Analysing the impact of field conditions, pitch features, and opponent strength on cricket performance: A machine learning approach

Rameshwari Lokhande*, Rawal Awale, Rahul Ingle

Electrical Engineering Department, Veermata Jijabai Technological Institute, Hanamant Ramchandra Mahajani Road, Matunga East, Mumbai, Maharashtra 400019, India

Received 8 April 2024; revised 25 April 2024; accepted 4 June 2024

Abstract:

Cricket, a sport that is beloved worldwide, requires a combination of expertise, and strategic intelligence. This exposition explores the study of cricket performance, specifically examining how factors such as playing circumstances, pitch dynamics, and the qualities of opponents affect the effectiveness of bowlers and the skill of hitters. The study tries to discover underlying patterns and relationships between these characteristics and player success by using rigorous statistical analysis and other machine learning techniques. Assessment criteria, including accuracy, mean absolute error (MAE), root mean square error (RMSE), and R2 scores, are used to measure the prediction effectiveness of the models. The findings highlight the significant influence of the quality of opponents, the features of the pitches, and the circumstances of the field on the performance of players. In addition, the analysis clarifies the predictive ability of several machine learning algorithms, highlighting Random Forest, XGBoost, and LightGBM as the most precise models. These discoveries provide useful knowledge for academics, educators, and cricket enthusiasts, enabling a better understanding of the various factors that influence player performance and promoting informed strategy discussions.

Keywords: **cricket analytics, machine learning, player performance analysis, prediction analysis, statistical modelling.**

Classification numbers: **1.2, 1.3**

1. Introduction

Cricket, often referred to as a gentleman's game, has captivated millions of fans worldwide through its unique blend of skill, strategy, and tradition. Originating in the 13th century, cricket's evolution from a rustic pastime to a global spectacle mirrors the broader growth of modern sports [1-3]. The transformation of cricket from a quaint English activity to an international phenomenon occurred over centuries, marked by significant milestones and innovations. A pivotal moment in this evolution was the formalisation of cricket's rules and regulations, which began in the 17th century and culminated in the establishment of the Marylebone cricket club (MCC) in 1787. The MCC's codification of cricket laws provided a standardised framework for the governance of the game, laying the foundation for its widespread popularity [4].

The 19th century saw the proliferation of cricket across the British Empire, as colonial administrators and military personnel introduced the sport to distant regions. Cricket became a symbol of British cultural influence, serving as a means of social integration and colonial identity. Its spread to countries such as India, Australia, and the West Indies facilitated cultural exchange and paved the way for the emergence of new cricketing powerhouses [1-3]. The dawn of the 20th century ushered in a new era for cricket, characterised by technological advancements, commercialisation, and globalisation. The advent of radio broadcasts, followed by television coverage, brought cricket into the homes of millions, transforming it from a local pastime into a mass entertainment spectacle. The inaugural Cricket World Cup in 1975 marked a watershed moment in the sport's history, signalling its transition into a global sporting phenomenon [1-3].

As cricket's popularity grew, so too did the scrutiny of player performance and team strategies. Analysts and statisticians began examining every aspect of the game, from batting averages to bowling strike rates, in search of patterns and insights. The advent of computer technology revolutionised cricket analysis, enabling researchers to process vast amounts of data and uncover hidden trends [5]. Of particular interest in cricket analysis is the prediction of player performance, which involves assessing various factors that influence individual and team outcomes. Elements such as pitch conditions, weather, opposition strength, and player form play crucial roles in determining match results. Historically, cricket pundits relied on intuition and experience to forecast outcomes, but the introduction of

*Corresponding author: Email: rameshwarilokhande@gmail.com

statistical modelling and machine learning has transformed predictive methodologies [6, 7].

Statistical tests, including hypothesis testing and regression analysis, form the foundation of cricket performance prediction. Researchers utilise historical data to identify correlations between different variables and develop predictive models capable of forecasting match results with a high degree of accuracy. Machine learning algorithms, such as random forests, support vector machines, and neural networks, have emerged as powerful tools in cricket prediction, capable of processing complex data sets and generating accurate forecasts [8-10].

This exposition explores the intricate domain of cricket performance to identify the numerous factors that contribute to player success on the field. By employing advanced machine learning methods and rigorous statistical analysis, we examine various aspects of player performance to gain a deeper understanding of their impact on the game. Our focus includes, but is not limited to, batting averages, bowling strike rates, fielding ability, pitch conditions, weather, match context, and opposition strength. Through a comprehensive analysis of these elements, we aim to elucidate the complex relationship between match dynamics, environmental factors, and individual capabilities.

Our study seeks to contribute to the ongoing discourse on cricket performance evaluation by offering an enhanced understanding of player effectiveness. Machine learning and rigorous statistical analysis have the potential to uncover complex patterns and relationships within large data sets, making them invaluable tools in cricket research. By systematically analysing player performance, pitch dynamics, and playing conditions, these analytical approaches enable researchers to identify key factors and their influence on match outcomes. Statistical techniques allow for the measurement of relationships between variables such as pitch conditions, weather, and opposition strength, revealing the intricate interactions at play.

Machine learning models, by leveraging historical data, provide powerful methods for modelling and predicting player success. These models can account for complex interdependencies among variables and incorporate nonlinear interactions, resulting in more accurate and robust performance forecasts. Overall, the integration of machine learning and statistical analysis in cricket research enhances our understanding of the multiple factors influencing player performance and match outcomes.

The prediction efficacy of the models created in this study is evaluated using several criteria, including R-squared (R^2) score analysis, mean absolute error (MAE), root mean squared error (RMSE), and mean squared error (MSE). RMSE and MAE provide insight into the accuracy of the model by measuring the average difference between predicted and actual values. MSE, by averaging the squares of the errors, reflects the overall variability in the prediction errors. Additionally, R^2 score analysis assesses the proportion of variance in the dependent variable explained by the independent variables. These evaluation criteria are considered significant as they provide comprehensive metrics for assessing the performance of predictive models, taking into account various aspects of predictive accuracy and reliability.

Through a detailed evaluation, researchers can determine how well their models capture performance variations based on different factors, allowing for informed decisions regarding model selection and improvement. Our study presents several new insights into the influence of factors such as opponents, pitch characteristics, and ground conditions. Specifically, we demonstrate that the calibre of opponents significantly impacts player performance, with stronger teams posing greater challenges to individual players. It is crucial to analyse a batsman's performance variation across different pitches, venues, and against varying opposition teams.

This study utilises data from Indian players, incorporating pitch data available from the pitch report. Variables such as the quality of opposition, venue, and pitch characteristics significantly influence player performance. Cricket matches are played on a wide range of surfaces, including grassy, flat, hard, plain, and soft pitches, each presenting different opportunities and challenges for players [11]. The various types of pitches are described below:

Grassy pitch: A pitch with substantial grass cover, benefiting fast bowlers as the ball tends to bounce and seam more on such surfaces.

Flat pitch: A level pitch with minimal grass or cracks, typically offering favourable conditions for batting with little assistance to bowlers.

Hard pitch: A compacted surface that increases ball speed and bounce, enabling batsmen to play with greater power and precision.

Plain pitch: A surface devoid of significant variations, presenting a neutral playing field.

Soft pitch: A softer, less compacted surface where the ball tends to slow down and lose speed, making shot-making more difficult.

By analysing how batsmen perform on different types of surfaces, we gain valuable insights into their versatility, shot-making abilities, and capacity for game-changing adjustments. This understanding is essential for enhancing player selection, game strategy, and training methods. The integration of statistical analysis with knowledge of pitch properties provides a deeper comprehension of cricket, facilitating more informed decision-making.

Cricket player performance is influenced by a myriad of factors, both intrinsic and extrinsic. In addition to technical skills and athleticism, psychological factors such as confidence, concentration, and temperament play a significant role in determining performance. Moreover, external factors, including pitch conditions, weather, and match context, profoundly impact player outcomes [12, 13]. One-day internationals (ODIs), test matches, and twenty-20 internationals (T20Is) represent the three primary formats of modern cricket, each characterised by unique challenges. ODIs are known for limited overs and high-scoring matches, while Test cricket is renowned for its strategic depth and endurance. T20 cricket, on the other hand, is fast-paced and explosive, appealing to a younger, more dynamic audience [14-16].

