
3. To analysis the algorithms and reduce the error rate based on real datasets.

IX. RESEARCH METHODOLOGY

The methodology of the research is easy to understand using a flow chat. Therefore a proper flow chart of our research methodology is provided. The basic strategy defined in this flow chart is as follows:

- Start with the collection of Dataset.
- Load the dataset in the desired WEKA tool by clicking the Preprocess tab in WEKA and then browse the CSV file of our dataset.
- Then tool will analyze the various attributes of dataset whether it is numeric or nominal. In case of numeric attribute the minimum and maximum value of that attribute is defined and in case of nominal attribute the frequency count of the data in that attribute is shown.
- After this the code of the proposed algorithm ENHALGO which is done in Net beans is fetched in the WEKA tool.
- After fetching apply the algorithm from the Classify tab on the processed dataset.
- Check the Classification results and output. Calculate the Correctly Classified Instances, Incorrect Classified Instances and sum of square error.
- Compare the results with the existing algorithms.

X. PROPOSED SYSTEM

The IPL is one of the most-attended cricket league in the world and it ranks sixth position on all sports leagues. Therefore the proposed system focuses on analyzing the IPL matches results by applying classification algorithm in data mining. Results must be greater than the existing system and accuracy will be compared to the above algorithms. The output will be of calculating the accuracy for the IPL match, after oversampling the imbalanced dataset.

XI. FLOW CHART OF PROPOSED TECHNIQUE

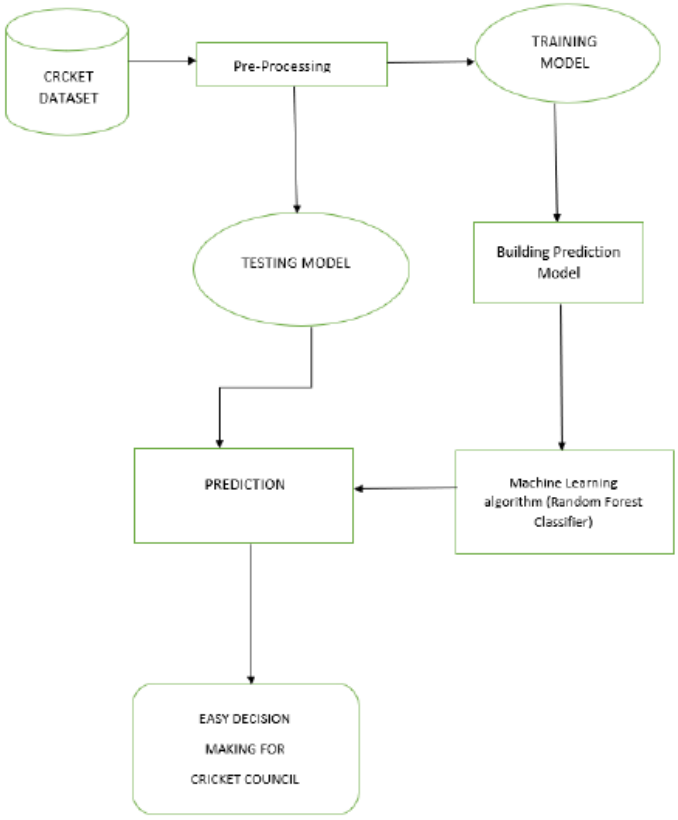

Fig. 1: Flow Chart of Proposed Technique

Results

Table 1: shows comparative results of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms on Correctly Classified Instances

Algorithms	Correctly Classified Instance
KNN	44.64
NBAYES	17.85
ENHALGO	55.35
RANDOM TREE	35.71

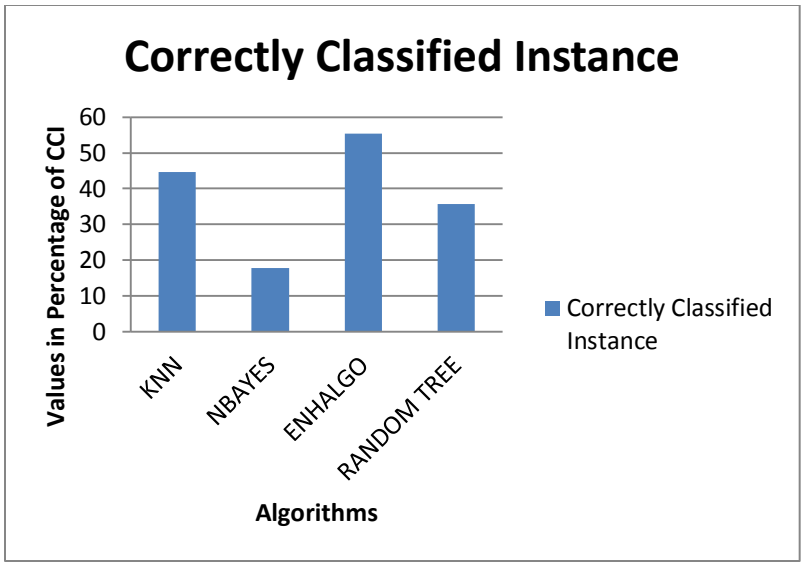

Fig. 2: Shows the percentage of Correctly Classified Instances of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms

