VFAST Transactions on Mathematics
http://vfast.org/journals/index.php/VTM@ 2025 ISSN(e): 2309-0022, ISSN(p): 2411-6343
Volume 13, Number 2, July-December, 2025 pp: 55-74

Predictive Modeling of T20 Opener Success in International Cricket Tournaments via Machine Learning Models

Syed Muhammad Arsalan [1] and **Syed Muhammad Zeeshan** [1*]

[1] Department of Statistics, Federal University of Minas Gerais (UFMG), Brazil;
[2] Department of Statistics, University of Peshawar, Pakistan

Keywords: *Game of cricket, International tournaments, T20 Opening batters, Machine Learning (ML) models, Prediction.*
Subject Classification: *62D05*

Journal Info:
Submitted:
August 02, 2025
Accepted:
December 19, 2025
Published:
December 31, 2025

Abstract In the game of cricket, particularly during high-stakes tournaments, players' performances have substantial consequences for their teams and energetic crowd. Predicting players' outcomes is often validated by experts' territory through mathematical and statistical models. However, due to the intricacies of cricket, player-related features in different sports cannot be evaluated comparatively. De spite these challenges, the rising utilization of Machine Learning (ML) models has proven crucial role in achieving precise predictions. In this research study, the ultimate aim was to predict the performance of T20 opening batters for upcoming T20 tournaments. Player records were compiled from ESPNcricinfo and Cricbuzz. Several ML models are implemented to predict players' outcomes. The analysis for this study was categorized into two cases: runs scored and strike rate, acknowledging both pre-match and all-match features. For predicting outcomes based on runs scored using pre-match features, Decision Tree and Naïve Bayes outperformed with an accuracy of 0.75, while for strike rate, K-Nearest Neighbor surpassed models with an accuracy of 0.68. Furthermore, assessing players' performance on runs scored using all-match features, Naïve Bayes and Support Vector Machine achieved exceptional accuracy of 0.98. For strike rate across all-match features, logistic regression beat the models with a leading accuracy of 0.98%.

***Correspondence Author Email Address:**
syedma@ufmg.br
DOI: 10.21015/vtm.v13i2.2207

1 Introduction

Among different types of sports, such as football, basketball, table tennis, volleyball, hockey, cricket, etc., cricket is considered one of the most prominent sports in the world. Specifically, in South Asian countries including, Pakistan, India, Bangladesh, and Sri Lanka. Outside South Asia, Australia, England, New Zealand, the West Indies, and South Africa are the top countries with leading cricket teams in the world. Cricket has three formats, and for all these formats, the cricket ground, the number of players' in each team, and the process of tossing a coin, such as the captains of both teams walking to the center of the ground and tossing a coin on the pitch, are the same. The captain who wins the toss has the right to choose whether to bat or bowl first. However, the way of playing in terms of overs, time, and innings differs. The three formats of cricket started at different times, i.e., test cricket was the first format, which started in 1871 and was first played at the Melbourne Cricket Ground (MCG) between England and Australia. Following Test cricket, another format of cricket, known as One-Day International (ODI) cricket, started in 1971, and the first official ODI cricket match was played between Australia and England at the MCG. The main difference between test and ODI cricket is that a test cricket match lasts a total of five days, with each team playing two innings and around 90 overs per day. Each day of test cricket includes two breaks: tea and lunch. On the other hand, ODI cricket consists of 50 overs per side and is completed on the same day, the match starts. Therefore, ODI cricket is a shorter format compared to test cricket. Because of this shorter format and quicker gameplay, the popularity of ODI cricket increased, which is why fans' interest rapidly shifted toward ODI cricket.

To cater to fans' interest in an even quicker and shorter format, ensuring they do not feel bored while maintaining the popularity of cricket, another format was introduced in 2007, known as T20 cricket. Officially, the first T20 cricket match took place on February 17, 2005, at Eden Park in Auckland, between Australia and New Zealand.

Nowadays, analyzing cricket has become crucial due to its immense popularity, with an estimated 2.5 billion fans [12]. Among different research fields, cricket has also become an optimistic research area. Fields such as survival analysis, operations research, economics, and game theory, etc., are being used to predict or evaluate the performance of teams and players [16]. For predicting team or player performance, researchers utilize details from past matches, batsmen's performance, and bowlers' performance, etc. The prediction of match outcomes or player performances is highly beneficial for teams as well as cricket fans worldwide. Most of the research studies in cricket have been on Test and ODI cricket because these two formats originated in the 18th and 19th centuries, respectively [17]. Therefore, many researchers have worked on different aspects of these two formats, including match results, player performance, and so on [4]. However, T20 cricket is more dynamic and started in the 20th century [32]. Due to this reason, it is thought-provoking to investigate research studies in T20 cricket.

In this scholarly work, the central focus was on using various machine learning models to predict T20 opening batters in international cricket tournaments. This is because, in T20 cricket tournaments, opening batsmen have a major responsibility to score runs with a good strike rate, especially during the powerplay [5]. Owing to the powerplay, which has an important impact on the entire match, the winning percentage in most matches increases when these opening batsmen hit boundaries and perform well during the powerplay [21] . Therefore, in T20 cricket, the performance of openers plays a key role, and predicting such performances for future tournaments is very crucial. This helps a cricket team and board members to carefully select opening batsmen, as they are the players who start the innings and may decide the outcome of the match. The primary questions we seek to investigate are:

- Are ML models capable of building accurate models to predict players' performance? If so, which model performs the best among several models?
- Are ML models suitable for identifying players' performance in T20 cricket tournaments? If so, which technique is the best in terms of Accuracy, Precision, Recall, and F1 score?

1.1 Hypothesis

This study hypothesizes that ML models can effectively predict T20 opener performance, with specific models demonstrating superior accuracy, precision, recall, and F1 score depending on the features (pre-match or all-match) and target variable (Runs scored or Strike rate). It also hypothesizes that all-match features will yield higher predictive results compared to pre-match features across all models.

2 Literature review

Various authors have made significant research contributions in analyzing and predicting cricket matches on different aspects, such as fielding, batting, and match outcomes, etc. Data mining techniques were used by [24] to evaluate the prediction of cricket match outcomes. They used different features of ODI cricket, which included home ground, toss decisions, player fitness, and other factors to predict ODI match results. Similarly, [11] used the SVM model of machine learning to predict the result of [7]. In their study, they investigated the accuracy of these techniques for obtaining the winning percentage of an ODI cricket match and developed a new tool known as the Cricket Outcome Predictor (COP).

Step-wise Multinomial Logistic Regression (SMLP) and the Naive Bayes classification model were used by [25] to classify all-rounders based on several factors, such as strike rate and economy rate. Additionally, [2] used a sequence of Multinomial Logistic Regression to forecast the result probabilities of test matches. From the results, they analyzed that the lead has a small effect on the match result. Furthermore, [9] used the SVM model to propose a supervised learning method with Radial Basis Function (RBF) kernels, linear and polynomial, to predict the outcomes of cricket matches and, by using performance statistics, outlined a player ranking system. Meanwhile, the use of an unsupervised machine learning approach that assists cycling specialists in key decisions about athlete selection, training regimens, and strategic preparation for track cycling was examined by [22]. Similarly, [13] collected the data of Indian Premier League (IPL) cricket matches spanning 10 years (2008–2017). The authors applied different probabilistic and statistical classification algorithms based on several features, such as Team 1, Team 2, match date, season, year, etc. From their study, they found the highest classification accuracy using the KNN method, achieving 62%. Moreover,[15] proposed a multivariate regression-based solution to influence the outcome of IPL matches. In their study, they used a total of six machine learning models to predict the results of each 2018 IPL match. The study revealed that the multilayer perceptron model outperformed all other models, achieving the highest accuracy of 71.66%.

