Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

Cricket Match Prediction Model based on 2023 World Cup Results

Dr Ajay Kumar[1*], Shashvat Priyam Khare[2], Richa Jadaun[3], Dr. T. Onima Reddy[4], Dr. Shailesh Kumar Singh[5], Dr Anurodh Sisodia[6]

[1]Sports Officer, Head of Physical Department, Govt. College Chinnor, Gwalior, Madhya Pradesh, India.
[2]PhD Scholar, Banaras Hindu University, Varanasi, Uttar Pradesh, India.
[3]PhD Scholar, DAVV Indore Madhya Pradesh, India
[4]Professor, Department of Physical Education, Banaras Hindu University, Varanasi, Uttar Pradesh, India.
[5]Assistant Professor, Lakshmibai National Institute of Physical Education Guwahati, Assam, India
[6] Professor, Director UGC-HRDC, LNIPE, Gwalior, Madhya Pradesh, India.

***Corresponding Author:** Dr Ajay Kumar
*Sports Officer, Head of Physical Department, Govt. College Chinnor, Gwalior, Madhya Pradesh, India

(Received: 22 November 2023 Revised: 30 November 2023 Accepted: 15 December 2023)

KEYWORDS

Cricket, 2023 World Cup, Prediction and Logistic Regression

ABSTRACT

The objective of this research was to establish a statistical model for forecasting the outcomes of One Day International matches in the ICC Cricket World Cup based solely on data from the match data. Such a predictive model could aid team strategies during intermissions. The study analysed data from 47 World Cup matches from 2023 tournament, excluding those decided by the Duckworth-Lewis method. The primary outcome measured was match results—win or loss. The analysis resulted in the creation of a logistic regression model to predict match outcomes based on weighted predictor variables. Key performance indicators were selected as predictors, including Toss, Opening Partnership Score, Run Scored in powerplay, Wicket lost in powerplay, Total No. of 4's in an inning and Total No. of 6's in an inning and Wickets Lost in an Inning. The methodology employed binary logistic regression for predicting outcomes. The research revealed that the logistic regression model developed was significant; however, of all the predictors considered, Total No. of 4's in an inning and Wickets Lost in an Inning was included in the model. The model was able to correctly predict 82.7% of the match results, which suggests that further refinement, including additional predictors, could enhance the accuracy and reliability of the predictions for future World Cup match outcomes.

INTRODUCTION

The use of data analysis in sports has become increasingly prevalent in recent years (Marcos et al., 2021). Teams, coaches, and organizations are now using statistical models and algorithms to analyse player performance, make strategic decisions, and gain a competitive edge. This trend has been fuelled by advancements in technology, which have made it easier to collect and analyse large volumes of data.(Mataruna-Dos-Santos et al., 2020) One area where data analysis has become particularly valuable is in the realm of talent identification and athlete development. Another area where data analysis has proved useful is in match prediction. By analysing historical data on team performance, player stats, and various external factors such as weather conditions and injuries, statisticians can create models that predict the outcome of sports matches with a reasonable degree of accuracy.

Statistical analysis for forecasting in sports has evolved rapidly, meeting demands from various stakeholders including gambling, coaches, and the media.(Du, 2018)(S et al., 2022) This evolution has allowed for a more precise and informed approach to predictions and decision-making in sports, ultimately contributing to the enhancement of the overall sporting experience.(S et al., 2022)

In addition to the use of statistical models and algorithms, data analysis in sports also involves the application of data mining techniques (Liu et al., 2023). These techniques involve extracting patterns and insights from large datasets to uncover hidden relationships and trends. By employing data mining techniques, coaches and athletes can analyse a vast amount of sports performance data to support decision-making more effectively. This can include identifying patterns in player performance, analysing opponent strategies, and detecting areas for improvement.(Machine Learning and Data Mining for Sports Analytics, 2020)

