A Score Prediction system for ODI Cricket matches using Machine learning Algorithms

Md Tahsin[1], Anika Tabassum Nafisa[2], Arefin Siddque[3], and Fazle Rifat Anonto[4]
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
[1]ID: 2020-2-60-112, [2]ID: 2019-3-60-072, [3]ID: 2019-3-60-122, [4]ID: 2020-1-60-238

***Abstract*—Cricket is one of the most followed sports in the world with massive fanbase. Among various categories of playing cricket, ODI (One Day Internationals) are played between international cricket teams and audience from different countries enjoys the game. As cricket fanatics, it is a common practice among audiences to anticipate match with outcomes, scores of a particular team or even performance of a player or team based on different factors. Moreover, the rapid increase in cricket fans and followers there are some significant consequences that are based on the outcome or scored runs of a game for stakeholders and other professionals involved. However, human predictions are not always efficient with respect to both resources and time. Therefore, a machine-driven system to predict the score of a match with high accuracy by assessing significant features can proved to be remarkably effective. Hence, this research proposes an automated and proficient score predicting approach using machine learning techniques such as- Linear Regression, Ridge Regression, Random Forest Regression Model, Naïve Bayes Regression model and ElasticNet.Here, Random Forest model provided with the highest R squared value among all the models which is 0.9985 and 0.9921 for train and test respectively. It also had the lowest Mean Squared Error and Root Mean Squared Error value than the other models.**

***Index Terms*—Score prediction , Linear regression, Random Forest regression,Ridge regression, ElasticNet regression, Gaussian Naive Bayes.**

I. Introduction

Cricket, a sport that is combined of bat and ball was originated in England back in the 17th century. However, the sport gained instantaneous popularity in the upcoming centuries among England and its colonies. In the year 1844, United states and Canada played the first ever International Cricket Match [1]. Furthermore, the sport has primarily three different formats named- ODI (One day International match), T20 match and Test match. ODI matches are played in a single day on 50 overs. The second one, T20 or Twenty20 is the newest addition which contains 20 overs. It was developed in 2006 and in the following year, India achieved victory in the inaugural T20 world championship [1]. Another criterion is Test match. The game contains overs ranging from 80 to 90 and the duration of the game is five days. It is named test matches as it tests the skills, endurance and temperament of the players in a five-day long game. Nonetheless, ODI cricket has limited numbers of overs that are played by both the teams. The format was initiated to provide the viewers with a concise and brief alternative to a five-day long test match. ODIs are formulated in such a way that it produces a result- Win or Loss at the end of each game. It is one of the most well-liked and exhilarating versions of the sport that is highly regarded and admired by cricket enthusiasts. Moreover, the game also evaluates a cricketer's strategy and decision-making abilities in a crucial environment. ODI cricket also employs batting and bowling powerplays that enables aggressive batting and scoring. In the Indian sub-continent, the passion and obsession over this sport is unfathomable. People from every age and walks of life take pleasure in watching the games as one. According to the ranking list of teams published by Firstpost, India, Pakistan and Sri Lanka are the top ranked south Asian teams in ODI cricket currently [2]. Due to the rapidly increasing hype of cricket, it is a regular practice for a cricket fan to predict the scores of a match based on different factors that might be affecting the game. However, there are still some scopes for error in mere human calculations. An automated system that is highly efficient and accurate to predict the score of an ODI cricket match before the game starts by analyzing crucial factors that have impacts in a game score can change the scenario for not only the cricket fans and viewers, it will also help the stakeholders and others as well. Machine learning algorithms can play a pivotal role in automation of foreseeing the match result by evaluating necessary features. The key contributions of this study are: • Developing an automated system to predict ODI match scored promptly and skilfully • Minimizing Mean Average Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R squared value for the prediction models. Following the introduction, other studies and researches connected with our study has been discussed briefly on the Related Works section. After that, data collection, data pre-processing, training models are presented in the Methodology section. Accuracy rate and performance evolution of the training models, comparison between the models are discussed in Result and Discussion section. Ultimately the paper is concluded in the Conclusions section.

II. Related Works

Owing to the expanding admiration for cricket, there has been many machine learning methods that has been proposed by researchers to automate both the match outcome and score

prediction system. The research reviewed in this paper can be categorized into two major categories- ODI match score prediction, T20 and IPL (Indian Premier League) match score prediction.