In the same way, [27] used several ensemble classifiers, statistical, and probabilistic classification algorithms to predict the highest accuracy while tackling ambiguity in results. In their study, all post-match features were excluded, and they only considered features available up to the toss of the match, as well as statistical career information—specifically, team strength, winning rate,consistency, and team rank, etc. As per their research, the highest prediction accuracy of 60.043% was achieved using the RF classifier. Similarly, [18] collected data from 4 seasons of the Pakistan Super League (PSL) to predict the winning team. In their research investigation, they implemented the ”PSL Eye” approach, which leverages neural networks (NNs) to predict the victorious team. ”PSL Eye” delivered an overall testing dataset accuracy of 82%. Analogously, [10] employed a supervised learning approach from a team composition-based viewpoint. According to their study's analysis, they found that the KNN algorithm generated better results compared to all competing algorithms.

Also, [26] used multiple machine learning models to predict the Men's ICC T20 World Cup. As outlined in their research, they collected data from matches played between 2005 and March 2020. Stemming from their work, they found that RF algorithms provide the highest prediction accuracy. Moreover, DT and multilayer perceptron networks (MLP) were executed by [14] to predict the final results of ODI cricket matches using pre- and post-match features. In their research study, they developed a cricket match outcome prediction system based on the obtained results.

Meanwhile, [29] used boosting algorithms in ML to predict the winning probability of a team and the runs scored at the end of the upcoming over. From their analysis, they achieved 84.4% accuracy in predicting the winning team and a mean square error (MSE) of 1.41 for runs scored.

Additionally, [23] applied various classifiers of machine learning in their research to maximize prediction accuracy in cricket. As interpreted from the study, the RF classifier was considered the best for both cases of study used in their research, while the other classifiers did not perform as well as RF. Meanwhile, [3] utilized a machine learning approach for sports outcome prediction. Within their study, they used several types of information, including results of previous matches, player performance statistics, and opposition team/player information. By considering this information, the main aim was to focus on artificial neural networks (ANN) to predict results in sports.

Furthermore, [30] considered the NB model of machine learning to speculate the winning team of an ODI cricket match. From the obtained results of their research study, 85.71% accuracy was achieved for winner prediction using the univariate attribute selection method. Similarly, in sports more independent features such as location, pitch information (e.g., green pitch, dry pitch, dusty pitch, hard pitch, flat pitch, wet pitch, dead pitch, and hybrid pitch), weather conditions, etc., were adopted by [31] and [19] to prophesy match results in sports analytics. Meanwhile, [8] used three machine learning algorithms in the study of correlation-based features because these correlated features are very necessary for good prediction. Jack Davis, Harsha Perera, and Tim B. Swartz [6] proposed a new metric "Expected Run Differential" for player assessment in T20 international cricket. In their research analysis, they suggest a method to evaluate player performance on two aspects: career performance and current form. From the output of their research objectives, they found that their metrics do not highly correlate with batting average, strike rate, bowling average, bowling economy, or ICC ratings.

Muthuswamy and Lam [20] used neural network (NN) approaches to predict bowlers performance in ODI cricket. In their study, backpropagation network (BPN) and radial basis function network (RBFN) were scrutinized to determine the effectiveness of Indian cricket team bowlers. From the interpretation of results, they concluded that the model performed more adeptly in pre- dicting runs scored rather than taking or losing wickets. Moreover, [28] implemented different classifiers of machine learning models in cricket to predict the winning team in the IPL. Some features, such as batsmen's statistics, team strength, and ground conditions, were selected in the data for analysis. From the results analysis of each model, the accuracies of Decision Tree (DT) (94.87%), Random Forest (RF) (80.76%), and XGBoost (94.23%) were obtained, with DT and XGBoost performing better than RF. Ahmed [1] used a data mining approach to predict the match result of the Pakistan cricket team in ODIs. Several machine learning algorithms were applied to the data to evaluate the ODI match outcomes. In their proposed work, a valuable accuracy of 80% was obtained, which performed better than the existing works in the literature.

3 Methodology

In this research study, machine learning (ML) models are used to predict the performance of T20 opening batsmen in international cricket tournaments. ML techniques play a vital role in sports analytics research. We have incorporated a total of 21 match features that impact player performance. In these 21 match features, player name, player team, opponent team, day/night matches, match types (such as Group, Super Eight, Knockout, and Final), toss win, toss decision, nationality of players, opponent key bowlers, player inning (e.g., first or second inning), ground, and date are the pre-match features. While runs scored, strike rate, mins (time spent at the crease), dismissal type (e.g., bowled, caught, lbw, retired hurt, and runout), player of the match, opponent bowling figures, balls faced, fours, and sixes are the post-match features. By using these match features, we evaluate the opening batters' performance on two aspects, which are Case I (Runs Scored) and Case II (Strike Rate). Additionally, from the collected data details, we eliminate incomplete information so that no missing observations exist in the dataset for further analysis.

To investigate this, several ML models, including Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Random Forest (RF), are tested experimentally to predict the opening batters performance. Moreover, the models' performances are then compared with each other in terms of accuracy, precision, recall, and F1 score. The detailed results and performance of these models are presented in the upcoming sections.

Below Figure 1, displays various steps of the conceptual proposed system for predicting T20 opening batsmens performance in international cricket tournaments.

4 Data and Results Analysis

4.1 Data

For predicting the T20 opening batters, data from the most recently played International Cricket Council (ICC) Cricket World Cup were used, which was hosted by the West Indies and the United States of America (USA) from 1st June to 29th June 2024 . The data were compiled from two official cricket websites: ESPNcricinfo and Cricbuzz. The complete dataset encompassed 51 opening batters who played 206 innings across 20 cricket teams that participated in the ICC T20 World Cup 2024. The tournament consisted of 55 matches, some of which were abandoned due to rain; therefore, incomplete information from those matches was eliminated from the data. The detailed information that directly impacts players' performance is explained earlier in Section 3, as part of the complete dataset. Summary of the players statistics is provided in the Table 1 below.

4.2 Results Analysis

In this section, the selected features for this research were analyzed using R-language software to predict T20 opening batters for the upcoming ICC T20 tournaments. The analysis of this research study consists of two parts: Pre-match Features Analysis and All-match Features Analysis, in which post-match features are considered along with pre-match features. Additionally, the two most important aspects: Case I (runs scored) and Case II (strike rate, SR) are considered as target variables to predict T20 opening batters for both Pre-match and All-match features respectively.

For Case I, we have mapped runs scored into two categories: runs scored in an inning 30 is considered good performance, and runs scored in an inning < 30 is considered bad performance. Similarly,

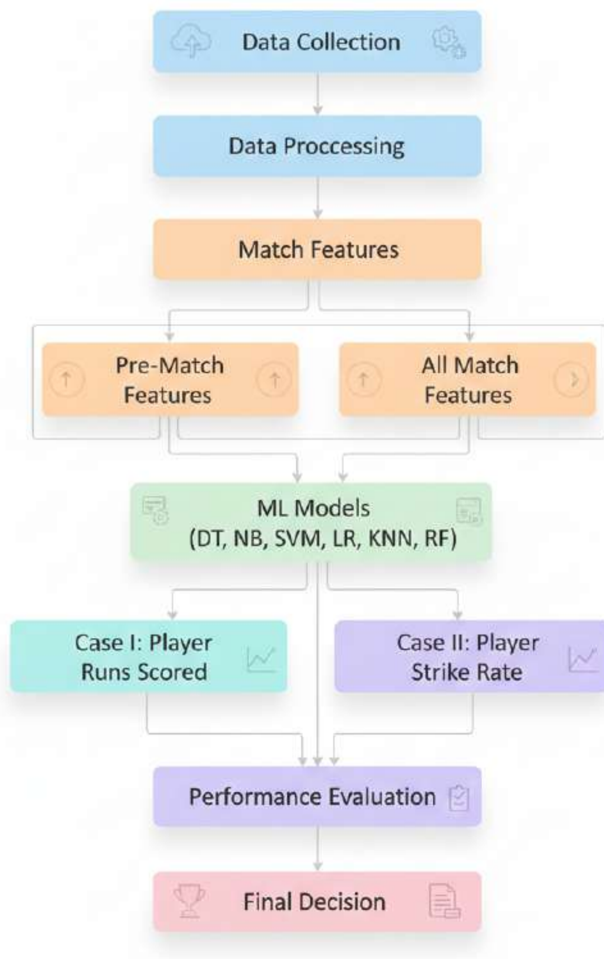


Figure 1. An innovative framework for predicting T20 opening batters performance

for Case II, SR in an inning 130 is categorized as good performance, while SR in an inning < 130 is considered bad performance. In T20 cricket, at least 30 runs with a good SR, without wasting balls, are necessary to benefit the team and increase the chances of winning the match. For prediction we split data into training (70%) and testing (30%) sets.

