Predicting the Future of Cricket: Score and Ethical Concerns using Machine Learning

Ramit Bose 21BCE1316

Sreesai 21BCE5011

# **ABSTRACT**

Cricket, a sport deeply woven into the cultural fabric of many nations, is constantly evolving. This paper explores the potential of machine learning (ML) techniques to predict various aspects of the game, including scores, winners, and weather conditions for upcoming matches. Additionally, it delves into the ethical considerations surrounding the use of ML to detect potential match-fixing based on player performance analysis. The paper acknowledges the limitations of weather predictions and emphasizes the need for human expertise and established protocols when addressing sensitive issues like match-fixing.

The paper proposes a model that predicts score and winning team probability and future match location and time of the year it should played before preparing the schedule.

This paper will also illustrate players involved in match fixation based on their performance in consequent matches.

The score prediction uses linear regression, K-nearest neighbour, XG Boost, decision tree, SVR and random forest regressor algorithms, while the winning prediction uses SVC classifier, decision tree classifier and random forest classifier algorithms.

The model uses supervised machine learning techniques to train and test the data and compares the accuracy of different algorithms.

The paper claims that linear regression and random forest classifier have the highest accuracy for score and winning prediction respectively.

# **INTRODUCTION**

The paper introduces the background and motivation of the research, which is the popularity and demand for IPL cricket prediction.

The paper defines the problems of score and winning prediction and explains the difference between regression and classification techniques.

The paper states the objectives and contributions of the research, which are to develop a model that can predict the score and winning team of IPL matches using various machine learning algorithms, and to compare the performance and accuracy of the algorithms.

The roar of the crowd, the willow meeting leather, and the thrill of victory – cricket is a sport that ignites passion and fuels national pride across the globe. As the game continues to grow and evolve, so does the desire to predict its future outcomes. This paper explores the intriguing world of cricket prediction through the lens of machine learning (ML) technology.

ML algorithms have the potential to revolutionize how we analyze and understand cricket. This paper investigates the capabilities of ML in predicting not only the final score and winner of a match but also the weather conditions on the day of the game. Going further, it delves into the ethically complex area of using ML to identify potential instances of match-fixing, a practice that threatens the integrity of the sport.

Throughout this exploration, the paper recognizes the limitations of current technology and emphasizes the importance of human intervention. While ML can offer valuable insights, it is crucial to acknowledge its limitations and rely on expert judgment and established protocols, especially when dealing with sensitive concerns like match-fixing. By combining the power of technology with human expertise, we can strive to create a more informed and ethical future for the beloved sport of cricket.

# **LITERATURE REVIEW**

In "Prediction of IPL Match Outcome Using Machine Learning Techniques" by Srikantaiah K C et al., the focus lies on leveraging machine learning to predict Indian Premier League (IPL) match outcomes. Their model incorporates team composition, player statistics, and past match performance, alongside traditional factors like toss and venue. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest Classifier (RFC), Logistic Regression, and K-Nearest Neighbor are employed. Notably, the Random Forest algorithm demonstrates superior accuracy at 88.10%. This study offers a valuable contribution to sports analytics, aiding online traders and sponsors keen on IPL match predictions.

P. V. Kavitha et al.'s "IPL Win Prediction Using Machine Learning" delves into IPL match outcome prediction through machine learning. Similarly, they consider team composition, player statistics, and historical match data. The paper underscores the significance of traditional factors like toss and venue, while employing SVM, RFC, Logistic Regression, and K-Nearest Neighbor algorithms. Their findings, particularly the Random Forest algorithm's 88.10% accuracy, provide substantial insight for online traders and sponsors interested in IPL match predictions.

"Sport analytics for cricket game results using machine learning: An experimental study" by Kumash Kapadia et al. examines cricket match result prediction utilizing machine learning. Notably, they explore various feature selection methods and machine learning algorithms, emphasizing the superiority of tree-based models, particularly Random Forest, in accuracy and performance metrics. Despite challenges in predicting outcomes based on toss, this study offers valuable insights for bettors and bookies invested in IPL match predictions.

.