In recent years, cricket analytics has undergone a paradigm shift with the advent of advanced metrics and data-driven decision-making. Analysts now employ sophisticated statistical models and machine learning algorithms to assess player performance, optimise team strategies, and predict match outcomes. These tools enable coaches, players, and administrators to make informed decisions based on objective data and empirical evidence [17, 18]. The application of predictive analytics in cricket extends beyond match outcomes to player selection, talent identification, and performance evaluation. By leveraging historical data and advanced modelling techniques, teams can identify emerging talent, optimise player combinations, and maximise their chances of success on the field. Predictive analytics also allows teams to adapt their strategies in real-time, responding to evolving match conditions and opponent tactics [19-21].

The future of cricket analytics holds immense promise, with ongoing advancements in technology and data science driving innovation and insight. As cricket becomes increasingly globalised and commercialised, the demand for sophisticated analytical solutions will continue to grow, presenting new opportunities and challenges for researchers, analysts, and practitioners [22, 23]. Cricket analytics represents a fascinating intersection of sport, statistics, and technology, with the potential to revolutionise the way the game is played, coached, and experienced. By harnessing the power of data and analytics, cricket stakeholders can unlock new insights, optimise performance, and enhance the overall quality of the sport. As the game continues to evolve in the 21st century, analytics will play an increasingly integral role in shaping its future trajectory and success [18, 24, 25].

Additionally, machine learning models have been used to forecast the performance of Indian players, considering variables such as field conditions, pitch characteristics, and opponent strengths [12, 13]. Performance metrics such as precision, accuracy, F1 score, and recall have been employed to evaluate the models' effectiveness in capturing performance differences based on these variables. This study investigates the influence of grounds, pitches, and opposition on the variability of bowler and batsman performance. The results reveal significant variations in wicket-taking performance based on these factors, supporting the alternative hypothesis. Furthermore, machine learning models demonstrate varying levels of accuracy and predictive power in forecasting player performance. Random forest, XGBoost, and LightGBM exhibit exceptional performance, while SVR and Decision Tree models show comparatively lower effectiveness. These findings underscore the importance of considering unique playing surfaces, pitch conditions, and opponent factors when evaluating and projecting player performance in cricket.

Machine learning and artificial intelligence have gained significant prominence in the railway transportation industry in recent years, particularly in predicting train delays. Several studies have attempted to solve this forecasting problem using different machine learning models. For example, on Serbian railways, support vector regression (SVR) was used to analyse train arrival delays and outperformed artificial neural networks (ANN) [26]. Statistical analysis has been utilised to predict running and dwell times [27], employing global prediction models such as LTS robust linear regression, regression trees (RT), and random forests (RF). The k-nearest neighbour (K-NN) and linear regression (LR) models have been used to determine off-peak and peak-hour dwell times, respectively [28]. Random forests (RF) have consistently outperformed extreme learning machines (ELM) and kernel. Regularised least squares (KRLS) in predicting train delays [29]. A deep extreme learning machine-based method for predicting train delays using real-time data has been presented [30]. Furthermore, methods based on neural networks such as backpropagation neural network (BPNN), wavelet neural network (WNN), and genetic algorithms (GA-BPNN and GA-WNN) have been employed to predict train arrival times [31]. Researchers have devised methods to predict running time, dwell time, train delays, and penalty costs by combining decision trees (DT) and random forest regression (RF) [32]. In other studies [33-35], Random forests have also demonstrated superior performance compared to other methods. Ensemble methods have been used to predict train delays up to 24 hours in advance [36]. Long short-term memory (LSTM) networks have shown improved performance over random forests and neural networks (NN) in predicting train arrival delays [37]. Gradient boosting regression trees (GBRT) have outperformed SVR and RF in predicting train delays [38]. A deep learning method called CLF-Net was developed for estimating train delays on China's high-speed rail lines by integrating 3D convolutional neural networks (CNN), LSTM, and fully connected neural network (FCNN) structures [39].

These machine learning methods demonstrate significant promise in improving the accuracy and ease of predicting train delays.

Significant progress has also been made in predicting air quality using machine learning techniques over the past few years. Researchers have investigated various forecasting algorithms for air pollutants, particularly PM2.5, in cities around the globe. J. Ma, et al. (2020) [40] predicted PM2.5 at the Beijing railway station using a spatial ensemble method. M. Asgari, et al. (2017) [41] employed a Hadoop architecture with multinomial naive Bayes and logistic regression to predict Tehran's air pollution, whereas M.Z. Joharestani, et al. (2019) [42] identified XGBoost as the most effective algorithm for predicting Tehran's air quality using geographical and satellite data. H. Karimian, et al. (2019) [43] utilised the Variance Inflation Factor (VIF) method and LSTM to forecast PM2.5 in Tehran, yielding promising results. In addition, other studies have investigated various machine learning models, such as random forest, support vector machine (SVM), ensemble methods, and deep learning approaches, to predict air pollutants in cities around the world. The objective of these efforts is to improve air quality forecasting, thereby enhancing public health and environmental management. As researchers continue to investigate and refine these AI-based techniques, the potential for more precise and effective air quality forecasts grows. These developments hold promise for addressing the challenges posed by air pollution and improving the overall quality of life in communities affected by poor air quality [40, 44-48].

2. Methods

2.1. Data wrangling

Data wrangling is a systematic method used to acquire, cleanse, and organise raw data relevant to cricket player performance. Its purpose is to forecast player success in the sport by extracting pertinent information from various sources, including match histories, player statistics, and match commentary. This acquired data is then cleaned, processed, and standardised to remove discrepancies and manage missing values. The resulting wrangled data is organised and prepared for further analysis, enabling the creation of prediction models to forecast player performance in cricket matches. The data for this study were collected from www.cricinfo.com. Batting contests from 14 January 2005 to 10 July 2017 were considered. Since Sachin Ramesh Tendulkar was the oldest player at that time, a record of each batsman's performance, innings by innings, was compiled, starting from his debut in an ODI on 18 December 1989. Bowling matches from 2 January 2000 to 10 July 2017 were also considered.

The batting parameters encompass several important components that shed light on a player's batting performance. The "average venue" field provides information about a player's performance under various playing conditions by displaying the average runs the player has scored at a particular ground or venue. The "ground" feature identifies the location or stadium where the innings were played. "Day," "month," and "year" provide a chronological reference by indicating the precise date of the innings. The field "Oppositionteam" displays the team against which the player competed, offering insight into their performance against different opponents. Additional batting characteristics include "innings runs score," which shows the player's overall contribution to the team's run total over the course of an innings. "Strike rate venue" calculates a player's scoring rate at a particular location and displays the number of runs scored per 100 balls faced. "Average opposition" showcases a player's batting average against a specific opposition team, highlighting their performance against different adversaries. "Innings balls faced" indicates the number of balls faced, representing the player's participation in the game, while "50s" denotes the number of times the player scored between 50 and 99 runs in an innings. The variables "average yearly" and "strike rate yearly" provide details on the player's average runs scored and batting strike rate per year, respectively, offering an overview of their consistency over time. The player's strike rate against a specific opposition team is displayed in the "strike rate opposition" field. The term "innings player" simply denotes the participant in the inning. Finally, "100s" shows how many times the player scored 100 or more runs in an inning, demonstrating their ability to convert promising starts into noteworthy results.