To evaluate performance, several machine learning models are implemented, including DT, NB, SVM, LR, KNN, and RF. Furthermore, we calculate metrics such as accuracy, precision, recall, and F1 score for each model to compare the results. The Formula's for the above-mentioned metrics are provided below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Table 1. Players' Statistics Summary

Player	Innings	Runs	Avg. Runs	Avg. SR	Total Mins	Avg. Mins	Fours	Avg. Fours	Sixes	Avg. Sixes
Rahmanullah Gurbaz	8	281	35.125	100.874	325	40.625	18	2.25	16	2.00
RG Sharma	8	257	32.125	140.004	225	28.125	24	3.00	15	1.875
TM Head	7	255	36.428	138.857	238	34.000	26	3.714	15	2.143
Q de Kock	9	243	27.000	116.041	257	28.556	21	2.333	13	1.444
Ibrahim Zadran	7	223	31.857	86.214	309	44.143	24	3.429	4	0.571
JC Buttler	7	214	30.571	145.769	199	28.429	22	3.143	10	1.429
PD Salt	7	188	26.857	171.786	200	28.571	16	2.286	10	1.429
DA Warner	7	178	25.429	131.067	204	29.143	17	2.429	11	1.571
V Kohli	8	151	18.875	76.785	201	25.125	8	1.000	7	0.875
J Charles	6	140	23.333	80.485	180	30.000	18	3.000	3	0.500
HG Munsey	4	124	31.000	134.020	142	35.500	8	2.000	9	2.250
AGS Gous	3	117	39.000	170.487	129	43.000	8	2.667	7	2.333
RR Hendricks	9	113	12.556	76.639	192	21.333	11	1.222	3	0.333
Mohammad Rizwan	4	110	27.500	97.300	185	46.250	5	1.250	3	0.750
Litton Das	4	93	23.250	96.883	175	43.750	9	2.250	2	0.500
A Johnson	3	89	29.667	123.207	111	37.000	12	4.000	4	1.333
MA Jones	4	89	22.250	117.433	98	24.500	8	2.000	5	1.250
BA King	5	86	17.200	129.664	132	26.400	13	2.600	2	0.400
SD Hope	2	82	41.000	105.125	93	46.500	4	2.000	8	4.000
MP O'Dowd	4	79	19.750	89.583	148	37.000	7	1.750	2	0.500

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where *TP* denotes True Positives, *TN* denotes True Negatives, *FP* denotes False Positives, and *FN* denotes False Negatives.

4.2.1 Pre-match features analysis

This section provides an extensive analysis of pre-match features. Pre-match features are those factors which considering prior to the commencement of the game. The pre-match features are mentioned in Section 3. Additionally, the result metrics are obtained for each ML model to compare their performance and predict players for upcoming ICC T20 events. Moreover, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are used to evaluate the best model in greater depth. ROC curves

Table 2. Players' Statistics Summary

Player	Innings	Runs	Avg. Runs	Avg. SR	Total Mins	Avg. Mins	Fours	Avg. Fours	Sixes	Avg. Sixes
TP Ura	4	16	4.00	46.110	65	16.25	1	0.25	0	0.00
Najmul Hossain Shanto	2	15	7.50	47.095	50	25.00	0	0.00	1	0.50
PR Stirling	3	12	4.00	45.423	44	14.667	1	0.333	0	0.00
S Ssesazi	4	9	2.25	66.665	30	7.50	2	0.50	0	0.00
RB Patel	2	6	3.00	105.000	37	18.50	1	0.50	0	0.00
Hazratullah Zazai	1	2	2.00	50.000	20	20.00	0	0.00	0	0.00
JP Kotze	1	0	0.00	0.000	2	2.00	0	0.00	0	0.00
RG Mukasa	2	0	0.00	0.000	3	1.50	0	0.00	0	0.00
Shayan Jahangir	1	0	0.00	0.000	1	1.00	0	0.00	0	0.00
Soumya Sarkar	1	0	0.00	0.000	2	2.00	0	0.00	0	0.00

are essential for assessing the performance of classification models, with FP on the x-axis and TP on the y-axis. Similarly, AUC plays a crucial role, summarizing the model's ability to differentiate between TP and FP . The gathered results and plots of Case I and Case II for pre-match feature analysis are provided below.

4.2.2 Case I

In this framework, the target variable runs scored is used for the pre-match features analysis to predict the T20 opening batter performance. As well, the aforementioned metrics in Subsection 4.2 for the model are acquired to determine the best ML technique. The detailed analysis for Case I is provided below in Table.3.

Table 3. Performance Metrics for Case I

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree (DT)	0.7540	0.7586	0.9777	0.8543
Naive Bayes (NB)	0.7540	0.7777	0.9333	0.8484
KNN	0.7213	0.7413	0.9555	0.8349
Support Vector Machine (SVM)	0.7377	0.7377	1.0000	0.8490
Logistic Regression (LR)	0.7377	0.7543	0.9555	0.8431
Random Forest (RF)	0.6885	0.7321	0.9111	0.8118

The performance metrics presented in Table 2 are evaluated to identify the most accurate model for Case I. The results show that DT and NB both excel in accuracy, achieving a value of 0.754,while SVM and LR follow with 0.7377. Conversely, RF demonstrates the lowest accuracy of 0.6885, making it the weakest predictive model. NB also leads in precision with 0.7777, outperforming all other models. LR and DT exhibit nearly identical precision values of 0.7543 and 0.7586 respectively, whereas SVM and RF deliver the lowest precision scores of 0.7377 and 0.7321, respectively. SVM stands out in recall with the highest value of 1, significantly surpassing the others, while KNN and LR achieve the same recall score of 0.955. DT achieves the highest F1 score (0.85), while other models show almost similar performance, except RF, which records the lowest F1 score of 0.8118. Figure. 2 provides a graphical overview of the performance

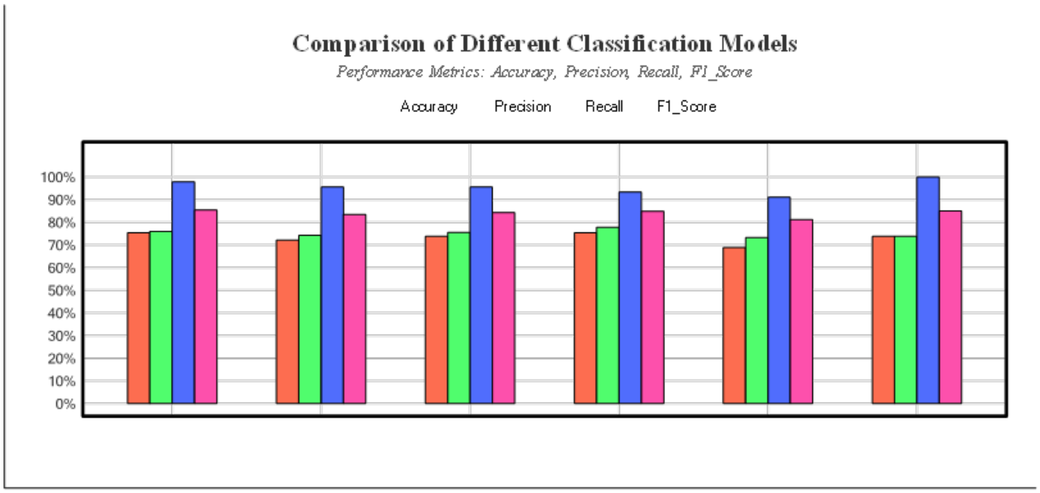


Figure 2. Comparing performance metrics across six ML models for Case I

metrics, complementing Table.2. The y-axis depicts metric percentages, and the x-axis lists the models. Each bar distinctly illustrates accuracy, precision, recall, and F1 score, enabling a detailed comparison of model performance.