Table 2: shows comparative results of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms on Incorrectly Classified Instances

Algorithms	Incorrect Classified Instance
KNN	55.35
NBAYES	82.14
ENHALGO	44.64
RANDOM TREE	64.28

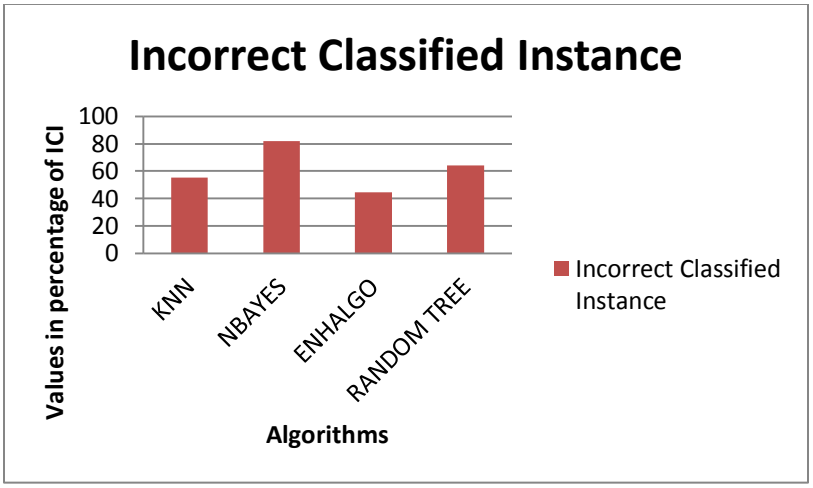

Fig. 3: Shows the percentage of Incorrect Classified Instances of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms

Table 3: shows comparative results of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms on Error Rate

	Error Rate
KNN	90.55
NBAYES	102.26
ENHALGO	85.29
RANDOM TREE	113.44

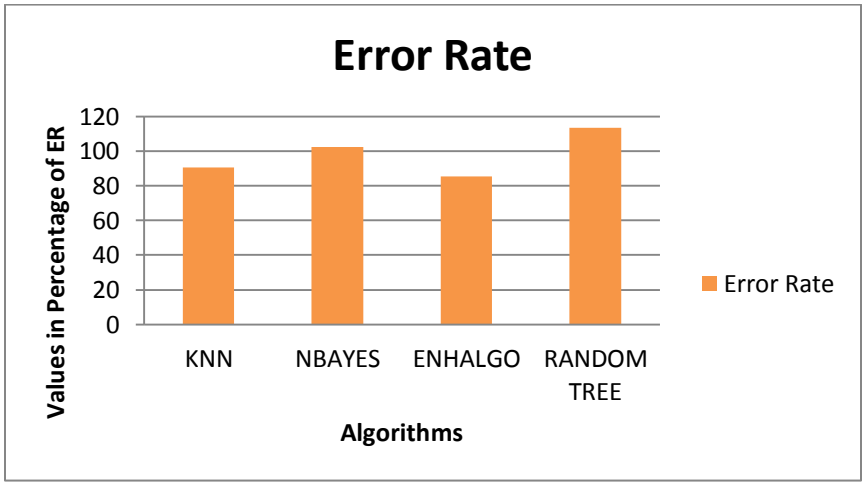

Fig. 4: Shows the percentage of Error Rate of KNN, Naive Bayes, Enhalgo and Random Tree Algorithms

Table 4: Result Analysis of different algorithms with ENHALGO

Algorithms	Correctly Classified Instance	Incorrect Classified Instance	Error Rate
KNN	44.64	55.35	90.55
NBAYES	17.85	82.14	102.26
ENHALGO	55.35	44.64	85.29
RANDOM TREE	35.71	64.28	113.44

