Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

A Hybrid Approach to T-20 Cricket Team Selection: Combining Probabilistic and Machine Learning Techniques

Muhammad Waqas
Department of Statistics, University of Peshawar, Pakistan.
Email: waqasuop1416@gmail.com

Qamruz Zaman (Corresponding Author)
Department of Statistics, University of Peshawar, Pakistan
Email: cricsportsresearchgroup@gmail.com

Imran ullah
Government college of Management Sciences Karak Pakistan .
Email: imran0310103@gmail.com

Fozia Mahsood
Government Girls Degree college Hangu Pakistan. Email: Foziamahsood786@gmail.com

Asma Shahnaz
Government Girls Post Graduate college Kohat Pakistan.
Email: Asmashahnaz99@gmail.com

Abstract
This study investigates the use of advanced machine learning models to improve player selection in T-20 cricket, focusing on both batsmen and bowlers. Performance thresholds, based on first-quartile metrics such as bowling economy, strike rate, batting averages, and boundary-hitting ability, were used to identify top-performing players. Exploratory data analysis highlighted key relationships between performance indicators and selection decisions. The study evaluates the predictive accuracy of four machine learning models: Random Forest, Neural Networks, Logistic Regression, and Naive Bayes. Random Forest outperformed all other models, achieving perfect classification accuracy, while Neural Networks and Logistic Regression also showed strong results. Naive Bayes, a probabilistic model, demonstrated lower accuracy but provided valuable insights into performance patterns. These results show how machine learning and probabilistic models can help build stronger T-20 cricket teams by focusing on consistent and impactful players.

Keywords: T-20 Cricket, Logistic Regression, Random Forest, Neural Networks, Naive Bayes. Random Forest

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

Introduction

Dream11 is a popular fantasy sports platform where users create virtual teams by selecting real-life players from sports like cricket, football, basketball, hockey, and baseball. Users earn points based on the players' actual performances in matches, such as runs, goals, wickets, and assists.[1]. Cricket data analysis has indeed become a crucial component of modern-day cricket, offering significant advantages for improving player performance and team strategies. With the rapid evolution of the game, analytics plays an essential role in helping teams stay competitive at the highest levels. [2]. The international T-20 format has become the most popular in cricket, offering a fast-paced 20-over structure that suits today's busy lifestyle. Events like the Indian Premier League (IPL) and Big Bash League (BBL) have further fueled its appeal. Each T-20 match delivers intense excitement, keeping fans on the edge of their seats, with many predicting the winner before or during the game, adding to the thrill. [3]. Each cricket team is a balanced mix of batsmen, bowlers, and all-rounders, with each player contributing to the team's success. Batsmen aim to score maximum runs, bowlers take wickets and restrict runs, while all-rounders contribute in both areas. A player's performance varies based on factors like the opponent, venue, current form, and past stats. The team management, coach, and captain analyze these factors to select the best playing XI for each match, aiming to predict and optimize player performance for that specific game. [4,5]. While many articles have used machine learning (ML) techniques in this domain to address the challenges mentioned, they have mostly relied on a categorical variable, such as win or loss, as the dependent variable. This approach simplifies the outcome into a binary classification, potentially overlooking the more intricate factors that could offer deeper insights into match performance and more accurate predictions. [6]. Machine learning approaches are increasingly being used by emerging classification methods that focus on significant features, such as weather, player positions, location, home team advantage, and toss decisions, which directly impact the final outcome of a cricket match. These factors are incorporated to estimate the likely match outcome more accurately [7]. The problem of team or playing eleven selections is challenging because each player, whether a batsman or bowler, has unique skills and capabilities. Comparing two players to determine who is the better choice for the Playing XI is difficult, as it involves evaluating various factors and performance metrics, making the decision complex and subjective. [8]. Many factors have been identified as important in affecting the outcome of cricket matches in previous studies. While their significance is often underestimated, they remain crucial and can immediately shift match results. [9]. Sanjaykumar, S identify which machine learning models exhibit superior predictive capabilities in the dynamic environment of T-20 cricket, with a particular focus on the high-pressure context of the World Cup. By analyzing and comparing the performance of these models, the study seeks to highlight their strengths, weaknesses, and potential areas for improvement. [10].Basit,,A.,Alv Different machine learning and data mining algorithms were applied for this prediction. The Naive Bayes algorithm, using 90% training data and 10% testing data, achieved an accuracy of 42.50%. The Decision Trees algorithm achieved 82.52%, while the Random Forest algorithm reached an accuracy of 90.88%. However, these results could potentially be improved by incorporating more advanced models, such as Recurrent Neural Networks (RNN) and Hidden Markov Models (HMM).[11]. Shilpi

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

Agrawal predicted the outcome of IPL matches using three machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, and CTree, based on the historical data available.[12]. Gholam R. Amin proposes a novel approach for measuring batting parameters in cricket by utilizing the OWA (Ordered Weighted Averaging) operator combined with regression methods to prioritize the most important batting parameters. [13].

This paper addresses the challenge of predicting match outcomes in T-20 cricket by analyzing the performance trajectories of batsmen and bowlers using machine learning techniques. By evaluating the skills and capabilities of players, the study aims to forecast how individuals will perform. The authors use performance data to identify an "ultimate T-20 team," composed of players whose combined abilities provide a competitive edge. This approach helps to predict match outcomes with greater accuracy, allowing teams and analysts to make informed decisions, refine strategies, and enhance performance in the fast-paced format of T-20 cricket.The data has been collected from the official CricInfo website, covering the period from 2005 to 2024.

Methods and materials

The methodology for this study involves defining the response variable as a **binary indicator**, determined by key performance metrics of bowlers and batters based on predefined selection criteria. The **independent variables** include essential performance indicators, and players are classified as selected (1) or non-selected (0) based on first-quartile thresholds. For **bowlers**, selection is based on maintaining a **bowling economy below 5**, a **bowling average under 35**, and a **wicket strike rate (WSR) below 39**. These thresholds ensure that only bowlers with **efficient run containment, effective wicket-taking ability, and optimal performance consistency** are considered. Similarly, for **batters**, selection depends on an **average score of at least 20**, a **strike rate of 120 or higher**, participation in at least **10 matches**, and an ability to hit a minimum of **10 boundaries (4s + 6s)**. These criteria highlight batters with a **consistent scoring ability, aggressive gameplay, and sufficient match experience**. Using these conditions, the dataset is processed to generate the binary response variable, which is then used in predictive modeling to analyze factors influencing player selection.

More explicitly, the response variable is a binary indicator created through independent variables by satisfying the following conditions.

$$S = \begin{cases} 1 & \text{if the player meets the criteria for selection} \\ 0 & \text{otherwise} \end{cases}$$

The selection criteria of a bowler are based on first-quartile thresholds for four key bowling metrics:

1. **Bowling Economy:** Should be less than 5.
2. **Bowling Average:** Should be less than 35.
3. **Bowling Wicket Strike Rate (WSR):** Should be less than 39.

The selection criteria of batters are based on first-quartile thresholds for four key bowling metrics:

1. **Average score:** Should be greater than or equal to 20.
2. **Strike rate:** Should be greater than or equal to 120.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

3. **Matches:** Should be greater than or equal to 10.
4. **Boundaries (4's + 6's):** should be greater than or equal to 10

Results and discussion

Exploratory Data analysis

The bar chart illustrates the distribution of player selection based on predefined performance thresholds for both bowlers and batters. The response variable is binary, where "1" represents players meeting the selection criteria and "0" represents those who do not. The significantly taller red bar (0) compared to the green bar (1) indicates that only a small proportion of players qualify under these stringent conditions. Bowlers are selected if their economy rate is below 5, bowling average is below 35, and wicket strike rate is below 39, while batters must have an average score of at least 20, a strike rate of 120 or more, at least 10 matches played, and a minimum of 10 boundaries (4s and 6s) combined.

