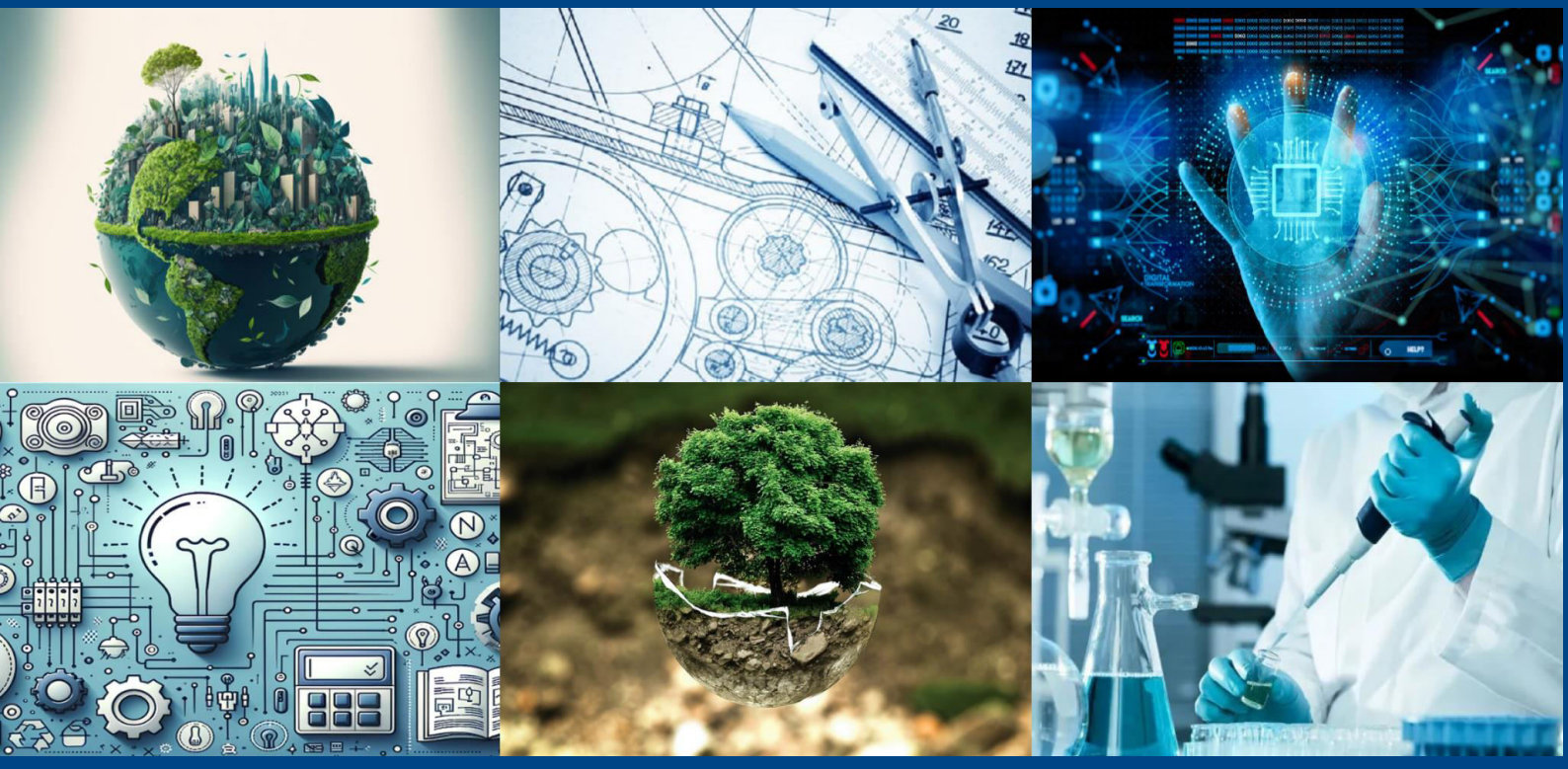
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 3, March 2025**

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803114

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Using Linear Regression to Predict Scores in IPL Matches

Jeevarathinam A[1], Dinesh Karthick P[2]

Assistant professor Department of Computer Science, Sri Krishna Arts and Science College, Tamil Nadu, India[1]

