Expert Systems With Applications 259 (2025) 125178

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Forecasting bowler performance in One-Day International cricket using Machine learning

Rameshwari Lokhande [*], Rawal N. Awale, Rahul R. Ingle

Veermata Jijabai Technological Institute, Mumbai, India

ARTICLE INFO

Keywords:
Machine learning
Artificial Intelligence
Prediction analysis
Cricket
Bowling performance

ABSTRACT

In cricket's dynamic environment, bowlers' performances are a key factor in determining success. In this dynamic sport, our study explores the field of bowler performance predictive analytics to enable decision-makers and strategists. Using advanced machine learning techniques, we provide a comprehensive analysis targeted at predicting and analyzing the complex factors impacting bowlers' performance on the field. This research methodology is based on the sophisticated use of machine learning algorithms to develop a reliable prediction model. Our study reveals an abundance of information about the complex interactions between these diverse factors and how they affect bowler performance. Specifically, we highlight the important impact that opposing dynamics and contextual factors like venue-specific performance trends play, emphasizing the necessity of flexible tactics that depend on contextual circumstances. Prediction methods have significant ramifications not only in cricket but also in other fields. They provide actionable insights for player selection, strategic planning, and ongoing performance evaluation, making them indispensable tools for cricketing companies. Moreover, our study broadens the scope of predictive analytics and holds potential for use in a variety of sports and sectors that depend on complex strategic decision-making. This research demonstrates the critical role that predictive analytics plays in cricket. It offers a rigorous model for predicting and understanding the complex dynamics of bowler performance, greatly enhancing strategic decision-making within the game and expanding its potential into other areas.

1. Introduction

Cricket has undergone a tremendous metamorphosis propelled by advanced analytics and data-driven insights in the last several years. It's impossible to overestimate the importance of bowlers in One-Day International (ODI) cricket, since they control match results by maintaining a careful balance between bat and ball. Through accurate bowler performance prediction, teams have actively pursued competitive advantages.(Bhattacharjee & Saikia, 2014; Chaudhary et al., 2019; Mittal et al., 2021; Passi & Pandey, 2018). This study aims to investigate how machine learning methods may be used to predict bowler performance in ODI cricket. We want to construct predictive models that can foretell a bowler's efficiency in particular match circumstances by utilising the large pool of historical cricket data that is readily available from reliable sources. We want to identify the characteristics that greatly influence bowler success and contribute to their performance in critical areas like wicket-taking ability, economy rate, and changes in bowling style by leveraging the power of machine learning algorithms(Hermanus H. Lemmer, 2008, 2014; McGrath et al., 2019). The gathering and thorough processing of cricket data, including in-depth statistics of bowlers from numerous ODI matches, constitutes the initial stage of this study. To ensure data accuracy and relevance, we carefully manipulate the data and engineer the features. Then, in order to create prediction models for predicting bowling performance, we use cutting-edge machine learning techniques. The relevance of this study rests in its potential to revolutionize cricket analytics by providing coaches, pickers, and teams with useful information that will help them make wise judgments. With the help of the established models, match strategies, the best bowling lineup, and the performance of specific bowlers in various match conditions can all be predicted. Our goal as we explore the complex field of cricket analytics is to close the gap between traditional cricket analysis and cutting-edge data-driven techniques. This study adds to the increasing body of knowledge in sports analytics and has applications not only to cricket but to other sports where the use of predictive modeling can improve performance and influence tactical choices. We anticipate a paradigm change in cricket analytics that will influence the

[*] Corresponding author.
E-mail addresses: ralokhande_p19@el.vjti.ac.in (R. Lokhande), rnawale@el.vjti.ac.in (R.N. Awale), rringle@el.vjti.ac.in (R.R. Ingle).

https://doi.org/10.1016/j.eswa.2024.125178
Received 10 March 2023; Received in revised form 22 July 2024; Accepted 21 August 2024
Available online 30 August 2024

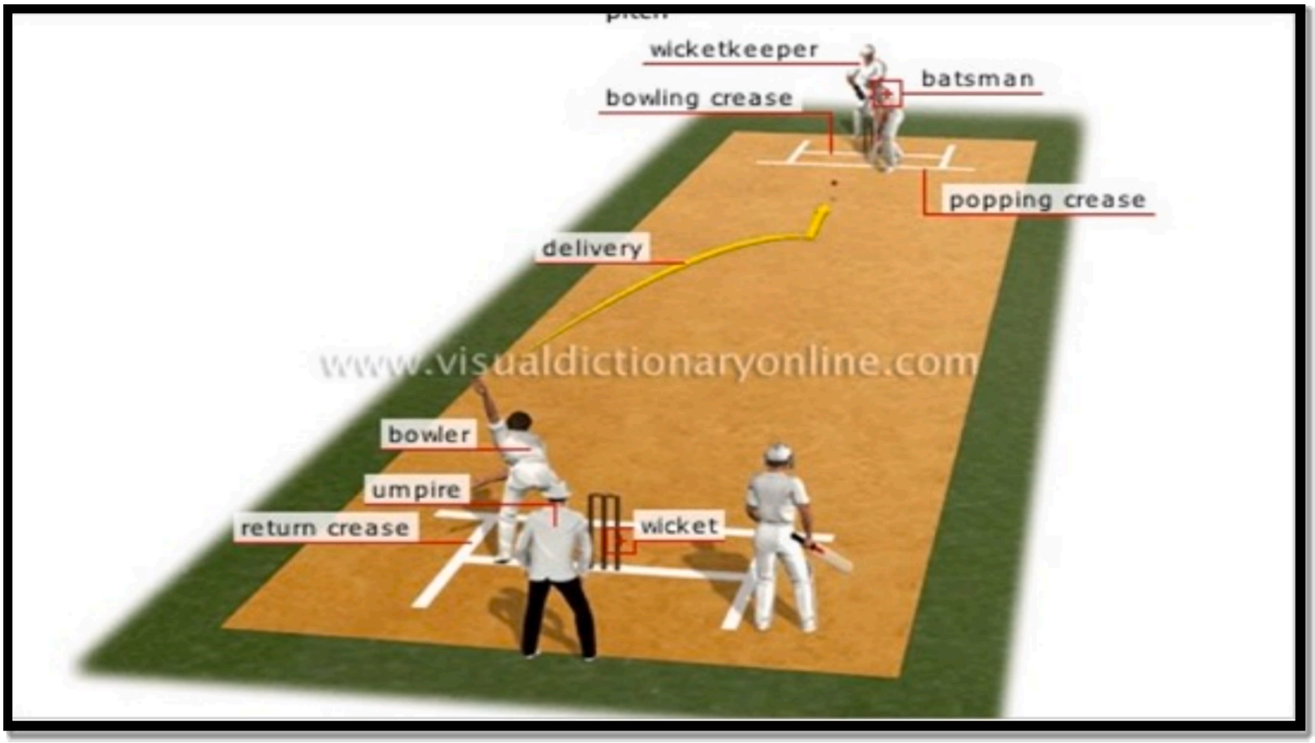

Fig.1.1. Block diagram.

course of the game by investigating the untapped potential of machine learning in predicting bowler performance. By analyzing the intricate interplay of various factors affecting bowler performance, we seek to provide cricket teams, coaches, and selectors with valuable insights for strategic planning and player selection(Bhattacharjee & Pahinkar, 2012; H.H. Lemmer, 2002).

Fig. 1.1 provides a detailed view of a cricket pitch from a bowler's perspective, highlighting key positions and markings crucial for delivering the ball. The bowler is depicted behind the bowling crease, ready to deliver the ball toward the batsman. The trajectory of the ball is shown with a yellow curved line, emphasizing its path from the bowler to the batsman. The bowling crease is a critical line where the bowler must release the ball before overstepping, while the return creases, which are perpendicular to the bowling crease, indicate the area within which the bowler's back foot must remain during delivery.

In addition to the bowling and return creases, the image also marks the popping crease, a line parallel to the bowling crease that the batsman must reach to be safe from being run out. The pitch, the central strip of the field where most of the action occurs, is 22 yards long and 10 feet wide. Other key elements include the wicket, which the bowler aims to hit, and the positions of the wicketkeeper and umpire, who play essential roles in the game. This illustration effectively captures the critical aspects a bowler must consider, such as the delivery path, positioning, and the importance of the various creases on the pitch.