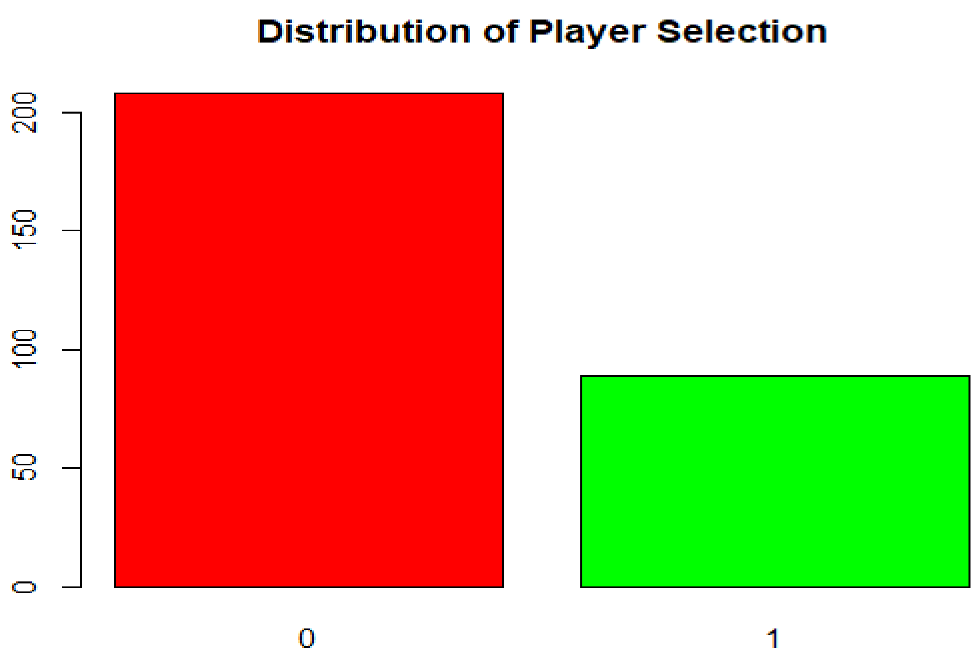

Figure 1 Bar plot of batter selection

Since these criteria are based on first-quartile thresholds, only the top 25% of players in each category are selected, leading to an imbalanced distribution where most players fail to meet at least one requirement.

The provided correlation matrix visualizes the relationships between various batting performance metrics, with the intensity and size of the circles indicating the strength and direction of correlations. Darker and larger blue circles represent strong positive correlations, while lighter or smaller circles indicate weaker relationships. Key insights include a strong positive correlation between **matches, innings, runs, and average score**, suggesting that players who play more matches tend to score more runs and have

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

higher averages.

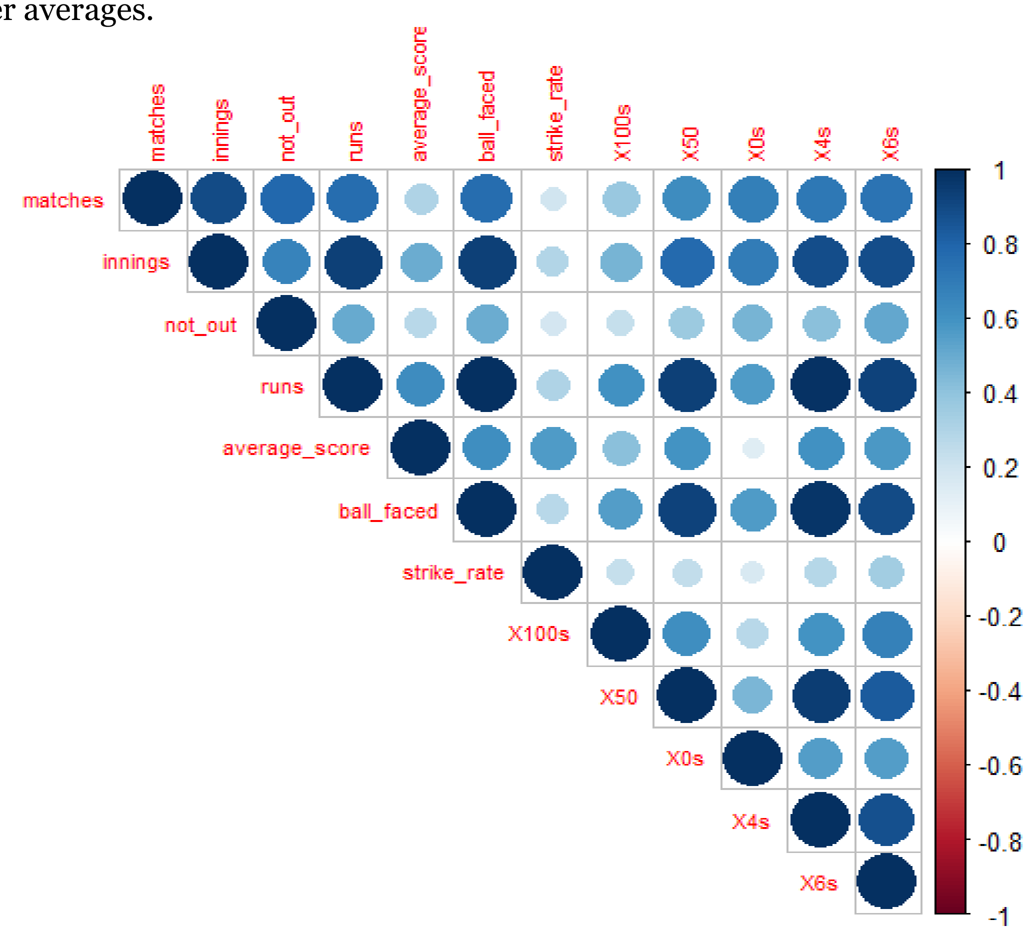

Figure 2 Correlation plot of study variables

Ball faced and strike rate also show a notable correlation, indicating that players who face more deliveries tend to have higher strike rates. Similarly, **boundaries (4s and 6s) are positively correlated with runs, strike rate, and centuries (100s)**, emphasizing their importance in high-scoring performances. The absence of strong negative correlations suggests that these batting metrics generally support each other rather than being inversely related. This analysis provides valuable insights into how different factors contribute to a batter's performance in T-20 cricket.