The bowling parameters, on the other hand, provide insightful data regarding a player's performance with the ball. The term "average venue" refers to a player's bowling average at a particular location, demonstrating their effectiveness in taking wickets at various grounds. "Balls bowled" refers to the total number of balls a player has bowled during an innings, providing information on their workload and participation in the match. To evaluate a player's performance in limiting the opposition's scoring, "runs conceded" displays the total number of runs the player has allowed. The "average opposition" option displays the player's bowling average against a certain opposition team, providing information about how they performed against different foes. The "4 wickets" statistic counts how many times they took four wickets during an innings, showcasing their ability to take multiple wickets in a single bowling attempt. The statistic "average yearly" shows the player's average annual wicket total, emphasising their consistency over time. Similar to the batting parameters, "ground" denotes the location or stadium where the bowling took place, while "year," "month," and "day" provide the precise date of the bowling performance. The "country" field indicates the player's country of birth, while the "bowlingstyle" field indicates their preferred bowling technique, such as fast or

spin. "Maidens bowled" displays how many maiden overs a player has bowled, demonstrating their capacity to apply pressure consistently and limit runs. The "FF" element denotes the player's economy rate, which is computed as runs allowed per over, providing information about how well they can limit the opposition's scoring. "Form" denotes the bowler's current performance level, while "5 wickets" lists the number of innings in which the player claimed five or more wickets, illustrating their propensity for significant bowling milestones. Indicating the player's participation in the game, "innings number" displays the number of the innings in which they bowled, and "wickets taken" displays the total number of wickets they claimed during that innings. Together, these batting and bowling metrics offer in-depth statistical data regarding a player's performance, strengths and weaknesses, consistency over time, and capacity to compete against different teams and in varied settings.

The proposed approach for predicting player performance in cricket using machine learning, as shown in Fig. 1, involves gathering comprehensive data from reliable databases using data preprocessing techniques to ensure consistency and quality.

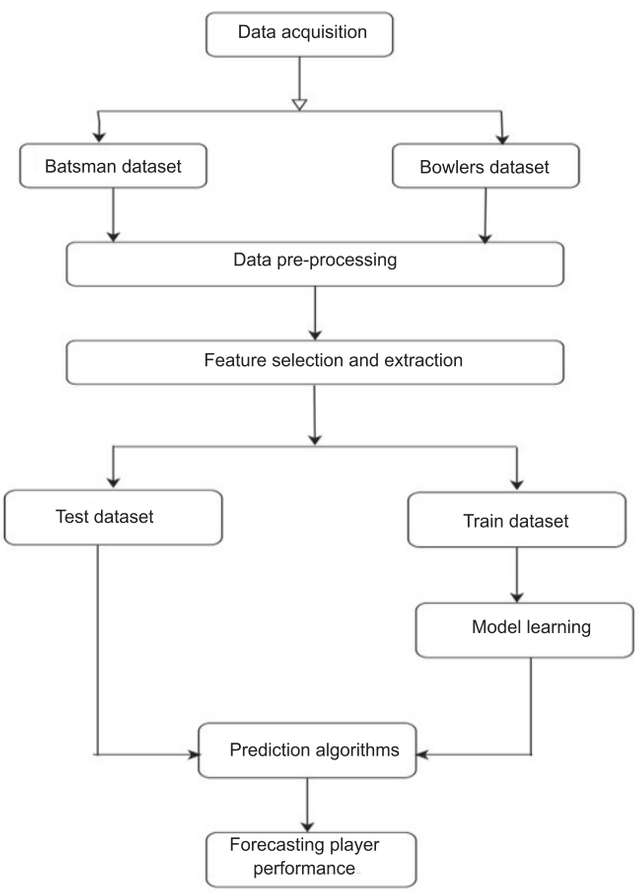

Fig. 1. Flow chart for player performance prediction in cricket.

In this work, machine learning methods such as random forest, decision tree, support vector machine (SVM), XGBoost, AdaBoost, and lightGBM have been demonstrated to be the most accurate models for forecasting cricket performance. These algorithms are distinguished by their ability to handle intricate, nonlinear relationships in the data and to forecast with high accuracy using past performance indicators. Random forest is renowned for its ensemble learning method, combining predictions from several decision trees to increase accuracy and reduce overfitting. Decision tree models are valuable for deciphering underlying trends in cricket performance statistics due to their ease of interpretation and visualisation. SVMs effectively capture nonlinear relationships and partition data points into discrete groups. XGBoost and AdaBoost employ boosting approaches, iteratively optimising prediction performance; lightGBM is particularly well-suited for real-time prediction tasks because of its efficient training speed and memory usage. These algorithms outperform conventional models by capturing complex patterns and relationships within cricket performance data, ultimately producing more precise forecasts.

Coaches, players, and analysts can also practically apply this information to optimise training plans, squad compositions, and game strategies, ultimately enhancing performance and competitiveness on the field. A better understanding of the various factors affecting cricket player performance can lead to more informed strategic discussions among players, coaches, and analysts. By identifying key success factors such as pitch conditions, opponent strength, and match context, stakeholders can tailor their plans to optimise performance outcomes. Training regimens designed by coaches can focus on specific skill sets and prepare players for a variety of playing environments. Players can utilise insights gained from data analysis to adjust their game plans and strategies, thereby improving their adaptability and efficiency on the field. Analysts can assist management and coaching staff in making informed decisions and recommendations that will guide team selections and strategic choices. Ultimately, teams of all skill levels benefit from a collaborative and data-driven approach to cricket strategy, fostered by a thorough understanding of player performance factors.

This research also involved conducting a thorough feature importance analysis to identify the salient player characteristics dictating performance in cricket. This entailed initially aggregating all pertinent player metrics, encompassing both batting and bowling parameters, for subsequent analysis. Machine learning algorithms such as random forest, decision trees, and gradient boosting were then trained using the compiled dataset. Following training, feature importance scores were computed to quantify

the contribution of each feature to the model's predictive efficacy. These scores were subsequently ranked, with higher values indicating a greater influence on player performance prediction. The findings of the feature importance analysis were then visualised using appropriate data visualisation techniques to facilitate comprehension and decision-making. This rigorous analysis provides crucial insights into the pivotal player characteristics driving performance in cricket, thereby informing feature selection for model refinement and directing efforts towards enhancing predictive precision.

2.2. Non-parametric models for predictive modelling

Non-parametric regression techniques such as support vector regression (SVR), random forest regressor, decision tree regressor, AdaBoost regressor, extreme gradient boosting (XGBoost), and light gradient boosting machine (lightGBM) were employed for predictive modelling. These models utilise various hyperparameters to optimise performance and accuracy. Hyperparameter optimisation techniques were applied to fine-tune model parameters and enhance predictive capabilities. The performance of each algorithm was evaluated based on metrics such as accuracy, root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2).

Support vector regression (SVR): SVR is a powerful regression approach that utilises support vector machines to forecast continuous outcomes. In the context of cricket player performance, SVR is used to model and predict various performance measures [20, 49, 50]. SVR functions by mapping the input features to a higher-dimensional space and finding a hyperplane that minimises the difference between expected and actual values. This hyperplane is adjusted by tuning hyperparameters such as 'C' for regularisation, 'kernel' for the type of kernel function used, and 'epsilon' for the error margin. For our study, we set C to 1.0, kernel to 'rbf' (Radial Basis Function), and epsilon to 0.1.

Random forest regressor: Random forest regressor is an ensemble learning technique that integrates multiple decision trees to produce precise predictions. The final prediction is derived by averaging the outputs of all the decision trees, each constructed using a different random subset of the data [51-53]. In the context of predicting player performance, the random forest regressor captures intricate correlations between features and performance measures. The hyperparameters we adjusted for the random forest regressor include n_estimators for the number of trees in the forest, max_depth for the maximum depth of each tree, min_samples_split for the minimum number of samples required to split an internal node, and min_samples_leaf for the minimum number of samples required to be at a leaf node. For our analysis, we specified the following parameters: n_estimators=100, max_depth=10, min_samples_split=2, and min_samples_leaf=1.

Decision tree regressor: Decision tree regressor is a non-parametric regression technique that creates a tree-like representation of decisions and their outcomes. Each leaf node of the tree indicates a predicted outcome, and each internal node represents a decision based on a particular attribute [54, 55]. In predicting cricket player performance, the decision tree regressor considers various characteristics. For the decision tree regressor, we set max_depth for the maximum depth of the tree, min_samples_split for the minimum number of samples required to split an internal node, and min_samples_leaf for the minimum number of samples required to be at a leaf node. For our study, we set the max_depth parameter to 5, min_samples_split to 2, and min_samples_leaf to 1.