The research begins with a meticulous data collection and preparation process, involving data wrangling, feature engineering, and handling missing values to ensure data integrity and completeness. We then delve into the realm of machine learning, leveraging algorithms like Decision Trees, Random Forests, Support Vector Regression, and Gradient Boosting, to develop accurate and robust predictive models. The potential impact of this research extends far beyond the cricket field. Forecasting bowler performance using machine learning techniques can serve as a paradigm for predictive analytics in other sports domains, guiding decision-making and performance optimization in various athletic disciplines. As cricket continues to evolve into a data-centric sport, our study adds to the growing body of knowledge in sports analytics, where data-driven insights and predictive modeling hold the key to success. By unraveling the complex patterns and factors influencing bowler performance in ODI cricket, we aim to revolutionize cricket analytics and contribute to the ongoing transformation of the game.

Ultimately, we envision our research advancing the understanding of sports performance prediction, benefitting not only cricket but also the broader realm of sports analytics and data-driven decision-making. This study represents a revolutionary advance in the use of machine learning to forecast bowler performance in One Day International cricket matches. By providing teams, coaches, and analysts with accurate information for strategic planning and well-informed decision-making, our technique transforms cricket analytics. This research has broad implications across several fields by combining modern analytics and machine learning. Especially, it explores previously uncharted territory by revealing bowler performance on specific venue and opponents—something that has never been done in previous research that has usually only included player forecasts based on conventional qualities. This effort seeks to improve team performance, modify player selection tactics, and adjust match dynamics outside of sports. Understanding these performance characteristics is important not just for cricket but also for other domains where using state-of-the-art analytics and machine learning techniques might impact decision-making approaches.

2. Literature review

Machine learning is an application of artificial intelligence that involves the use of algorithms to learn patterns and insights from data without being explicitly programmed. Its main objective is to address real-world problems by making predictions or identifying patterns based on past data. This technology is so ubiquitous that we might not even realize its presence in everyday life. By utilizing mathematical models, heuristic learning, information acquisition, and decision-making trees, machine learning provides the power to make accurate predictions, gain recognition, and create resilient systems. In the field of sports analytics,

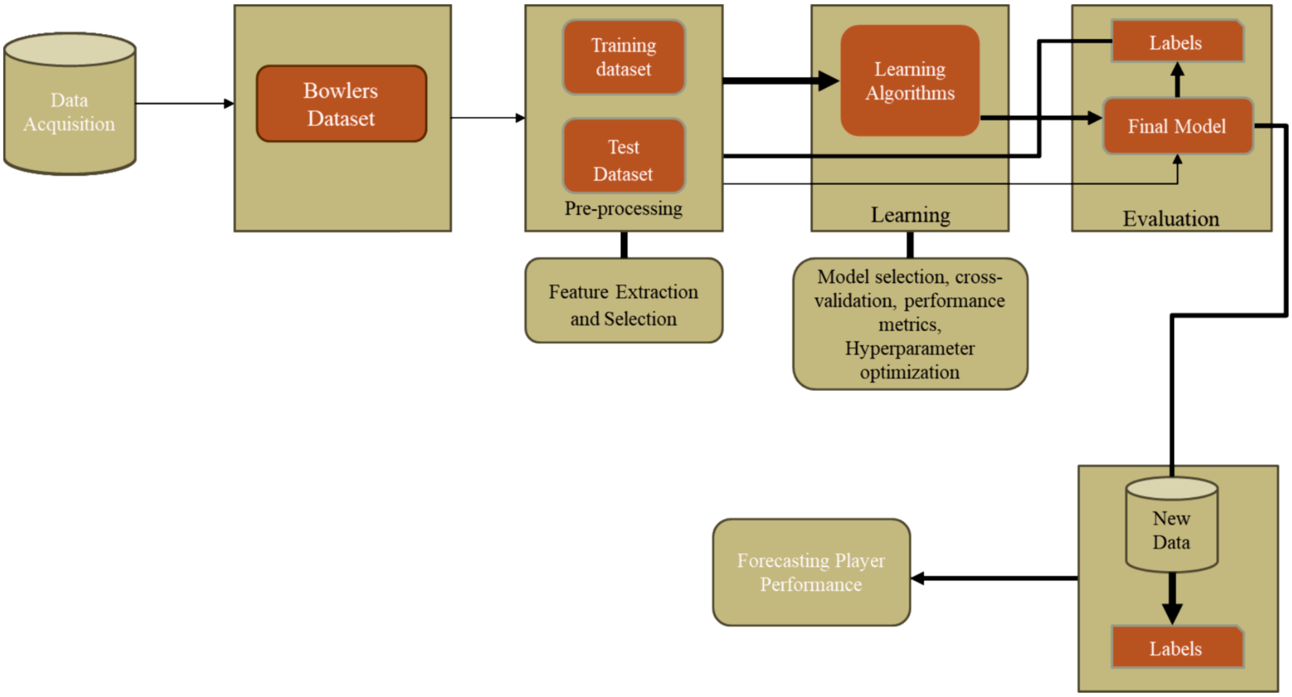
Data Acquisition
Bowlers Dataset
Training dataset
Test Dataset
Pre-processing
Learning Algorithms
Learning
Labels
Final Model
Evaluation
Feature Extraction and Selection
Model selection, cross-validation, performance metrics, Hyperparameter optimization
New Data
Labels
Forecasting Player Performance

Fig.3.1. Block diagram.

historical data, and player attributes are commonly used to assess and predict players' performance. In cricket, for instance, factors such as batting average, strike rate, number of centuries, as well as a player's position, batting or bowling hand, and previous performances, are all taken into account when predicting their ability and performance in an upcoming match(Bunker & Thabtah, 2019; Kapadia et al., 2019; Nevill et al., 2008). By analyzing this data, teams can make informed decisions about team selection and strategy. Machine learning algorithms can be applied to this data to build predictive models that can forecast player performance with greater accuracy(Kumar & Roy, 2018; Mittal et al., 2021; Singh & Kaur, 2017). Various data points are considered to assess a player's performance, including their average, strike rate, number of hundreds, batting/bowling hand, position, and past performance against specific teams and at certain locations. All of these factors are taken into account to project their performance in upcoming matches. In recent years, cricket match outcomes have been extensively studied, with focus on factors such as home advantage, past performances, and player form. Traditional models typically predict outcomes based on pre-match data or static player metrics. However, this study introduces a dynamic model that predicts match outcomes during live play. By considering factors like wickets fallen, venue, and target score, the model provides real-time predictions for both innings using a combination of Linear Regression, Q-Learning, and Naïve Bayes classifiers. This approach offers a more accurate, in-progress prediction compared to previous static models (Lokhande and Chawan (2018), Lokhande et al. (2018)). Machine learning techniques can be applied to analyze this data and make accurate predictions. Forecasting bowlers' performances using machine learning and data mining is a promising approach in cricket. The use of advanced statistical techniques and algorithms can help to analyze the large amounts of data available and make accurate predictions about individual player performances. In this particular study, the researchers compared the performance of four supervised machine learning methods to forecast bowlers' performances in a specific match. This indicates that machine learning can be used to supplement traditional methods of player selection and provide more accurate and data-driven decision-making in cricket. Fast bowlers, fast-medium bowlers, and medium-fast bowlers are compared(Malhotra & Krishna, 2018). They demonstrate that fast bowlers are superior to the other bowling categories. They suggest a dynamic bowling rate (DBR), similar to the Combined bowling rate (CBR). Bowler's average of harmonic mean, a strike rate of bowler, and the economy rate are used to calculate DBR (Bhattacharjee & Pahinkar, 2012). use a linear programming technique to present a data envelopment analysis (DEA) approach. They rate players, such as batsmen and bowlers, based on statistics from the IPL(Indian premier league) 4(Mukherjee, 2014). The output ranges from 0 to 1, with a number near to 1 indicating a high chance of winning and a value close to 0 indicating a higher chance of losing. The next section explains one such metric that may be used to assess a bowler's performance. The fourth part examines the variables that may influence bowler performance in one-day international cricket.

2.1. Performance metrics of the bowlers

Bowlers' performance has traditionally been measured using several metrics such as bowlers' average, economy rate, and strike rate.

(i) Bowling average: The total number of runs given by bowler per wicket.

$$\text{Bowling average} = \text{AVG}_{\text{BOWL}} = \frac{\text{Number of runs given by bowler}}{\text{Number of wickets taken by bowler}} = \frac{\text{runs}}{\text{wickets}} \tag{1}$$

Where r = number of runs given by bowler w = total number of wickets taken by the bowler

(ii) Economy Rate: This metric represents the average number of runs a bowler has given up each over. This rate is crucial in determining a bowler's success, especially in limited-overs cricket.

$$\text{Economy Rate} = \text{ECON}_{\text{BOWL}} = \frac{\text{Number of runs given}}{\text{Total number of balls bowled}} \times 6 \tag{2}$$

Where b is the total number of balls bowled by the bowler.