Sports analytics have become increasingly popular and intriguing, attracting the attention of both researchers and professionals (Egidi & Ntzoufras, 2020). This

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

interest has led to the development and application of various statistical models, including logistic regression, to predict outcomes in team sports. Logistic regression, a commonly used modelling approach in sports prediction, is particularly effective when the outcome of interest is the win-draw-loss scenario. By utilizing logistic regression, analysts are able to analyse the correlates of sport participation, and accurately predict the outcome of a football match or any other team sport.(South & Egros, 2020)(Tsokos et al., 2018) This powerful modelling technique estimates the probabilities using a logistic function, allowing analysts to make informed predictions based on the influencing data points. It has also been noted that logistic regression models have been successfully used in the prediction methodology for events such as the NCAA men's basketball tournament, demonstrating the value of quality data over complex modelling techniques. The application of logistic regression in sports analytics continues to expand, contributing to the growing body of research and professional interest in this field.(Lopez & Matthews, 2015)(Lopez & Matthews, 2014) With its versatility and ability to estimate probabilities using a logistic function, logistic regression continues to be at the forefront of sports prediction, enhancing the accuracy of forecasts and contributing to the expanding body of research in this domain.(Verma, 2016)(Willoughby, 2002)

Cricket is a sport that has evolved over the years, and with the rise of one-day international matches, the game has become even more competitive. The prediction of scores and winning chances in cricket has become a topic of interest for researchers.

This integration of advanced techniques in cricket analysis has the potential to greatly improve decision-making processes for teams, coaches, and cricket enthusiasts alike. By leveraging machine learning and data analytics, researchers can now predict the winner of a cricket match with greater accuracy. This improved accuracy in prediction allows teams to strategize more effectively, make informed team selection decisions, and adapt their gameplay based on the predicted outcome of a match.

Tejinder Singh et al proposed a model that uses two methods to predict the score and winning chances in cricket. The first method predicts the score of the first innings based on the current run rate, venue of the match, number of wickets fallen, and the batting team using Linear Regression Classifier.(Singh et al., 2015)

In addition to this, various statistical calculations and machine learning techniques have been applied to understand the potential predictors of game results.(Kamble, 2021)(Singh et al., 2020) The advantage of playing at the home ground, the toss result, the decision of batting or fielding first, and the game format have been identified as crucial factors in predicting match outcomes. Furthermore, researchers like Constantinou et al and Kapadia et al have developed probabilistic models and utilized machine learning techniques such as logistic regression to predict match outcomes. These models and techniques take into account the binary nature of the outcome (win or loss) and analyze the relationship between various independent variables and the dependent variable (result of the match). The purpose of the study was to develop a model to predict the outcome of ICC Cricket World cup matches on the basis of match data.

The methodology of the "Predicting the outcome of ICC cricket world cup matches" study included several distinct steps designed to develop a predictive model for match outcomes. Here's a summary of the methodology based on the provided excerpts:

MATERIAL AND METHODS

The study collected data from the 2023 ICC Cricket World Cup tournament. The researcher recorded data from 47 World cup matches, although they excluded 1 match from the analysis due to match resolved by the Duckworth-Lewis Method. The dependent variable for the study was the Match Outcome (Win/Loss). For independent variables, the researchers selected various cricketing metrics from the match, including Toss, Opening Partnership Score, Run Scored in powerplay, Wicket lost in powerplay, Total No. of 4's in an inning Total No. of 6's in an inning and Wickets Lost in an Inning which are explained as follows.