AdaBoost regressor: AdaBoost (adaptive boosting) is an ensemble learning method that combines multiple weak learners to produce a strong predictive model. In the context of cricket player performance, AdaBoost is used to enhance the performance of individual weak models and increase forecast precision [51, 53, 56]. We adjusted two hyperparameters for AdaBoost: n_estimators for the number of weak learners (decision trees) in the ensemble, and learning_rate for the contribution of each weak learner to the outcome of the prediction. To determine the best settings for our study, we experimented with a range of n_estimators values (50, 100, and 200) and learning_rate values (0.1, 0.5, and 1.0).

Extreme gradient boosting (XGBoost): Extreme gradient boosting, or XGBoost, is another ensemble learning technique that combines multiple weak models to produce a powerful prediction model. XGBoost is renowned for handling large datasets and intricate relationships with accuracy and efficiency [42, 51, 54, 57]. In the context of predicting player performance, XGBoost captures complex patterns and interactions within the data. The hyperparameters n_estimators (number of weak models in the ensemble), learning_rate (contribution of each weak model), and max_depth (maximum depth of each weak model) were adjusted. To find the optimal configuration for our investigation, we experimented with various values of n_estimators (50, 100, 200), learning_rate (0.1, 0.5, 1.0), and max_depth (3, 6, 9).

Light gradient boosting machine (Light GMB): LightGBM is a gradient-boosting framework that employs a tree-based learning technique. It is designed to be efficient in terms of memory and training speed, making it suitable for large datasets. In the context of cricket player performance, LightGBM effectively captures complex relationships between features and performance measures [49, 51, 58, 59]. We modified the following hyperparameters for LightGBM: num_leaves for the maximum number of leaves

in a tree, learning_rate for the contribution of each tree, and n_estimators for the total number of trees in the ensemble. To determine the best parameters for our study, we tried a variety of num_leaves (31, 50, 100), learning_rate (0.05, 0.1, 0.2), and n_estimators (100, 200, 300) values.

By adjusting these hyperparameters, we aimed to identify the optimal configuration for each algorithm, resulting in better performance and more precise forecasts of cricket player performance. The chosen hyperparameter values were based on experimentation and empirical data from prior research, considering the trade-off between model complexity and performance.

The aggregation of comprehensive cricket player data from reliable sources included match statistics, player profiles, and performance metrics. This raw data underwent meticulous preprocessing steps to ensure consistency, integrity, and suitability for analysis. Feature engineering techniques were then applied to extract relevant information from the dataset, transforming raw attributes into meaningful predictors of player performance. Subsequently, machine learning models-including random forest, decision trees, and gradient boosting-were trained using the pre-processed dataset to predict player performance. Model performance was rigorously evaluated using appropriate metrics. Moreover, cross-validation techniques were employed to assess model generalisation and mitigate overfitting. The predictive models were then fine-tuned through hyperparameter optimisation to enhance their predictive efficacy. Finally, the performance of the optimised models was assessed on unseen data to gauge their real-world applicability and effectiveness. This systematic approach ensures the development of accurate predictive models capable of providing valuable insights into cricket player performance for stakeholders such as team management, selectors, and strategists.

3. Results

An analysis of batsman and bowlers performance variability: ground, pitch, and opposition. In this study, we aim to investigate the variability in the performance of both batsmen and bowlers concerning different grounds, pitches, and opposition teams. The null hypothesis suggests that there is no significant difference in the performance metrics across various conditions, while the alternative hypothesis argues for the presence of substantial disparities based on these factors. Statistical tests were conducted, yielding p-values to assess the significance of these differences. A p-value less than 0.05 indicates rejection of the null hypothesis, implying significant variations in performance. Figs. 2 and 3 illustrate the statistical test results for both bowler and batsman performance across different conditions, respectively.

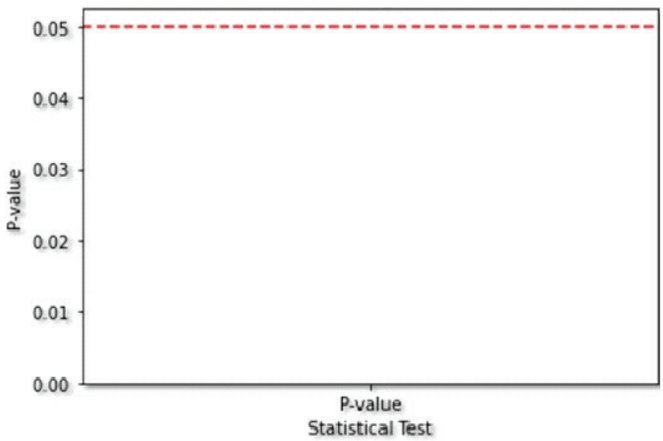

Fig. 2. Statistical test for bowler performance on different grounds, pitches, and against different opposition.

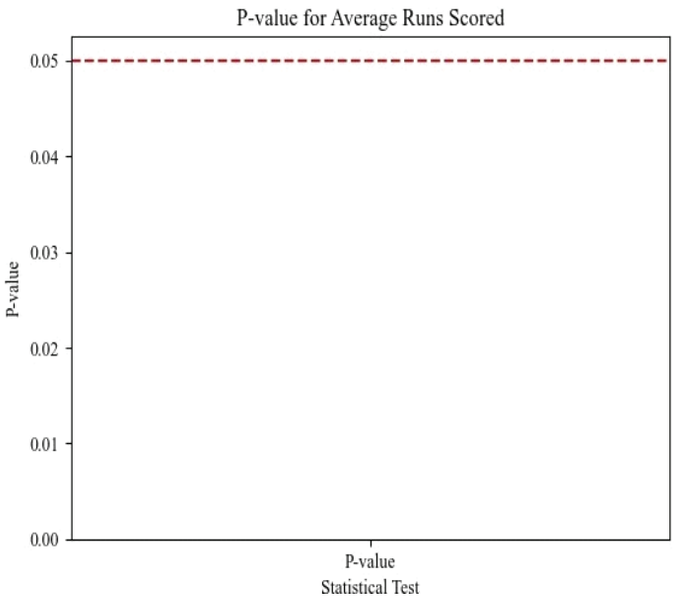

Fig. 3. Statistical test for batsman performance on different grounds, pitches, and against different opposition.

Table 1 presents the performance metrics of various machine learning algorithms used for forecasting bowler performance. These algorithms include decision tree, random forest, SVR (support vector regression), AdaBoost, LightGBM, and XGB (extreme gradient boosting). Each algorithm's accuracy, RMSE, MAE, and R^2 score are provided, offering insights into their effectiveness in predicting player performance.

The findings indicate that while decision tree and random forest algorithms provide reasonable accuracy in predicting bowling performance, random forest outperforms the decision tree in terms of RMSE, MAE, and R^2 score. Conversely, SVR's negative R^2 score suggests it fails to capture the underlying patterns in the data, resulting in

subpar performance. AdaBoost, LightGBM, and XGB demonstrate superior predictive power with higher accuracies and R^2 scores, making them more suitable for accurately forecasting bowling performance.

Table 1. Model performance metrics for bowler.

Model	Accuracy	RMSE	MAE	R^2
Decision tree	0.852075	0.305931	0.147925	0.542835
Random forest	0.919707	0.187015	0.080293	0.829165
SVR	0.738119	0.459590	0.261881	-0.031732
Adaboost	0.821608	0.318849	0.178392	0.503412
LightGBM	0.984394	0.067300	0.015606	0.977876
XGB	0.995257	0.018557	0.004743	0.998332

Similarly, Table 2 presents the performance metrics of machine learning algorithms for forecasting batsman performance. The algorithms and metrics are consistent with those used for bowler performance analysis.

Table 2. Model performance metrics for batsman.