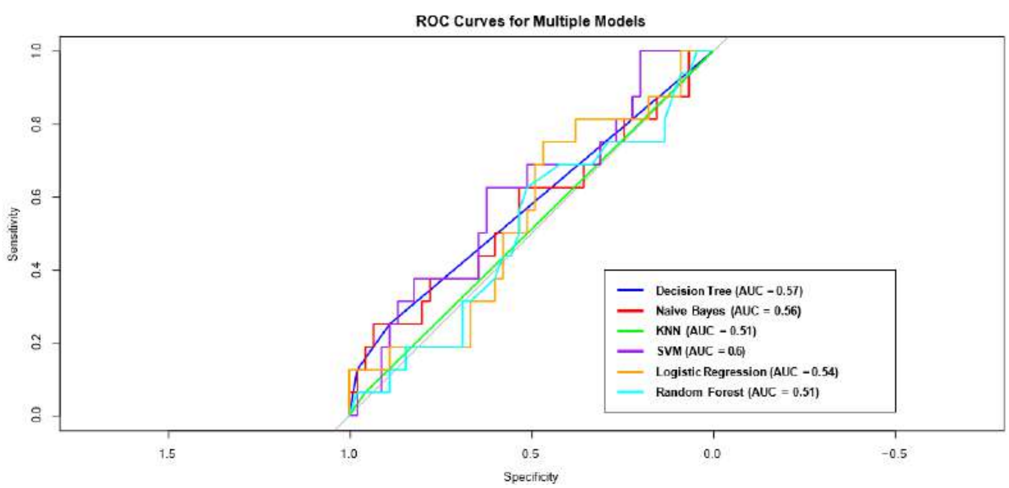


Figure 3. Comparing ROC curves across six ML models for Case I

4.2.3 Case II

In light of this framework, we used SR as a target variable for the pre-match feature analysis. SR is crucial in T20 cricket due to limited overs and quick gameplay. The metrics highlighted in Subsection 4.2 are obtained for the prediction of T20 opening batters, based on the results of several ML methods. The analysis for Case II is as follows.

Table 3 outlines the performance metrics used to identify the most effective model for Case II. According to the results, KNN achieves a better accuracy of 0.6885 and a precision of 0.7058 Compared to the others. Meanwhile, SVM and RF both have an accuracy of 0.6557, performing well after KNN. The re-

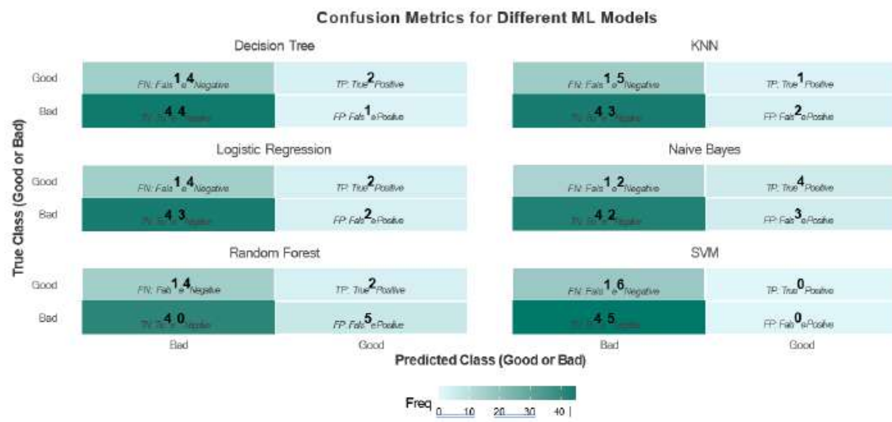


Figure 4. Comparing confusion metrics across six ML models for Case I

Table 4. Performance Metrics for Case II

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree (DT)	0.5573	0.6511	0.700	0.6746
Naive Bayes (NB)	0.5573	0.6326	0.775	0.6966
KNN	0.6885	0.7058	0.900	0.7912
Support Vector Machine (SVM)	0.6557	0.6557	1.0000	0.7920
Logistic Regression (LR)	0.6065	0.6481	0.875	0.7446
Random Forest (RF)	0.6557	0.6862	0.875	0.7692

maining methods exhibit lower accuracy compared to KNN, SVM, and RF. Moreover, SVM achieves a Recall score of 1, clearly outperforming the other predictive models. Similarly, in the F1 score column, KNN and SVM achieve 0.7912 and 0.7920 respectively, making them the top model in terms of F1 score, while the rest of the models perform approximately equal. Additionally, Figure 5 provides a more comprehensive visualization of metrics for all models and summarizes the revelation of table 3. The metrics results in figure offering a clearer view of the performance metrics.

Figure 6 represents the ROC curves for each ML technique for the situation of Case II. The curves clearly indicate that RF is the best method, with the strongest AUC value of 0.67, demonstrating the best performance, followed by KNN with an AUC value of 0.59. These two models significantly outperform the others. Similarly, SVM has an AUC value of 0.58, which is not as good as KNN and RF but is better than LR, DT, and NB. Moreover, LR and NB offer minimal performance, with AUC values of 0.53 and 0.52, respectively. Within the entire set of approaches, DT has the weakest AUC value of 0.50, equivalent to random guessing. Moreover, Figure 7 demonstrates the use of data from the confusion matrix to assess the capability of each model. The previously stated metrics in Subsection 4.2 are procured to identify the superior technique that best predicts T20 opening batters. In the figure, the true classes are plotted on the y-axis, while the predicted classes are shown on the x-axis. The figure also highlights TP, FP, TN, and FN, providing a detailed breakdown of each model's results as they relate to the metrics in Table 3. By analyzing these values, the comparative effectiveness of each model for prediction can be evaluated.