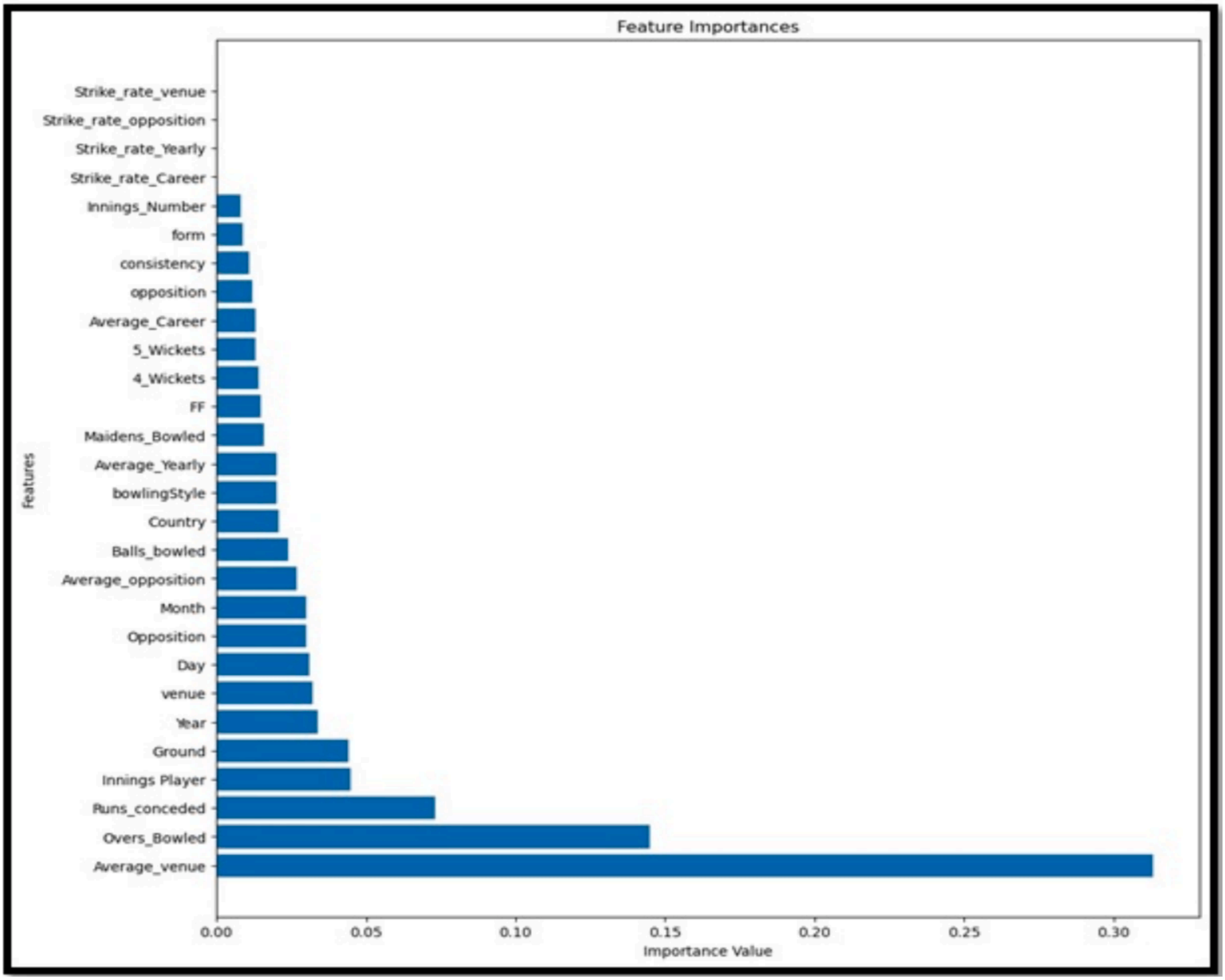

Fig.3.2. Feature Importance for Bowlers.

$$\text{Bowlers Strike Rate} = \text{SR}_{\text{BOWL}} = \frac{\text{the total number of balls delivered}}{\text{Number of wickets taken}} = \frac{\text{balls}}{\text{wickets}} \tag{3}$$

(iii) Bowler's Strike Rate: To get a wicket, a bowler must bowl a certain number of balls. In mathematics, the number of deliveries delivered to the total number of wickets taken is the ratio.

However, it is commonly acknowledged that such statistics have significant limits in measuring a player's genuine ability. It is also difficult to mix conventional measurements since they are in distinct units of measurement. All of these constraints in measuring cricketer performance have been thoroughly examined. Thus, a statistic termed the combined bowling rate (CBR) is used to assess bowler performance (Koulis et al., 2014). This metric is used to evaluate bowlers' performance by combining the three standard metrics stated above. Lemmer (Bhattacharjee & Pahinkar, 2012; Hermanus H. Lemmer, 2014) attempted to aggregate these three rates into a single index. He utilized the harmonic mean of I (ii), and (iii) as a model (iii). In the case of rates and ratios, the harmonic mean is a suggested averaging measure. However, for a harmonic mean, all of the values to be averaged must have the same numerator. The numerator in both I and (ii) is 'Total runs conceded.' It was suggested that (iii) be worded as follows:

$$\text{SR}_{\text{BOWL}} = \frac{\text{balls}}{\text{wickets}} = \frac{\text{runs} * \text{balls}}{\text{runs} * \text{wickets}} = \frac{\text{runs}}{\frac{\text{runs}*\text{wickets}}{\text{balls}}} \tag{4}$$

As a result, the CBR is calculated as follows:

$$\text{CBR} = \text{Harmonic mean}(\text{AVG}_{\text{BWL}}, \text{ECON}_{\text{BWL}}, \text{SR}_{\text{BWL}}) = \frac{3}{\frac{\text{wickets}}{\text{runs}} + \frac{\text{runs}*\text{wickets}}{\text{runs}*\text{balls}} + \frac{\text{balls}}{6\text{runs}}} = \frac{3\text{r}}{\text{wickets} + \frac{\text{runs}*\text{wickets}}{\text{balls}} + \frac{\text{balls}}{6}} \tag{5}$$

It should be emphasized that the lower the CBR number, the better the bowler.

3. Data wrangling

The process of developing precise and dependable prediction models for the purpose of forecasting cricket player performance using machine learning is illustrated in Fig. 3.1. Comprehensive player performance data will be gathered in the stage from reliable cricket sources such as cricinfo.com. Important player metrics like strike rates, economy rates, bowling averages, and more will be included in this collection. Several methods of data pretreatment will be used to guarantee the dataset's consistency and quality. These methods will deal with problems like values in duplicate entries and data format standardization for uniformity. To increase the predictive models' efficacy, a feature significance analysis will also be carried out. The models can concentrate on these important elements by determining the player traits that have a significant impact on performance through this study. Moreover, tweaking the hyperparameters is crucial to improving the machine learning models. It is possible to greatly increase each model's accuracy and generalization capacity by carefully choosing and adjusting a set of hyperparameters. This painstaking hyperparameter tuning guarantees that the models are customized to offer player performance forecasts.

In our study, we collected a comprehensive dataset of 870 bowlers from the renowned cricket statistics website, https://www.espncricinfo.com/ from 2000 to 2019. The dataset comprises around 100,000 data points, encompassing various attributes related to bowling performance. However, before conducting our analysis, we recognized the importance of data cleaning to ensure the accuracy and integrity of our results. During the data cleaning phase, we focused on removing any inconsistencies, duplicates, and irrelevant data. Additionally, we handled missing values and performed feature engineering to enhance the dataset's quality and suitability for further analysis. We meticulously imputed missing data using appropriate techniques, ensuring minimal information loss and preserving the dataset's integrity. To enhance the predictive power of our model. Robust feature engineering approaches were also used to improve the quality and relevance of the dataset for further research. The results obtained from several machine learning algorithms will be provided in more depth throughout the training phase, providing a thorough comparison and explanation of their advantages and disadvantages in terms of accurately forecasting bowling performance. By directly connecting these methods to well-established sports analytics theories, we have reinforced the rationale behind our selected methodology and demonstrated its applicability to cricket analytics. Additionally, the methods used to reduce overfitting will be highlighted, along with how regularization and hyperparameter tweaking support the robustness and generalizability of the model. Interestingly, out of a starting set of 26 features, the study focused on the top 15 features that had a substantial impact on a bowler's performance. Additionally, the dataset—which was originally provided in YAML format was converted into CSV format using the NumPy and Pandas libraries to enable smooth data processing and modification, guaranteeing the dataset's appropriateness for analysis and integrity. For feature selection, Random Forest—a robust ensemble learning technique—was used because of its capacity to handle high-dimensional data, control multicollinearity, and pinpoint the most significant predictors. By combining many decision trees, this technique makes it possible to rank the features according to how much they improve prediction accuracy. When deciding which features to include in a tree, Random Forest determines how important a feature is based on how much each variable reduces impurity or increases information gain. Features that are more essential are those that reliably improve prediction accuracy over a larger set of trees.