Algorithm	Accuracy	RMSE	MAE	R^2 Score
Decision tree	0.735	0.473	0.265	0.831
Random forest	0.813	0.342	0.187	0.912
SVR	0.278	1.146	0.722	0.011
Adaboost	0.706	0.502	0.294	0.810
XGB	0.953	0.178	0.047	0.976
LightGBM	0.931	0.193	0.069	0.972

The results, as shown in Table 1 that the decision tree algorithm achieves an accuracy of 85.2%, with an RMSE of 0.3059 and an MAE of 0.1479. However, its R^2 score indicates that it can only explain 54.3% of the variance in bowling performance. In contrast, the Random Forest algorithm performs better in terms of accuracy (91.9%), RMSE (0.1870), MAE (0.0803), and R^2 score (82.9%). The SVR algorithm demonstrates poorer performance with an accuracy of 27.8% and a very low R^2 score, suggesting a lack of fit to the data. Adaboost, lightGBM, and XGB exhibit even higher accuracy and better predictive power, with XGB performing exceptionally well with an accuracy of 93.1% and an outstanding R^2 score of 97.2.

For further analysis, we utilised various machine learning models to forecast batsman and bowler performance, considering factors such as field conditions, pitch type, and opposition team. By evaluating the predictive power of these models using performance metrics such as accuracy, precision, recall, and F1 score, we could determine how effectively they represent differences in performance based on these variables. The results supported the alternative hypothesis, indicating significant differences in performance across different conditions.

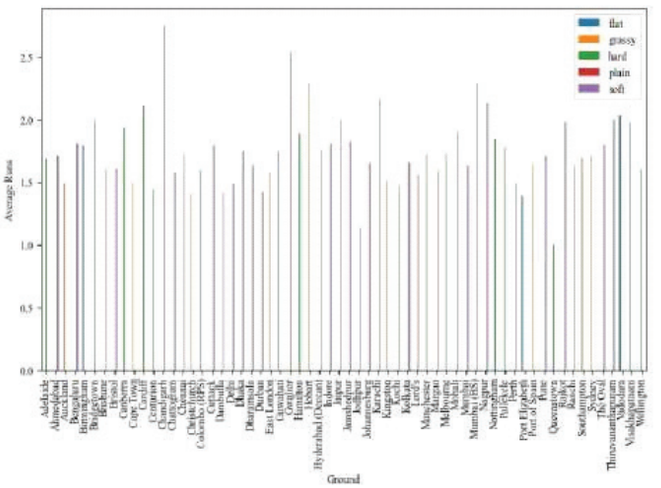

Fig. 4. Runs scored by batsman on different grounds and pitches.

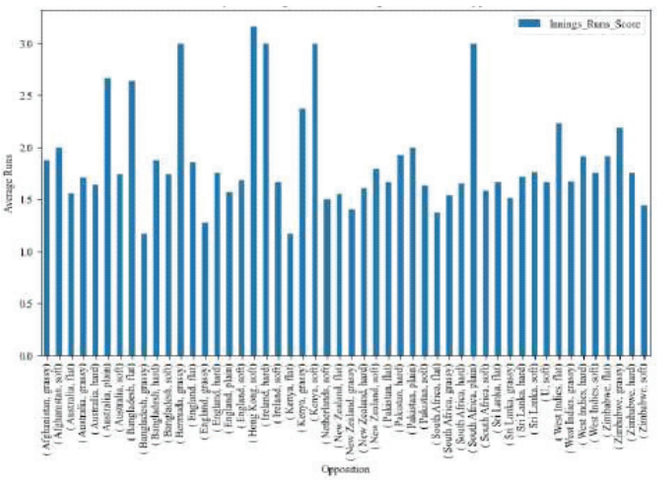

Fig. 5. Runs scored by batsman against different opposition and on different pitches.

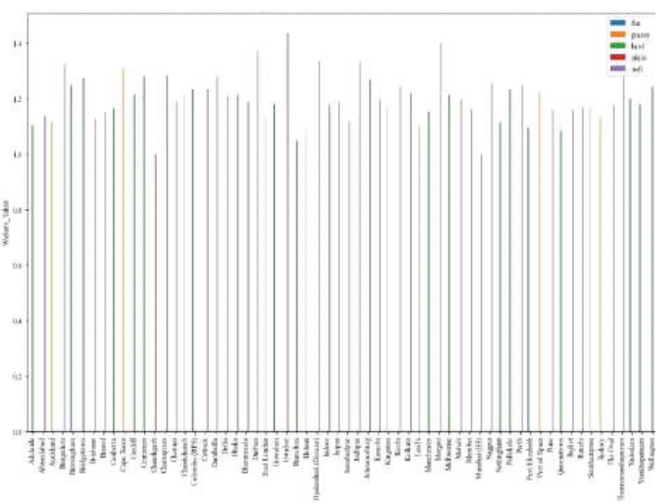

Fig. 6. Wickets taken by bowlers on different grounds and pitches.

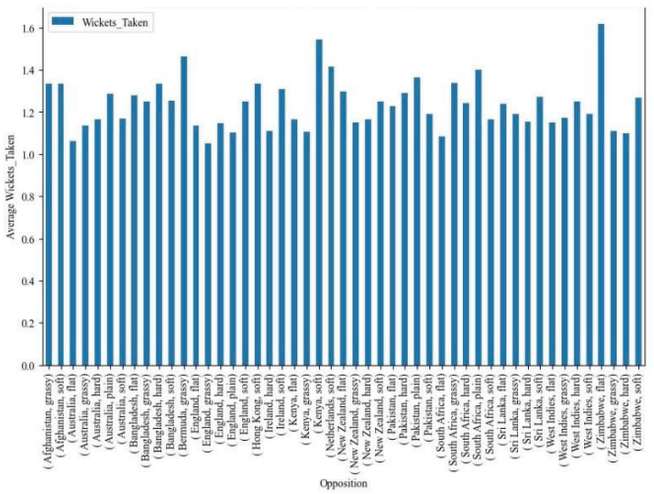

Fig. 7. Wickets taken by bowlers against different opposition and pitches.

Figures 4 and 5 display the runs scored by batsmen on different grounds and pitches, and against different opposition teams and on various pitches, respectively. Similarly, Figs. 6 and 7 illustrate the wickets taken by bowlers on different grounds and pitches, and against different opposition teams and on various pitches, respectively. These figures further emphasise the impact of ground conditions, pitch types, and opposition teams on batsman and bowler performance.

In summary, our analysis demonstrates significant variability in player performance concerning different ground conditions, pitch types, and opposition teams. Machine learning algorithms, particularly random forest, XGB, and lightGBM, prove to be effective tools for forecasting player performance under diverse conditions. Understanding and accounting for these factors are essential for comprehensive player performance analysis and prediction.

4. Discussion

The investigation focused on evaluating the efficacy of various machine learning algorithms in predicting batter performance based on factors such as ground conditions, pitch characteristics, and opposition teams. The outcomes of both statistical analyses and machine learning models offer valuable insights into the significance of these variables and the accuracy of the algorithms employed. The results of the statistical tests revealed significant variations in the average runs scored on different grounds and pitches, indicating the substantial influence of venue and pitch properties on a batsman's scoring potential. Figs. 8 and 9 illustrate the actual and predicted runs and wickets for different algorithms,