A. ODI match score prediction

Singh and Singla [12] suggested an approach to predict score of an ODI match utilizing factors other than the current run rate. Using Linear Regression method, they predicted score of a match assessing the matches played between 2002 and 2014 for each team individually. Following that in the year 2021, Kamble and Koul [3] proposed a method to predict the score of an ODI match using Random Forest Classifier. The system analyzed the last 5 years data to predict the score of a match. Their used features were- present over, current run scored, no of wickets, runs created by the striker and runs created by the non-striker.

B. T20 and IPL (Indian Premier League) match score prediction

In the same year, another method was suggested for T20 match score and winning team prediction [4]. T20 match results from the year 2005-2021 was analyzed for the method. Later on, Thorat and Buddhivant [9] suggested a method to anticipate IPL (Indian Premier League) match score using Linear Regression, Random Forest Regression and Lasso Regression. Best performance was acquired by Linear Regression model with a Mean Average Error (MAE)- 12.114, Root Mean Square Error (RMSE)- 15.864 and R squared value- 0.752. However, in [5], Linear Regression Method was used for the same purpose. Mean Average Error (MAE) and Root Mean Square Error (RMSE) of the model was 12.118 and 15.8432 respectively considering 15 features altogether while developing the system. Similar results were also found in a research conducted in 2022 by Ahmed and Sareen [10]. Their technique used both Linear Regression and Ridge Regression where Linear Regression presented the best outcome. Afterwards, cricket score prediction system for IPL score and win were evaluated by Mozar and More [6] in 2022. Lasso Regression algorithm was used to predict the match score. Nevertheless, another IPl score predicting system was proposed using algorithms like KNN Regressor, Linear Regressor, Decision Tree Regressor, Random Forest Regressor and lastly XGBoost Regressor concurrently [8]. In the data pre-processing step, outdated teams were not considered to increase model's efficiency. Among the algorithms implemented, XGBoost showed the best performance with Mean Average Error (MAE)- 16.60, Mean Squared Error (MSE)- 467.50, Root Mean Squared Error (RMSE)- 21.67 and Mean Squared Log Error (MSLE)- 0.02. During the same year, IPL data from 2008-2019 was analyzed to present a method to predict the match outcome using Support Vector Machine, Decision Tree and Random Forest model [7]. All models provided with almost 88However, using a deep learning method for IPL score prediction, achieved Mean Squared Error (MSE) was 9.315 and 11.857 for train and test data respectively [11]. The model implemented two hidden layers to reduce the number of neurons.

III. Methodology

In order to ensure an efficient score predicting model, the data must go through certain steps like data cleaning and data pre-processing. Afterwards, based on the features extracted from the data different machine learning algorithms are employed and lastly the outcome from each algorithm is evaluated.

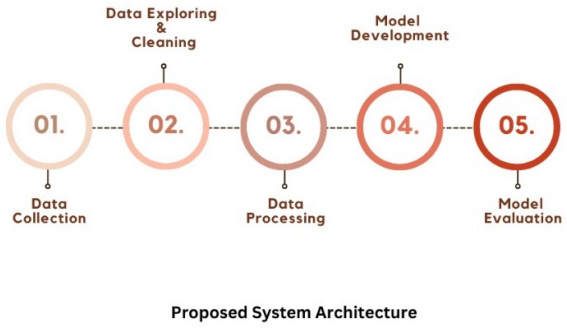

Fig. 1. System work Flow Chart

A. Data Collection

The dataset for this research was collected from GitHub. It contained the ODI match information of 21 countries between the year 2006 and 2017. It consists of 350899 rows and 15 columns. The following columns are included: • Team: The bowling and batting teams. • Venue: The location where the game was played. • Batsman: Name of the Batsman • Bowler: Name of the Batsman • Overs: The quantity of bowled overs. • Runs: The total number of runs the batting team has scored. • Striker and non-Striker: • Wickets: The total number of wickets the batting team has lost. • Runs in last 5 overs • Wicket in last 5 overs • Mid: Match id • Date: match date • Total: final score

B. Data pre-processing

To train and test the model features named: runs, wicketsLeft, ballsLeft, runsLast5overs, year, batting team, bowling team, run rate have been taken.

