Predicting Top-Order Dismissals in T20 Internationals: A Machine Learning Analysis of Bowler Type Effectiveness

Harshdeep Singh[1], Monika[2] and Veenu Mangat[3]
University Institute of Engineering & Technology, Panjab University, Chandigarh, India

Abstract

In cricket, it is often observed that some batsmen consistently struggle against certain types of bowlers. This study examines how different patterns in data reveal the influence of a bowler's type on the likelihood of dismissing a batsman. We focus on the top two top-order batsmen from four major cricketing nations other than India. South Africa, England, Australia, and New Zealand are considered to analyze how they have historically performed against various bowling styles, including pace, spin, and variations in arm and delivery. This research will help strategists and coaches to make informed decisions, such as selecting the right bowler for key matchups or adjusting field placements based on predicted outcomes. By combining traditional cricketing knowledge with data-driven insights, this research presents a fresh perspective on how analytics can improve decision-making in modern cricket. The data for this study was collected from datasets of previous ICC T20 World-Cups, recording bowlers and dismissal outcomes for each batter. Using machine learning techniques we have built a predictive model that estimates the likelihood of dismissal based on the bowler's style and batsman who is at the batting crease. This model identifies meaningful patterns, showing that certain bowlers consistently outperform others against specific batsmen.

Keywords: Cricket performance modeling; Machine learning; Bowler classification; Cricket analytics; T20 dismissals

1. Introduction

Cricket, often referred to as a gentleman's game, has evolved tremendously since its origins in 16th-century England. What began as a rural pastime played on village greens has transformed into one of the world's most strategic and passionately followed sports. Over the centuries, cricket spread globally through British colonial influence, taking firm roots in countries like India, Australia, South Africa, the West Indies, and New Zealand. Today, it enjoys an almost religious following in nations like India, where the sport is not just a game but a deeply embedded part of culture and national identity.

Modern cricket is played in three primary formats: Test matches, One Day Internationals (ODIs), and Twenty20 (T20). Test cricket is the oldest and most traditional form of the game. It is played over five days, with each team getting two innings to bat and bowl. The format is known for its emphasis on patience, technique, and endurance. It is considered the ultimate test of a cricketer's skill and mental strength. ODIs are limited-overs matches where each team plays a maximum of 50 overs. Introduced in the 1970s, this format brought a more time-bound, fast-paced version of the game compared to Test matches. ODIs require a balanced approach between aggressive play and strategic planning, making it popular for global tournaments like the ICC World Cup. T20 is the shortest and most explosive format of cricket, with each team limited to 20 overs. Launched in the early 2000s, it quickly gained popularity for its fast-paced action and entertainment value. T20 matches typically last about three hours and are known for big hitting, innovative bowling, and thrilling finishes. This format has brought a new dimension to the sport and is the primary focus of this research[1].

This research focuses exclusively on the T20 format, where matches are limited to 20 overs per side, and every ball can be a potential game-changer. In this high-pressure environment, understanding player weaknesses and exploiting them becomes crucial. Core objectives of our research study are:

- To analyze the influence of bowler types, classified by bowling style and arm orientation on the probability of dismissing top-order batsmen in T20 cricket.
- To identify and examine the top two top-order batsmen from four major cricketing nations: South Africa, England, Australia, and New Zealand.
- To collect and curate structured datasets capturing the bowler types faced by each selected batsman along with their corresponding dismissal outcomes.
- To apply machine learning algorithms for modelling the relationship between bowler types and dismissal patterns for each batsman.
- To predict the most effective bowler types against each selected top-order batsman based on the outcomes of the trained models.

2. Literature Review

In the modern era of T20 cricket, where matches are often decided in a matter of a few overs, the importance of smart, data-driven decision-making has grown tremendously. Teams are continually seeking ways to gain even the smallest tactical advantage, particularly when it comes to neutralizing top-order batsmen who often set the tone for an entire innings. Summary of state-of-the-art studies on match prediction is shown in table 1.