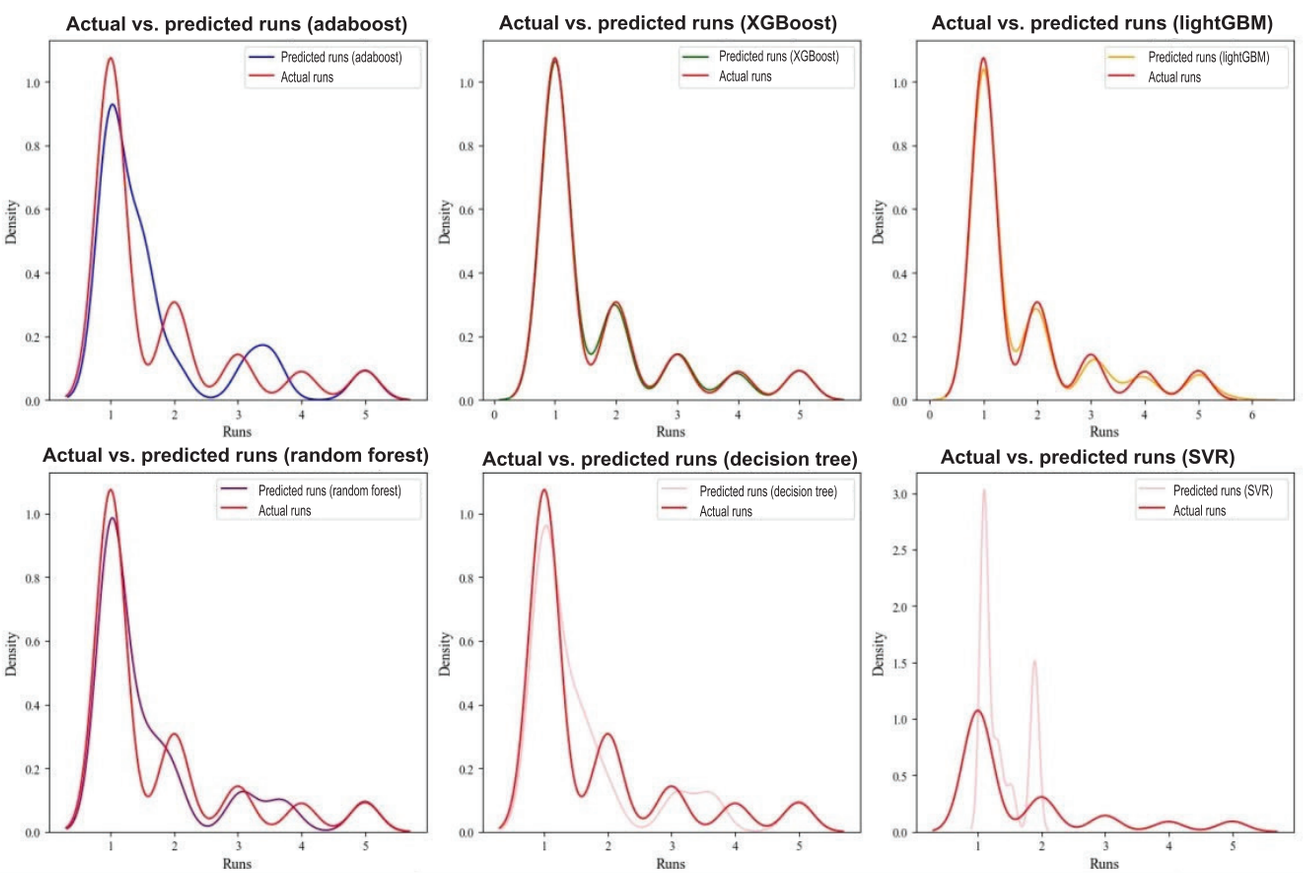

Fig. 8. Actual vs. predicted runs for different algorithms.

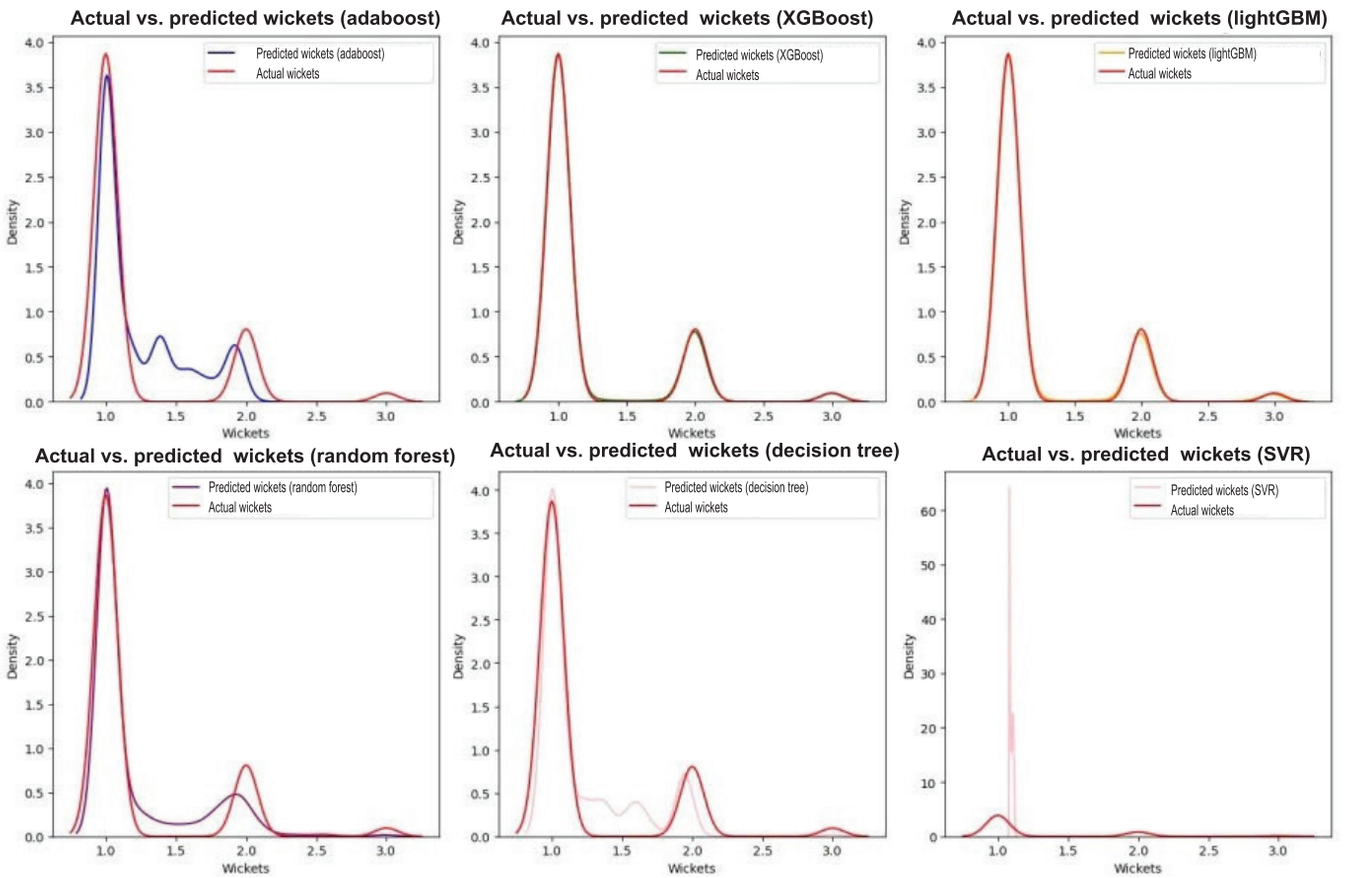

Fig. 9. Actual vs. predicted wickets for different algorithms.

underscoring the importance of considering these variables when assessing and forecasting batter performance.

Several machine learning techniques, including decision tree, random forest, support vector regressor (SVR), AdaBoost, XGBoost, and lightGBM, were employed to further analyse and validate these findings. The performance of these algorithms as predictors was evaluated based on metrics such as accuracy, root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2) scores. The decision tree algorithm achieved an accuracy score of 73.5%, indicating its ability to accurately forecast the outcome of batsman performance in 73.5% of cases. However, the RMSE and MAE values were relatively high, suggesting a significant level of prediction error. The decision tree method accounted for 83.1% of the variance in batting performance, as evidenced by an R^2 score of 0.831. In comparison, the random forest demonstrated superior performance with an accuracy of 81.3% and lower RMSE and MAE values than the decision tree method. Notably, the random forest algorithm explained approximately 91.2% of the variance in batting performance, as indicated by an R^2 score of 0.912. These results highlight the superior predictive accuracy of the random forest algorithm compared to the decision tree method.