(i)Initially the dataset was assessed for null values. However, the data set used did not contain any null values. (ii)From the "date" column year was extracted a feature. (iii)Only the consistent teams playing ODI cricket match was considered for both batting and bowling team. (iv) after counting the number of venues, the top venues(for feature) will be updated the data frame. (iv)From the existing columns, features like "wicketLeft", "ballLeft" and "runrate" were calculated and added as features. (v)One hot encoding method was applied for columns- 'batTeam', 'bowlTeam', 'venue', 'runrate', 'batsman' and 'bowler' to convert the value to numeric. Finally,

TABLE I
RELATED WORK TABLE

Reference	Title	Year	Features/Data Used	Models
[3]	Cricket Score Prediction Using Machine Learning	2021	present over, current runs scored, no of wickets, runs created by the striker, runs created by the non-striker	Naive Bayes
[4]	T20 cricket match score and winning team prediction using machine learning techniques	2021	Team Name 1, Team Name 2, Venue, Toss, Decision and Time, Pitch Conditions, Temperature, Humidity and Precipitation, team1 batting average, team2 batting average, team1 bowling average, team2 bowling average, team1 economy and team2 economy	Random Forest algorithm, Decision Tree algorithm
[5]	Building an IPL Score Predictor – End-To-End ML Project	2021	15 different features	Linear Regression
[7]	Super Predictor of Indian Premier League (IPL) using Various ML techniques with help of IBM Cloud	2022	IPL data from 2008 to 2019 is used for the player analysis	Support Vector Machine, Decision Tree and Random Forest
[8]	IPL Score Prediction Using Machine Learning	2021	IPl scores dataset with 15 different features	Linear Regression
[9]	CRICKET SCORE PREDICTION	2021	runs scored, overs bowled, wickets taken	Linear regression , Random forest regression , Lasso regression
[10]	First Inning Score Prediction of an IPL Match Using Machine Learning	2022	IPL dataset taken from Kaggle	Linear Regression, Ridge Regression
[11]	Score and winning prediction in cricket through data mining	2015	non-curtailed ODI matches played between 2002 and 2014 of every team independently	Linear Regression

the features that are utilized for model training and testing are- runs, wicketsLeft, ballsLeft, runsLast5, year,batTeam, bowlTeam, batsman, bowler,venue, runrate, total(target). The subsequent action is to train a machine learning model using the chosen features. For ODI run prediction, this research focuses on employing machine learning algorithms named as-Linear regression, Polynomial regression, Random Forest and Naive Bayes. The data was split into 80-20 ratio for training and testing purpose. Objective of the model is to efficiently predict score of the game after 50 overs.

C. Training Models

1) Linear Regression : The "linear regression" statistical method uses one or more independent variables to predict continuous target variables. It determines the best-fit regression line to lessen the difference between the projected and actual values. The correlation between two variables are measured using linear regression. Using this modelling technique, a dependent variable is predicted based on one or more independent variables. [13] Formula for Simple Linear regression is,

yi(Predicted)=$\sum_{i=0}^{n}(\beta$xi) + ϵ

here, b = Slope of the line, a = Y-intercept of the line , xi = Independent variables, yi = Dependent variable, β= Slope coefficient for each independent variable, $\epsilon = Model'serrorterm.$

(ii) Ridge Regression: The technique used to analyze multicollinearity in multiple regression data is called ridge regression. It works best when there are more predictor variables in a data collection than there are observations. Ridge Regression also known as L2 regularization is a method employed on the linear regression approach in order to combat multicollinearity and overfitting. In real applications, the ridge estimator is a linear combination of the coefficient of least squares regression of the explanatory variables. [14]

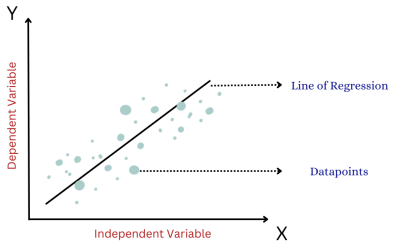

Fig. 2. linear regression

2) Gaussian Naive Bayes: The naive Bayes learning scheme, despite being straightforward, excels at most classification problems and is frequently noticeably more accurate than more complex models. Even though the probabilities it generates can be off, it frequently gives the right class the highest probability. Implying that, it may only perform well in scenarios when the output is categorical. However, forecasts in domains where a numeric value is expected are more suscepti-

ble to incorrect probability estimations, it will be fascinating to observe how it performs in these domains.[15] Every potential value inside the target range is given a probability through Naive Bayes. After that, the resultant distribution is reduced to only one forecast.