Table 1: Summary of key studies on match prediction, player analytics, and AI in cricket

Ref	Research Focus	Domain / Format	Method(s)	Key Finding	Limitation
[2]	Match score prediction	ODI cricket	Linear Regression, Random Forest	Regularized Linear Regression outperformed Random Forest on prediction error	Relies on numerical data only; excludes pitch conditions and player form as contextual variables
[3]	Shot selection prediction (death overs)	T20 cricket	Ensemble learning, SHAP values (XAI)	Interpretable models using SHAP values enhanced tactical decision-making insights for teams	Manual feature extraction from ball-by-ball data limits scalability
[4]	Player performance prediction (multi-sport)	Cricket + other sports	Linear Regression, Naïve Bayes	Generalized model applicable across multiple sports including cricket	Lacks cricket-format-specific customization; reduced predictive depth for T20 and ODI
[5]	IPL player auction price analysis	IPL (T20)	Regression, Decision Tree	Strong relationship identified between past performance metrics and player auction cost	Ignores qualitative factors such as brand value and team-specific auction strategies

Ref	Research Focus	Domain / Format	Method(s)	Key Finding	Limitation
[6]	ICC T20 World Cup winner prediction	International T20	Random Forest, SVM, KNN, Voting classifiers	Random Forest-based ensemble achieved the best predictive performance in comparative analysis	Uses only pre-tournament data; no live or in-match updating capability
[7]	Match outcome prediction via hierarchical feature engineering	Cricket (general)	Naïve Bayes, hierarchical feature engineering	Hierarchical feature engineering yielded notable improvement in predictive performance	Complex feature engineering demands significant domain expertise, limiting reproducibility
[8]	Ball-by-ball match outcome prediction	Cricket (general)	LSTM, GRU networks (Deep Learning)	Deep learning models effectively captured temporal dependencies with competitive accuracy	Requires large datasets and substantial computational resources

Researchers have focused on ODI match score prediction using regression-based machine learning models. Their study found that Linear Regression with appropriate regularization outperformed Random Forest in terms of prediction error. However, the approach relied heavily on numerical match data and did not incorporate contextual variables such as pitch conditions or player form variations[2].

In literature, authors have explored shot selection prediction during death overs in T20 cricket using ensemble learning and explainable AI techniques such as SHAP values. The study demonstrated that interpretable models can enhance decision-making insights for teams. Nevertheless, manual feature extraction from ball-by-ball data limits scalability[3]. A generalized player performance prediction model applicable across multiple sports, including cricket. Using Linear Regression and Naïve Bayes classifiers, was proposed in literature[4]. However, the lack of cricket-format-specific customization reduced the model's predictive depth for T20 and ODI formats. Authors have analyzed IPL player auction prices using regression and decision-tree models, identifying a strong relationship between past performance and player cost[5]. While insightful, the study did not incorporate qualitative factors such as brand value or team-specific auction strategies.

A comprehensive study to predict the ICC Men's T20 World Cup 2020 winner using multiple machine learning algorithms, including Random Forest, SVM, KNN, and Voting classifiers has been conducted by the researchers[6]. Their comparative analysis revealed Random Forest-based ensembles as the most effective. However, the model relied solely on pre-tournament data and lacked live updating capabilities. Researchers have introduced a hierarchical feature engineering approach for cricket match outcome prediction[7]. Using Naïve Bayes classifiers, the study achieved notable improvements in predictive performance. However, the complex feature engineering process required significant domain expertise, limiting reproducibility.

Authors have applied deep learning techniques such as LSTM and GRU networks for ball-by-ball cricket match outcome prediction[8]. The models captured temporal dependencies effectively and produced competitive accuracy. Nonetheless, the approach required large datasets and extensive computational resources.

3. Methodology

The methodology adopted for this study was carefully designed to align with the specific goals of analyzing batsman dismissals in T20 International cricket, with a special focus on predicting how different types of bowlers impact the dismissal rate of top-order batsmen. As shown in Fig. 3.1, the approach involved multiple stages: dataset selection, consolidation, filtering, manual enrichment, and cleaning, followed by preparation of a final dataset suitable for predictive modelling.

3.1 Selection of Format and Relevance to T20 World Cup

Given the upcoming ICC T20 World Cup, the study was designed exclusively around the T20 format. Unlike Test or ODI cricket, T20 demands faster decision-making and more aggressive play, often turning the outcome of a game within just a few deliveries. Hence, building a model that can predict dismissal probabilities in this format has practical, real-time utility for coaching staff and analysts. The T20 format's pace also makes it an ideal candidate for machine learning applications due to the density of key events per unit time (balls or overs).