The selection of machine learning algorithms for this study was based on their proven success in handling regression tasks and their ability to capture complex relationships in data. The random forest regressor, known for its ensemble approach, was expected to perform well due to its capability to capture intricate interactions between features. Conversely, the decision tree regressor, despite its simpler model structure, was anticipated to excel in predicting performance based on various factors. The SVR algorithm, specifically designed for regression tasks, was evaluated based on its ability to handle non-linear relationships effectively. Similarly, AdaBoost, XGBoost, and lightGBM, renowned for their boosting techniques, were expected to exhibit comparable accuracy and predictive capabilities. Through statistical analysis and comparison of performance metrics, including accuracy, RMSE, MAE, and R^2 scores, this study aimed to demonstrate the significant impact of ground conditions, pitch characteristics, and opposition teams on batsman performance. By elucidating these differences and their implications, the study sought to provide insights into the most effective methods for predicting performance in diverse cricketing scenarios.

5. Conclusions

This study underscores the pivotal role played by contextual factors such as the playing field, pitch conditions, and the quality of opposition teams in accurately predicting cricket player performance. We conducted an evaluation of various machine learning algorithms, including decision tree, random forest, SVR, AdaBoost, XGBoost, and lightGBM. Notably, random forest emerged as the most promising model, achieving an impressive accuracy of 81.3%. Moreover, it effectively explained 91.2% of the variance in batting performance. This superiority over the decision tree model, which exhibited an accuracy rate of 73.5%, was further substantiated by its smaller RMSE and MAE values, indicating fewer prediction errors. The potential applications of sports analytics and data science extend beyond cricket, with these models offering valuable insights for other sports as well. The outcomes of this study lay a strong foundation for the development of more intricate prediction models and data-driven insights, benefiting cricket players, teams, and enthusiasts alike.

CRediT author statement

Rameshwari Lokhande: Conceptualisation, Methodology, Data collection and analysis, Interpretation of the result, Original draft preparation, Revision; Rawal Awale: Interpretation of the result, Revision; Rahul Ingle: Revision.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] R. Chaudhary, S. Bhardwaj, S. Lakra (2019), "A DEA model for selection of Indian cricket team players", *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates, pp.224-227, DOI: 10.1109/AICAI.2019.8701424.

[2] R.A. Stretch, R. Bartlett, K. Davids (2000), "A review of batting in men's cricket", *J. Sports Sci.*, **18(12)**, pp.931-949, DOI: 10.1080/026404100446748.

[3] S.E. Arefin, T.A. Heya, D.M. Zaber (2021), "Predictive analysis of Chikungunya", *arXiv*, DOI: 10.48550/arXiv.2101.03785

[4] R.P. Bunker, F. Thabtah (2019), "A machine learning framework for sport result prediction", *Appl. Comput. Informatics*, **15(1)**, pp.27-33, DOI: 10.1016/j.aci.2017.09.005.

[5] H.H. Lemmer (2014), "Perspectives on the use of the combined bowling rate in cricket", *International Journal of Sports Science & Coaching*, **9(3)**, pp.513-523, DOI: 10.1260/1747-9541.9.3.513.

[6] A. Malhotra, S. Krishna (2018), "Release velocities and bowler performance in cricket", *J. Appl. Stat.*, **45(9)**, pp.1616-1627, DOI: 10.1080/02664763.2017.1386772.

[7] A. Malhotra, S. Krishna (2018), "Release velocities and bowler performance in cricket", *Journal of Applied Statistics*, **45(9)**, pp.1616-1627, DOI: 10.1080/02664763.2017.1386772.

[8] K.P. Jayalath (2017), "A machine learning approach to analyze ODI cricket predictors", *J. Sport. Anal.*, **4(1)**, pp.73-84, DOI: 10.3233/JSA-17175.

[9] T. Allen, O.F. Brault, D. James, et al. (2014), "Finite element model of a cricket ball impacting a bat", *Procedia Eng.*, **72**, pp.521-526, DOI: 10.1016/j.proeng.2014.06.090.

[10] H. Mittal, D. Rikhari, J. Kumar, et al. (2012), "A study on machine learning approaches for player performance and match results prediction", *arXiv*, DOI: 10.48550/arXiv.2108.10125.

[11] Z. Ahmed (2023), "5 types of pitches in cricket, kind of pitch can lead to thrilling performance by batsman or bowler", *TeamCric*, https://teamcric.com/5-types-of-pitches-in-cricket, accessed 1 May 2024.

[12] A. Bandulasiri, T. Brown, I. Wickramasinghe (2016), "Factors affecting the result of matches in the one day format of cricket", *Oper. Res. Decis.*, **4(4)**, pp.21-32, DOI: 10.5277/ord160402.

[13] A.I. Anik, S.Yeaser, A.G.M.I. Hossain, et al. (2018), "Player's performance prediction in ODI cricket using machine learning algorithms", *4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, pp.500-505, DOI: 10.1109/CEEICT.2018.8628118.

[14] G.D.I. Barr, B.S. Kantor (2004), "A criterion for comparing and selecting batsmen in limited overs cricket", *Journal of The Operational Research Society*, **55(12)**, pp.1266-1274, DOI: 10.1057/palgrave.jors.2601800.

[15] H. Lemmer (2022), "The combined bowling rate as a measure of bowling performance in cricket", *South African J. Res. Sport. Phys. Educ. Recreat.*, **24(2)**, pp.37-44, DOI: 10.4314/sajrs.v24i2.25839.

[16] D.G. Pahinkar, J. Srinivasan (2010), "Simulation of reverse swing of the cricket ball", *International Journal of Sport Science and Engineering*, **4(1)**, pp.53-64.

[17] M. Shanthi, L.S. Sarah (2008), "Bowler performance prediction for one-day international cricket using neural networks", *IIE Annu. Conf. Expo,* pp.1391-1395.

[18] H. Saikia, D. Bhattacharjee, U.K. Radhakrishnan (2016), "A new model for player selection in cricket", *International Journal of Performance Analysis in Sport*, **16(1)**, pp.373-388, DOI: 10.1080/24748668.2016.11868893.

[19] G.D. Sharp, W.J. Brettenny, J.W. Gonsalves, et al. (2011), "Integer optimisation for the selection of a twenty 20 cricket team", *J. Oper. Res. Soc.*, **62(9)**, pp.1688-1694, DOI: 10.1057/jors.2010.122.

[20] M.W. Ahmad, J. Reynolds, Y. Rezgui (2018), "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees", *J. Clean. Prod.*, **203**, pp.810-821, DOI: 10.1016/j.jclepro.2018.08.207.

[21] S.N. Omkar, R. Verma (2023), "Cricket team selection using genetic algorithm", *Int. Congr. Sport. Dyn.*, pp.1-3.

[22] A.J. Lewis (2005), "Towards fairer measures of player performance in one-day cricket", *J. Oper. Res. Soc.*, **56(7)**, pp.804-815, DOI: 10.1057/palgrave.jors.2601876.

[23] H. Saikia, D. Bhattacharjee, U. Krishnan (2016), "A new model for player selection in cricket", *International Journal of Performance Analysis in Sport*, **16(1)**, pp.373-388, DOI: 10.1080/24748668.2016.11868893.

[24] M.G. Jhanwar, V. Pudi (2016), "Predicting the outcome of ODI cricket matches: A team composition based approach", *Conference: Machine Learning and Data Mining for Sports Analytics, ECML-PKDD'16*.

[25] T.B. Swartz, P.S. Gill, D. Beaudoin, et al. (2006), "Optimal batting orders in one-day cricket", *Comput. Oper. Res.*, **33(7)**, pp.1939-1950, DOI: 10.1016/j.cor.2004.09.031.

[26] N. Markovi´c, S. Milinkovi´c, K.S. Tikhonov, et al. (2015), "Analysing passenger train arrival delays with support vector regression", *Transportation Research Part C: Emerging Technologies*, **56**, pp.251-262, DOI: 10.1016/j.trc.2015.04.004.