Student III B.Sc. Department of Computer Science, Sri Krishna Arts and Science College, Tamil Nadu, India[2]

ABSTRACT: The Indian Premier League (IPL) has transformed T20 cricket, blending entertainment with strategic gameplay. Accurately predicting the first-inning score is crucial for teams, coaches, analysts, and fantasy cricket participants. This study investigates the application of machine learning techniques in forecasting first-inning scores based on historical IPL data. The dataset included extensive match records, player performance metrics, pitch conditions, and environmental factors such as temperature and humidity. Four machine learning models — Linear Regression, Random Forest Regression, Gradient Boosting, and XGBoost — were applied to evaluate predictive accuracy. Each model was trained and tested using carefully pre-processed data, ensuring balanced feature selection and optimized hyperparameters. The results indicated that XGBoost achieved the highest prediction accuracy due to its robust handling of non-linear relationships and feature interactions. Gradient Boosting followed closely, demonstrating strong performance in capturing data patterns. Conversely, Random Forest Regression and Linear Regression exhibited comparatively lower accuracy, suggesting limitations in handling complex cricket dynamics. The study emphasizes the potential of machine learning in enhancing sports analytics, providing actionable insights for strategic planning, player performance assessment, and real-time decision-making during IPL matches. Future research could further improve model accuracy by integrating real-time data feeds, player fitness conditions, and dynamic match scenarios.

KEYWORDS: IPL, First Inning Score Prediction, Machine Learning, Regression Analysis, Sports Analytics, Predictive Models

I. INTRODUCTION

Cricket, particularly the T20 format, has evolved into a data-rich sport where insights derived from analytics play a crucial role in enhancing performance and strategic decision-making. Among various T20 leagues worldwide, the Indian Premier League (IPL) stands out as one of the most popular and competitive platforms, drawing immense global attention. Predicting the first-inning score in IPL matches has emerged as a vital aspect of sports analytics, with applications extending to team strategies, match commentary, fantasy cricket platforms, and betting markets. Accurate score predictions enable teams to make informed decisions regarding batting approaches, field placements, and bowling strategies, ultimately influencing match outcomes. Consequently, this area has garnered growing interest from researchers aiming to develop effective prediction models using machine learning techniques.

The prediction of a first-inning score is complex due to the dynamic nature of T20 cricket, where several unpredictable factors can influence the outcome. Variables such as team performance trends, individual player form, pitch conditions, and environmental factors like temperature and humidity can significantly impact the total score. Moreover, factors such as the toss outcome, match venue, and opposition strength further add to the intricacies of score prediction. The challenge lies in capturing these diverse influences effectively and translating them into accurate score estimates.

Machine learning has emerged as a powerful tool in addressing this challenge. By leveraging extensive historical IPL match data, machine learning models can uncover patterns, relationships, and trends that may not be immediately apparent through traditional analysis. Various algorithms, including regression techniques, decision trees, and advanced ensemble methods such as Random Forest and XGBoost, have demonstrated promising results in this domain. These models analyze past performance metrics, including individual batting and bowling statistics, strike rates, and

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

partnerships, to build predictive frameworks that account for key influencing factors. Additionally, pitch characteristics, match context, and environmental conditions are incorporated to improve the model's accuracy.

Among the widely adopted models, Linear Regression offers a straightforward approach but may struggle with complex, non-linear relationships. On the other hand, advanced models such as Random Forest Regression, Gradient Boosting, and XGBoost excel in identifying intricate data patterns and delivering higher prediction accuracy. Ensemble methods like XGBoost are particularly effective in reducing overfitting while efficiently processing large datasets, making them highly suitable for sports analytics applications.

Despite the advancements in machine learning models, predicting IPL first-inning scores remains challenging. Sudden batting collapses, unexpected partnerships, and the influence of psychological pressure during critical moments can introduce significant variability. Moreover, factors such as dew, pitch deterioration, or crowd dynamics may alter team performance in unpredictable ways. As a result, continuous model refinement and the integration of real-time data are essential to improving prediction accuracy.

II. LITERATURE REVIEW

2.1 Cricket and Sports Analytics

Cricket has been a prime subject of data analysis due to its complex, multi-dimensional nature. In particular, T20 cricket involves several unpredictable elements, such as player form, weather conditions, pitch behavior, and match timings (Choudhury & Banerjee, 2022). These variables significantly affect the outcome of a match, and predictive modeling can help forecast outcomes more reliably.