Variable	Explanation
Toss	The toss refers to the pre-match event where the captains of the two cricket teams flip a coin, and the winning captain gets the opportunity to choose whether to bat or bowl first. This variable is crucial as it can impact the team's strategy, particularly in terms of adapting to the playing conditions, setting a target, or chasing a total.
Opening Partnership Score	The opening partnership score is the cumulative score of the first two batsmen of a team who open the innings. This metric provides insights into the team's initial batting performance, setting the tone for the rest of the innings. A strong opening partnership is often associated with a solid foundation, while a low score may indicate early struggles.
Run Scored in Powerplay	The powerplay is the initial phase of an innings where fielding restrictions are in place, typically the first Ten overs in limited-overs formats. Run scored in the powerplay indicates how well a

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

	team capitalizes on the fielding restrictions. A high score in the powerplay suggests aggressive batting and a potentially advantageous position.
Wickets Lost in Powerplay	This variable represents the number of wickets a team loses during the powerplay overs. Losing wickets in the powerplay can hinder a team's ability to build a substantial total, as it often disrupts the flow of the innings. Researching this metric helps assess a team's stability and resilience at the beginning of an innings.
Total No. of 4's in an Inning	This metric signifies the total number of times a batsman hits the ball along the ground and scores four runs. A high number of fours suggests good shot placement and the ability to find gaps in the field. It reflects the team's ability to accumulate runs consistently without taking too many risks.
Total No. of 6's in an Inning	Similar to fours, this variable represents the total number of times a batsman clears the boundary and scores six runs. The number of sixes in an inning is an indicator of aggressive and boundary-clearing batting. It often contributes significantly to the team's overall run-scoring rate.
Wickets Lost in an Inning	This is the total number of wickets a team loses in the entire innings. Monitoring wickets lost provides insights into a team's batting performance and its ability to sustain partnerships. Fewer wickets lost generally indicates a more stable and controlled innings, while losing wickets frequently may suggest struggles or collapses.

Statistical Analysis:

Binary Logistic Regression was the primary statistical technique used to predict the outcome of a match (Win/Loss). This method is suitable for binary (two-outcome) dependent variables and can handle both categorical and continuous independent variables. The analysis culminated in the development of a logistic regression model. The significance of the predictor variables was determined, and they were numerically weighted to predict the match outcome.

RESULTS

The Hosmer and Lemeshow test was the first output of this logistic regression analysis, as shown in Table No. 1.

Step	Chi-square	df	Sig.
1	8.473	5	.132
2	5.216	8	.734

Table No. 1 Hosmer and Lemeshaw Test

The Hosmer and Lemeshaw test are used to examine if the generated logistic model is efficient in predicting the occurrence of the dependent variables. This test is used to assess the logistic model's goodness of fit. Hosmer and Lemeshow suggested using chi-square statistics to assess the model's efficiency. The model is judged excellent if the chi-square is negligible. (Kumar, 2023)

The p value linked with the chi-square is 0.132 and 0.734 for the first and second model respectively, which is larger than.05. The overall model is statistically efficient, ($\chi 2$ (8) = 5.126, p >.05) since the p-value associated with chi-square in Table1 is 0.940 in third model, which is greater than 0.05, it is insignificant and can be interpreted that the model is efficient.

Observed			Predicted		
			Result		Percentage Correct
			Winning	Loosing	
Step 1	Result	Winning	35	12	74.5
		Loosing	6	41	87.2
	Overall Percentage				80.9
Step 2	Result	Winning	38	9	80.9
		Loosing	8	39	83.0
	Overall Percentage				81.9
a. The cut value is .500					

Table No. 2 Classification Table.

The classification table compares predicted group membership based on the logistic model to the actual known group membership or percentage accuracy in classification (PAC). The Table No 2 which shows the observed and the predicted values of the dependent variable in the second model. Sensitivity (i.e., true

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

positives), which is the percentage of cases that had the observed characteristic which were correctly predicted by the model. The sensitivity of this model was 80.9%. Specificity (i.e., true negatives), which is the percentage of cases that did not have the observed characteristic and were also correctly predicted as not having the observed characteristic The specificity of this model was 83.0%.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	79.382[a]	.418	.558
2	67.403[a]	.488	.651
a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.			