The relative relevance of characteristics as examined by the Random Forest method is displayed in the Fig. 3.2. It presents the importance of every predictor variable in affecting cricket players' performance, which helps identify the critical elements determining player success. Through this analysis, we identified the top features that significantly influenced a bowler's performance. These features included Average_venue, Balls_bowled, Runs_conceded, Average_opposition, 4_Wickets, Average_Yearly, Innings Player, Ground, Average_Career, Year, Month, Day, Country, bowlingStyle, Maidens_Bowled, FF, form, 5_Wickets, Innings_Number, and Wickets_Taken are explained as follows.

- **Average_venue:** This feature represents the average performance of the bowler at a particular cricket venue. It indicates how well the bowler performs on average at different grounds.
- **Balls_bowled:** This feature denotes the total number of deliveries bowled by the player during matches. It reflects the bowler's workload and the number of opportunities they had to take wickets.
- **Runs_conceded:** This feature represents the total number of runs given away by the bowler during their bowling spells. It reflects the bowler's ability to contain the opposition batsmen.
- **Average_opposition:** This feature signifies the average performance of the bowler against different opposition teams. It provides insights into how the bowler's performance varies against various batting line-ups.
- **4_Wickets:** This feature records the number of matches in which the bowler took four wickets in a single innings. It reflects the bowler's ability to take crucial wickets in a match.
- **Average_Yearly**: This feature represents the bowler's average performance in a calendar year. It allows analysis of the bowler's yearly consistency and performance trends.
- **Innings Player:** This feature denotes the number of matches in which the bowler played an innings. It reflects the bowler's participation in matches.
- **Ground:** This categorical feature represents the cricket ground where the matches were played. It provides information on the specific venues where the bowler performed.
- **Average_Career:** This feature indicates the overall career average of the bowler. It reflects the bowler's performance across their entire cricket career.
- **Year, Month, Day:** These features represent the year, month, and day when the matches were played. They allow the analysis of any temporal patterns in the bowler's performance.
- **Country:** This categorical feature denotes the country where the matches were played. It provides information about the countries where the bowler has performed.
- **BowlingStyle:** This feature describes the style of bowling used by the player (e.g., fast bowler, spin bowler). It reflects the bowler's preferred technique.
- **Maidens_Bowled:** This feature records the number of maidens (overs with no runs conceded) bowled by the player. It indicates the bowler's ability to create pressure on the batsmen.
- **FF:** Represents number of innings a player has taken more than 3 wickets.
- **Form:** This feature represents the form or recent performance of the bowler. It allows the analysis of how recent performances impact the bowler's overall performance.
- **Innings_Number:** This feature denotes the innings number in a match where the bowler played. It provides information on the bowler's position in the game.
- **Wickets_Taken:** This feature records the total number of wickets taken by the bowler in all matches. It is the target variable for our performance prediction model.

4. Methodology

4.1. Analysing the performance of predictive statistical models

In our study, we employed various regression algorithms, including Decision Tree Regressor, Random Forest Regressor, SVR (Support Vector Regressor), Adaboost Regressor, XGBoost, and LightGBM, to predict bowler performance. These algorithms were selected based on their widespread use and proven effectiveness in handling regression tasks, making them suitable candidates for predicting numerical values such as bowler performance metrics. To evaluate the performance of our models, we utilized several key metrics: MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R2 (R-squared). Each metric serves a specific purpose in assessing the quality and accuracy of our predictions. The mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors that is, the average squared difference between the estimated values and the actual value. RMSE, on the other hand, calculates the average difference between the predicted and actual values, penalizing larger errors more than smaller ones. This metric allows us to understand how well our predictions align with the ground truth values. MAE complements RMSE by providing an average of the absolute errors, making it less sensitive to outliers and providing a more robust representation of prediction accuracy. Finally, R2 evaluates the goodness of fit of our regression models, taking into account the number of features and avoiding overfitting.

By utilizing these evaluation metrics, we aimed to gain a comprehensive understanding of the performance of our models and identify the most accurate and reliable predictor for bowler performance. This approach ensures that our study's findings and predictions are statistically sound and useful for cricket teams, coaches, and selectors in optimizing their decision-making processes. We pre-processed the cricket match data before beginning the analysis, resolving missing values, scaling features, and, if needed, converting categorical variables. By doing this, the data were appropriately cleaned and readied for model training. We next used the pre-processed data to train each regression model. In order to produce precise predictions regarding player performance, the models underwent training where they analyzed past cricket match data for patterns. During the training phase, our approach involved partitioning the dataset into a 70–30 split, allocating 70 % for training the models and reserving 30 % for testing their performance. With this split, a significant amount of the data may be used for model training, while a distinct section is kept for assessing the model's generalization to new data. The 70–30 ratio creates a compromise between making the most of training data to improve model learning and making sure there is a sufficient test set to reliably assess model performance. We used a 5-fold cross-validation technique to analyze the robustness of the model and mitigate any overfitting. With this approach, the training set is divided into five equal-sized subsets; four of these subsets are used for training, while the fifth subset is used for validation. These subsets are rotated over several iterations such that every part is used for training and validation. By providing a more thorough assessment of the model's performance across different data subsets, the 5-fold cross-validation provides insight on stability and consistency.

We used the standard scaler to reconcile the data and reduce scale differences across features. By standardizing numerical characteristics to have a mean of zero and a standard deviation of one, this preprocessing step helps the model to converge. It prevents larger-scale factors from having an excessive impact. Additionally, to solve class imbalance and ensure fair representation of all classes, we employed binning techniques. To transform categorical data into a numerical format that could be used as model input, these attributes had to be encoded for algorithms to handle them efficiently. We used regularization approaches like L2 regularization in our modelling procedure to combat overfitting tendencies.

In addition, we controlled model complexity and monitored performance on validation data to minimize overfitting by employing strategies like hyperparameter adjustment and early halting during model training. Together, these methods attempted to balance generalization and model complexity, reducing the possibility of overfitting to the training set and assuring stable model performance. The models were assessed using a variety of measures, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2). These measures gave information about the model predictions' accuracy and goodness of fit in relation to the actual player performance. We also gave consideration to any overfitting or underfitting problems that can impair the models' performance throughout the research. This allows us to make sure that the models capture patterns in the data while also generalizing to new and unexplored data.

- **Decision Tree Regressor:** Decision trees divide the data into subsets depending on a set of features, and at the leaf nodes, they generate predictions(Banga et al., 2021). The prediction $\hat{y}$ from a decision tree can be calculated as follows given a feature vector x:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{6}$$

Where:

y_i is the target value of the training sample at leaf node i.
N is the total number of training samples at the leaf node.

- **Random Forest Regressor:** Random Forest is an ensemble method that combines multiple decision trees to make more accurate predictions(Xiong et al., 2021). The prediction $\hat{y}$ from a Random Forest Regressor is calculated as the average prediction of all decision trees in the forest:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \tag{7}$$

Where:

$\hat{y}_i$ is the prediction of the i-th decision tree in the Random Forest.
n is the total number of decision trees in the Random Forest.

- **Support Vector Regressor (SVR):** SVR is a kernel-based regression algorithm that aims to find a hyperplane in a higher-dimensional space that best approximates the relationship between input features and target values. The prediction $\hat{y}$ from an SVR can be calculated as:

$$\hat{y} = w^T.x + b \tag{8}$$

Where:

w is the weight vector.
x is the input feature vector.
b is the bias term.

- **AdaBoost Regressor:** AdaBoost is an ensemble method that combines multiple weak learners (usually decision trees) to create a strong learner. The prediction $\hat{y}$ from an AdaBoost Regressor is a weighted sum of the predictions from the weak learners:

$$\hat{y}_{(x)} = \sum_{i=1}^{M} a_i \hat{y}_{(x)} \tag{9}$$

Where:

$\hat{y}_{(x)}$ is the final predicted value.
M is the number of weak learners.
a_i is the weight of the iii-th weak learner.
$\hat{y}_{(x)}$ is the prediction of the iii-th weak learner.