[27] P. Kecman, R.M. Goverde (2015), "Predictive modelling of running and dwell times in railway traffic", *Public Transport*, **7**, pp.295-319, DOI: 10.1007/s12469-015-0106-7.

[28] D. Li, W. Daamen, R.M. Goverde (2016), "Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station", *Journal of Advanced Transportation*, **50(5)**, pp.877-896, DOI: 10.1002/atr.1380.

[29] L. Oneto, E. Fumeo, G. Clerico, et al. (2016), "Advanced analytics for train delay prediction systems by including exogenous weather data", *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp.458-467, DOI: 10.1109/DSAA.2016.57.

[30] L. Oneto, E. Fumeo, G. Clerico, et al. (2018), "Train delay prediction systems: A big data analytics perspective", *Big Data Research*, **11**, pp.54-64, DOI: 10.1016/j.bdr.2017.05.002.

[31] Y. Liu, T. Tang, J. Xun (2017), "Prediction algorithms for train arrival time in urban rail transit", *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp.1-6, DOI: 10.1109/ITSC.2017.8317609.

[32] A. Lulli, L. Oneto, R. Canepa, et al. (2018), "Large-scale railway networks train movements: A dynamic, interpretable, and robust hybrid data analytics system", *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp.371-380, DOI: 10.1109/DSAA.2018.00048.

[33] C. Jiang, P. Huang, J. Lessan, et al. (2019), "Forecasting primary delay recovery of high speed railway using multiple linear regression, supporting vector machine, artificial neural network and random forest regression", *Canadian Journal of Civil Engineering*, **46(5)**, DOI: 10.1139/cjce-2017-0642.

[34] M. Arshad, M. Ahmed (2021), "Train delay estimation in Indian railways by including weather factors through machine learning techniques", *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, **14(4)**, pp.1300-1307, DOI: 10.2174/2666255813666190912095739.

[35] Z.C. Li, C. Wen, R. Hu, et al. (2021), "Near-term train delay prediction in the Dutch railways network", *International Journal of Rail Transportation*, **9(6)**, pp.520-539, DOI: 10.1080/23248378.2020.1843194.

[36] R. Nair, T.L. Hoang, M. Laumanns, et al. (2019), "An ensemble prediction model for train delays", *Transportation Research Part C: Emerging Technologies*, **104**, pp.196-209, DOI: 10.1016/j.trc.2019.04.026.

[37] W. Mou, Z. Cheng, C. Wen (2019), "Predictive model of train delays in a railway system", *Proc. of The 8th International Conference on Railway Operations Modelling Analysis (Rail-Norrk¨oping)*, pp.913-929.

[38] R. Shi, J. Wang, X. Xu, et al. (2020), "Arrival train delays prediction based on gradient boosting regression trees", *Proceedings of The 4th International Conference on Electrical and Information Technologies for Rail Transportation (EITRT) 2019: Rail Transportation Information Processing and Operational Management Technologies,* Springer, pp.307-315, DOI: 10.1109/ITSC45102.2020.9294365.

[39] P. Huang, C. Wen, L. Fu, et al. (2020), "A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems", *Information Sciences*, **516**, pp.234-253, DOI: 10.1016/j.ins.2019.12.053.

[40] J. Ma, Z. Yu, Y. Qu, et al. (2020), "Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: A case study of Shanghai", *Aerosol Air Qual. Res.*, **20(1)**, pp.128-138, DOI: 10.4209/aaqr.2019.08.0408.

[41] M. Asgari, M. Farnaghi, Z. Ghaemi, et al. (2017), "Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster", *Proceedings of The 2017 International Conference on Cloud and Big Data Computing*, pp.89-93, DOI: 10.1145/3141128.31411.

[42] M.Z. Joharestani, C. Cao, X. Ni, et al. (2019), "$PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data", *Atmosphere*, **10(7)**, DOI: 10.3390/atmos10070373.

[43] H. Karimian, Q. Li, C. Wu, et al. (2019), "Evaluation of different machine learning approaches to forecasting $PM_{2.5}$ mass concentrations", *Aerosol and Air Quality Research*, **19(6)**, DOI: 10.4209/aaqr.2018.12.0450.

[44] J.K. Deters, R. Zalakeviciute, M.G. Zalez, et al. (2017), "Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters", *Journal of Electrical and Computer Engineering*, DOI: 10.1155/2017/5106045.

[45] I.M.Gad, H. Doreswamy, K.S Harishkumar, et al. (2020), "Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models", *Procedia Computer Science*, **171**, pp.2057-2066, DOI: 10.1016/j.procs.2020.04.221.

[46] R.O. Sinnott, Z. Guan (2018), "Prediction of air pollution through machine learning approaches on the cloud", *IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, pp.51-60, DOI: 10.1109/BDCAT.2018.00015.

[47] M. Lee, L. Lin, C.Y. Chen, et al. (2020), "Forecasting air quality in Taiwan by using machine learning", *Scientific Reports*, **10(1)**, DOI: 10.1038/s41598-020-61151-7.

[48] A. Masood, K. Ahmad (2010), "A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches", *Procedia Computer Science*, **167**, pp.2101-2110, DOI: 10.1016/j.procs.2020.03.258.

[49] X. Yu, Y. Wang, L. Wu, et al. (2020), "Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting", *J. Hydrol.*, **582**, DOI: 10.1016/j.jhydrol.2019.124293.

[50] Y.G. Wang, J. Wu, Z.H. Hu, et al. (2023), "A new algorithm for support vector regression with automatic selection of hyper parameters", *Pattern Recognit.*, **133**, DOI: 10.1016/j.patcog.2022.108989.

[51] A. Banga, R. Ahuja, S.C. Sharma (2021), "Performance analysis of regression algorithms and feature selection techniques to predict $PM_{2.5}$ in smart cities", *Int. J. Syst. Assur. Eng. Manag.*, **14**, pp.732-745, DOI: 10.1007/s13198-020-01049-9.

[52] J. Xiong, S.Q. Shi, T.Y. Zhang (2021), "Machine learning of phases and mechanical properties in complex concentrated alloys", *J. Mater. Sci. Technol.*, **87**, pp.133-142, DOI: 10.1016/j.jmst.2021.01.054.

[53] W. Yuchi, E. Gombojav, B. Boldbaatar, et al. (2019), "Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city", *Environ. Pol.*, **245**, pp.746-753, DOI: 10.1016/j.envpol.2018.11.034.

[54] H. Laifa, R. Khcherif, H.H.B. Ghezalaa (2021), "Train delay prediction in Tunisian railway through lightGBM model", *Procedia Comput. Sci.*, **192**, pp.981-990, DOI: 10.1016/j.procs.2021.08.101.

[55] W. Dai, T.S. Brisimi, W.G. Adams, et al. (2015), "Prediction of hospitalization due to heart diseases by supervised learning methods", *Int. J. Med. Inform.*, **84(3)**, pp.189-197, DOI: 10.1016/j.ijmedinf.2014.10.002.

[56] M.Y. Shams, O.M. Elzeki, L.M.A. Magd, et al. (2021), "Hana: A healthy artificial nutrition analysis model during COVID-19 pandemic", *Comput. Biol. Med.*, **135**, DOI: 10.1016/j.compbiomed.2021.104606.

[57] A. Shehadeh, O. Alshboul, R.E.A.M. Look, et al. (2020), "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, lightGBM, and XGBoost regression", *Autom. Constr.*, **129**, DOI:10.1016/j.autcon.2021.103827.

[58] T. Thongthammachart, S. Araki, H.S. Madera, et al. (2022), "Incorporating light gradient boosting machine to land use regression model for estimating NO_2 and $PM_{2.5}$ levels in Kansai region, Japan", *Environ. Model. Softw.*, **155**, DOI: 10.1016/j.envsoft.2022.105447.

[59] D.N. Wang, L. Li, D. Zhao (2022), "Corporate finance risk prediction based on LightGBM", *Inf. Sci. (Ny)*, **602**, pp.259-268, DOI: 10.1016/j.ins.2022.04.058.