Table No. 3 Model Summary

The explained variation is calculated using the Cox & Snell R Square and the Nagelkerke R Square techniques, as shown in the table above. Pseudo R^2 values are the name given to these numbers (and will have lower values than in multiple regression). According to the table no. 3 above, the second model in table no. 3 can explain between 48.8% and 65.1% percent of the variation on match result based on selected variables.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Wickets Lost in an Inning	.833	.178	22.028	1	.000	2.300
	Constant	-6.874	1.597	18.540	1	.000	.001
Step 2[b]	Total No. of 4's in an inning	-.147	.048	9.358	1	.002	.863
	Wickets Lost in an Inning	.843	.174	23.491	1	.000	2.324
	Constant	-3.445	1.639	4.418	1	.036	.032
a. Variable(s) entered on step 1: Wickets Lost in an Inning.							
b. Variable(s) entered on step 2: Fours.							

Table No. 4 Variables in the Equation

The most significant table displays the value of regression coefficients B (odd ratio), which are required to establish the logistic regression equation for predicting dependent variables from independent variables. The Table No 4 which shows the value of regression coefficients B, Wald statistic, and odd ratio Exp (B) for each variable in all three models. The B coefficients are used to develop the logistic regression equation for predicting the dependent variable from the independent variables. These coefficients are in log odds units. Thus the logistic regression equation in the second is:-

$\mathbf{Log}\frac{p}{1-p}$ = -3.445 +0.843 **(**Wickets Lost in an Inning**)**– 0.147 **(**Total No. of 4's in an inning**)**

Where, *p* is the probability of losing a match (reference variable). The dependent variable in the logistic regression is known as logit(p) which is equal to:

$\mathbf{Log}\frac{p}{1-p}$

The logistic regression equation estimates provided above describe the relationship between the independent and dependent variables, where the dependent variable is on a logit scale. These estimates reveal how much of an increase (or reduction, depending on the sign of the coefficient) in the estimated log chances of " losing a match"=1 would be anticipated by a one-unit rise (or decrease) in the predictor, assuming all other predictors remain constant. (Verma, 2016)

In order to make the regression coefficients B more understandable, they are changed into odd ratios equal to Exp (B). Table no. 4's leftmost column has these odds ratios. Only three variables, out of six, can substantially predict match on the basis of selected variables, as shown in Table 4. The Wickets Lost in an Inning in this model has a higher odd ratio of 2.324, making it the most relevant predictor in match results. With Exp (B) = 2.324, the Wald statistics for Wickets Lost in an Inning, were revealed to be significant (p=0.00). When the odds ratio is greater than one, it means that increasing the independent variable increases the probabilities of the dependent variable odds. That is, if the Wickets Lost in an Inning is increased by one unit, the likelihood of having losing a match increase by 1.324 (2.324-1.00) times or vice versa, assuming all other variables stay unchanged.

With Exp (B) = 0.863, the Wald statistics for Total No. of 4's in an inning, were revealed to be significant (p=0.02). When the odds ratio is less than one, it means that increasing the independent variable decreases the probabilities of the dependent variable odds. That is, if the Total No. of 4's in an inning is increases by one unit, the likelihood of losing decreases by 0.137 (0.863-1.00)

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

times or vice versa, assuming all other variables stay unchanged.

DISSCUSSION ON FINDINGS

Cricket, with its various formats such as One Day International matches, the test format, and the Twenty20 format, presents a challenging and dynamic environment for players. In recent years, the popularity of the game has skyrocketed, leading to an increase in the number of matches played at international levels and the evolution of the game itself.