- **XGBoost Regressor:** XGBoost is a gradient-boosting algorithm that builds an ensemble of weak learners (usually decision trees) sequentially. The prediction $\hat{y}$ from an XGBoost Regressor is the sum of the predictions from all the weak learners:

$$\hat{y} = \sum_{i=1}^{n} \hat{y}_i \tag{10}$$

Where:

$\hat{y}$ is the final predicted value.
$\hat{y}_i$ is the prediction of the i-th weak learner.

n is the number of trees (or models).

- **LightGBM Regressor:** LightGBM is another gradient boosting algorithm that uses a novel decision tree algorithm to improve training speed and efficiency(Banga et al., 2021). The prediction $\widehat{y}$ from a LightGBM Regressor is the sum of the predictions from all the decision trees:

$$\widehat{y}(x) = \sum_{i=1}^{M} f_i(x) \tag{11}$$

Where:

$f_i(x)$ is the prediction of the iii-th tree for input xxx.

M is the total number of trees.

These formulas give a basic understanding of how each regression model works in predicting target values based on input features. The actual implementation of these models may involve additional parameters, optimizations, and regularization techniques to achieve better performance.

4.2. Performance metrics for bowler's performance prediction

For the prediction of bowlers' performance, different metrics are used and are explained as follows(Laifa et al., 2021; Shams et al., 2021).

- **MSE:** The average squared difference between expected and actual values is what the MSE calculates. Without taking into account the direction of the mistakes, it offers a measurement of the total forecast error. MSE is more sensitive to outliers or significant differences between predicted and actual values because it assigns greater weight to bigger mistakes. It is calculated using the following formula:

$$MSE = \frac{1}{n}{*}(y\_pred - y\_actual)^2 \tag{12}$$

- **Root Mean Squared Error (RMSE):** RMSE is a metric used to evaluate the performance of regression models. It measures the average magnitude of the errors between predicted values and actual values. It is calculated by taking the square root of the mean of the squared differences between the predicted and actual values. RMSE is represented by the following formula:

$$RMSE = \sqrt{\frac{1}{N}{*}(y\_pred - y\_actual)^2} \tag{13}$$

- **Mean Absolute Error (MAE):** MAE is another metric used for evaluating the performance of regression models. It measures the average absolute difference between the predicted values and the actual values. It is calculated by taking the mean of the absolute differences between the predicted and actual values. MAE is represented by the following formula:

$$MAE = \left(\frac{1}{n}\right){*}\leqslant\left|y_{pred} - y_{-actual}\right| \tag{14}$$

where n is the number of data points, and y_{pred} and y_{actual} refer to the predicted and actual values of the target variable, respectively.

- **R-squared (R2):** The percentage of the variance in the dependent variable that can be predicted from the independent variables in a regression model is depicted by the R-squared (R2) score, sometimes referred to as the coefficient of determination. It is a useful indicator for assessing how well a regression model fits the data. The R2 score ranges from 0 to 1. The model performs as poorly as a horizontal line through the mean of the data when R2 = 0, which means that it does not explain any of the variance in the dependent variable. R2 = 1 denotes that the model properly fits the data points and fully accounts for all the variance in the dependent variable. If the model performs worse than a horizontal line, the R2 score can potentially be negative. The formula to calculate the R2 score is:

$$R^2 = 1\frac{sum\,squared\,regression\,(SSR)}{total\,sum\,of\,squares\,(SST)} \tag{15}$$

Table 1
Performance metrics of all algorithms without hyperparameter tuning for Bowling.

Algorithm	MSE	MAE	R2 Score	RMSE
Decision Tree	0.255252	0.032062	0.935318	0.0893
Random Forest	0.296413	0.142773	0.914295	0.09513
SVR	0.44886	0.632456	0.059431	0.4337
Adaboost	0.826626	0.567359	0.336939	0.32432
XGB	0.097323	0.1255241	0.828152	0.2474
Light GBM	0.750895	0.503104	0.452866	0.27966

Table 2
Performance metrics of all algorithms with hyperparameter tuning for Bowling.

Algorithm	RMSE	MAE	R2 Score	MSE
Decision Tree	0.088371	0.006768	0.960316	0.08787
Random Forest	0.094949	0.035956	0.954188	0.0950
SVR	0.433784	0.222177	0.043804	0.4337
Adaboost Regressor	0.317692	0.173780	0.487124	0.3199
XGB	0.073661	0.016700	0.972427	0.07366
LightGBM	0.139975	0.065212	0.900437	0.1399

5. Result

In this study, we utilized GridSearchCV as a pivotal technique for refining hyperparameters, a crucial step in optimizing machine learning models. It systematically explores diverse hyperparameter combinations within specified ranges to determine the optimal set that maximizes model performance. The goal of this approach is to identify the optimal hyperparameter values, which have a substantial impact on a model's performance and usefulness, in order to increase accuracy and generalizability. Different algorithms were used in our predictive modeling, and in order to improve their performance, each algorithm needed a different set of hyperparameters. For example, 'n_estimators' determined the number of trees,'max_features' controlled the consideration of features during node splits,'max_depth' limited the depth of trees,'min_samples_split' determined the minimum samples required for a split, and'min_samples_leaf' indicated the minimum samples for a leaf node in the Random Forest model. The parameters "max depth," which controls the tree depth, "min samples split," and "min samples leaf" regulate splits and leaf node samples in decision trees. Support Vector Regression (SVR) relied on factors such as 'gamma' influencing kernel influence, 'C' for regularization, and 'kernel' for data mapping. Adaboost used the variables "n_estimators" and "learning_rate" to calculate the proportion of weak learners in the predictions. 'subsample' was utilized for subsampling,'max_depth' for tree depth, and 'learning_rate' for boosting in XGBoost. The learning_rate,''max_depth,' 'feature_fraction,' and'min_data_in_leaf' were used by LightGBM for different controls. These hyperparameters were essential for maximizing the generalization and performance of the model. GridSearchCV thoroughly investigates preset values to make it easier to choose the best possible combination. By fine-tuning parameters, this approach greatly enhances models and makes it possible for them to more successfully capture complex data patterns.

The performance evaluation of several machine learning methods used to forecast bowling performance is displayed in the tables. Table 1, which is untuned for hyperparameters, shows the predictive power of each method. The Decision Tree method performs admirably; its robust

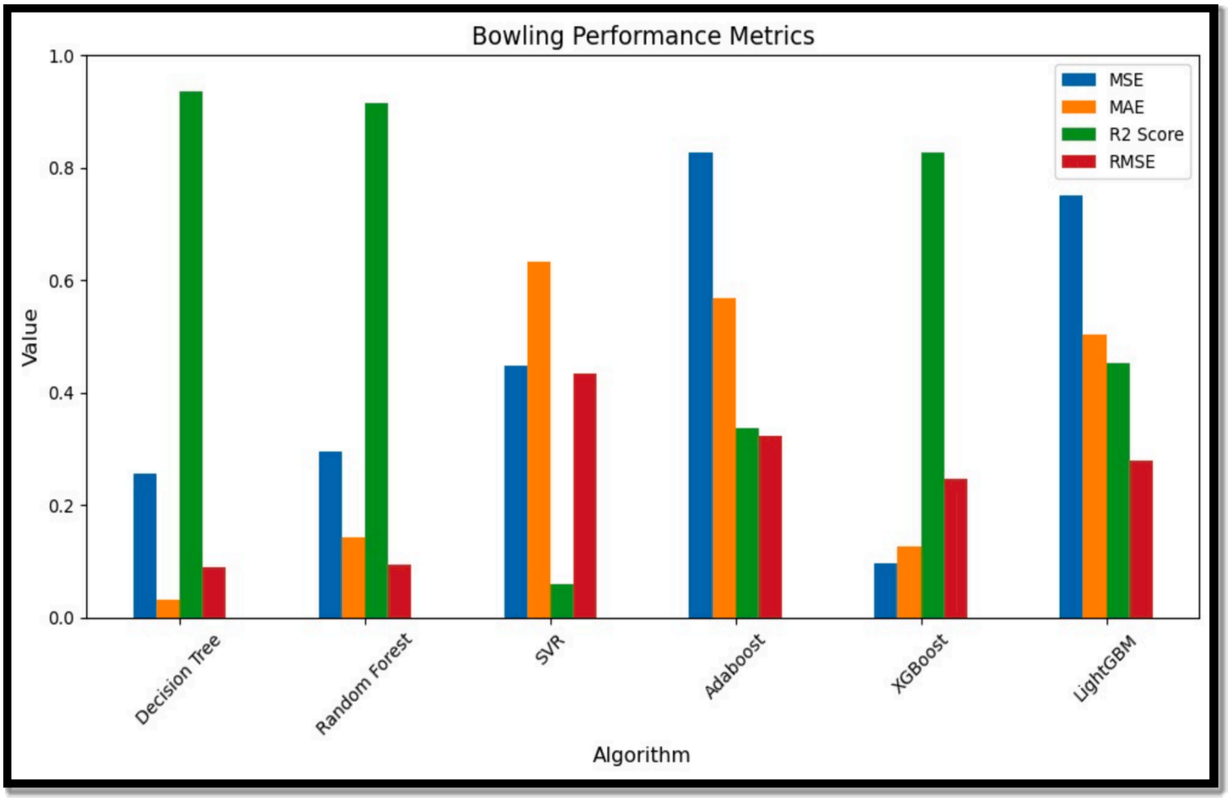

Fig 4.1. Performance metrics of all algorithms without hyperparameter tuning for Bowling.