The scatter plot visualizes the relationship between **strike rate** and **runs scored**, categorized by **selection status** (0 and 1). Players marked with **0 (red dots)** were not selected based on predefined performance thresholds, while those marked with **1 (blue dots)** met the selection criteria. The distribution indicates that selected players tend to have both **higher strike rates and higher total runs**, suggesting that strike rate plays a crucial role in the selection process.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

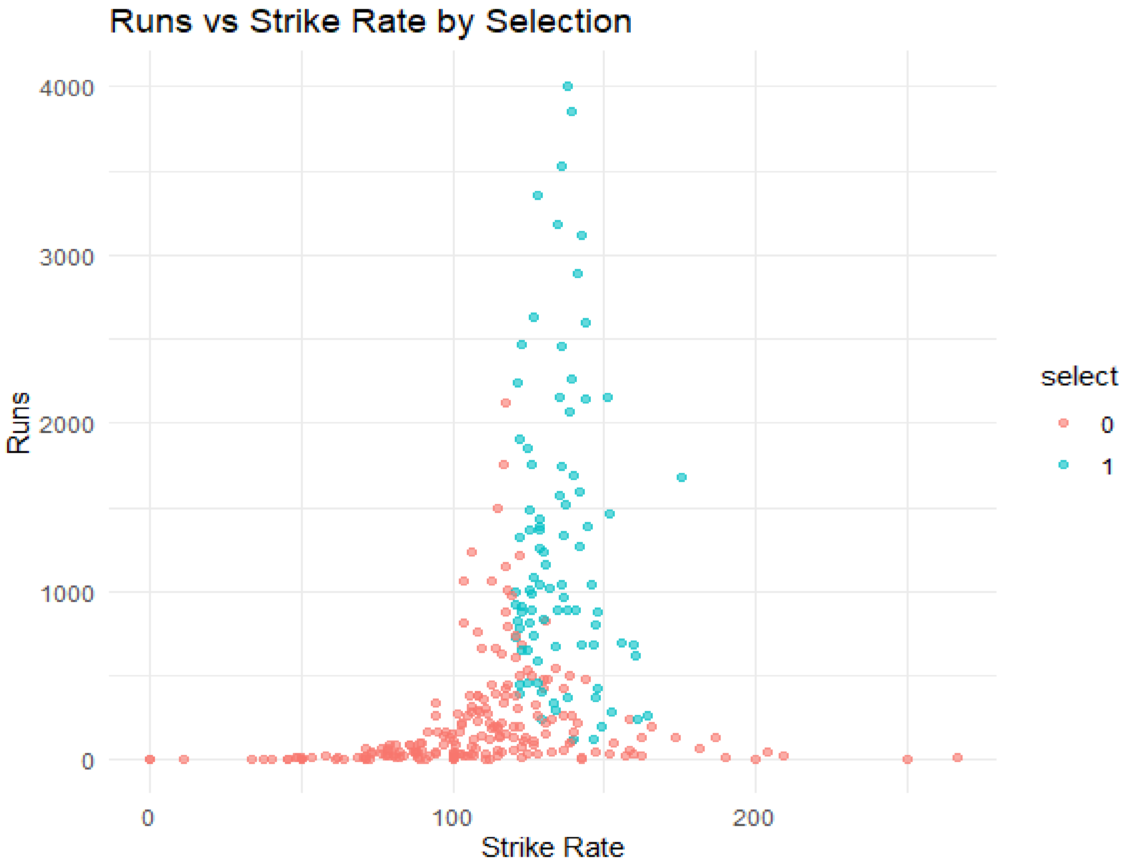

Figure 3 Scatterplot of runs against strike rate by player selection.
Most non-selected players cluster at lower strike rates and lower run totals, reinforcing the idea that higher-scoring and more aggressive batters are preferred. Additionally, the spread of selected players at higher strike rates and runs shows that elite batters consistently maintain a high scoring rate. However, some players with exceptionally high strike rates but lower run totals still fall into both categories, indicating that other performance factors might influence selection.
The chart titled "Distribution of bowler selection" illustrates the binary response variable for bowlers based on specific selection criteria. The binary indicator "0" (represented by the red bar) denotes bowlers who did not meet the thresholds, while "1" (represented by the green bar) signifies bowlers who satisfied the stringent criteria. The significantly larger red bar indicates that the majority of bowlers failed to meet the first-quartile thresholds for bowling economy (less than 5), bowling average (less than 35), and wicket strike rate (less than 39).

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

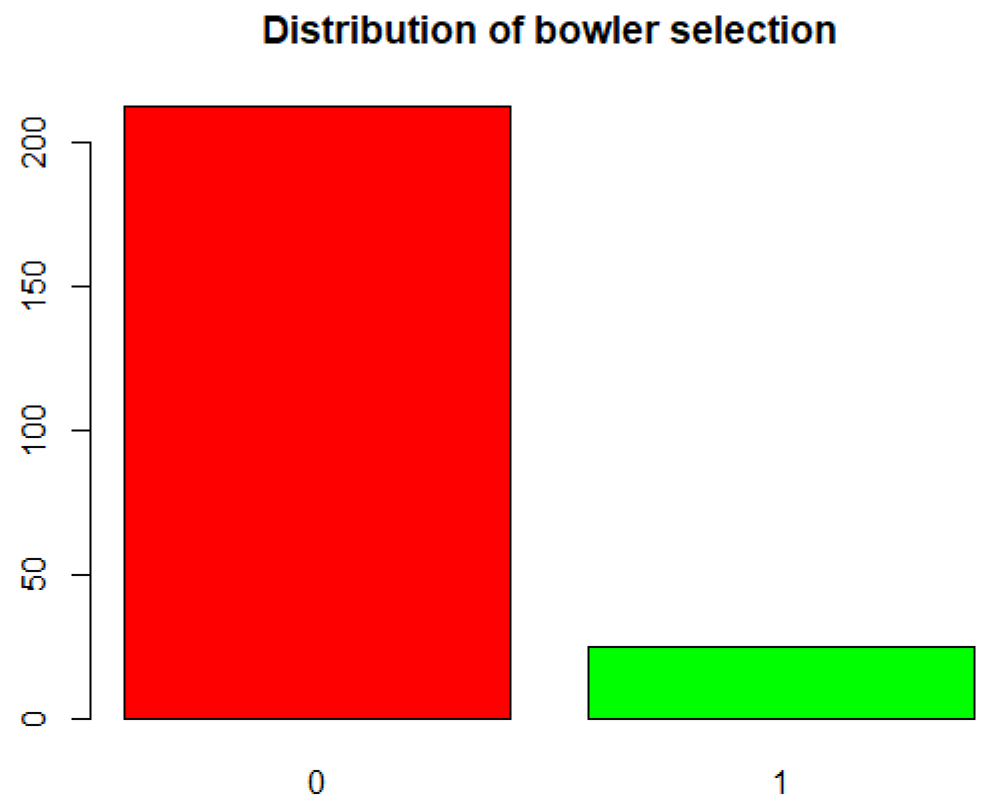

Figure 4 Bar plot of bowler selection.

This stark imbalance suggests that the defined benchmarks are highly selective, resulting in only a small proportion of bowlers being classified as "selected." The chart effectively highlights the rigorous nature of the criteria and its impact on bowler selection.

The scatter plot titled "Bowling Average vs Bowling Economy" illustrates the relationship between two key bowling metrics: bowling average (x-axis) and bowling economy (y-axis). Each blue dot represents an individual bowler, with the majority of the points clustered in the lower-left quadrant, indicating bowlers with both low bowling averages (below 50) and low bowling economy rates (below 6). This suggests that many bowlers maintain efficient performance in terms of both restricting runs and taking wickets.