The significance of hitting boundaries can be observed in the One-Day International version of cricket, which has further highlighted the need for aggressive batting strategies. The higher intensity and fast-paced nature of this format have placed greater importance on hitting boundaries, with teams looking to maximize their run-scoring opportunities within a limited number of overs. This research statistically investigated the importance of scoring boundaries in ODI format, examining its influence on team tactics, player effectiveness, and match results. The ability to hit boundaries in cricket is important for several reasons: 1. It allows for quick runs: Hitting boundaries helps in scoring runs quickly, especially in limited-overs formats like T20 and ODI cricket (Mathankar et al., 2022). 2 It puts pressure on the opposition: Scoring boundaries not only increases the team's run rate but also creates pressure on the opposing team. This makes it harder for the opposing bowlers to maintain control and execute their plans effectively. (Chowdhury et al., 2020) 3. It builds momentum: Hitting boundaries can create a positive momentum for the batting team. It boosts the confidence of the batsmen and demoralizes the opposition, leading to a stronger position for the batting team. 4. It demoralizes the opposing team: When boundaries are hit consistently, it can lead to a demoralizing effect on the opposition. They may feel disheartened and lose confidence, which can ultimately impact their performance on the field. 5 Hitting boundaries is a key factor in winning matches because it helps in achieving higher scores. Boundary hitting is a crucial aspect of cricket as it allows for quick runs, puts pressure on the opposition, builds momentum, demoralizes the opposing team, and ultimately contributes to higher scores. (Raizada, 2018) Therefore, the ability to hit boundaries plays a significant role in determining the outcome of a cricket match.

In the game of cricket, maintaining wickets is crucial for a team's success and ultimately winning the match. Wickets act as a measure of a team's batting strength and stability. The more wickets a team loses, the more pressure it puts on the remaining batsmen to score runs and keep the scoreboard ticking. (Allsopp & Clarke, 2004)Losing wickets puts a team in a vulnerable position as it reduces the number of batsmen available to continue scoring runs. It also allows the opposition team to gain an upper hand by targeting new batsmen who may take some time to settle in. Furthermore, losing wickets can also disrupt the flow of the batting innings and make it difficult to build partnerships. Additionally, losing wickets often leads to a decrease in the run rate as batsmen need time to adjust and rebuild the innings. (Lohawala & Rahman, 2018)By keeping wickets intact, a team can maintain stability and control over the game. These aspects have been extensively studied in the context of various cricket competitions, with scholars identifying the trade-offs between aggressive batting run rates and wicket loss. It's clear that losing wickets can have a domino effect on the overall performance of the team, making it crucial for teams to strategize and prioritize the preservation of wickets as a key component of their game plan.

India's 2023 World Cup journey was a rollercoaster, a surging wave of dominance cresting in a heartbreaking finale. An unprecedented 9-0 group stage romp, powered by Kohli's rediscovered roar and a bowling tsunami led by Bumrah and Shami, left the world spellbound. The semi-final was a symphony of runs, Kohli's record ton and Iyer's double centuries harmonizing beautifully, while Shami's 7-wicket spell silenced the doubters. But the final tide turned, the top-order faltered, and only Kohli's lone warrior century stood against a resilient Australia. The trophy slipped away, leaving behind a bittersweet taste. Yet, India's World Cup journey wasn't just about the ending. It was a story of resurgence, of rising stars, and of a bowling attack that sent shivers down spines. It was a reminder of India's immense potential, a promise etched in the sand for the next tide to carry forward.

CONCLUSION

In conclusion, the use of data analysis in sports has become increasingly prevalent in recent years, with the integration of advanced techniques such as machine learning, data mining, and statistical models. These techniques have enabled researchers to predict match outcomes with greater accuracy, identify key performance indicators, and make informed decisions based on data-driven insights. This study revealed the odds of winning a match and emphasized hitting more boundaries and losing less wickets in an inning. As technology continues to advance, we can expect to see further developments in the field of sports data analysis, ultimately leading to a more competitive and exciting sporting landscape.