Figure 5. Comparing performance metrics across six ML models for Case II

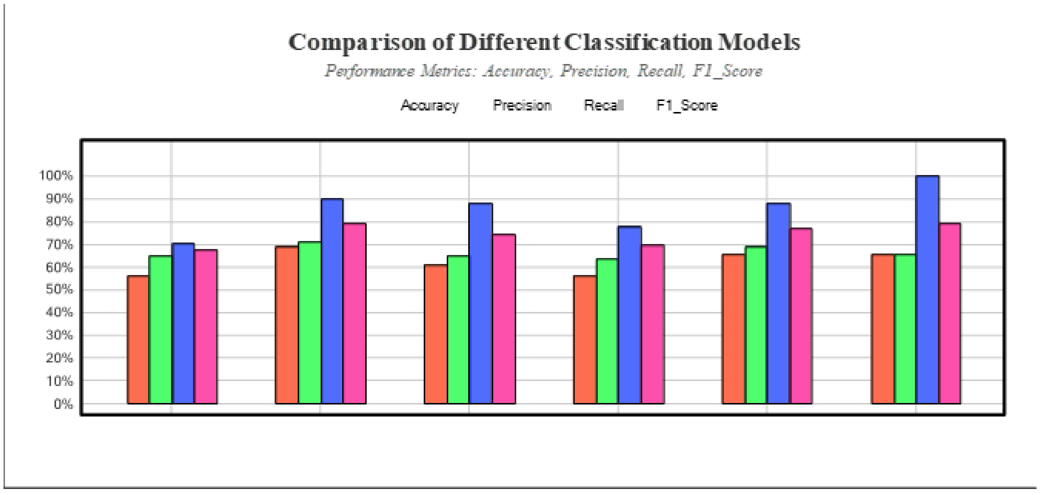
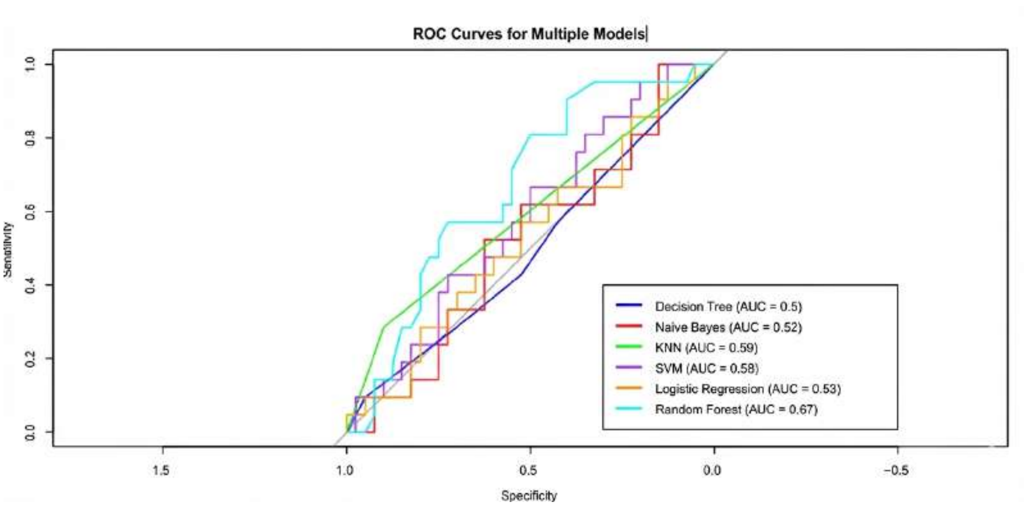


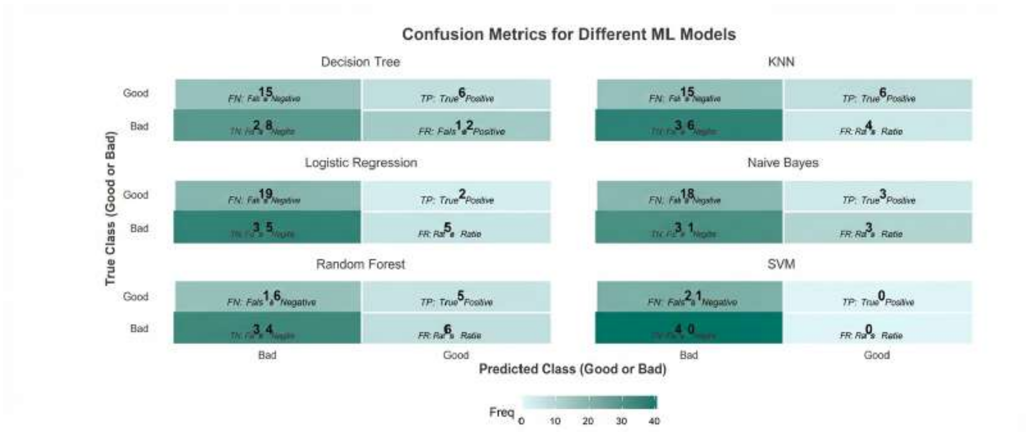
Figure 6. Comparing ROC curves across six ML models for Case II



4.3 All-match features analysis

This section demonstrates an in-depth analysis of all-match features, which consider both factors such as prior to the commencement of the game and continuing until after the outcome of the match. The detailed information of all-match features is outlined formally in Section 3. Whereas the detailed execution of performance metrics is derived for the models. In addition to performance metrics, the ROC and AUC are utilized to analyze the best model by comparing AUC and ROC curves. Furthermore, for metrics results, the confusion matrix is secured in which TP, FP, TN, and FN values are clearly mentioned for each classification model. Based on confusion matrix details, metrics score can be obtained for valid comparison. While the results of metrics are in Tables 4 and 5, similarly, separate figures are extracted for confusion matrix, ROC, and AUC respectively. The detailed analysis of all-match features for both Case I and Case II is exhibited below.

Figure 7. Comparing confusion metrics across six ML models for Case II



4.3.1 Case I

For this scenario, runs scored is employed as the target variable for all-match feature analysis. A variety of predictive models are implemented to obtain performance metrics, including ROC and AUC curves, to compare the results. Additionally, the confusion matrix is evaluated, where TP, FP, TN, and FN are calculated to derive the metric results. Below is a comprehensive analysis of this section.

Table 5. Outcome Metrics for Case I

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree (DT)	0.9672	0.9574	1.0000	0.9782
Naive Bayes (NB)	0.9836	1.0000	0.9777	0.9887
KNN	0.9344	0.9183	1.0000	0.9574
Support Vector Machine (SVM)	0.9836	0.9782	1.0000	0.9890
Logistic Regression (LR)	0.9180	0.9545	0.9333	0.9438
Random Forest (RF)	0.9672	0.9574	1.0000	0.9782

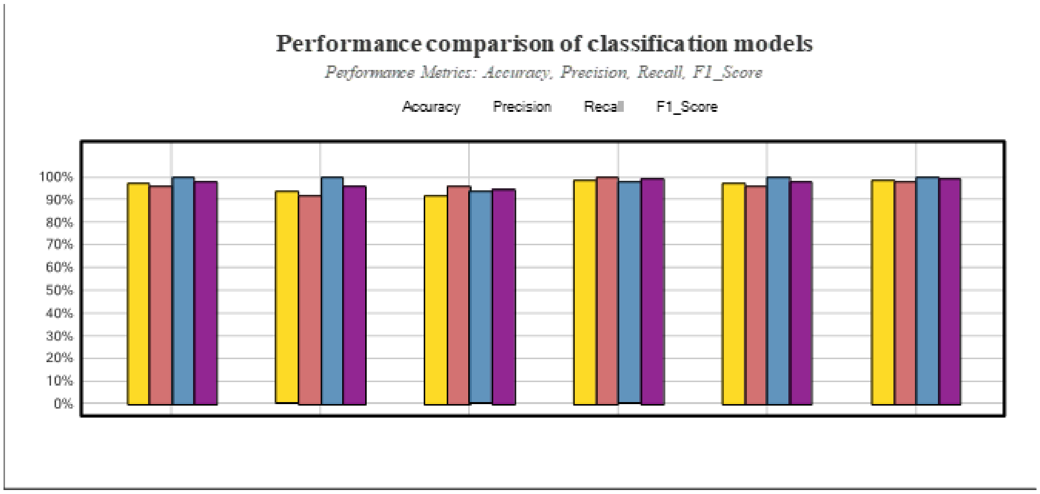
As indicated in Table 4, the performance metrics are derived for Case I using the selected classification models. Based on the results in the table, it is noticeable that NB and SVM each achieved an impressive accuracy of 0.9836, making them the most accurate models for predicting player performance. Following these, DT and RF attained an accuracy of 0.9672. In comparison, LR produced an accuracy of 0.9180, which is relatively lower than the others.