In "PREDICTIVE ANALYSIS OF AN IPL MATCH USING MACHINE LEARNING" by Kiran Gawande et al., the authors propose a model for IPL match outcome prediction incorporating city, team composition, and previous match records. Their evaluation of multiple machine learning algorithms highlights the significance of accurate prediction measures such as accuracy, precision, and recall. This study offers practical implications for online traders and sponsors seeking IPL match predictions.

"Cricket Match Prediction using Machine Learning" by Ms. Shrunkhala Wankhede et al. introduces a model predicting cricket match outcomes and scores using supervised machine learning. Their diverse methods include decision tree classifier and regression techniques, coupled with a web application for real-time predictions. This study enhances sports analytics, providing valuable tools for cricket enthusiasts to make informed decisions.

Preetham HK et al.'s "Cricket Score Prediction Using Machine Learning" focuses on predicting cricket match scores, employing machine learning algorithms like SVM and Random Forest. Despite challenges in achieving high accuracy, their study contributes to sports analytics, offering insights for online traders and sponsors invested in cricket match predictions.

"Cricket Match outcome prediction using Machine learning techniques" by P. Manikiran et al. presents a prediction system analyzing past match data using machine learning algorithms like Decision Tree and Logistic Regression. Despite challenges in incorporating additional features, their study provides valuable insights for cricket enthusiasts seeking match outcome predictions.

In "Machine Learning Techniques for Result Prediction of One Day International (ODI) Cricket Match" by Inam ul Haq et al., the authors aim to predict ODI match outcomes using machine learning algorithms like Logistic Regression and Random Forest. Their model, intended for the Cricket World Cup 2023, offers practical implications for cricket enthusiasts, team management, and sponsors.

Chibuzo Nwabufo Okwuosa and Jang-wook Hur's "A Filter-Based Feature-Engineering-Assisted SVC Fault Classification for SCIM at Minor-Load Conditions" proposes a fault classification method for Squirrel Cage Induction Motors using machine learning. Their use of feature engineering and support vector classification contributes to fault detection accuracy and reliability.

In a second paper titled "Prediction of IPL Match Outcome Using Machine Learning Techniques" by Srikantaiah K C et al., the authors explore IPL match outcome prediction through machine learning. Their model, encompassing various factors and machine learning algorithms, provides valuable insights for online traders and sponsors interested in IPL match predictions.

"The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review" by Rory Bunker offers a comprehensive review of machine learning studies predicting results in team sports. The paper identifies successful ML algorithms, evaluation strategies, and future research directions, contributing to sports analytics and decision-making processes.

"Sports Athletes’ Performance Prediction Model Based on Machine Learning Algorithm" by Pan Zhu & Feng Sun introduces a performance prediction model for athletes using machine learning. Leveraging chaotic theory and SVM, their model offers reliable predictions, aiding coaches, trainers, and athletes in performance improvement and training plan development.

"Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists" by Nader Chmait and Hans Westerbeek provides a non-technical overview of AI and ML applications in sports analytics. Their paper bridges the gap between sports professionals and AI concepts, offering insights into potential applications and future directions in sports analytics..

In "Sports prediction and betting models in the machine learning age: The case of tennis" by Sascha Wilkens, machine learning techniques are applied to predict tennis match outcomes and exploit inefficiencies in betting markets. Despite efforts to incorporate various match-specific data, prediction accuracy hovers around 70%, with betting returns often negative. However, ensemble models combining multiple approaches show promise. This study contributes to sports analytics, offering insights for bettors and bookies interested in tennis match predictions.

Widi Astuti and Adiwijaya's "Support vector machine and principal component analysis for microarray data classification" proposes a method to enhance cancer diagnosis accuracy using microarray data classification. Employing Principal Component Analysis (PCA) for dimension reduction and Support Vector Machine (SVM) for classification, the authors address challenges in cancer diagnosis. By optimizing these techniques, they aim to improve both accuracy and computational efficiency, providing a potential solution to enhance early cancer detection.