3.2 Data Sourcing from CricSheet

The raw match data was sourced from CricSheet, a trusted and widely-used open data resource that provides ball-by-ball data for international and franchise cricket[9]. The initial step involved downloading datasets that specifically included T20 International matches played after 2012. This time range was deliberately chosen to ensure that the data reflects contemporary player forms, strategies, and team compositions. Including outdated matches would have diluted the relevance of the predictions, as player skills, roles, and bowling styles evolve significantly over time. Each downloaded match file represented an individual game. These were then merged programmatically into a unified dataset that captured all necessary interactions between batsmen and bowlers across matches.

3.3 Inclusion of Emerging Player Data

In addition to standard international data, a focused effort was made to supplement the dataset with information about Jake Fraser-McGurk, a promising top-order batsman from Australia. This step was motivated by the recent retirement of veteran opener David Warner, creating a potential vacancy at the top of Australia's batting lineup. Including Fraser-McGurk ensured that the model considers potential future scenarios and remains relevant for upcoming tournaments.

3.4 Filtering and Preprocessing

Following data aggregation, the dataset was filtered down to eight specific top-order batsmen from four countries widely regarded as T20 powerhouses: New Zealand, South Africa, Australia and England. The players included were Finn Allen and Devon Conway from New Zealand, Quinton De Cock and Reeza Hendricks from South Africa, Travis Head and Jake Fraser-McGurk from Australia, and Jos Buttler and Jonny Bairstow from England. These players were selected based on their consistent appearances in top-order positions, their importance in international squads, and their impact in recent T20 series.