Furthermore, NB is the only model to achieve a remarkable precision of 1.000, making it the most precise model among the set, while SVM ranks second with a precision of 0.9782. Similarly, DT, LR, and RF yielded precision values of 0.9574, 0.9545, and 0.9574, respectively, which are nearly equivalent.

In terms of recall, all models performed perfectly except for LR, which has a score of 0.9333, the lowest recall value compared to the others. In contrast, SVM and NB secured F1 scores of 0.9890 and 0.9887, respectively, while the remaining models exhibited nearly identical F1 scores, with values close to one another.

Upon closer examination of Figure 8, which visualizes the metrics of all models and outlines the results from Table 4, it provides a solid foundation for comparing their performance. The *y – axis* of the figure represents the percentage values of performance metrics, while the *x – axis* lists the model names.

Figure 8. Comparing the effectiveness of six ML models using performance metrics for Case I



Bars corresponding to accuracy, precision, recall, and F1 score effectively illustrate the metrics and values for each model, facilitating a comprehensive comparison.

As indicated in Table 4, the performance metrics are derived for Case I using the selected classification models. Based on the results in the table, it is noticeable that NB and SVM each achieved an impressive accuracy of 0.9836, making them the most accurate models for predicting player performance. Following these, DT and RF attained an accuracy of 0.9672. In comparison, LR produced an accuracy of 0.9180, which is relatively lower than the others.

Furthermore, NB is the only model to achieve a remarkable precision of 1.000, making it the most precise model among the set, while SVM ranks second with a precision of 0.9782. Similarly, DT, LR, and RF yielded precision values of 0.9574, 0.9545, and 0.9574, respectively, which are nearly equivalent.

In terms of recall, all models performed perfectly except for LR, which has a score of 0.9333, the lowest recall value compared to the others. In contrast, SVM and NB secured F1 scores of 0.9890 and 0.9887, respectively, while the remaining models exhibited nearly identical F1 scores, with values close to one another.

Upon closer examination of Figure 9, which visualizes the metrics of all models and outlines the results from Table 4, it provides a solid foundation for comparing their performance. The *y*-axis of the figure represents the percentage values of performance metrics, while the *x*-axis lists the model names. Bars corresponding to accuracy, precision, recall, and F1 score effectively illustrate the metrics and values for each model, facilitating a comprehensive comparison.

Model comparison through ROC curves

Figure 9 displays the ROC and AUC curves, effectively demonstrating the performance of various classification models. As shown in the figure, the AUC values for SVM, NB, LR, and RF are 1. A detailed analysis revealed that this perfect score across four models is likely caused by feature leakage, where input variables such as minutes, fours, sixes, and balls faced are strongly correlated with runs scored. This correlation enables the models to predict the target variable perfectly, rather than learning general patterns. In this scenario, these models are deemed the most suitable for predicting the performance of opening batters

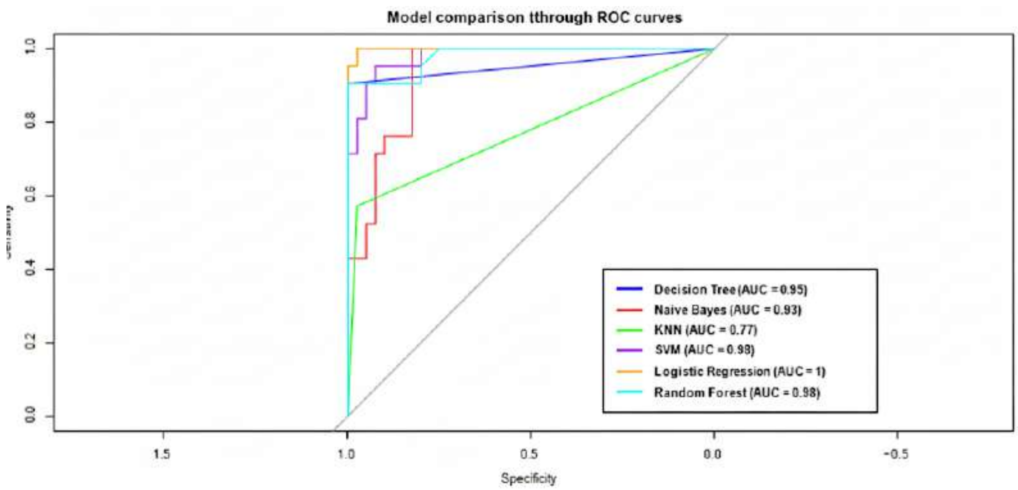


Figure 9. Case I analysis for comparing ROC curves from six different models

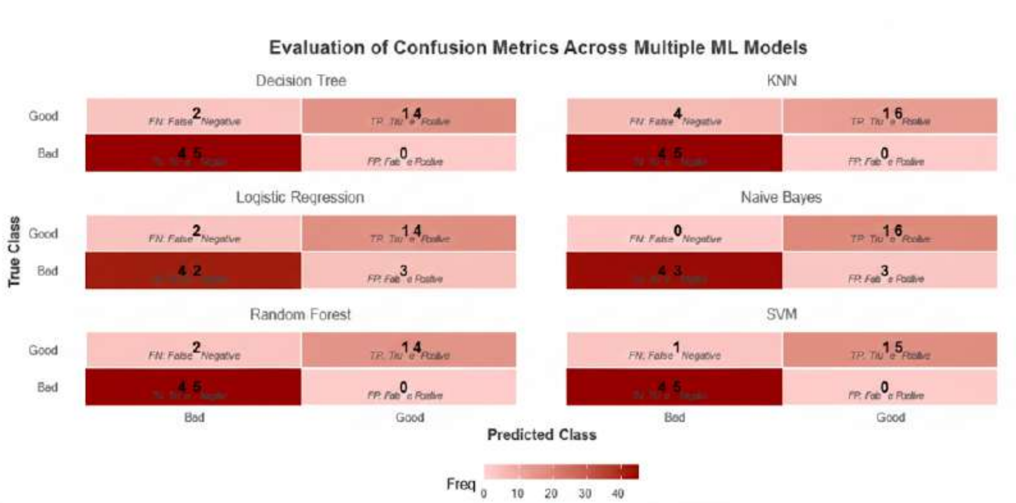


Figure 10. Case I analysis for comparing confusion metrics across six ML models

---

based on their AUC values, while their overall performance evaluation is detailed in Table 4. Similarly, DT and KNN achieved AUC values of 0.94 and 0.88, respectively, which are the lowest among the models considered.

Additionally, as depicted in Figure 10, the confusion metrics for each model are provided, displaying the counts of TN, FN, TP, and FP. These results enable straightforward evaluation, facilitating a comparison of the methods and identifying the most accurate model for predicting player performance.

4.4 Case II

In the context of all-match features, the target variable SR is considered essential for predicting the outcome of the best model. Identifying the most accurate model for player performance is critical for making valid and reliable predictions. In this analysis, we evaluated the performance metrics of each model and generated ROC and AUC curves to assess their effectiveness. Additionally, the confusion metrics were analyzed to interpret the classification metrics, highlighting the results of each model in predicting the capabilities of opening batters. The following analysis provides a comprehensive review of Case II.