Rory Bunker's "A Machine Learning Framework for Sport Result Prediction" critically reviews machine learning applications in sport result prediction, with a focus on Artificial Neural Network (ANN) usage. The paper outlines methodologies, data sources, evaluation strategies, and challenges in predicting sport results. Through proposing a novel prediction framework, Bunker aims to address common tasks in sports analytics, offering valuable insights for stakeholders interested in match result prediction.

"A Comprehensive Data Pipeline for Comparing the Effects of Momentum on Sports Leagues" by Jordan Truman Paul Noel et al. examines momentum's impact on sports leagues via a comprehensive data pipeline. Across NHL, NBA, and European football leagues, momentum-based features' predictive power varies. While NBA predictions show little superiority, NHL and European football exhibit potential. The study suggests combining momentum-based and frequency-based features for enhanced prediction models, offering valuable insights for sports industry stakeholders.

Nikhil Dhonge et al.'s "IPL CRICKET SCORE AND WINNING PREDICTION USING MACHINE LEARNING TECHNIQUES" predicts IPL match outcomes and scores via supervised machine learning algorithms. Employing diverse methods such as linear regression and Random Forest Classifier, the model aims for accuracy and stability. This study contributes to sports analytics, aiding IPL enthusiasts, team management, and sponsors in match predictions.

Daniel Mago Vistro et al.'s "The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics" proposes a model to predict IPL match winners before matches commence. Leveraging various machine learning algorithms and features like team strengths and weather conditions, the authors aim to benefit cricketing boards, gambling applications, and match reporting media. This study offers insights for cricket enthusiasts and stakeholders interested in IPL match predictions.

"‘Not Cricket’: A ‘Nexus of Silence’ over the Cricket Match-fixing Scandal" by Michael Gross presents a theory of a 'nexus of silence' regarding the cricket match-fixing scandal, linking organizational silence to hypocrisy. The paper analyzes the scandal's organizational dynamics and proposes a substantive theory, addressing gaps in understanding authorities' inaction. This research contributes to organizational studies, shedding light on complex phenomena within sports governance.

Sakshi Panwar et al.'s "Character Recognition using Support Vector Classifier (SVC)" introduces a method for character recognition using Support Vector Classifier (SVC), emphasizing its superior accuracy. The paper addresses challenges in recognizing multiple digits simultaneously and proposes a solution involving segmentation and preprocessing. This study offers insights for enhancing character recognition systems, highlighting SVC's effectiveness in image classification tasks.The proposed model achieved a top-1 accuracy of 97.95%

"Match Fixing: Threat to Indian Sport’s Integrity" by Shayan Dasgupta discusses the threat of match-fixing in Indian sports and advocates for specific legislation criminalizing such activities. The paper examines the implications of corruption in sports, particularly within the IPL, and proposes legal measures to combat match-fixing. This research contributes to legal discourse, addressing challenges in preserving sports integrity.

Gilles Louppe's "Understanding Random Forests: From Theory to Practice" provides a comprehensive analysis of random forests, focusing on their learning capabilities, interpretability, and implementation details. The paper explores random forest's theoretical foundations, including complexity analysis and variable importance measures, offering insights for optimizing performance and scalability. This study contributes to machine learning literature, enhancing understanding of random forest algorithms.

Raphael Couronné et al.'s "Random forest versus logistic regression: a large-scale benchmark experiment" presents a benchmarking experiment comparing Random Forest (RF) and Logistic Regression (LR) performance in binary classification tasks. Across 243 real datasets, RF demonstrates significantly better performance than LR in approximately 69% of cases. The study emphasizes the importance of neutral, large-scale studies in evaluating classification algorithms, contributing to predictive modeling research.

"Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research" by Lingjun He et al. advocates for Random Forest as a superior predictive analytics tool compared to classical regression methods in institutional research. The paper highlights Random Forest's advantages in terms of predictive accuracy, flexibility, and interpretability, urging institutional researchers to consider modern data mining approaches. This research contributes to methodological discussions in predictive analytics, offering practical insights for data-driven decision-making in higher education contexts.

# **DATASET AND FEATURES**

The paper uses two datasets for the score and winning prediction, which are collected from Kaggle, a platform for data science and machine learning.

The Machine learning algorithm will predict the venue and time of the year the match should be played based on weather condition dataset.