3) Random forest regression: RDF works by constructing a multitude of trees during training, each using a random subset of data and features. Each tree in a random forest depends on the values of a random vector that was sampled randomly and with the same distribution for all the trees in the forest. As the number of trees in a forest increases, the generalization error converges to a limit. The strength of each individual tree in the forest and the correlation between them determine the generalization error of a forest of tree classifiers. [16]

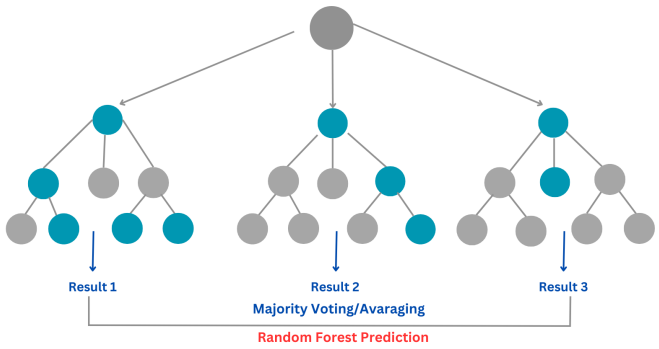

Fig. 3. Random Forest Prediction

4) ElasticNet: ElasticNet, a regression model combines L1 (Lasso) and L2 (Ridge) regularization methods. It operates by including a penalty component in the equation for linear regression, which promotes feature selection (Lasso) and decreasing coefficient magnitudes (Ridge). According to real-world data and a simulated study, the elastic net frequently outperforms the lasso while enjoying a similar degree of sparsity of representation. Additionally, the elastic net promotes a grouping effect, where strongly linked predictors are more likely to be included in the model or excluded from it as a group. When there are many more predictors (p) than observations (n), the elastic net is especially helpful. In the p n instance, however, the lasso is not a very effective tool for variable selection.[17]

IV. Result and Analysis

Here the implemented models are evaluated based on Train and Test Accuracy, Mean Squared Error, Mean Average Error, R squared Value and Root Mean Squared Value. Among the employed models, Here's a brief interpretation of these metrics: Train Accuracy and Test Accuracy indicate how well the model performs on the training and test datasets, respectively. Higher values are better. Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values. Lower MSE values indicate better model performance. R-squared is a measure of how well the model explains the variance in the data. Values closer to 1 are better, indicating that the model explains a large portion of the variance. Mean Absolute Error (MAE) measures the average absolute difference between the predicted and actual values. Lower MAE values are better. Root Mean Squared Error (RMSE) is the square root of the MSE and provides a similar measure of prediction error. Lower RMSE values are better. Linear Regression:

TABLE II
Linear Regression

Train Accuracy	Test Accuracy	MSE	R^2	MAE	RMSE
0.8423	0.8401	527.9906	0.8401	16.2880	22.97804

Train Accuracy and Test Accuracy are quite close, indicating that the model is not significantly overfitting or underfitting.The R-squared ($R\hat{2}$ value of 0.8401 suggests that the model explains approximately 84% of the variance in the target variable.The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values are moderate, indicating a reasonable predictive performance.

TABLE III
Ridge Regression

Train Accuracy	Test Accuracy	MSE	R^2	MAE	RMSE
0.8336	0.8369	549.2157	0.8336	16.8834	23.4353

Similar to Linear Regression, Train and Test Accuracy are close, suggesting a balanced model. The R-squared value (0.8336) is slightly lower than Linear Regression, indicating a slightly worse fit. The MAE and RMSE values are slightly higher compared to Linear Regression, suggesting slightly worse predictive accuracy.

TABLE IV
Gaussian Naive Bayes

Train Accuracy	Test Accuracy	MSE	R^2	MAE	RMSE
0.8933	0.8818	549.2157	0.8336	16.8834	23.4353

The model shows the highest Test Accuracy (0.8818) among the models, indicating good classification performance. However, it's important to note that this model seems to have the same MSE, $R\hat{2}$ MAE, and RMSE values as Ridge Regression, which suggests it might be better suited for classification tasks rather than regression tasks.

TABLE V
Random Forest

Train Accuracy	Test Accuracy	MSE	R^2	MAE	RMSE
0.8351	0.81632	690.0888	Train: 0.9977 Test: 0.98750	19.685	26.2695415

Train Accuracy is very high (0.8351), which could be a sign of overfitting, but the Test Accuracy (0.81632) is also reasonably high. The model's R-squared values for both train and test datasets are exceptionally high (0.9985 and

0.9921, respectively), indicating that the model explains almost all of the variance in the target variable. The MAE and RMSE values are significantly lower than the other models, suggesting that Random Forest provides the best predictive performance in terms of regression accuracy. Train Accuracy

TABLE VI
ELASTICNET

Train Accuracy	Test Accuracy	MSE	R^2	MAE	RMSE
0.8351	0.81632	25.7645	Train: 0.9985 Test: 0.9921	1.5291	5.0758

and Test Accuracy are reasonably high, indicating that the model is performing well in terms of classification accuracy. The R-squared values for both train and test datasets are exceptionally high (Train: 0.9985, Test: 0.9921). These high R-squared values suggest that the ElasticNet model explains almost all of the variance in the target variable, which is a strong indicator of a well-fitted model for regression tasks. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values are significantly lower than the other models, indicating that ElasticNet provides very accurate predictions with low prediction errors.