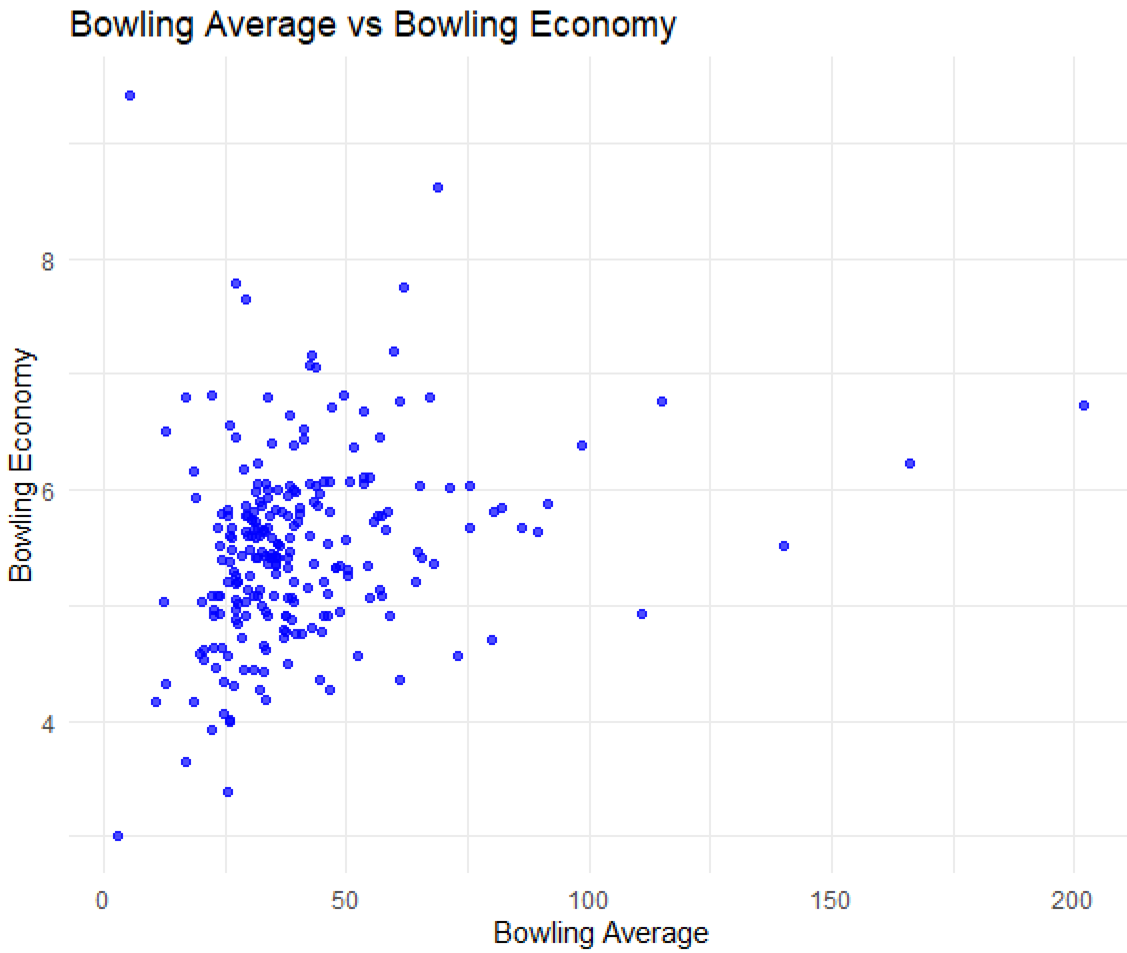

Figure 5 Scatter plot of bowling economy against bowling average.

However, there are some outliers with high bowling averages (exceeding 100) and higher economy rates, indicating less effective performance. The overall pattern suggests a loose positive correlation, where higher bowling averages are often associated with slightly higher economy rates. This chart provides insight into the distribution of bowler performance, emphasizing that most bowlers fall within a competitive range for these metrics.
The scatter plot titled "Strike Rate vs Wickets per Match" explores the relationship between bowling strike rate (x-axis) and wickets per match (y-axis). The majority of data points are concentrated near lower strike rates (below 50) and a wide range of wickets per match, with some reaching over 200. This suggests that bowlers with low strike rates (i.e., taking wickets more frequently) tend to deliver better performances, achieving higher wickets per match. Conversely, as strike rates increase beyond 50, there is a noticeable decline in wickets per match, with very few bowlers achieving significant numbers in this range.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

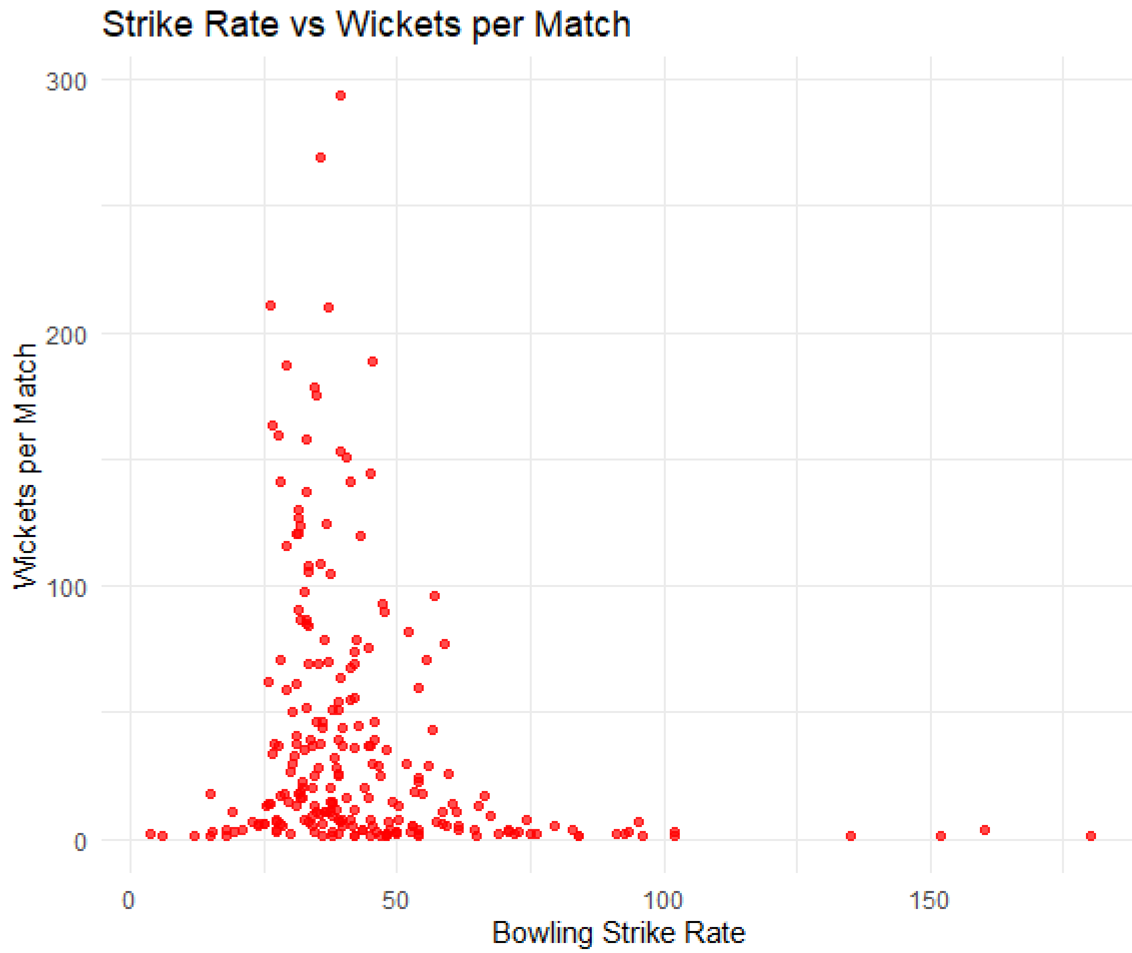

Figure 6 Scatter plot of wickets per match against bowling strike rate.

The plot highlights that efficient bowlers, who maintain lower strike rates, are key contributors to wicket-taking success. A few outliers with extreme values in both metrics may indicate exceptional performances or anomalous data points.