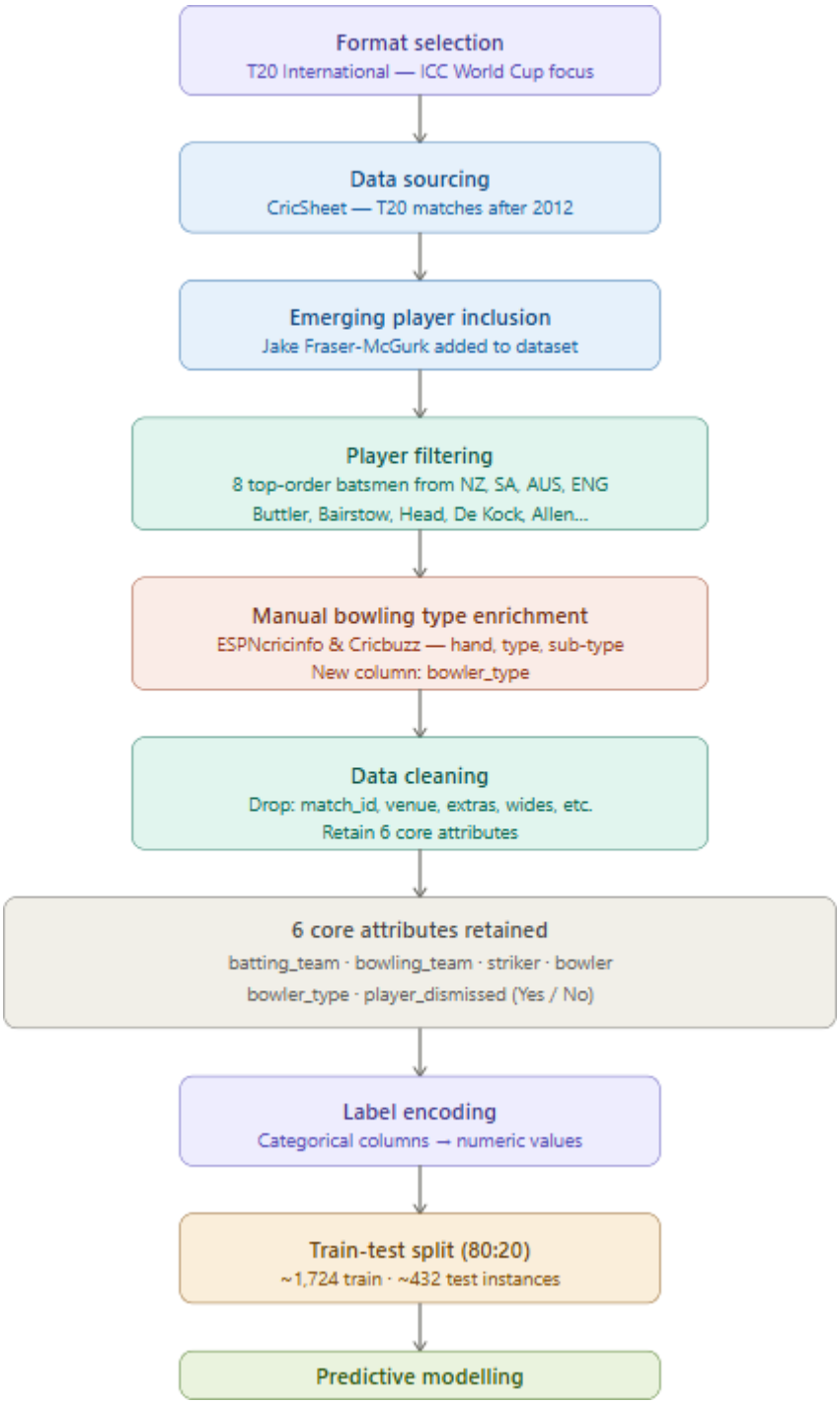

Fig. 3.1 Methodology Flowchart for T20 Batsman Dismissal Prediction

3.5 Manual Extraction of Bowling Types

One of the key contributions of this research was the manual enrichment of the dataset with bowling style information, which was not available in the raw data. To do this, each bowler faced by the selected

batsmen was individually researched using reliable cricket databases such as ESPNcricinfo and Cricbuzz. For each bowler, attributes such as bowling hand (left-arm/right-arm), type (fast, medium, spin), and sub-category (leg spin, off spin, orthodox spin, etc.) were identified and entered into a new column called bowler_type. This was an intensive process, but essential for establishing the relationship between bowling style and dismissal probability. The inclusion of this manually verified data significantly enhanced the analytical value of the dataset and set the foundation for our predictive modelling.

3.6 Data Filtering and Cleaning

To streamline the dataset and remove irrelevant information, several non-essential columns were dropped. Match Metadata columns removed included: match_id, start_date, venue, innings, and ball_no. Auxiliary Data columns removed included: non_striker, runs_off_bat, extras, wides, and noballs.

After cleaning, the dataset retained the following 6 core attributes:

1. batting_team – Team to which the batsman belongs
2. bowling_team – Team to which the bowler belongs
3. striker – Name of the batsman
4. bowler – Name of the bowler
5. bowler_type – Type of bowler
6. player_dismissed – Contains two values: Yes or No

3.7 Label Encoding

Because models require numerical data, the following categorical columns were encoded using Label Encoding:

- batting_team → batting_team_encoded
- bowling_team → bowling_team_encoded
- striker → striker_encoded
- bowler → bowler_encoded
- bowler_type → bowler_type_encoded
- player_dismissed (YES/NO) → target (1/0)

3.8 Splitting the Dataset

A standard 80:20 train–test split was used. Training samples: approximately 1,724 instances. Testing samples: approximately 432 instances. This split ensures generalization and avoids overfitting. This clean, focused dataset laid the groundwork for deeper analysis and predictive modelling.

4. Predictive Models and Metrics Used

Multiple machine learning models are employed in this study to ensure a comprehensive and unbiased evaluation of batsman dismissal prediction in T20 cricket. Since dismissal events are influenced by complex, non-linear interactions among player attributes, match conditions, and bowler characteristics,

using a single algorithm may not adequately capture all underlying patterns. The inclusion of diverse models allows comparison across different learning strategies and helps identify the most effective approach for this classification task.

i. Logistic Regression

Logistic Regression is a widely used statistical classification technique for binary outcome prediction problems. It models the probability of occurrence of an event by fitting data to a logistic function and estimating the relationship between independent variables and the target variable. In classification tasks, Logistic Regression applies the sigmoid activation function to map predicted values between 0 and 1, enabling probabilistic interpretation of outcomes. Due to its simplicity, interpretability, and computational efficiency, Logistic Regression is extensively applied in predictive analytics. In the context of player dismissal prediction, Logistic Regression helps in understanding how different match and player-related features influence the probability of a batsman being dismissed[10].

ii. Random Forest Classifier

Random Forest is an ensemble machine learning method that constructs multiple decision trees during training and outputs the class predicted by the majority of trees. Each tree is trained on a bootstrap sample of the data, and at each split, only a random subset of features is considered. This randomness increases generalization and reduces the risk of overfitting, a common issue in classification tasks involving high-cardinality categorical variables such as bowler names and team identities[11].