V. CONCLUSION

Primary objective of the research was to develop an automated score predictor for ODI cricket matches. With the aim of doing so, match information of previous 11years from 2006 to 2017 is evaluated using various techniques. Various branches of data science have been employed for data preprocessing, data visualization, data preparation and lastly feature extraction. Furthermore, machine learning techniques are implemented to ensure accurate and efficient score prediction. In this study, highest accuracy in score prediction with lowest error was achieved with Random Forest Model. However, the proposed method can be more accurate by considering features like pitch quality, humidity and temperature.

VI. REFERENCES

1) *History of cricket*. Wikipedia, The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/wiki/Historyofcricket. [Accessed: September 17, 2023].
2) Firstpost. "Cricket Teams Ranking." 2023. https://www.firstpost.com/firstcricket/teams-ranking.
3) R. R. Kamble, N. Koul, K. Adhav, A. Dixit, and R. Pakhare, "Cricket Score Prediction Using Machine Learning," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 1S, 2021, pp. 23-28, doi: file:///D:/projecttest/1546-Article
4) K.A.D.A Pramoda, "T20 cricket match score and winning team prediction using machine learning techniques," *UCSC*, doi: https://dl.ucsc.cmb.ac.lk/jspui/bitstream/123456789.pdf.
5) Analytics Vidhya, "Building an IPL Score Predictor: End-to-End ML Project," Oct. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/building-an-ipl-score-predictor-end-to-end-ml-project/. [Accessed: September 17, 2023].
6) Omkar Mozar, Soham More, Shubham Nagare, Prof. Nileema Pathak, "Cricket Score and Winning Prediction," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 4, 2022, pp. 1031-1022, doi: https://www.irjet.net/archives/V9/i4/IRJET-V9I4181.pdf.
7) K Rushikesh Reddy, Prem Sai A, Chandrakanth V, Shiva Sumanth Reddy, "Super Predictor of Indian Premier League (IPL) using Various ML techniques with help of IBM Cloud," *IJRASET*, 2022, doi: https://doi.org/10.22214/ijraset.2022.43654.
8) https://medium.com/mlearning-ai/ipl-score-prediction-using-machine-learning-f6587e353532.
9) Prasad Thorat, Vighnesh Buddhivant, Yash Sahane, "CRICKET SCORE PREDICTION," *International Journal Of Creative Research Thoughts*, vol. 9, 2021, pp. g169-g175, doi: https://ijcrt.org/papers/IJCRT2105677.pdf.
10) Raja Ahmed, Prince Sareen, Vikram Kumar, Rachna Jain, Preeti Nagrathe, Ashish Gupta, Sunil Kumar Chawla, "First Inning Score Prediction of an IPL Match Using Machine Learning," *AIP Conference Proceedings*, 2555, doi: https://doi.org/10.1063/5.0108928.
11) https://www.geeksforgeeks.org/ipl-score-prediction-using-deep-learning/.
12) Score and winning prediction in cricket through data mining (Tejinder Singh; Vishal Singla; Parteek Bhatia).
13) Khushbu Kumari, Suniti Yadav, Department of Anthropology, University of Delhi, New Delhi, India.Linear Regression Analysis Study (20).
14) Mohamad Shariff, Nurul Sima & Duzan, H., "An Application of Proposed Ridge Regression Methods to Real Data Problem," *International Journal of Engineering & Technology*, vol.7, pp. 106-108, doi: https://www.researchgate.net/publication/332676351.
15) Eibe Frank, Len Trigg, Holme (Naive Bayes for Regression), https://www.researchgate.net/publication/2360319.
16) Eibe Frank, Len Trigg, Geoffrey Holmes, and Ian Witten, "Naive Bayes for Regression," Machine Learning, vol. 32, no. 3, 1998, pp. 267-286, doi: 10.1023/A:1007465528199.
17) Hui Zou, Trevor Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Volume 67, Issue 2, April 2005, Pages 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.