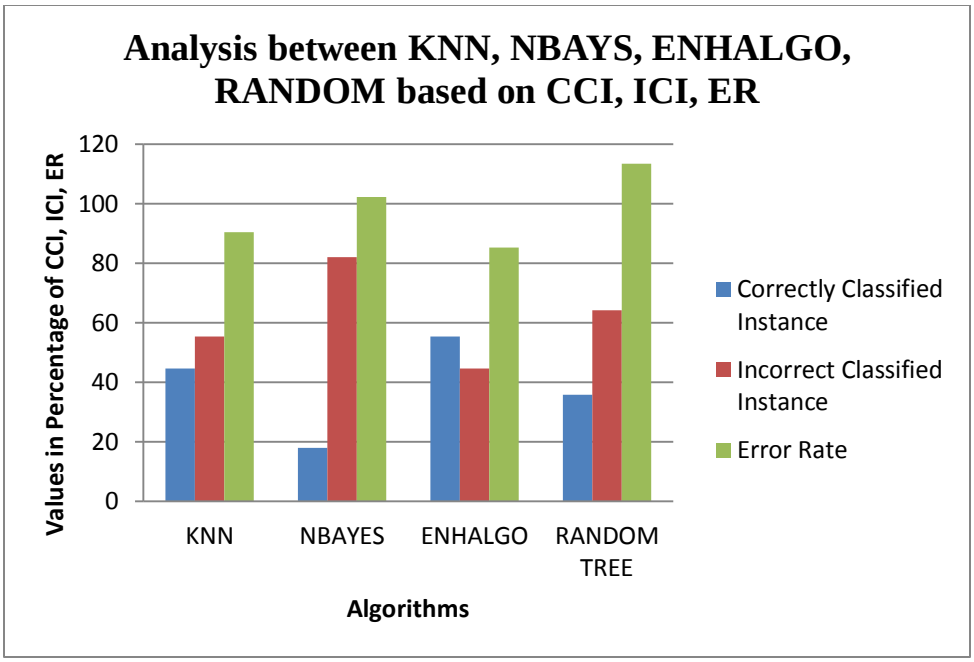

Figure 5: shows the Result Analysis of different algorithms with ENHALGO

CONCLUSIONS AND FUTURE SCOPE

Applying data mining algorithms for analysing cricket sports by considering historical game data, players performance, natural parameters, pre-game conditions and other features is beneficial for multiple stakeholders. In a dynamic format like T20, where the situation in a game changes on every ball, it becomes challenging to predict the match outcome. For predicting the final outcome of a T20 cricket match, we have investigated data mining technology for the possibility of improving the prediction rate of the results of matches. We have formulated the problem in two scenarios, named for the most influential features, firstly the Home Team features set and secondly Toss Winner decision features set. By analysing the results achieved using four different data mining techniques on 2 years' T20 matches, the model built on Toss related features generates slightly better results than Home Advantage in terms of the evaluation measures used (Correctly Classified Instances, Incorrect Classified Instances and Error Rate) etc). Particularly, ENHALGO outperformed the other algorithms when processing the Toss Winner feature set by deriving higher accuracy predictive models than Decision Trees models. On the other hand, the results derived from ENHALGO on Toss Decision subset are not promising due to the class independence assumption of the algorithm. This study is beneficial to team managers and scholars interested in cricket data analytics.

REFERENCES

[1]. Kumash Kapadia, Hussein Abdel-Jaber, Fadi Thabtah, Wael Hadi, "Sport analytics for cricket game results using machine learning: An experimental study", Applied Computing and Informatics Emerald Publishing Limited 2210-8327, 2019.

[2]. Ruchitha M, Dr. Preeti Savant, "IPL Winner Prediction Using Ml Algorithms", International Journal of Engineering Applied Sciences and Technology, 2022, Vol. 6, Issue 9, ISSN No. 2455-2143, Pages 104-107.

[3]. Aman Sahu, Devang Kaushik, A. Meena Priyadharsini, Predictive Analysis of Cricket, Turkish Journal of Computer and Mathematics Education Vol.12 No.6 (2021).

[4]. Miss. Amruta Gujar, Prof. N. G. Pardeshi, "Review On An Sentiment Analysis And Pr. Edicting Winner For Indian Premier League Using Machine Learning Technique", International Research Journal of Modernization in Engineering Technology and Science Volume:02/Issue:06/June-2020.

[5]. Shimona.S, Nivetha.S, "Analyzing IPL Match Results Using Data Mining Algorithms", International Journal of Scientific & Engineering Research Volume 9, Issue 3, March-2018.

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

[7]. Amlan Ghosh, Abhirup Sinha, Pritam Mondal, Anusree Roy and Pritilata Saha, "Indian Premier League Player Selection Model Based on Indian Domestic League Performance",IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021.