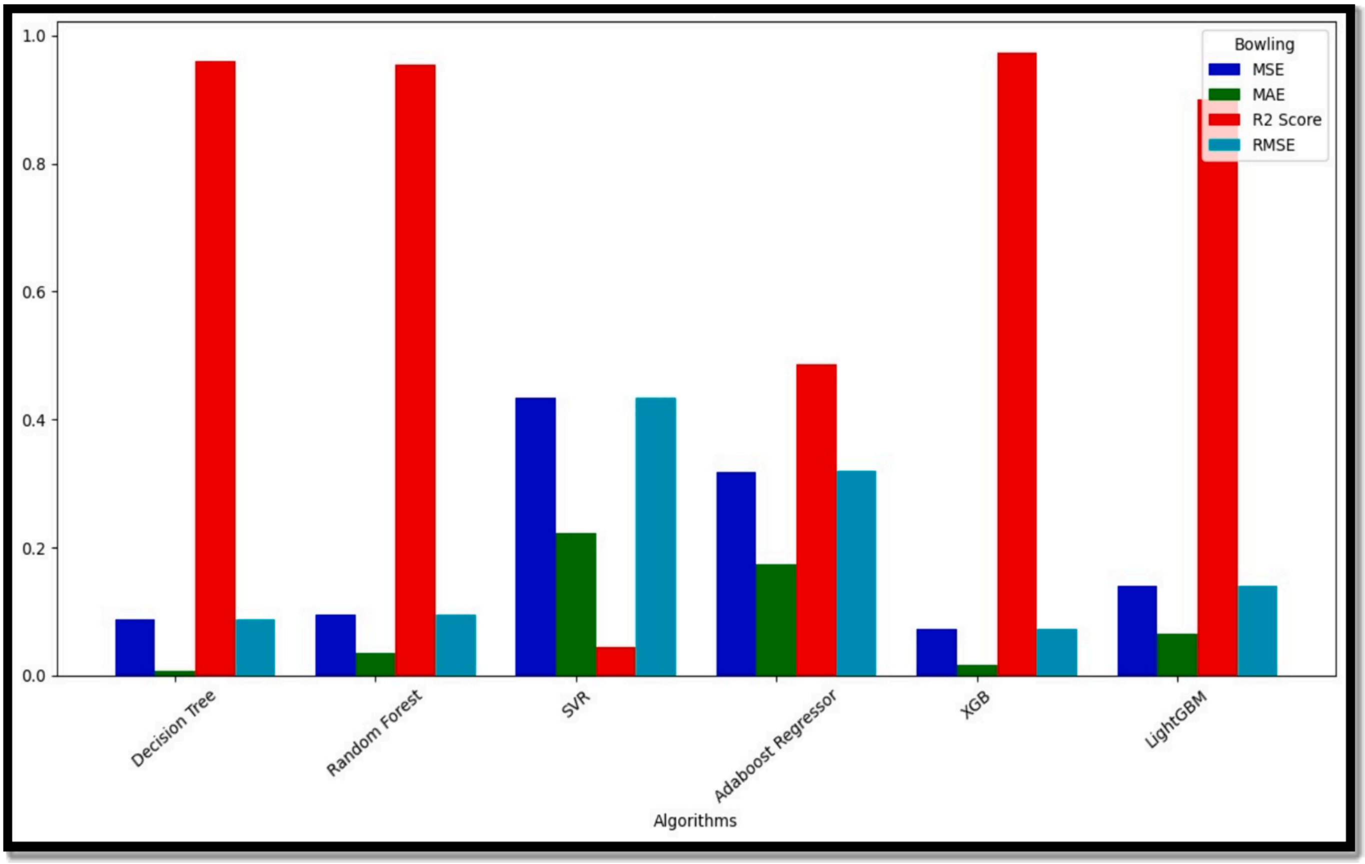

Fig 4.2. Performance metrics of all algorithms with hyperparameter tuning for Bowling.

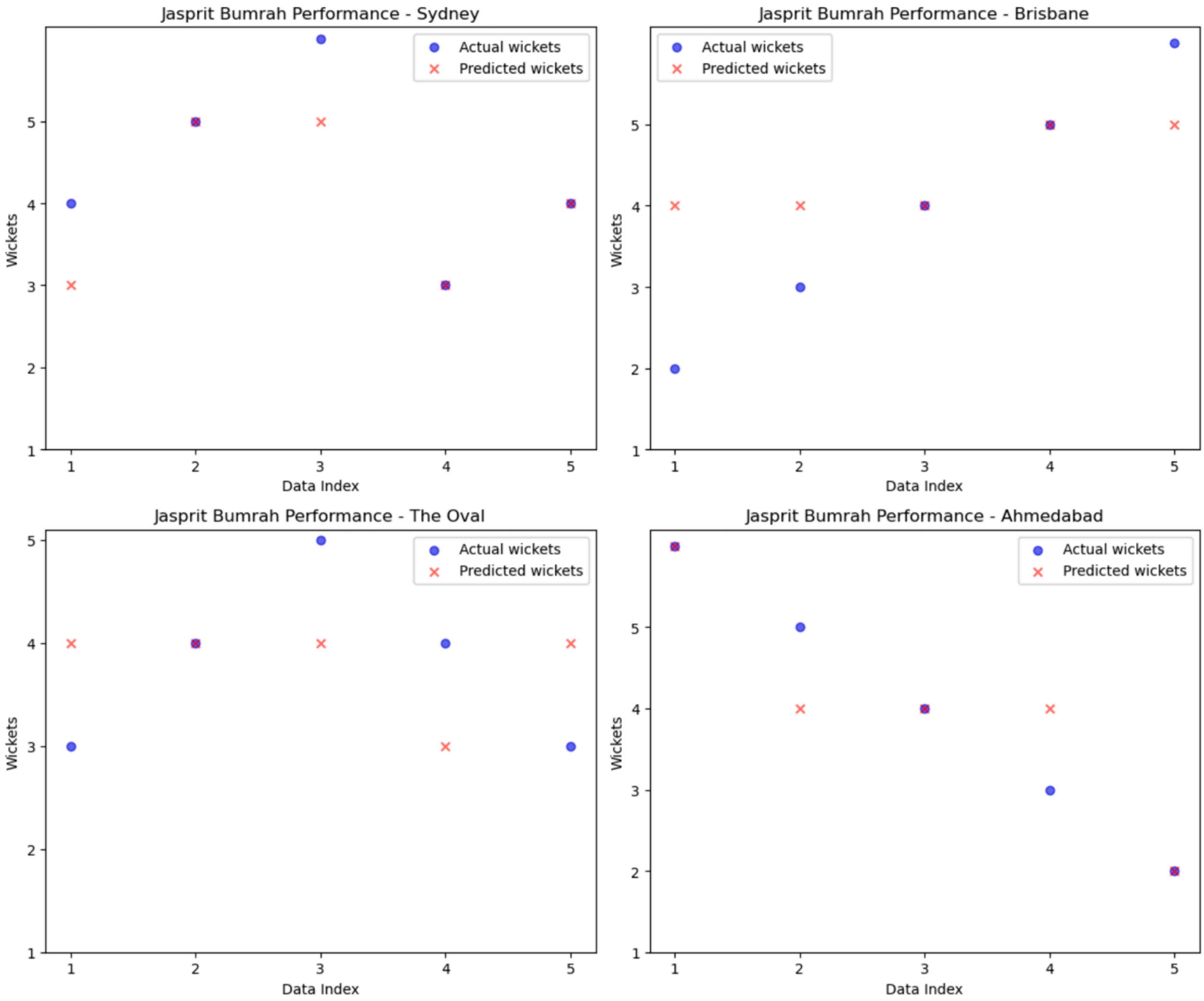

Fig 4.3. Jasprit Bumrah's performance versus Australia on various fields.