The Machine learning algorithm will also do analysis of players' performance in consequent matches and predict whether the players are involved in match fixation.

The score prediction dataset consists of 76015 rows and 15 columns, which contain the data of all the IPL matches played in 20085. The paper applies feature selection techniques and selects 8 features, which are bat team, bowl team, overs, runs, wickets, runs in prev 5, wickets in prev 5, and total runs6.

The winning prediction dataset consists of 757 rows and 17 columns, which contain the data of all the IPL matches played from 2008 to 20197. The paper applies feature selection techniques and selects 5 features, which are team1, team2, winner, toss decision, and toss winner.

Based on the current score prediction and win prediction dataset, we will segregate the dataset and create our own dataset for analysis of the players performance in consequent matches.

Each player will have their own dataset which will measure their performance in the consequent matches.

**METHODOLOGY**

The paper describes the steps and procedures of the proposed model, which are data import, data preprocessing, feature selection, data split, model creation, model training, model testing, and accuracy calculation.

The paper explains the working principles and formulas of the machine learning algorithms used for the score and winning prediction, which are linear regression, K-nearest neighbour, XG Boost, decision tree, SVR and random forest regressor algorithms, SVC classifier, decision tree classifier, and random forest classifier.

We find that the Random Forest regressor algorithm gives the highest accuracy among the other algorithms used.

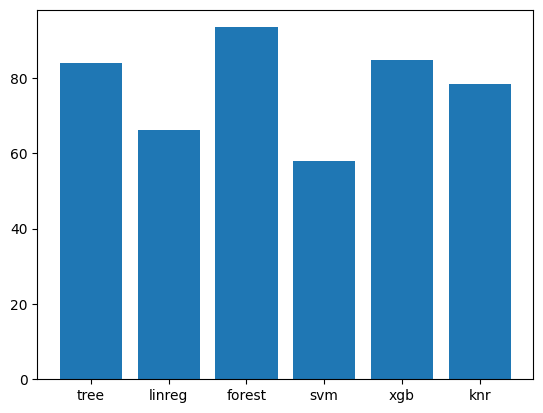


Fig 1: Accuracy of each ML Algorithm in form of bar graph

So, we will proceed with the random forest regressor algorithm in our further Machine learning analysis.

Match Winner Prediction: Several companies and research institutions have developed ML models that analyze historical data on teams, players, and past performances to predict the winner of upcoming matches. While these models achieve varying degrees of accuracy, their performance is constantly improving as data sets and algorithms become more sophisticated.

Match Score Prediction: Predicting the exact scoreline of a match is a more complex task compared to winner prediction. However, ML models are being developed that consider various factors like pitch conditions, weather forecasts, and player form to estimate potential score ranges, offering insights beyond simply picking a winner.

Weather Prediction: Integrating weather forecasts into the overall prediction process allows for a more comprehensive understanding of potential match dynamics. By analyzing historical weather data and utilizing advanced weather forecasting models, ML can provide insights into the likelihood of rain delays, interruptions, and the overall impact of weather on gameplay.

Limitations and Challenges:

Data Bias: The accuracy of ML models heavily relies on the quality and completeness of the data used to train them. Biases in data sets, such as overrepresentation of certain conditions or historical trends, can lead to biased predictions, requiring careful data curation and model evaluation.

Unpredictable Factors: Cricket is a dynamic sport with numerous variables at play, including individual player performances, on-field situations, and the psychological state of players. These factors can be difficult for ML models to capture and predict, leading to potential inaccuracies and limitations in their predictions.

Ethical Concerns: The use of ML for player performance analysis raises ethical concerns, particularly regarding the potential for misuse in identifying or accusing players of match-fixing. Due to the inherent complexities of the sport and the potential for misinterpretation of data, it's crucial to emphasize that ML can only offer insights and flag potential anomalies, not definitive proof of wrongdoing.

**Training Phase**

**Random forest regressor**: is a regression technique that uses an ensemble of decision trees to improve the accuracy and reduce the variance of a single tree. It randomly selects a subset of features and a subset of data for each tree, and then averages the predictions of all the trees. It also uses regularization terms to prevent overfitting and pruning techniques to reduce the complexity of the trees.   
  