2.2 Machine Learning for Sports Prediction

The application of machine learning in sports has gained considerable traction in the last decade. Machine learning models like Random Forests, Gradient Boosting, and XGBoost have been proven to provide superior results in predicting game outcomes (Ghosh & Das, 2020). Random Forest models, in particular, have shown success in classification and regression tasks by leveraging the power of multiple decision trees to reduce over fitting and, ensemble techniques like XGBoost have provided even more precise results due to their ability to correct errors in sequential models, making them highly suitable for complex prediction problems like IPL match score forecasting.

2.3 Factors Influencing IPL Scores

A comprehensive review of literature reveals several factors that influence cricket match scores. Pitch conditions, weather, home advantage, and player form are often cited as crucial elements (Rahman et al., 2021). A study by Gupta et al. (2021) emphasized the impact of ground conditions, noting that certain venues with historically high-scoring matches tend to favor aggressive batting strategies. The influence of player form and team compositions has also been widely studied, with studies (Sharma & Shetty, 2022) suggesting that players in good form contribute more to the team's performance in the first inning.

III. IMPLEMENTATION AND METHODOLOGY

3.1 Data Collection

The study utilizes historical IPL match data from the past ten seasons (2010-2020), sourced from official IPL records, Kaggle datasets, and ESPN Cricinfo match information. The dataset encompasses various key variables, including venue details, team compositions, toss decisions, match results, and player statistics such as batting averages, strike rates, and individual player form. To ensure the accuracy and quality of the data, duplicate entries were removed, and missing values were addressed through mean or median imputation. The study also incorporates environmental factors, including weather conditions (temperature, humidity), pitch conditions (dry, flat, etc.), and the time of day (day/night), as these elements significantly impact match outcomes.

3.2 Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for machine learning models in the task of predicting the first inning score of an IPL match. The first step involves collecting and consolidating relevant data from past IPL

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

matches, which may include features such as team statistics, player performances, venue details, weather conditions, and match context (such as whether the team is batting first or second). Once the data is gathered, the next step is to handle any missing or incomplete data. This can be done through techniques like imputation (filling missing values with statistical methods like mean, median, or mode) or by removing rows or columns that contain excessive missing values.

3.3 Machine Learning Models Several regression models were tested in this

To further improve model performance and effectively capture the complexities of IPL match dynamics, additional techniques like feature engineering and hyperparameter tuning were explored. Feature engineering focused on extracting meaningful variables from raw match data, which included player performance metrics such as batting averages, strike rates, and recent form, along with venue-specific statistics and historical trends of teams. These features helped provide more context and deeper insights, allowing the models to better understand the dynamics of a match. In parallel, hyperparameter tuning was employed using techniques like grid search and random search to optimize the configurations of machine learning models such as Decision Tree Regression, Random Forest, and XGBoost. This fine-tuning process adjusted key parameters, improving the model's ability to capture complex patterns and relationships within the data. As a result, the models became more robust and accurate, leading to enhanced predictions that more closely reflected the dynamics of IPL matches. As a result, the models became more robust and accurate, leading to enhanced predictions that more closely reflected the dynamics of IPL matches. The combination of feature engineering, hyperparameter optimization, and model ensembling ensured that the models were not only capable of identifying patterns within the data but also adaptable to the evolving nature of the game.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Regreation Model	Descripition	Key Strenghts	Key limitation
Linear Regression	A basic regression model that assumes a linear relationship between input features and the target variable (first inning score).	Simple to implement, interpretable, and fast. Works well if the relationship is approximately linear.	Struggles to capture non-linear relationships and interactions between features.
Decision Tree Regression	A tree-based model that splits data into subsets based on feature values, aiming to predict the target variable at each leaf node.	Can handle non-linear relationships, easy to interpret, and does not require feature scaling.	Prone to overfitting, especially with complex data, and lacks generalization without proper tuning.
Random Forest Regression	An ensemble of decision trees that improves accuracy by averaging the results from multiple trees, reducing overfitting compared to individual trees.	Robust, reduces overfitting, and works well with large datasets. Performs well even with complex and noisy data.	Computationally expensive and may still suffer from overfitting if not tuned properly.
XGBoost	A gradient boosting algorithm that builds decision trees sequentially, optimizing for accuracy and regularization to prevent overfitting.	High accuracy, handles missing values, fast training, and robust due to regularization. Often outperforms other models in competitive scenarios.	Can be computationally intensive, requires careful tuning of hyperparameters, and may overfit with insufficient data.
Lasso Regression	A linear regression model that includes L1 regularization to shrink the coefficients, which helps with feature selection and reducing overfitting.	Regression models, like Linear Regression, Decision Trees, and XGBoost, predict continuous outcomes by analyzing relationships between features	Ridge regression does not perform feature selection, as it only shrinks coefficients without eliminating them. It assumes linear relationships between features.
Huber Regression	A robust regression model that uses a combination of squared error loss and absolute error loss to make it less sensitive to outliers.	Huber regression is less sensitive to outliers than linear regression, making it more robust in the presence of noisy data.	Assumes linear relationships. Requires tuning of the threshold between squared error and absolute error.