The heatmap depicts correlations among cricket bowling performance metrics. Strong positive correlations exist between "Total overs bowled," "Innings bowled," and "Matches played," indicating their alignment. "Bowling economy" and "Bowling average" are positively correlated, suggesting that higher economy rates often lead to worse averages. Milestones like "5-wicket haul" and "4-wicket haul" correlate positively with opportunities such as overs and innings bowled.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

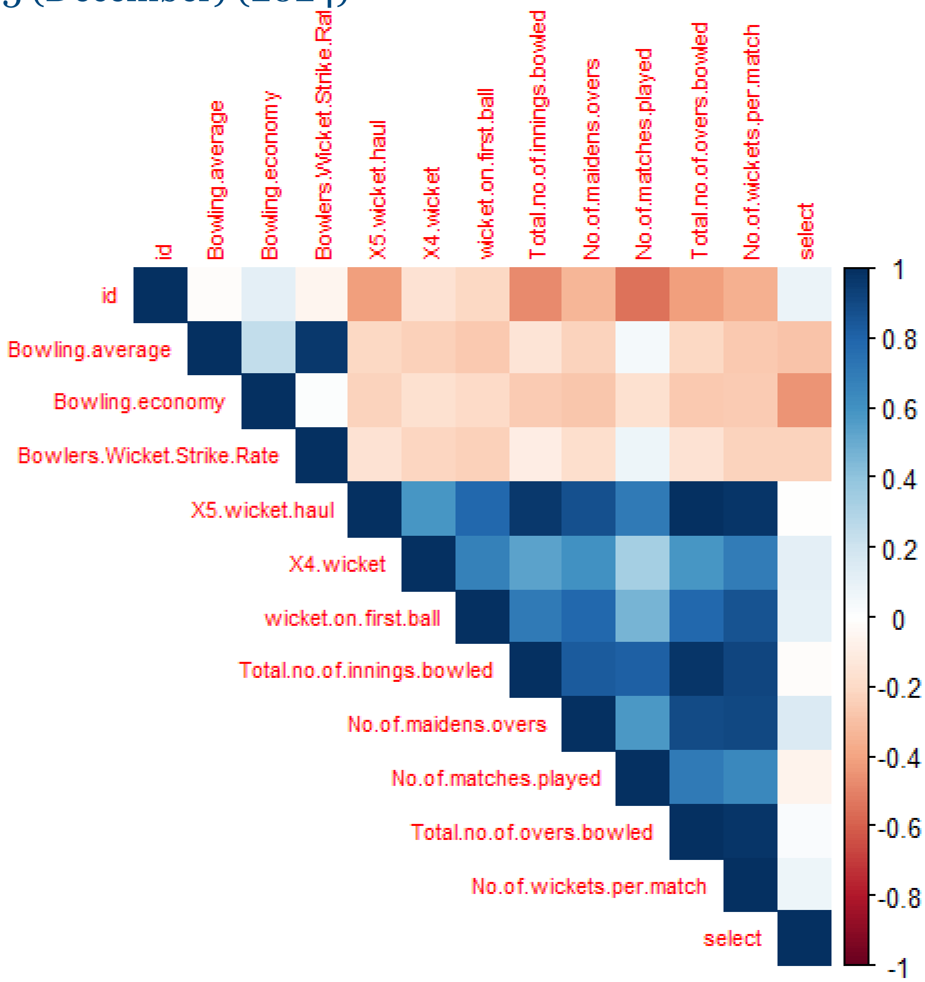

Figure 7 Correlation plot of numeric features of data set.

Negative correlations, such as between "Bowling average" and "Wickets per match," highlight that lower averages are associated with higher wicket-taking efficiency. This analysis reveals key relationships among bowling metrics for performance evaluation.

The logistic regression summary table provides insights into the factors influencing batter selection in cricket. Significant predictors include **Innings** ($p < 0.001$), **Not out** ($p < 0.01$), **Runs** ($p < 0.001$), **High score** ($p < 0.05$), **Balls faced** ($p < 0.001$), **100s** ($p < 0.01$), and **0s** ($p < 0.05$). These variables significantly impact the likelihood of selection. Negative coefficients for **Innings**, **Balls faced**, and **100s** suggest a decreasing likelihood of selection with higher values, while positive coefficients for **Not out**, **Runs**, **High score**, and **0s** indicate an increasing likelihood.

Model Summaries

Table 1: Summary of logistic regression for batter selection

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-1.934	1.791	-1.080	0.280	

Matches	-0.026	0.053	-0.492	0.623	
Innings	-0.760	0.226	-3.360	0.001	***
Not out	0.731	0.264	2.769	0.006	**
Runs	0.126	0.029	4.372	0.000	***
High score	0.061	0.027	2.288	0.022	*
Average score	-0.117	0.093	-1.256	0.209	
Ball faced	-0.107	0.024	-4.365	0.000	***
Strike rate	-0.015	0.018	-0.855	0.393	
100s	-7.024	2.239	-3.137	0.002	**
50s	-0.453	0.347	-1.305	0.192	
0s	0.739	0.314	2.356	0.018	*
4s	-0.003	0.050	-0.068	0.946	
6s	-0.027	0.069	-0.387	0.699	

Sig. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Other variables, such as **Matches**, **Strike rate**, **50s**, **4s**, and **6s**, are not statistically significant ($p > 0.05$), implying limited influence on selection. This analysis highlights the critical performance metrics for batter selection decisions.

The logistic regression summary for bowler selection highlights several significant predictors. Key variables influencing selection include **Bowling average** ($p < 0.05$), **Bowling economy** ($p < 0.05$), **Bowlers' Wicket Strike Rate** ($p < 0.05$), **5-wicket haul** ($p < 0.05$), **4-wicket haul** ($p < 0.05$), **No of maidens overs** ($p < 0.05$), and **No of wickets per match** ($p < 0.05$). Positive coefficients for **Bowling average**, **No of maidens overs**, and **No of wickets per match** suggest that better performance in these metrics increases the likelihood of selection. Negative coefficients for **Bowling economy**, **Wicket strike rate**, **5-wicket haul**, and **4-wicket haul** indicate that higher values in these metrics (often associated with poorer performance) decrease selection chances.

Table 2: Summary of logistic regression for bowler selection

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	92.870	45.940	2.022	0.043	*
Bowling average	1.826	0.866	2.109	0.035	*
Bowling economy	-16.733	8.330	-2.009	0.045	*
Bowlers Wicket Strike Rate	-1.920	0.907	-2.116	0.034	*
5 wicket haul	-0.051	0.023	-2.222	0.026	*
4 wicket	-7.316	3.103	-2.358	0.018	*
wicket on first ball	-0.687	0.668	-1.028	0.304	
Total no of innings bowled	0.235	0.162	1.450	0.147	
No of maidens overs	0.203	0.103	1.972	0.049	*

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

No of matches played	-0.003	0.026	-0.119	0.905	
Total no of overs bowled	0.053	0.035	1.531	0.126	
No of wickets per match	1.016	0.445	2.284	0.022	*

Sig. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Other variables, such as **Wicket on first ball**, **Innings bowled**, and **Overs bowled**, are not statistically significant ($p > 0.05$). This analysis underscores the importance of efficiency and consistency in bowler performance for selection.

The Random Forest variable importance table highlights the most influential factors for bowler and batter selection based on the **Mean Decrease Gini** values. For **bowlers**, the most critical variable is **Bowling economy** (15.851), followed by **Bowling average** (10.789) and **Bowlers' Wicket Strike Rate** (7.358). Other variables, such as **No of maidens overs**, **5-wicket haul**, and **Total no of overs bowled**, have relatively lower importance, indicating a lesser contribution to the model.