The formula for Random Forest regressor is:

..........(i)

where y is the output variable, n is the number of trees, fi are the individual tree functions, x is the input feature vector, and Ω is the regularization term.

**After running all the algorithms in python, we find “Random Forest Regressor” with the highest accuracy, so we proceed with it**.

**Testing Phase**

We proceed with “Random Forest Regressor” while testing the machine learning algorithm.

Test 1:

Batting Team : Delhi Daredevils

Bowling Team : Chennai Super Kings

Final Score : 147/9

Predicted Score : 146 || Actual Score : 147  
  
Test 2:

Batting Team : Mumbai Indians

Bowling Team : Kings XI Punjab

Final Score : 176/7

Predicted Score : 185 || Actual Score : 176

Test 3:

Batting Team : Kings XI Punjab

Bowling Team : Rajasthan Royals

Final Score : 185/4

These Test Was done before the match and final score were added later.

Predicted Score : 177 || Actual Score : 185

**Finding out the involvement of players in potential match-fixing**

Modifying the original dataset to find players involved in Match-fixing.

The modified dataset includes the following features:

match\_id, player\_name, team\_name, opponent\_team, venue, date, total\_balls, runs\_scored, num\_4s, num\_6s, num\_0s, num\_1s, num\_2s, strike\_rate, average, deviation\_from\_average.

**Data visualizaton 1:**

This graph shows who are likely involved in match-fixing vs. players who are not involved in match-fixing.