iii. XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is an optimized implementation of gradient boosting machines designed for speed, efficiency, and high accuracy. It builds an ensemble of weak decision tree learners, with each new tree correcting errors from previous iterations using gradient descent. It includes features like regularization, parallel tree learning, and handling missing values, making it one of the highest-performing algorithms for structured/tabular datasets[12].

iv. Support Vector Classifier (SVC)

Support Vector Classifier is a supervised machine learning algorithm that constructs an optimal decision boundary (hyperplane) to separate different classes in the feature space. It maximizes the margin between data points of different classes, thereby improving classification reliability. SVC can handle both linear and nonlinear classification problems through the use of kernel functions such as linear, polynomial, and radial basis function kernels. In this study, SVC is applied to distinguish between dismissal and non-dismissal events based on multiple performance and contextual features[13].

v. Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming that features are independent and normally distributed. It estimates the probability of each class by calculating the likelihood of the data belonging to a particular class and selecting the class with the highest posterior probability. Despite its simplicity, Gaussian Naive Bayes is computationally efficient and performs well on large datasets. In this research, Gaussian Naive Bayes is used as a baseline probabilistic model to

evaluate and compare dismissal prediction performance against more complex machine learning algorithms[14].

Metrics Used

Precision, recall, and F1 score are essential evaluation metrics in classification problems because accuracy alone can often give a misleading picture of a model's performance, especially when the dataset is imbalanced.
Precision: Precision measures how many of the predicted positive cases are actually correct. Mathematically, it is represented as follows [15]:

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

Where,
TP: True Positives
FP: False Positives

Recall (Sensitivity): Recall measures how many actual positive cases were correctly identified. Mathematically, it is represented as follows [15]:

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

Where,

TP: True Positives
FN: False Negatives

F1 Score: F1 Score is the harmonic mean of precision and recall. It is computed as [15]:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

5. Results and Discussions

The predictive performance of all classification models was primarily assessed using overall accuracy, as it provides a direct and interpretable measure of how correctly the models classify batsman dismissal outcomes. Accuracy was computed as the proportion of correctly classified instances across both dismissal and non-dismissal classes.

5.1 Discussion on Accuracy Dominance

As shown in Fig 3.2, we get the idea that XGBoost classifier came out to be the winner in the game of predicting accurately. It achieved highest accuracy score of 0.91, followed by Random Forest, the runners up. These two models outperformed the rest in terms of model accuracy comparision. Logistic Regression achieved an accuracy score of 0.53, SVC achieved 0.61 and Gaussian NB managed to achieve 0.60, this shows that these models were not as reliable as XGBoost and Random Forest in terms of Model Accuracy.

In contrast, Support Vector Classifier (SVC) and Gaussian Naive Bayes achieved moderate accuracies of 0.61 and 0.60, respectively. These models demonstrate limited effectiveness in modeling the underlying

data distribution. The Logistic Regression model performed the worst, with an accuracy of 0.53, suggesting that linear decision boundaries are insufficient for this classification task. Overall, the results highlight that ensemble-based methods (XGBoost and Random Forest) provide substantial improvements in predictive performance compared to traditional linear and probabilistic approaches.