Table 3: Random Forest variable importance for bowler and batters selection

Variable (Bowlers)	Mean Decrease Gini	Variable (Batters)	Mean Decrease Gini
Bowling economy	15.851	Average score	21.490
Bowling average	10.789	Strike rate	20.034
Bowlers Wicket Strike Rate	7.358	Runs	15.036
No of maidens overs	1.752	6s	14.984
5 wicket haul	1.720	High score	14.851
Total no of overs bowled	1.512	4s	11.822
No of matches played	1.402	Ball faced	8.811
Total no of innings bowled	1.212	50s	5.441
No of wickets per match	1.179	Innings	5.136
wicket on first ball	0.865	Matches	2.735
4 wicket	0.426	0s	1.942
***	***	Not out	1.901
***	***	100s	0.172

For **batters**, the key factors are **Average score** (21.490) and **Strike rate** (20.034), followed by **Runs** (15.036) and **6s** (14.984). Variables like **High score**, **4s**, and **Ball faced** also play notable roles, whereas **100s**, **Not out**, and **0s** have minimal impact. Overall, performance efficiency and consistency are the dominant contributors for both bowlers and batters in the selection process.

The confusion matrix results in Table 4 reveal that Random Forest outperformed all other models, achieving perfect classification for both bowlers and batters, with no false positives or false negatives. Neural Networks also performed well, especially for batters, where it had no false positives and minimal false negatives. Logistic Regression demonstrated reasonable accuracy but was less effective compared to Neural Networks and Random Forest, with a small number of false positives and negatives. Naive Bayes showed the highest classification errors, with notable false positives and false negatives for both bowlers and batters.

Table 4: A confusion matrix of testing is set by each model

Methods	Predicted	Reference (Bowlers)		Reference (Batters)	
		0	1	0	1
Naïve Bays	0	192	10	193	20
	1	20	15	15	69
Neural Networks	0	210	7	208	13
	1	2	18	0	76
Logit	0	209	3	203	9
	1	3	22	5	80
Random Forest	0	212	0	208	0
	1	0	25	0	89

Overall, Random Forest emerged as the most accurate model for both classifications, followed closely by Neural Networks.

The testing set accuracy measurements in Table 5 show that Random Forest achieved perfect scores across all metrics (accuracy, sensitivity, specificity, and kappa) for both bowlers and batters, highlighting its exceptional predictive performance.

Table 5: Testing set accuracy measurements of each model

Role	Methods	Accuracy	Sensitivity	Specificity	Kappa
Bowlers	Neive Bayes	0.873	0.906	0.600	0.430
	Neural Networks	0.962	0.991	0.720	0.780
	Logit	0.975	0.986	0.880	0.866
	Random Forest	**1.000**	**1.000**	**1.000**	**1.000**
Batters	Neive Bayes	0.882	0.928	0.775	0.715
	Neural Networks	0.956	1.000	0.854	0.891
	Logit	0.953	0.976	0.899	0.886
	Random Forest	**1.000**	**1.000**	**1.000**	**1.000**

Neural Networks also performed strongly, with high accuracy and kappa values, particularly for batters where it achieved perfect sensitivity. Logistic Regression displayed comparable accuracy and kappa values to Neural Networks but with slightly lower

sensitivity and specificity.
The confusion matrix plots compare the classification performance of four machine learning models—Naive Bayes, Neural Network, Logistic Regression, and Random Forest. Random Forest demonstrates flawless classification with zero misclassifications, achieving perfect predictions for both classes (true positives and true negatives). The Neural Network also performs strongly, with a few false negatives and false positives, indicating its high accuracy but slightly lower performance compared to Random Forest. Logistic Regression shows solid performance, with minimal false negatives and false positives, though slightly more errors than the Neural Network.

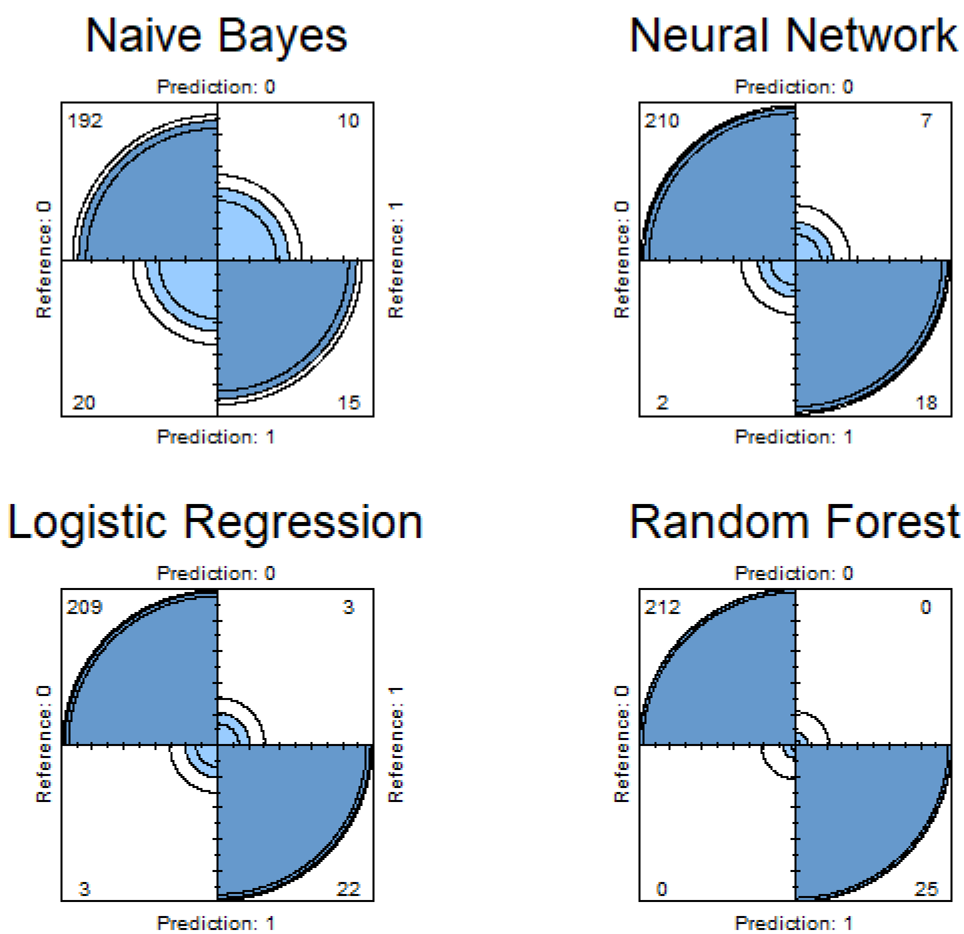

Figure 8 Plot of confusion matrix for each model for bowler selection.
In contrast, Naive Bayes has a higher number of misclassifications, including false positives and false negatives, reflecting its relatively lower accuracy. Overall, Random Forest is the most effective model, followed by Neural Network and Logistic Regression, with Naive Bayes being the least accurate among the four.
The confusion matrix plots illustrate the performance of four models (Naive Bayes, Neural Network, Logistic Regression, and Random Forest) in predicting the classes of bowlers and batters. Random Forest demonstrates perfect classification with no misclassifications (all

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

true positives and true negatives), indicating its superior accuracy. The Neural Network performs well with minimal errors, showing a small number of false negatives and false positives. Logistic Regression also performs strongly but has slightly more false negatives compared to the Neural Network. In contrast, Naive Bayes exhibits a relatively higher number of misclassifications, particularly in false positives, suggesting it is the least accurate model among the four.