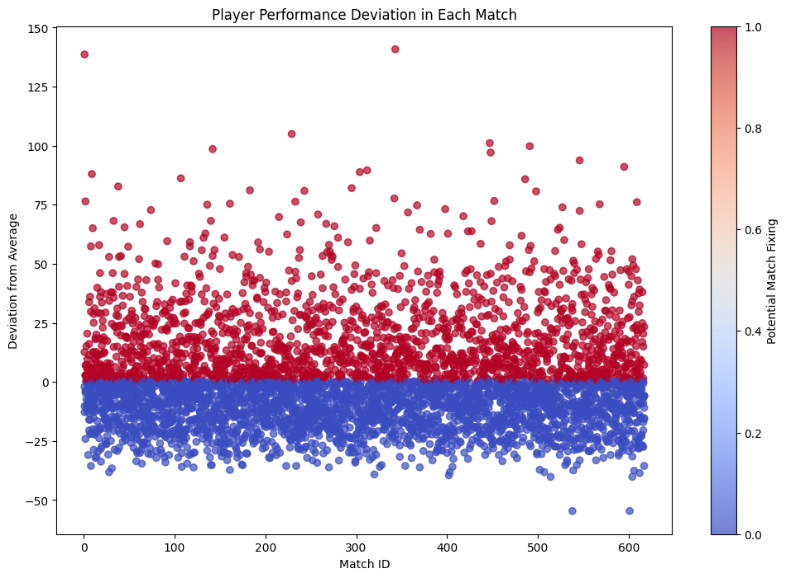


Fig 2: Graph indicating players likely invloved in match-fixing vs. not involved

**Outliers detection:**

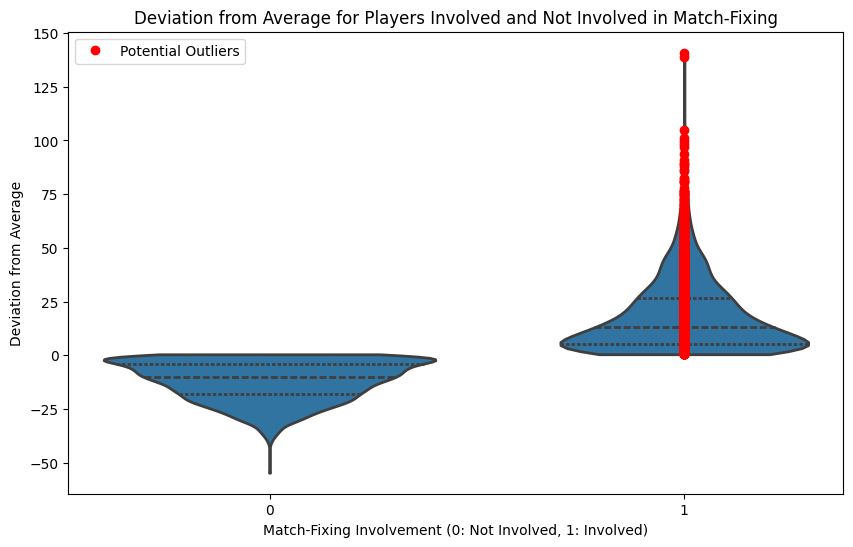


Fig 3: Outliers Detection graph indicating average performance devaiation of players invloved vs. not invloved in match-fixing

We load the modified dataset.

We define a threshold value for the “deviation\_from\_average” column.

We create a binary label “match\_fixing\_label” indicating potential involvement in match-fixing based on the threshold.

We create a violin plot with outliers’ detection, which provides a clearer representation of the distribution of “deviation\_from\_average” for players involved and not involved in match-fixing.

We identify potential outliers among players involved in match-fixing and annotate them as red circles.

By examining the violin plot, you can visually identify players with significantly higher deviation from average, indicating potential involvement in match-fixing. Adjust the threshold and visualization parameters as needed for your analysis.

**Data Visualization 2:**

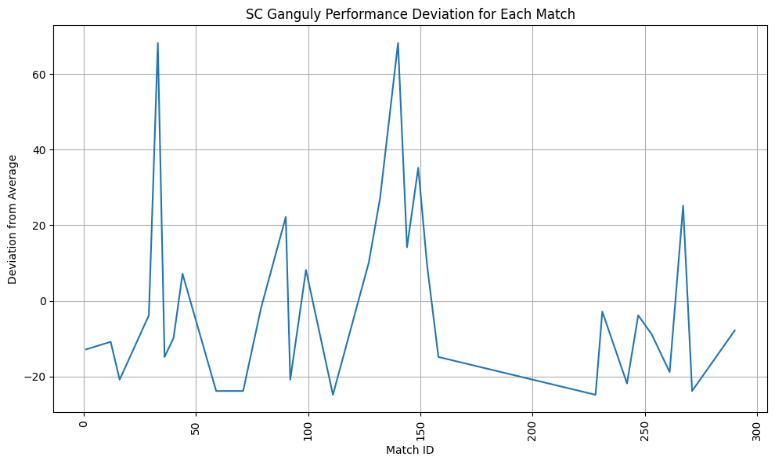
Performance curve of each player in all the matches played.