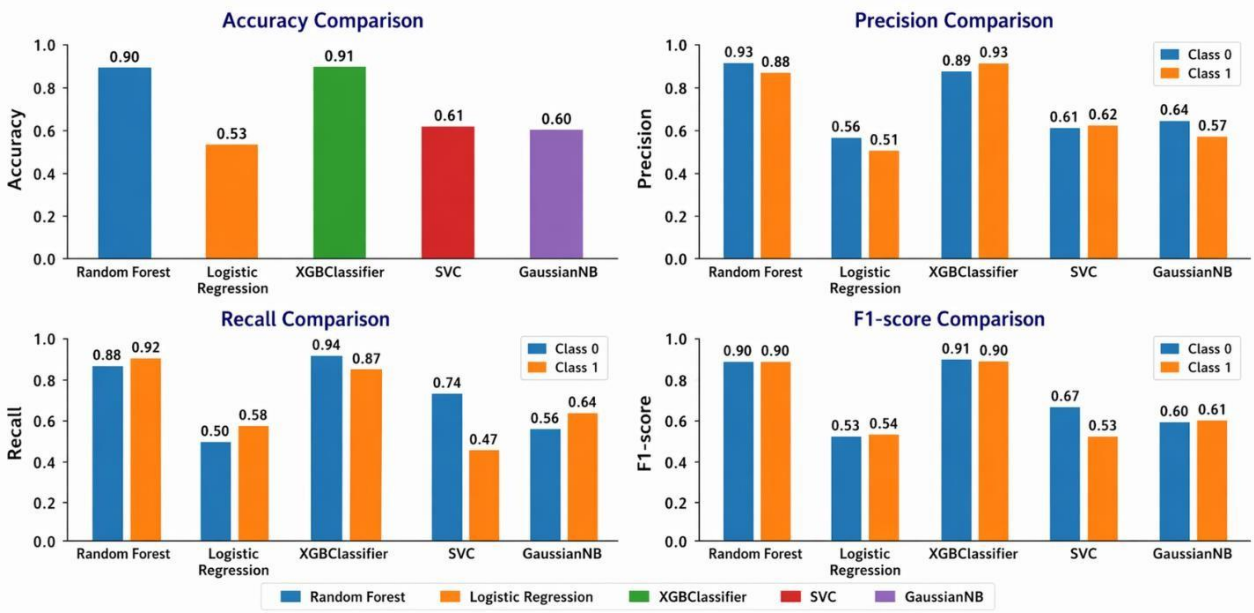

Fig. 3.2 Comparative Analysis of Results

5.2 Class-wise Performance Interpretation

Accuracy scores are largely driven by the correct classification of Class 0 (Not Dismissed) instances, which form the majority of the dataset and Class 1 (Dismissal). Claasification report for Random Forest, Logistic Regression, XG-Boost and SVC is shown in table 2.

Table 2: Classification Report for All Models

Algorithm	Class	Precision	Recall	F1-Score
Random Forest	0	0.93	0.88	0.90
Random Forest	1	0.88	0.92	0.90
Logistic Regression	0	0.56	0.50	0.53
Logistic Regression	1	0.51	0.58	0.54
XG-Boost	0	0.89	0.94	0.91
XG-Boost	1	0.93	0.87	0.90
SVC	0	0.61	0.74	0.67

Algorithm	Class	Precision	Recall	F1-Score
SVC	1	0.62	0.47	0.53
Gaussian Naive Bayes	0	0.64	0.56	0.60
Gaussian Naive Bayes	1	0.57	0.64	0.61

We have observed that Random Forest exhibits a balanced performance for the two classes (0 and 1), Precision scores for class 0 and 1 were 0.93 and 0.88 which indicate a strong classification capability. Equal F1 score assures that the model was unbiased towards any class. XGboost however shows a different story when compared to the first. Despite having similar F1 score to Random Forest, Precision score for class 0 is 0.89 and for class 1 is 0.93. Thus, accuracy for Class 1 i.e. dismissal is higher for XGBoost when compared with Random Forest.

On the contrary, Logistic Regression, SVC and Gaussian NB fail to achieve the performance level of the other two. Logistic Regression achieves an accuracy score of 0.56 for class 0 and 0.51 for class 1 with F1 score of 0.53 and 0.54 respectively,indicating that the model was least reliable among all 5. Similar case can be seen with Gaussian NB. Gaussian Nb manages only 0.64 accuracy for class 0 and 0.57 accuracy for class 1 with F1 score of 0.60 and 0.61 respectively, concluding that it also fails to be robust, unbiased and reliable for the prediction

SVC shows an interesting trend which can be observed by comparing the Recall Values of the two classes. SVC scores an accuracy of 0.61 for class 0 and 0.62 for class 1. However, if we observe the Recall values of the two, for class 0 it manages a value of 0.74 whereas for class 1 it has a value of 0.47 which shows that the model is biased towards class 0 i.e. no dismissal which is also our majority class.

6. Conclusion and Future Scope

In conclusion, the present study on cricket analytics incorporating bowling type demonstrates that ensemble learning techniques, particularly Extreme Gradient Boosting and Random Forest, significantly outperform traditional machine learning models in terms of predictive accuracy, class balance, and overall robustness. The superior performance of these models indicates their strong capability to capture complex, non-linear relationships inherent in cricket datasets, such as variations in bowling styles, match situations, and player performance dynamics. While simpler models like Logistic Regression, Support Vector Classifier, and Gaussian Naive Bayes provide baseline insights, they fall short in effectively modeling the intricate interactions between features, thereby limiting their predictive reliability. The findings reinforce the importance of selecting advanced ensemble techniques when dealing with sports analytics problems where contextual and dynamic variables play a crucial role.