R-squared score of 0.935 indicates its significant predictive power. In comparison, the Random Forest model exhibits noteworthy performance in MSE and RMSE measures despite having a lower R2 value. Support vector regression, or SVR, has a lower R2 value, which suggests a worse fit, but it still exhibits a considerable level of predictive power. Adaboost and Light GBM provide mediocre performance in all measures, however XGBoost has a very high R2 score that suggests a strong fit to the data. Table 2, on the other hand, shows the algorithms' performance after hyperparameter modification and demonstrates improvements. The Decision Tree approach shows uniform improvement in performance measures. After adjustments of hyperparameter tunning, Random Forest shows slight gains in MAE and RMSE.

Performance of SVR shows little variation after adjustment. Adaboost, however, shows considerable improvements across a range of measures, especially RMSE and R2 score. XGBoost shows significant gains in every measure following hyperparameter adjustment. Last but not least, LightGBM shows modest gains in RMSE and R2 score after the tuning procedure. The majority of models have greatly improved their fits for the bowling performance prediction task as a result of the hyperparameter tuning phase. These tables highlight the significance of optimizing model parameters, showing significant improvements in model performance and predicted accuracy across several methods. Figs. 4.1 and 4.2 visually compare the performance of various algorithms through graphical representations, offering a comprehensive analysis of their predictive capabilities for bowling performance.

In a comprehensive analysis utilizing the potent XGBoost algorithm for predictive modeling, we delved into the performance of the prolific bowler, Jasprit Bumrah, against Australia. The results, depicted in Fig. 4.3, unveiled a nuanced landscape of prediction accuracy across different cricket grounds. Our focus on employing the XGBoost algorithm aimed to unravel the intricate patterns of Bumrah's performance, particularly in terms of wicket-taking abilities, shedding light on the algorithm's efficacy in providing reliable forecasts. The visual representation in Fig. 4.3 offers a compelling narrative of the variable degrees of prediction accuracy observed at distinct locations. Notably, the analysis showcased the algorithm's impressive precision in predicting Bumrah's real wicket-taking performances. Brisbane emerged as a standout venue where the model exhibited remarkable consistency, accurately estimating Bumrah's actual wickets in every match. This pattern of reliable predictions suggests a robust and trustworthy approximation for Bumrah's performance in the Brisbane cricket ground. Contrastingly, Oval presented a different scenario, with significant variations in individual games exposing a degree of forecast unreliability. Although the model generally followed a similar trend to the actual results, the discrepancies highlighted the need for refinement, especially when predicting Bumrah's performance in matches held at The Oval. The unpredictability observed in these individual games emphasizes the complex nature of cricket analytics and the challenges associated with capturing the dynamics of specific playing conditions.

Ahmedabad, on the other hand, demonstrated a varied pattern in the

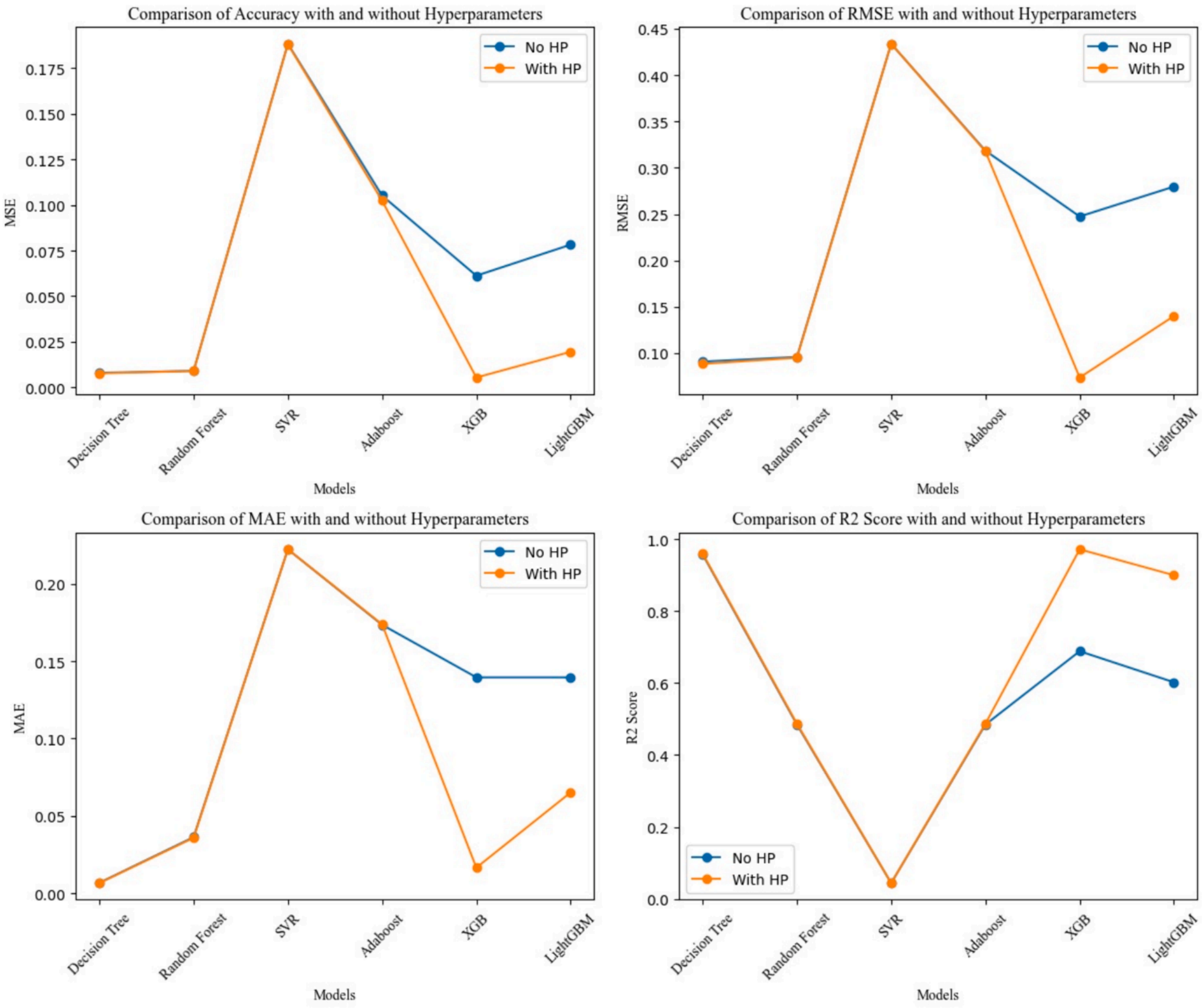

Fig 5.1. Model comparison with and without hyperparameter tuning for bowling.

model's predictions. While some games exhibited correct estimates, noticeable differences surfaced in others, pointing towards the model's need for improvement to achieve more uniform accuracy across all situations in Ahmedabad. This variation underscores the importance of refining the algorithm to account for location-specific nuances and opponent-specific elements. The analysis further emphasized the significance of considering opponent and location-specific factors in predictive modeling. The distinct performance variations observed ground-by-ground underscore the multifaceted nature of a bowler's performance, influenced by the unique playing circumstances and competitor matches. The need for a nuanced approach to predictive modeling becomes evident, necessitating adjustments to the algorithm to enhance its adaptability to diverse cricketing environments. To provide a more granular understanding, Fig. 4.3 illustrates the ground-specific data points on the x-axis and the corresponding number of wickets on the y-axis. This visual representation allows for a detailed examination of the predictive model's accuracy at each location, enabling stakeholders to discern trends, identify areas of improvement, and refine strategies for enhanced performance prediction. In conclusion, the analysis not only delves into the intricacies of predictive modeling using the XGBoost algorithm but also underscores the importance of tailoring such models to the specific challenges posed by different cricket grounds, opponents, and playing conditions.