REFERENCES

1. Allsopp, P., & Clarke, S R. (2004, September 24). Rating Teams and Analysing Outcomes in One-Day and Test Cricket. Journal of the Royal Statistical

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(6), 3017-3022 | ISSN:2251-6727

Society, 167(4), 657-667. https://doi.org/10.1111/j.1467-985x.2004.00505.x

2. Du, X. (2018, January 1). The Application of Big Data Technology in Competitive Sports Research. Communications in computer and information science, 466-471. https://doi.org/10.1007/978-981-13-0896-3_46
3. Egidi, L., & Ntzoufras, I. (2020, September 1). A Bayesian Quest for Finding a Unified Model for Predicting Volleyball Games. https://scite.ai/reports/10.1111/rssc.12436
4. Kamble, R R. (2021, April 11). Cricket Score Prediction Using Machine Learning. Turkish Journal of Computer and Mathematics Education, 12(1S), 23-28. https://doi.org/10.17762/turcomat.v12i1s.1546
5. Kumar, A. (2023, January 1). Non-Invasive estimation of muscle fiber type using ultra Sonography. https://doi.org/10.22271/kheljournal.2023.v10.i1b.2758
6. Liu, J., Hsu, M., Lai, C., & Wu, S. (2023, June 27). Using Video Analysis and Artificial Intelligence Techniques to Explore Association Rules and Influence Scenarios in Elite Table Tennis Matches. https://scite.ai/reports/10.21203/rs.3.rs-3078938/v1
7. Lohawala, N., & Rahman, M A. (2018, August 22). Are strategies for success different in test cricket and one-day internationals? Evidence from England-Australia rivalry1. Journal of sports analytics, 4(3), 175-191. https://doi.org/10.3233/jsa-180191
8. Lopez, M J., & Matthews, G J. (2014, November 30). Building an NCAA mens basketball predictive model and quantifying its success. arXiv (Cornell University). https://arxiv.org/abs/1412.0248
9. Lopez, M J., & Matthews, G J. (2015, January 1). Building an NCAA men's basketball predictive model and quantifying its success. Journal of Quantitative Analysis in Sports, 11(1). https://doi.org/10.1515/jqas-2014-0058
10. Machine Learning and Data Mining for Sports Analytics. (2020, January 1). Communications in computer and information science. https://doi.org/10.1007/978-3-030-64912-8
11. Marcos, S U., Rodríguez-Rodríguez, R., Alfaro-Saiz, J., Carballeira, E., & Marcos, M U. (2021, April 7). Improving on Half-Lightweight Male Judokas' High Performance by the Application of the Analytic Network Process. https://scite.ai/reports/10.3389/fpsyg.2021.621454
12. Mataruna-Dos-Santos, L J., Faccia, A., Helú, H M., & Khan, M S. (2020, May 15). Big Data Analyses and New Technology Applications in Sport Management, an Overview. https://doi.org/10.1145/3437075.3437085
13. S, S K., HV, P., & Nandini, C. (2022, September 14). A Survey on the application of Data Science And Analytics in the field of Organised Sports. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2209.07528
14. Singh, S., Aggarwal, Y., & Kundu, K. (2020, June 18). Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction using Machine Learning. International journal of computer applications, 176(32), 46-51. https://doi.org/10.5120/ijca2020920388
15. Singh, T P., Singla, V., & Bhatia, P. (2015, October 1). Score and winning prediction in cricket through data mining. https://doi.org/10.1109/icscti.2015.7489605
16. South, C., & Egros, E. (2020, February 27). Forecasting college football game outcomes using modern modeling techniques. Journal of sports analytics, 6(1), 25-33. https://doi.org/10.3233/jsa-190314
17. Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G A., & Király, F J. (2018, August 1). Modeling outcomes of soccer matches. Machine Learning, 108(1), 77-95. https://doi.org/10.1007/s10994-018-5741-1
18. Verma, J P. (2016, April 11). Logistic Regression for Developing Logit Model in Sport. , 293-318. https://doi.org/10.1002/9781119206767.ch11
19. Willoughby, K A. (2002, May 1). Winning Games in Canadian Football: A Logistic Regression Analysis. College Mathematics Journal, 33(3), 215-220. https://doi.org/10.1080/07468342.2002.11921944