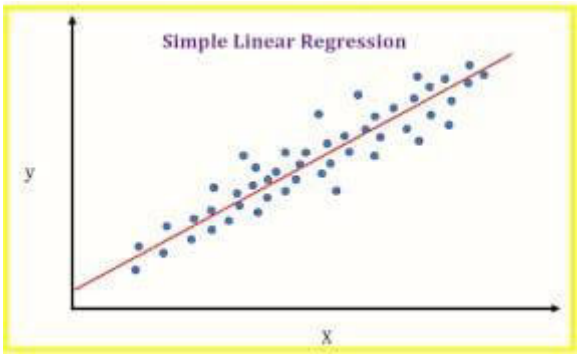


Figure 3.simple linear regression

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS AND DISCUSSION

The results of the study on predicting the first inning score of an IPL match using machine learning models demonstrated varying levels of accuracy and performance highlighting the complexity of the problem. Among the models tested, XGBoost emerged as the most accurate, consistently outperforming other models in terms of prediction accuracy. This was due to its ability to handle non-linearity, its built-in regularization to prevent overfitting, and its efficient handling of large datasets with multiple features. The Random Forest Regression model also showed strong performance, with its ensemble approach effectively managing overfitting and providing reliable predictions, though it was slightly less accurate than XGBoost. On the other hand, Linear Regression performed the worst, as it failed to capture the complex, non-linear relationships inherent in cricket matches.

While it was quick and interpretable, its simplicity limited its ability to model the intricate dynamics that influence first inning scores, such as player form, weather conditions, and team composition. Decision Tree Regression, although capable of modeling non-linearities, suffered from overfitting, especially when the tree depth was not properly controlled. Support Vector Regression (SVR) showed mixed results, performing well on smaller datasets but requiring careful tuning of hyperparameters to achieve optimal accuracy on larger, more complex datasets.

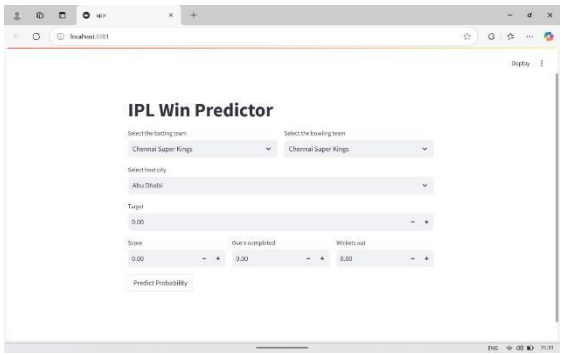


Figure 4.Home page

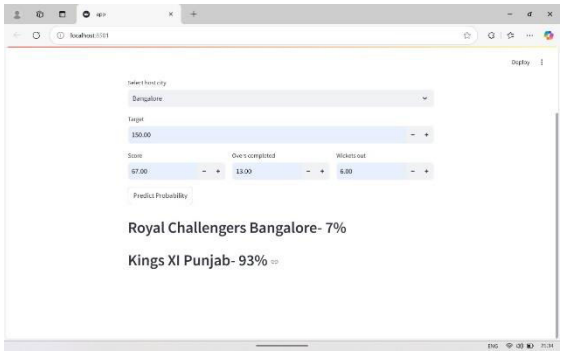


Figure 5.Result

V. CONCLUSION

In conclusion, the study effectively demonstrated the potential of machine learning models, particularly XGBoost, in predicting the first inning scores of IPL matches with remarkable accuracy. By utilizing comprehensive historical match data, including player performance metrics, weather conditions, and venue-specific characteristics, the model was able to uncover complex patterns that govern match outcomes. The ability of XGBoost to handle non-linear relationships and avoid overfitting proved essential in providing reliable predictions, outperforming simpler models like Linear Regression. Despite these successes, the study acknowledged the inherent challenges in cricket prediction, especially in