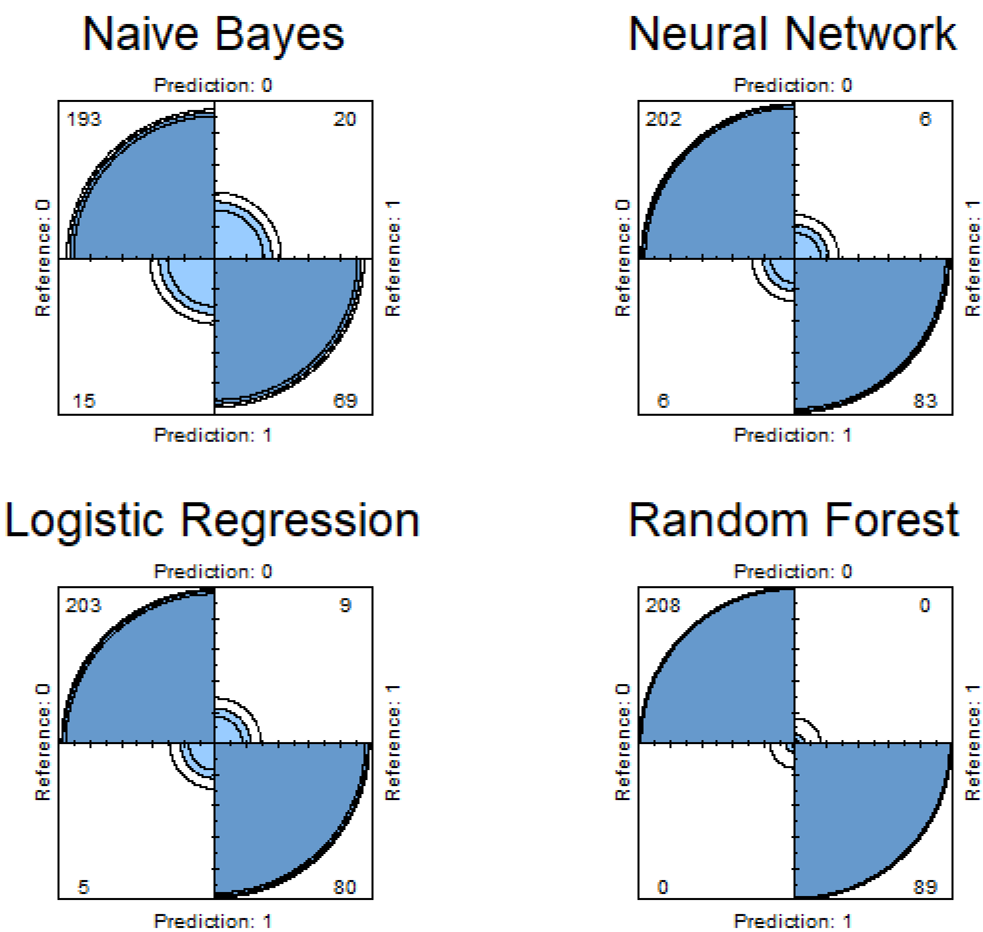

Figure 9 Plot of confusion matrix for each model for batter selection.

Overall, Random Forest stands out as the most effective model for prediction, followed by Neural Network and Logistic Regression, with Naive Bayes lagging behind.

In the ROC curve the selection of bowlers is based on predefined performance thresholds for key bowling metrics, ensuring that only the most effective bowlers are chosen. The selection criteria prioritize **bowling economy (less than 5), bowling average (less than 35), and wicket strike rate (WSR less than 39)**, reflecting a preference for bowlers who are both economical and capable of taking wickets efficiently. The analysis likely reveals that selected bowlers (category **1**) have significantly better control over runs conceded and a higher wicket-taking ability compared to non-selected bowlers (category **0**). Non-selected bowlers tend to have higher economy rates, weaker averages, and a longer interval between wickets, making them less effective in a match scenario.

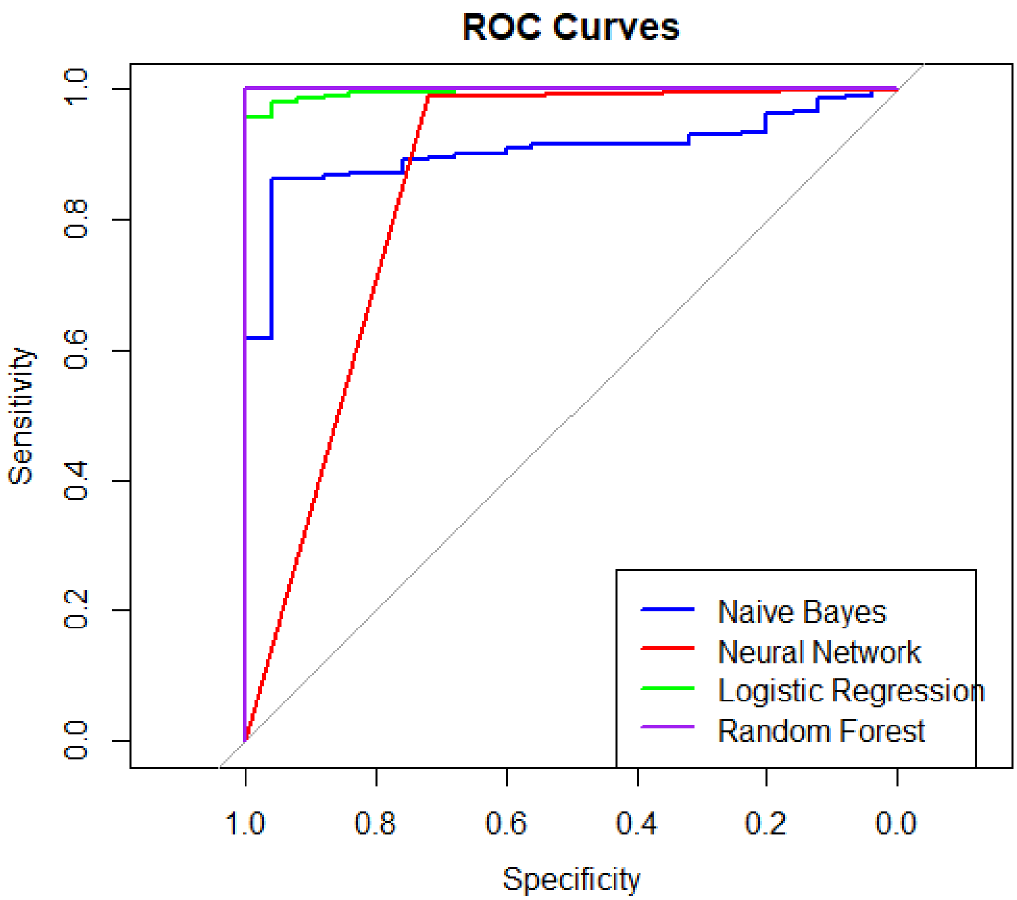

Figure 10 ROC curve of each model for bowler selection.

This selection process ensures that only the most impactful bowlers, capable of restricting runs and taking wickets consistently, are included in the final squad, optimizing overall team performance.