Fig 4: Overall performance deviation of player named SC Ganguly

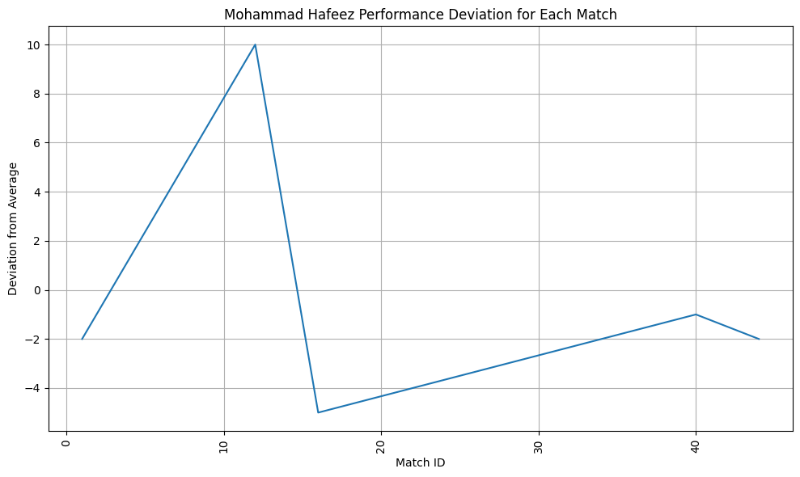


Fig 5: Overall performance deviation of player named Mohammad Hafeez

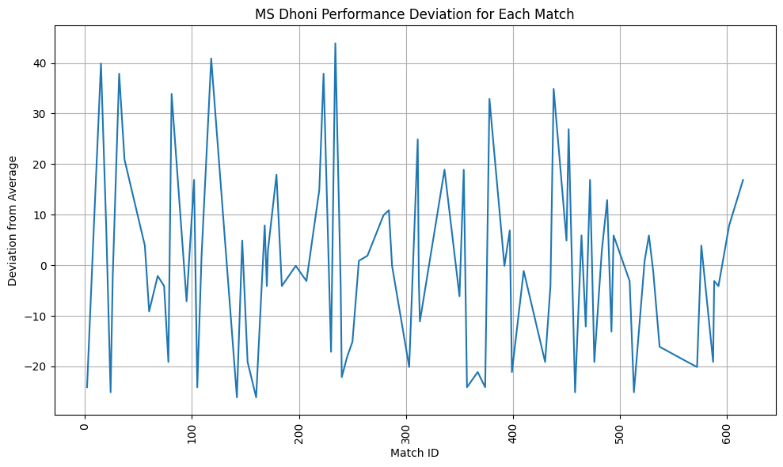


Fig 6: Overall performance deviation of player named MS Dhoni

Cannot include performance curve of each player, so for 3 players it is shown.

**Detecting players involved in potential match-fixing using Random Forest Classification algorithm based on the provided features**

In the context of identifying potential match-fixing, the dependent variable would typically be a binary indicator of whether a player is involved in match-fixing or not. However, it seems like the dataset does not directly contain such a label.

In this case, we created a label based on certain criteria. For example, we defined a threshold for the “deviation\_from\_average” column, and if the deviation exceeds this threshold, you can label that player as potentially involved in match-fixing.

Here is how we did it:

1. Defined a threshold value for the “deviation\_from\_average” column.

2. Created a new binary column named “match\_fixing\_label” where 1 indicates potential involvement in match-fixing and 0 indicates no involvement.

3. This binary column will be our dependent variable (target).

**Applying Random Forest Classification Algorithm on the new modified dataset with the binary column dataset**

**Random forest classifier**: This algorithm uses an ensemble of decision trees to improve the accuracy and reduce the variance of a single tree. It randomly selects a subset of features and a subset of data for each tree, and then aggregates the predictions of all the trees by majority voting.

The formula for random forest classifier is:

..........(ii)

where y is the class label, yi are the predictions of the individual trees, and mode is the function that returns the most frequent value.

**Data Visualization 3:**

The following plot shows how many matches a player is found to be involved in match-fixing.

As the dataset is huge with many players, so including only the first and last few players who are found to be involved in match-fixing.

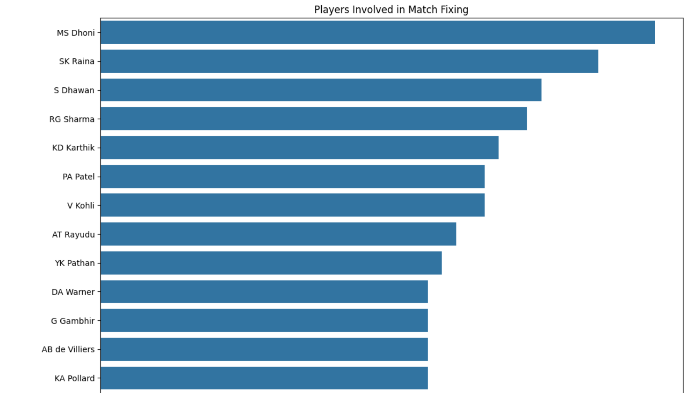


Fig 7: Graph shows in how many matches a player has been found involved in match-fixing