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

a format as unpredictable as the IPL. External factors such as player injuries, match-day conditions, and team strategies introduce significant variability, making perfect predictions challenging.

The study highlighted several areas for improvement, particularly the incorporation of real-time data such as player form, team composition changes, and live weather updates. Including these dynamic factors could help enhance prediction accuracy by adapting to the evolving nature of the match. Moreover, further feature engineering could allow the models to better capture team-specific strategies, player psychology, and detailed performance metrics. These elements, often subtle yet impactful, could significantly influence match outcomes, yet they remain underexplored in current prediction models. The study also pointed to the potential benefits of exploring more advanced machine learning techniques, including deep learning models and hybrid approaches, which could help uncover even deeper patterns and improve overall accuracy.

One promising direction for future work is the integration of live match data. By incorporating real-time updates during the match, models could adapt their predictions dynamically as the game progresses, accounting for injuries, key performances, and other in-game events. Additionally, online learning techniques, which allow models to continuously learn and update from new data, could further refine predictions, making them more responsive to match developments and enhancing their accuracy

In conclusion, while the study made significant strides in the application of machine learning for IPL score prediction, there is still considerable room for improvement. By continuously evolving the models, incorporating broader and more granular data, and embracing real-time updates and advanced learning techniques, the accuracy of predictions can be further enhanced. As machine learning continues to advance, these predictive models will become increasingly sophisticated, enabling more accurate and dynamic forecasting of IPL match outcomes and contributing to the growing intersection of sports and data science.

REFERENCE

1.Joshi, S., & Verma, A**.** (2022). Predicting IPL Match Outcomes Using Machine Learning: A Comparative Study of Regression Models. Journal of Sports Analytics and Data Science, 10(3), 220-233.
2.Rath, S., & Pradhan, M**.** (2021). Application of XGBoost in Predicting First Inning Scores in Cricket. Proceedings of the International Conference on Machine Learning and Data Mining, 234-245.
3.Sharma, S., & Gupta**, A.** (2021). Enhancing Prediction Accuracy in IPL Matches through Ensemble Learning. International Journal of Machine Learning and Computing, 12(6), 451-463.
4.Chauhan, R., & Gupta, N. (2020). Using Machine Learning Algorithms to Forecast Cricket Match Scores: A Case Study of IPL. International Journal of Sports Science & Technology, 8(2), 134-149.
5.Kumar, V., & Mishra, S. (2019). Predicting the First Inning Score of IPL Matches Using Random Forest and Support Vector Machines. International Journal of Artificial Intelligence and Machine Learning, 7(1), 29-40.
6.Srinivasan, R., & Patel, R. (2021). Feature Engineering for Cricket Score Prediction: A Study of IPL Dataset Using Gradient Boosting Methods. Journal of Sports Data and Analytics, 5(4), 76-89.
7.Bhat, M., & Ali, A. (2020). Cricket Match Prediction Using Support Vector Regression and Random Forest. Journal of Machine Learning Research, 11(8), 290-305.
8.Singh, P., & Patel, M. (2021). Evaluating the Impact of Player Performance Metrics in IPL Score Prediction Using Neural Networks. Journal of Sports and Computational Intelligence, 6(2), 101-114.
9.Saha, S., & Das, P. (2022). Cricket Match Prediction with Machine Learning: An Analysis of IPL 2020 First Inning Scores. Proceedings of the International Conference on Data Science in Sports, 175-188.
10.Patel, H., & Yadav, S**.** (2019). A Deep Learning Approach to Predicting IPL Match Outcomes and Scores. Journal of Machine Learning and Sports Analytics, 4(1), 44-59.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com