Table 6. Outcome Metrics for Case II

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree (DT)	0.9672	0.9523	1.0000	0.9756
Naive Bayes (NB)	0.8032	0.8684	0.8250	0.8461
KNN	0.8360	0.8125	0.9750	0.8863
Support Vector Machine (SVM)	0.9180	0.9069	0.9750	0.9397
Logistic Regression (LR)	0.9836	0.9756	1.000	0.9876
Random Forest (RF)	0.9672	0.9523	1.0000	0.9756

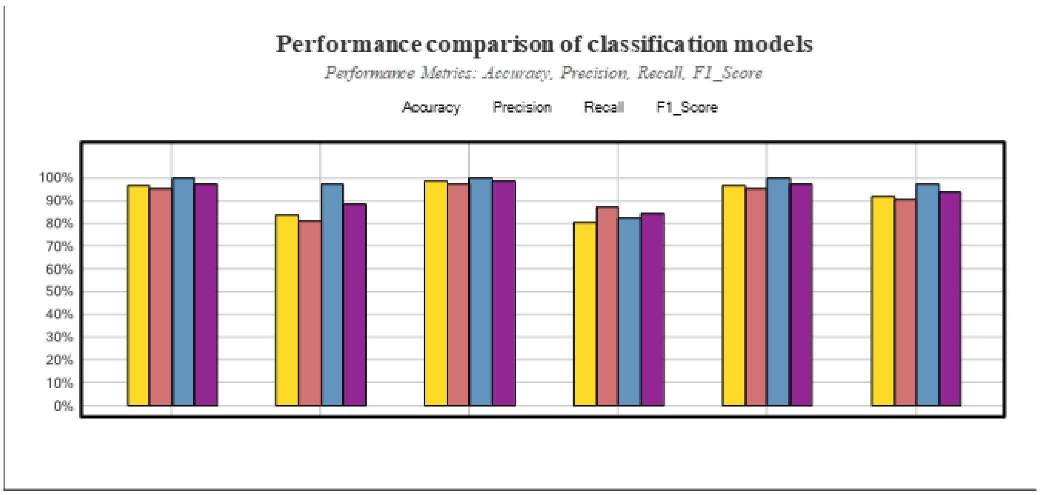


Figure 11. Comparing the effectiveness of six ML models using performance metrics for Case II

As shown in Table 5, the computed evaluation metrics for Case II using the classification models are presented. It is evident that LR achieved exceptional performance across all metrics, establishing itself as the top-performing predictive model with the highest accuracy, precision, and F1 score values of 0.9836, 0.9756, and 0.9876, respectively. Additionally, it achieved a Recall score of 1, which matches the performance of DT and RF in this metric.

Similarly, DT and RF followed LR with an accuracy of 0.9672, precision of 0.9523, a recall score of 1, and F1 scores of 0.9756. Notably, the prediction performance of both DT and RF is identical across all metrics for the all-match feature analysis, though this trend differs in the pre-match analysis. Moreover, in terms of the F1 score, NB recorded the lowest value of 0.8461, reflecting a relatively unsatisfactory performance among the models considered. Furthermore, KNN achieved the lowest accuracy of 0.8032 and a precision of 0.8125. In contrast, SVM performed well after LR, DT, and RF, achieving precision and recall values of 0.9069 and 0.9750, respectively.

Figure 12 graphically represents the metrics results across the models, complementing the findings from Table 5 and serving as a valuable tool for comparing the outcomes. The y-axis in the figure portrays the performance metric percentages, while the x-axis lists the models. The bars effectively visualize accuracy, precision, recall, and F1 score, facilitating a thorough evaluation of model performance.

The ROC and AUC curves are depicted in Figure 12. Based on the AUC values for each model,

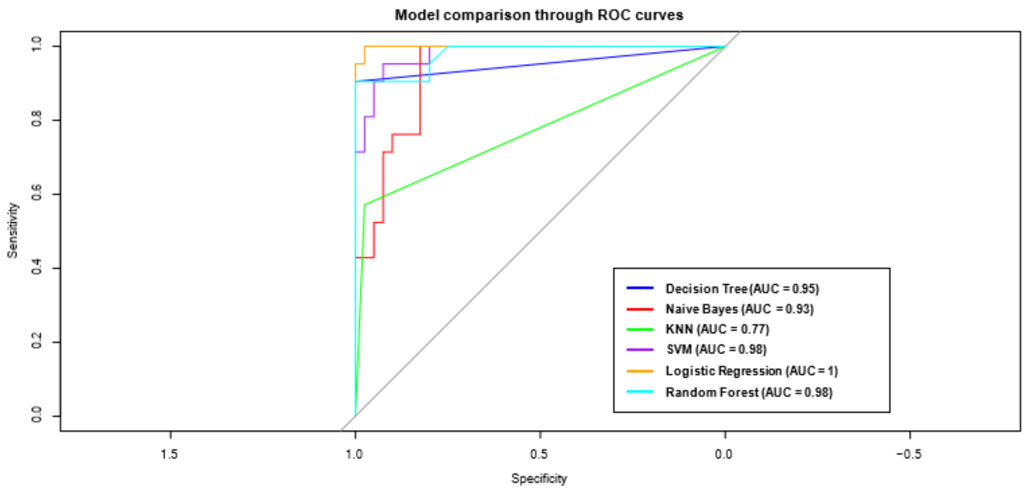


Figure 12. Case II analysis for comparing ROC curves from six different models

LR stands out with an impressive AUC value of 1, indicating a flawless model and making it the best-performing model for predicting outcomes. Similarly SVM and RF both scoring an exceptional AUC value of 0.98, delivers nearly equivalent predictive accuracy. DT also secures high reliability, reflected in its AUC of 0.95. And NB scoring an AUC of 0.93, demonstrate good potential, slightly behind the top three models in terms of accuracy. Furthermore, KNN is the only classification model to record the lowest AUC (0.77), reflecting weaker performance but still reasonable. Collectively LR is recommended for scenarios requiring accurate prediction, with SVM and RF are reliable alternatives.

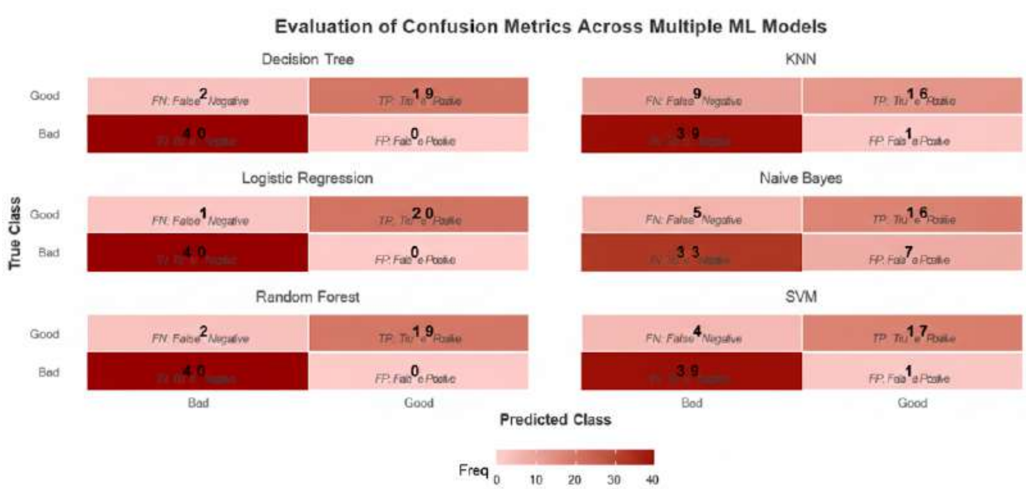


Figure 13. Case II analysis for comparing confusion metrics across six ML models

As shown in Figure 13, the confusion metrics for all the respective models are provided with in-depth details, showcasing the distribution of the true class on the *y*-axis and the predicted class on the *x*-axis. The figure also highlights the counts of TP, FP, TN, and FN. These counts serve as a clear breakdown of each model's metrics results, as shown in Table 6, enabling a thorough analysis of model accuracy and capability. By examining the values within these metrics, the pros and cons of each method can also be assessed. Moreover, based on the matrix results, it can be determined which model demonstrates the most effective approach for predicting player outcomes.