6. Discussion

The obtained results underscore the critical role of hyperparameter tuning in optimizing the performance of non-parametric models for predicting bowling outcomes in cricket. The meticulous tuning of hyperparameters for algorithms such as AdaBoost, XGBoost, Decision Tree, Support Vector, Random Forest, and LightGBM regressions aims to enhance accuracy and predictive power, contributing to the field of sports analytics. The comparison of predictive models with and without hyperparameters provides valuable insights into the technical and practical significance of hyperparameter tuning specifically for bowling performance prediction. Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) scores were employed for evaluation, showcasing the alignment of findings with research objectives. Results for models without hyperparameters, including Adaboost, XGBoost, Random Forest, Decision Tree, and Support Vector Regression (SVR), illustrated varying predictive accuracy. Decision Tree exhibited the lowest RMSE, while Light GBM showed the highest. Despite generally low RMSE values, R2 scores indicated varying degrees of explained variance in bowler performance, ranging from 0.043 to 0.688 as shown in Fig. 5.1. Notably, Decision Tree and XGBoost demonstrated higher R2 scores, signifying better explanatory power. Upon the introduction of hyperparameters, models such as Random Forest exhibited improved predictive accuracy, with lower RMSE values, indicating enhanced performance. R2 scores also increased, with XGBoost showing the highest R2 score, signifying improved variance explanation. The findings align with the research objectives by demonstrating the impact of hyperparameter tuning in refining the predictive accuracy of bowling performance models.

The significance of hyperparameters in predictive modeling for bowling performance is evident in the improved performance metrics, including MAE, R2 score, and RMSE, for models with hyperparameters. This improvement underscores the importance of optimizing model parameters for precise prediction of bowler performance. Cricket teams, coaches, and analysts can leverage this information for talent identification, strategy formulation, and team composition specifically tailored to bowling strengths. The enhanced ability to explain variance in bowler performance, as indicated by higher R2 scores with hyperparameters, contributes significantly to advancing sports analytics in the field of cricket. Predictive modeling, especially with hyperparameter tuning, emerges as a transformative tool for strategic decision-making and performance optimization in cricket.

7. Conclusion

This research underscores the exceptional predictive capabilities of XGBoost in forecasting cricket bowler performance across diverse parameters, encompassing opponent analysis and varying playing fields. The superiority of the XGBoost approach lies in its profound understanding of the intricate factors influencing bowler performance against a spectrum of oppositions and settings. By unveiling these complexities, the study paves the way for groundbreaking developments in the application of sophisticated analytics, offering valuable insights for astute decision-making not only in sports but also in various disciplines. The in-depth investigation into every aspect of bowlers contributes to a comprehensive understanding of their diverse performance ranges. XGBoost's utilization of sophisticated analytics not only illuminates the nuances of bowler performance but also establishes a foundation for meticulous and insightful forecasts across multiple domains. The study's future directions aim to fortify prediction models by incorporating real-time data and accounting for seasonal variations, enhancing flexibility and precision in forecasts. This technological advancement holds strategic depth, extending its impact beyond the realm of cricket and into various fields. In delving deeper into the practical implications of our research, particularly concerning strategic decision-making in cricket, our findings empower teams, coaches, and analysts with a nuanced understanding of bowler dynamics. This knowledge translates into more informed choices regarding team composition, bowling rotations, and match strategies. The predictive prowess of XGBoost offers a competitive advantage, allowing teams to adapt their tactics based on anticipated bowler performances against specific opponents and in different playing conditions.

As for future work, our focus shifts towards the integration of real-time data, enabling more dynamic and up-to-the-minute predictions. Additionally, accounting for seasonal variations in player performance will be crucial for ensuring the adaptability and robustness of our models. These advancements not only elevate the predictive precision within the realm of cricket but also hold promise for broader applications in sports analytics and decision-making across diverse domains. In conclusion, this research, driven by the capabilities of XGBoost, not only advances the understanding of cricket bowler performance but also sets the stage for a new era in predictive analytics. The study's implications extend far beyond the cricket field, shaping the future of decision-making processes in various disciplines. The fusion of technology and sports analytics offers a pathway to strategic depth, enriching our understanding and influencing decision-making practices across multiple fields.

Declaration of competing interest

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.

Data availability

The authors do not have permission to share data.

References

Banga, A., Ahuja, R., & Sharma, S. C. (2021). Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities. *International Journal of Systems Assurance Engineering and Management.*. https://doi.org/10.1007/s13198-020-01049-9

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Chaudhary, R., Bhardwaj, S., & Lakra, S. (2019). A DEA Model for Selection of Indian Cricket Team Players. *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, 224–227. https://doi.org/10.1109/AICAI.2019.8701424.

Lokhande, R., & Chawan, P. M. (2018). Live cricket score and winning prediction. *International Journal of Trend in Research and Development, 5*(1), 30–32.

Lokhande, R. A., Chawan, P. M., & Pramila, S. (2018). Prediction of Live Cricket Score and Winning. *International Journal of Trend in Research and Development.*

Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2019). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics, xxxx.*. https://doi.org/10.1016/j.aci.2019.11.006

Bhattacharjee, D., & Pahinkar, D. G. (2012). Analysis of Performance of Bowlers using Combined Bowling Rate. *International Journal of Sports Science and Engineering, 06* (03), 184–192.

Bhattacharjee, D., & Saikia, H. (2014). On performance measurement of cricketers and selecting an optimum balanced team. *International Journal of Performance Analysis in Sport, 14*(1), 262–275. https://doi.org/10.1080/24748668.2014.11868720

Koulis, T., Muthukumarana, S., & Briercliffe, C. D. (2014). *A Bayesian stochastic model for batting performance evaluation in one-day cricket., 10*(1), 1–13. https://doi.org/10.1515/jqas-2013-0057

Kumar, S., & Roy, S. (2018). *Score Prediction and Player Classification Model in the Game of Cricket Using Machine Learning., 9*(8), 237–242.

Laifa, H., Khcherif, R., Ghezalaa, H. H., & Ben.. (2021). Train delay prediction in Tunisian railway through LightGBM model. *Procedia Computer Science, 192*, 981–990. https://doi.org/10.1016/j.procs.2021.08.101

Lemmer, H. H. (2002). The combined bowling rate as a measure of bowling performance in cricket. *South African Journal for Research in Sport, Physical Education and Recreation, 24*(2), 37–44. https://doi.org/10.4314/sajrs.v24i2.25839

Lemmer, H. H. (2008). An analysis of players' performances in the first cricket Twenty20 world cup series. *South African Journal for Research in Sport, Physical Education and Recreation, 30*(2), 71–77. https://doi.org/10.4314/sajrs.v30i2.25990

Lemmer, H. H. (2014). Perspectives on the use of the combined bowling rate in cricket. *International Journal of Sports Science and Coaching, 9*(3), 513–523. https://doi.org/10.1260/1747-9541.9.3.513

Malhotra, A., & Krishna, S. (2018). Release velocities and bowler performance in cricket. *Journal of Applied Statistics, 45*(9), 1616–1627. https://doi.org/10.1080/02664763.2017.1386772

McGrath, J. W., Neville, J., Stewart, T., & Cronin, J. (2019). Cricket fast bowling detection in a training setting using an inertial measurement unit and machine learning. *Journal of Sports Sciences, 37*(11), 1220–1226. https://doi.org/10.1080/02640414.2018.1553270

Mittal, H., Rikhari, D., Kumar, J., & Singh, A. K. (2021). *A study on Machine Learning Approaches for Player Performance and Match Results Prediction*. 1–7.

Mukherjee, S. (2014). Quantifying individual performance in Cricket - A network analysis of batsmen and bowlers. *Physica A: Statistical Mechanics and Its Applications, 393*, 624–637. https://doi.org/10.1016/j.physa.2013.09.027

Nevill, A., Atkinson, G., & Hughes, M. (2008). Twenty-five years of sport performance research in the Journal of Sports Sciences. *Journal of Sports Sciences, 26*(4), 413–426. https://doi.org/10.1080/02640410701714589

Passi, K., & Pandey, N. (2018). Increased Prediction Accuracy in the Game of Cricket Using Machine Learning. *International Journal of Data Mining & Knowledge Management Process, 8*(2), 19–36. https://doi.org/10.5121/ijdkp.2018.8203

Shams, M. Y., Elzeki, O. M., Abouelmagd, L. M., Hassanien, A. E., Elfattah, M. A., & Salem, H. (2021). HANA: A Healthy Artificial Nutrition Analysis model during COVID-19 pandemic. *Computers in Biology and Medicine, 135*(April), Article 104606. https://doi.org/10.1016/j.compbiomed.2021.104606

Singh, S., & Kaur, P. (2017). IPL Visualization and Prediction Using HBase. *Procedia Computer Science, 122*, 910–915. https://doi.org/10.1016/j.procs.2017.11.454

Xiong, J., Shi, S. Q., & Zhang, T. Y. (2021). Machine learning of phases and mechanical properties in complex concentrated alloys. *Journal of Materials Science and Technology, 87*, 133–142. https://doi.org/10.1016/j.jmst.2021.01.054