Looking ahead, the scope of this research can be significantly expanded beyond the current T20-based analysis. Future studies can apply these models to other cricket formats such as One Day Internationals (ODIs) and Test matches, where game dynamics, player strategies, and temporal dependencies differ

substantially, offering richer avenues for analysis. Additionally, these models can be integrated into real-world applications such as performance prediction systems for leagues like the Indian Premier League, where team management and analysts can leverage predictive insights for player selection, match strategy, and opposition analysis. With the increasing adoption of data-driven decision-making in sports, such models also hold potential for use in upcoming global tournaments, including the ICC Men's T20 World Cup 2028 and even the Cricket at the 2028 Summer Olympics, where cricket is expected to gain broader international visibility. Furthermore, future work may involve incorporating real-time data streams, deep learning architectures, and advanced feature engineering techniques to enhance predictive accuracy and adaptability. Thus, this research not only establishes a strong analytical foundation but also opens pathways for scalable, real-world deployment of machine learning in modern cricket analytics.

Declaration: The authors declare that they have no relevant financial or non-financial/ competing interests to disclose in any material discussed in this article

References

1. D. Sudarsanan, R. Karthikeyan, and S. Balaji, "Prediction of ICC T20 Cricket World Cup 2024 Using Machine Learning Techniques," International Journal of Advanced Research in Computer Science, vol. 15, no. 2, pp. 55–62, 2024.
2. M. S. Hossain, A. Rahman, and T. Ahmed, "One Day International Cricket Match Score Prediction Using Machine Learning Approaches," International Journal of Data Analytics and Artificial Intelligence, vol. 3, no. 1, pp. 25–34, 2024.
3. D. K. Sahoo, S. Mohanty, and P. Das, "Prediction of Shot Selection of a Batsman in the Death Over of a T20 Cricket Match Using Machine Learning," Procedia Computer Science, vol. 225, pp. 120–129, 2024.
4. M. Rajeswari and R. Srinivasan, "Player Performance Prediction in Sports Using Machine Learning," International Journal of Engineering Research & Technology (IJERT), vol. 13, no. 1, pp. 98–103, 2024.
5. P. Nagaraj, K. R. Prasad, and S. Reddy, "IPL Players Cost and Pay Prediction Using Machine Learning Techniques," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 9, no. 2, pp. 345–352, 2023.
6. S. Singh, Y. Aggarwal, and K. Kundu, "Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction Using Machine Learning," International Journal of Computer Applications, vol. 176, no. 32, pp. 46–52, June 2020.
7. A. Kampakis and W. Thomas, "Predicting the Outcome of English County Twenty Over Cricket Matches Using Machine Learning," Journal of Quantitative Analysis in Sports, vol. 11, no. 4, pp. 161–173, 2015.
8. A. Kaluarachchi and A. Aparna, "Cricket Match Outcome Prediction Using Machine Learning," International Journal of Computer Science and Information Technologies, vol. 7, no. 2, pp. 565–570, 2016.
9. https://cricsheet.org
10. A. Satwani, R. Patel, and M. Shah, "Live Cricket Predictions for Runs and Win Using Machine Learning," International Journal of Computer Applications, vol. 185, no. 7, pp. 14–21, 2024.
11. S. Kumar, A. Verma, and R. Mehta, "Analysis and Winning Prediction in T20 Cricket Using Machine Learning," International Journal of Creative Research Thoughts (IJCRT), vol. 11, no. 4, pp. 890–896, 2023.
12. S. Sharma and P. Gupta, "Cricket Analytics: ODI Match Score Prediction and Resource Metric Modeling Using Machine Learning," International Journal of Data Science and Analytics, vol. 10, no. 3, pp. 211–220, 2022.

13. R. Goel, S. Gupta, and A. Jain, "Ball-by-Ball Cricket Match Outcome Prediction Using Deep Learning," Proceedings of the IEEE International Conference on Computational Intelligence and Communication Technology (CICT), Ghaziabad, India, 2018, pp. 1–6.

14. R. Passi and N. Pandey, "Data Analysis and Machine Learning in Cricket," International Journal of Sports Science and Engineering, vol. 11, no. 3, pp. 123–130, 2017.

15. N. Passi, S. Jain, and A. Mishra, "Player Performance Evaluation in Cricket Using Random Forest Algorithm," International Journal of Computer Applications, vol. 167, no. 5, pp. 10–15, 2017.