5 Conclusion

Currently, cricket stands out as one of the leading sports worldwide. In cricket, winning a match relies on the key roles played by each player. In every format of cricket, a player's performance is scrutinized through various metrics. Specifically, in T20 cricket, a fast-paced and highly competitive format, a batsman's performance is evaluated based on their ability to accumulate runs score at a high strike rate, while a bowler's performance is primarily measured by their capacity to take wickets while conceding minimal runs. Therefore, predicting such player contributions is critical to determining the match outcome. For achieving high predictive accuracy, one of the emerging sectors is machine learning (ML) models. Consequently, in this research study, we implemented numerous ML models to predict the performance of T20 opening batters. Two distinct cases were analyzed in the study: runs scored and strike rate, considering both pre-match and all-match features. Through the analysis, we determined that predicting performance for pre-match features based on runs scored yielded the highest accuracy of 0.7540 with NB and DT. Additionally, NB achieved the highest precision of 0.7777, surpassing all other models. Similarly, for strike rate under pre-match features, KNN demonstrated superior accuracy and precision of 0.6885 and 0.7058, respectively. Coupled with all-match features, SVM and NB delivered remarkable accuracy of 0.9836 for runs scored, with NB being the only model to achieve perfect precision of 1. For strike rate with all-match features, LR attained the highest accuracy (0.9834) and precision (0.9756), showcasing exceptional performance. To summarize, NB, DT, and KNN are the most appropriate models for evaluating player outcomes based on pre-match features to predict future tournaments. Meanwhile, SVM, NB, and LR are the most effective models for analyzing player success using all-match features for predictive analysis.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information

The authors received no funding from any source, including university projects or external grants, for the completion of this research.

Author Contributions

All authors contributed equally to this research work and participated in various aspects of the study according to the defined objectives. Each author was involved in the design, analysis, interpretation, and writing of the manuscript.

Data Availability Statement

The availability of data is mentioned within the manuscript. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request, subject to the journal's and reviewers' requirements.

Research Involving Human and/or Animals

This research did not involve any human participants or animals. The data were collected from publicly available official cricket websites, and were used solely for the purpose of this research.

Informed Consent

Informed consent was not applicable, as the study did not involve human participants, interviews, or personal/private data.

References

[1] Ahmed, W. (2015). A multivariate data mining approach to predict match outcome in one-day international cricket. Master's thesis, Karachi Institute of Economics and Technology, Pakistan.

[2] Akhtar, S. and Scarf, P. (2012). Forecasting test cricket match outcomes in play. *International Journal of Forecasting*, 28(3):632–643.

[3] Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33.

[4] Caso, S., Furley, P., and Jordet, G. (2025). Using video-notational analysis to examine soccer players' behaviours. *International Journal of Sport and Exercise Psychology*, pages 1–21.

[5] Chathurangi, A., Silva, R., Withanage, N., and Jayasinghe, C. (2025). Impact ranking methodologies in limited-overs cricket: A systematic review of performance metrics. *International Journal of Sports Science & Coaching*, 20(3):1307–1319.

[6] Davis, J., Perera, H., and Swartz, T. B. (2015). Player evaluation in twenty20 cricket. *Journal of Sports Analytics*, 1(1):19–31.

[7] Gunn, S. R. (1997). Support vector machines for classification and regression. Technical report, Citeseer.

[8] Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

[9] Jayanth, S. B., Anthony, A., Abhilasha, G., Shaik, N., and Srinivasa, G. (2018). A team recommendation system and outcome prediction for the game of cricket. *Journal of Sports Analytics*, 4(4):263–273.

[10] Jhanwar, M. G. and Pudi, V. (2016). Predicting the outcome of odi cricket matches: A team composition based approach. In *MLSA@PKDD/ECML*, volume 78.

[11] Kampakis, S. and Thomas, W. (2015). Using machine learning to predict the outcome of english county twenty over cricket matches. *arXiv preprint arXiv:1511.05837*.

[12] Kamruzzaman, M. (2024). Technological impact on fan engagement and revenue generation: Context of bangladesh cricket.

[13] Kapadia, K., Abdel-Jaber, H., Thabtah, F., and Hadi, W. (2022). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*, 18(3/4):256–266.

[14] Kumar, J., Kumar, R., and Kumar, P. (2018). Outcome prediction of odi cricket matches using decision trees and mlp networks. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 343–347. IEEE.

[15] Lamsal, R. and Choudhary, A. (2018). Predicting outcome of indian premier league (ipl) matches using machine learning. *arXiv preprint arXiv:1809.09813*.

[16] Lockwood, C., Soublière, J.-F., and Gehman, J. (2025). The cogs and wheels of cultural entrepreneurship: The becoming of new possibilities. *Academy of Management Annals*, (ja):annals–2023.

[17] Lokhande, R., Awale, R. N., and Ingle, R. R. (2025). Forecasting bowler performance in one-day international cricket using machine learning. *Expert Systems with Applications*, 259:125178.

[18] Mahmood, T., Riaz, M., Nasir, M., Afzal, U., and Siddiqui, M. H. (2021). Psl eye: Predicting the winning team in pakistan super league (psl) matches. *KIET Journal of Computing and Information Sciences*, 4(2):13–13.

[19] Munir, F., Hasan, M. K., Ahmed, S., and Quraish, S. M. (2015). *Predicting a T20 cricket match result while the match is in progress*. PhD thesis, Brac University.

[20] Muthuswamy, S. and Lam, S. S. (2008). Bowler performance prediction for one-day international cricket using neural networks. In *IIE Annual Conference Proceedings*, page 1391. Institute of Industrial and Systems Engineers (IISE).

[21] November, R. V., Cai, H., Taliep, M. S., Nyirenda, C., and Leach, L. L. (2025). Identification of key performance indicators for t20—a novel hybrid analytical approach. *Applied Sciences*, 15(12):6483.

[22] Ofoghi, B., Zeleznikow, J., Dwyer, D., and Macmahon, C. (2013). Modelling and analysing track cycling omnium performances using statistical and machine learning techniques. *Journal of Sports Sciences*, 31(9):954–962.

[23] Passi, K. and Pandey, N. (2018). Increased prediction accuracy in the game of cricket using machine learning. *arXiv preprint arXiv:1804.04226*.

[24] Pathak, N. and Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of odi cricket. *Procedia Computer Science*, 87:55–60.

[25] Saikia, H. and Bhattacharjee, D. (2011). On classification of all-rounders of the indian premier league (ipl): A bayesian approach. *Vikalpa*, 36(4):51–66.

[26] Singh, S., Aggarwal, Y., and Kundu, K. (2020). Quantitative analysis of forthcoming icc men's t20 world cup 2020 winner prediction using machine learning. *International Journal of Computer Applications*, 975:8887.

[27] Tripathi, A., Islam, R., Khandor, V., and Murugan, V. (2020). Prediction of ipl matches using machine learning while tackling ambiguity in results. *Indian J. Sci. Technol*, 13(38):4013–4035.

[28] Vistro, D. M., Rasheed, F., and David, L. G. (2019). The cricket winner prediction with application of machine learning and data analytics. *International Journal of Scientific & Technology Research*, 8(09):21–22.

[29] Weeraddana, N. and Premaratne, S. (2021). Unique approach for cricket match outcome prediction using xgboost algorithms. *Journal of Theoretical and Applied Information Technology*, 99(9):2162–2173.

[30] Wickramasinghe, I. (2020). Naive bayes approach to predict the winner of an odi cricket game. *Journal of Sports Analytics*, 6(2):75–84.

[31] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2005). *Practical machine learning tools and techniques*, volume 2, pages 403–413. Elsevier, Amsterdam, The Netherlands.

[32] Zaman, Q., Khattak, M. I., Nawaz, S., Khan, M. U., and Sahib, G. (2025). A cluster analysis of the performance of allrounders in t-20 international cricket. *Dialogue Social Science Review (DSSR)*, 3(1):520–544.