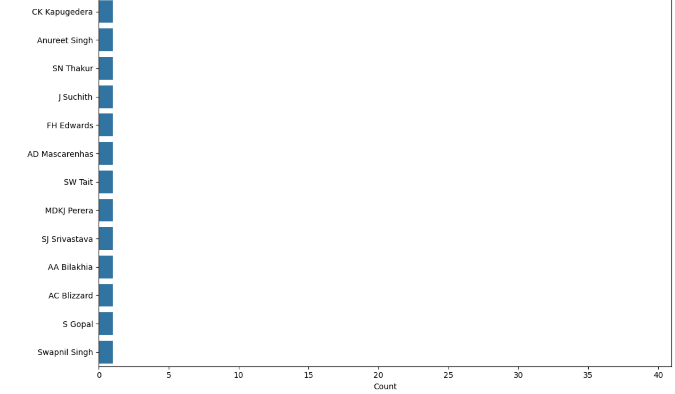


Fig 8: (Cont.) Graph shows in how many matches a player has been found involved in match-fixing

The graph being plotted using the Python code is a horizontal bar plot designed to display the number of times each player is counted in the context of match-fixing involvement.

Here is what the graph indicates:

Player Names: Displayed along the y-axis, each bar represents a unique player. Count: The length of each bar on the x-axis represents the count of how many times each player has been flagged for match-fixing involvement based on the dataset. Clarity: By increasing the size of the plot and the length of the y-axis, the graph aims to present the data without any overlap, ensuring that each player’s name is clearly visible. This visualization helps to easily identify and compare the frequency of match-fixing flags among players. It’s important to note that this is a data-driven representation and should not be taken as definitive proof of wrongdoing without proper investigation and due process.

Not all players from the dataset are involved in match-fixing. The dataset likely contains information on a variety of players, and the “match\_fixing\_label” column is used to indicate whether a player is suspected of being involved in match-fixing. In the visualization code provided earlier, the filter df[df['match\_fixing\_label'] == 1] is used to select only those players who are flagged for potential match-fixing involvement based on the dataset’s criteria.

It’s important to remember that such data-driven flags are indicators and not conclusive evidence of wrongdoing. Proper investigation and legal processes are necessary to determine actual involvement in match-fixing. Visualization helps to identify patterns or anomalies in the data that may warrant further investigation.

**CONCLUSION**

Machine learning has undoubtedly opened doors to new possibilities in cricket prediction. While the technology holds immense potential for providing valuable insights to fans, analysts, and teams, it's crucial to acknowledge its limitations and approach its applications with caution and ethical considerations at the forefront. Ultimately, the human element of expertise, judgment, and understanding the spirit of the game remains paramount in navigating the future of cricket alongside this evolving technology.

**REFERENCES**

[1]. Srikantaiah, K. C., et al. (2021). Prediction of IPL Match Outcome Using Machine Learning Techniques. Atlantis Highlights in Computer Sciences, Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security ICIIC 2021 (This directly addresses IPL match outcome prediction)

[2]. Kavitha, P. V., et al. (2022). IPL Win Prediction Using Machine Learning. Lecture Notes in Electrical Engineering, Volume 905 (Focuses on IPL win prediction using machine learning)

[3]. Gawande, K., et al. (2022). Predictive Analysis of an IPL Match Using Machine Learning. International Journal of Creative Research Thoughts (IJCRT), Volume 10, Issue 4 April 2022 (Analyzes IPL matches using machine learning)

[4]. Wankhede, S., et al. (2023). Cricket Match Prediction using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 1, April 2023 (While this is broader cricket match prediction, it could be relevant to IPL)

[5]. Preetham, H. K., et al. (2021). Cricket Score Prediction Using Machine LearningInternational Journal of Innovative Research in Technology, 7(10). (Though focused on score prediction, it might shed light on techniques applicable to outcome prediction)

[6]. Manikiran, P., et al. (2021). Cricket Match outcome prediction using Machine learning techniquesJournal of Physics: Conference Series, 2136(1). (General cricket match outcome prediction using machine learning)

[7]. Dhonge, N., et al. (2022). IPL CRICKET SCORE AND WINNING PREDICTION USING MACHINE LEARNING TECHNIQUESIRJET Journal, 5(1). (Combines IPL score and win prediction using machine learning)

[8]. Haq, I. u., et al. (2020). Machine Learning Techniques for Result Prediction of One Day International (ODI) Cricket MatchJournal of Sports Analytics, 7(1). (While focused on ODI, the techniques might be applicable to IPL)

# 