The ROC curve analysis for each model used in **batter selection** provides insights into their predictive performance in distinguishing between selected and non-selected batters. The **area under the ROC curve (AUC)** serves as a key metric, indicating the discriminatory power of each model. A higher AUC value, closer to **1.0**, signifies a model with strong predictive ability, while an AUC near **0.5** suggests a model with weak or random classification.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

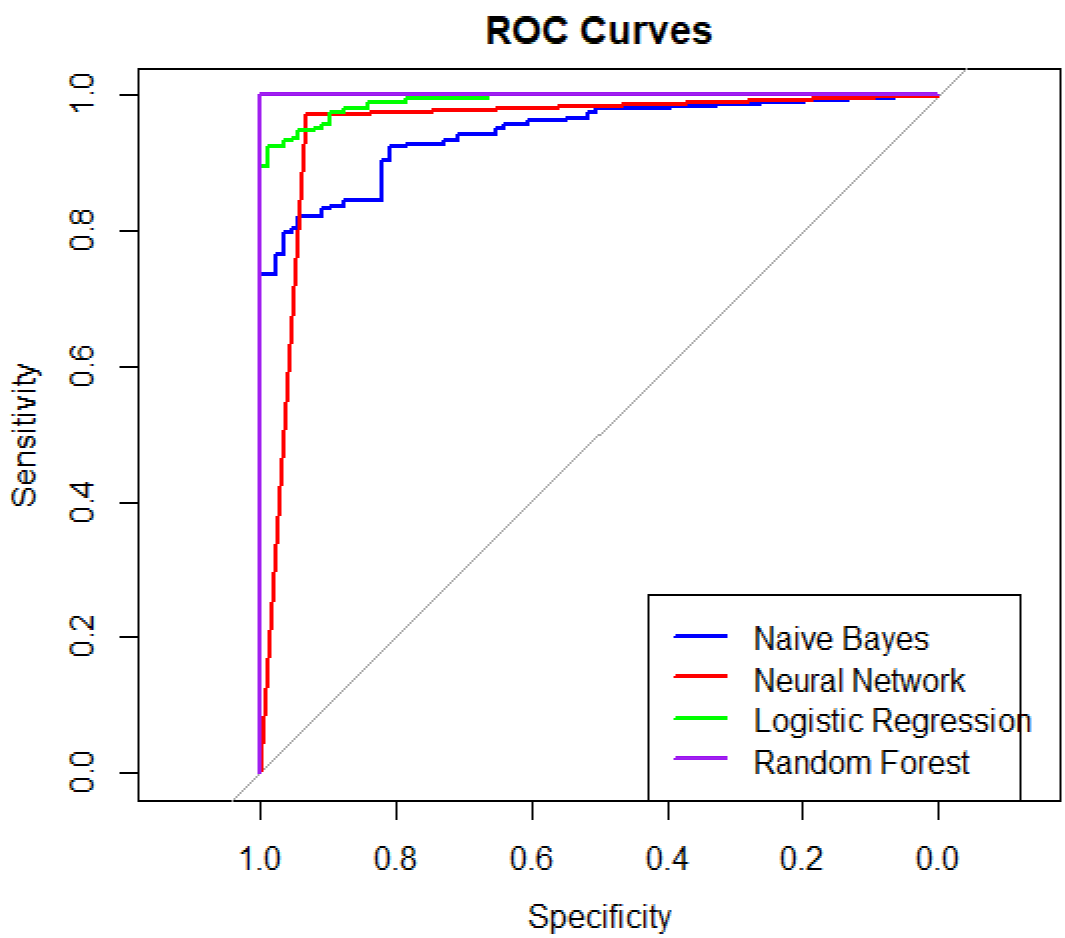

Figure 11 ROC curve of each model for batter selection.

In this context, models with **higher AUC values** effectively capture the critical features of batter selection, such as **average score, strike rate, number of matches played, and boundary-hitting ability (4s and 6s)**, ensuring accurate classification. Conversely, models with **lower AUC values** may struggle to distinguish between selected and non-selected batters due to overlapping performance metrics. By comparing ROC curves, the most optimal model for batter selection can be identified, ensuring a more reliable and data-driven approach in filtering high-impact players.

Conclusion

The analysis of player selection based on predefined performance thresholds for both batters and bowlers highlights key insights into the factors influencing selection decisions in cricket. The use of first-quartile thresholds for bowling metrics and batting performance led to an imbalanced distribution, with a small proportion of players meeting the stringent criteria. For batters, factors such as runs, strike rate, innings played, and high score were significant predictors, while for bowlers, metrics like bowling economy, average, strike rate, and wickets per match were crucial for selection.The exploratory data analysis revealed strong correlations between performance metrics, indicating that consistent and efficient players tend to perform better. The logistic regression and Random Forest models identified significant predictors and demonstrated that performance efficiency, in terms of both batting and bowling, was the dominant factor in selection.The Random Forest model outperformed all other models, achieving perfect classification for both batters and bowlers. Neural Networks also performed well, particularly for batters, while Naive Bayes showed the highest classification errors. These results suggest that advanced machine learning techniques, particularly Random Forest, are highly effective in predicting player selection based on the defined performance criteria.

References

[1]. Himagireesh, C., PV, P., & Taj, T. (2023). Selection of dream-11 players in T-20 cricket by using TOPSIS method. *Management Science Letters*, *13*(4), 257-264.

[2]. Wickramasinghe, I. (2022). Applications of Machine Learning in cricket: A systematic review. *Machine Learning with Applications*, *10*, 100435.

[3]. Chakraborty, S., Mondal, A., Bhattacharjee, A., Mallick, A., Santra, R., Maity, S., & Dey, L. (2024). Cricket data analytics: Forecasting T-20 match winners through machine learning. *International Journal of Knowledge-based and Intelligent Engineering Systems*, (Preprint), 1-20.

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

[5]. Kota, J. S., & Vayelapelli, M. (2020). Predicting the Outcome of a T-20 Cricket Game Based on the Players' Abilities to Perform Under Pressure. *IEIE Transactions on Smart Processing & Computing*, *9*(3), 230-237.

[6]. Puram, P., Roy, S., Srivastav, D., & Gurumurthy, A. (2023). Understanding the effect of contextual factors and decision making on team performance in Twenty20 cricket: an interpretable machine learning approach. *Annals of Operations Research*, *325*(1), 261-288.

[7]. Ashraf, J., Bhatti, S., & Talpur, S. Predicting the Best Team Players of Pakistan Super League using Machine Learning Algorithms. *International Journal of Computer Applications*, *975*, 8887.

[8]. Prakash, C. D. (2016). A new team selection methodology using machine learning and memetic genetic algorithm for ipl-9. *Int. Jl. of Electronics, Electrical and Computational System IJEECS ISSN*.

[9]. Azam, S. N., Haider, B., Hassan, S., & Shakoor, F. (2024). Prediction Of Outcomes Of Extra Deliveries In T-20I Cricket By Using Regression And Various Machine Learning Models.

[10]. SANJAYKUMAR, S., NATARAJAN, S., LAKSHMI, P. Y., ORHAN, B. E., LOBO, J., GEANTĂ, V. A., & NÉMETH, Z. (2024). Predictive analytics for identifying most valuable players in the 2024 T-20 cricket world cup: A comparative study of support vector regression Kernels. *Journal of Physical Education & Sport*, *25*(10).

[11]. Basit, A., Alvi, M. B., Jaskani, F. H., Alvi, M., Memon, K. H., & Shah, R. A. (2020, November). ICC T-20 Cricket World Cup 2020 winner prediction using machine learning techniques. In *2020 IEEE 23rd International Multitopic Conference (INMIC)* (pp. 1-6). IEEE.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)
[12]. Agrawal, S., Singh, S. P., & Sharma, J. K. (2018, November). Predicting results of Indian premier league T-20 matches using machine learning. In *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 67-71). IEEE.
[13]. Amin, G. R., & Sharma, S. K. (2014). Measuring batting parameters in cricket: A two-stage regression-OWA method. *Measurement*, *53*, 56-61.