¹ Stats & Stumps: Using Machine Learning to Predict T20I Matches ² with Player and Venue Data

Archith Sharma¹

¹Texas Academy of Mathematics and Science

Abstract

Cricket is gaining popularity worldwide rapidly, and at the front is the newest format of the game, Twenty20 Internationals (T20I), and big data. This project attempts to predict cricket match outcomes 7 using player-level performance metrics and machine learning models. A dataset of 1,029 T20I matches 8 was analyzed, with player-level features engineered from batting and bowling statistics such as runs, q strike rate, boundaries, wickets, economy rate, and maiden overs. These features were normalized by 10 ground-specific scoring rates to account for venue effects. Four modeling approaches were compared: a 11 simple heuristic based on total player impact, logistic regression with regularization, random forests, and 12 13 support vector machines (SVMs). All models were implemented in R using packages such as glmnet and e1071. Logistic regression achieved the highest test accuracy of 70.24%, balancing predictive performance 14 with model interpretability. The final model was used to generate win probabilities for both past and 15 unseen matches, including the 2024 T20 World Cup Final (India vs South Africa) and a March 2025 16 match between New Zealand and Pakistan. These predictions reflect tight contests and demonstrate 17 the model's ability to quantify match uncertainty. Overall, the project illustrates how venue-adjusted, 18 player-level impact metrics can enable robust and interpretable match outcome prediction and player 19 comparison in international T20 cricket. 20

21 Keywords:

3

4

5

T20I, Cricket, Machine Learning, Player Impact, Cricket Analytics, Venue Adjustment, Regression
 ²³

²⁴ GitHub Code for Data Processing, Figure Generation and Analysis, Modeling:

²⁵ https://github.com/ArchithSharma/CricketPredictions

26 Contents

27	1	Intr	oduction	3						
28	2	4 Materials and Methods								
29		2.1	Description of Data	4						
30		2.2 Derivation of Impact Factor								
31			2.2.1 Definition of Bat and Bowl Impact	5						
32			2.2.2 ANOVA and Histogram Validation of Bat/Bowl Similarity	6						
33		2.3	Determining Impact Factors to be used in Prediction	7						
34			2.3.1 Player Ratings and Impact Calculation	7						
35			2.3.2 Ground Buffs (Venue Factors)	7						
36			2.3.3 Weighting by Player Ranking	7						
37			2.3.4 Impact Per Game Adjustments to be used for prediction	8						
38			2.3.5 Calculation of Total Player Impact	8						
39		2.4	Modeling	8						
	2	Dog	ulto	0						
40	J	2 1	Pleyer and Venue vigualizations	9						
41		0.1 2.0	Model Posulta	9 10						
42		3.2 Model Results								
43	4	Dise	cussion	10						
44		4.1	Analysis of Players and Venues	10						
45			4.1.1 Player Analysis: Virat Kohli	10						
46			4.1.2 Player Analysis: Ajantha Mendis	10						
47			4.1.3 Player Analysis: Glenn Maxwell	11						
48			4.1.4 Venue Analysis: England vs. Subcontinent Venues	11						
49			4.1.5 Team Rankings Based on Recent Matches	12						
50		4.2	Analysis of Models	12						
51	5	Conclusions 13								
52	6	Acknowledgements 13								
53	33 Keterences 14									
54	$_{44}$ Appendices									
55	\mathbf{A}	A Player and Venue Visualizations, Code Used for Analysis and Figure Generation 16								
56	A.1 Player and Venue Diagrams									
57	A.2 Code									
58	в	B Glossary of Cricket Terms Used 18								

⁵⁹ 1 Introduction

⁶⁰ Cricket is a bat-and-ball sport that originated in England and has grown into one of the most popular ⁶¹ sports worldwide, particularly in countries like India, Australia, South Africa, and Pakistan. Played between ⁶² two teams of 11 players each, the game involves batting, bowling, and fielding, with the primary objective ⁶³ being to score more runs than the opposing team. Cricket is played in different formats, ranging from the ⁶⁴ traditional five-day Test matches to the fast-paced Twenty20 (T20) format.

T20 cricket, the shortest and most explosive format of the game, revolutionized the sport with its fastpaced action and entertainment value. Each team gets a maximum of 20 overs (120 balls) to bat, encouraging aggressive stroke play, quick scoring, and strategic bowling. Since its official introduction in 2003, T20 cricket has gained massive popularity, leading to global tournaments such as the ICC T20 World Cup and franchisebased leagues such as the Indian Premier League (IPL) and the Big Bash League (BBL). With its thrilling finishes, power-hitting, and emphasis on adaptability, T20 cricket has attracted a new generation of fans while maintaining the essence of the sport.

The rise of predictive analytics in cricket has closely coincided with the rapid growth of the T20 format. With T20 matches fast-paced and often decided by fine margins, teams, analysts, and betting markets have increasingly turned to data-driven approaches to gain a competitive edge. The shorter format requires quick decision making, making real-time data analysis and predictive modeling crucial to optimize team strategies, player selection, and match predictions.

In this paper, a unique approach will be taken that focuses on a new impact factor (IF) of players for both batting and bowling, in addition to consideration of match venues. These factors will then be used to train 3 types of models commonly used in sports prediction: Regularized logistic regression, random forest classifier, and support vector machines (SVMs), with high test set accuracies of 67 - 70%. (Vistro et al., 2019)

The paper is organized as follows. Section 2 (Materials and Methods) will explore the Cricinfo dataset and derive the impact factor used for batting and bowling, along with a ground-based scaling to account for different venues. Afterwards, multiple models are tested on the adjusted comprehensive impact factors. Section 3 shows the accuracy percentages of the models along with its prediction on select matches. Section 4 and 5 analyze the impact factor for different players and venues, followed by a discussion of the models and some conclusions/limitations. There are two appendices, the first containing figures generated that are used in the analysis and the code used (A), and the second containing a definition of common cricket terms used throughout the paper and the test-playing nations list (B).

⁹⁰ 2 Materials and Methods

All data pre-processing, analysis and modeling was performed in R. (R Core Team, 2024) See Appendix A.2
to locate the code used on GitHub and regenerate the figures.

93 2.1 Description of Data

^	Date 🗧	Player ÷	Country ÷	Runs 🗧	NotOut 🌻	Minutes 🗧	BallsFaced [©]	Fours ÷	Sixes 🌻	StrikeRate
1	2018-07-03	AJ Finch	AUS		FALSE					226.3158
2	2019-02-23	Hazratullah Zazai	AFG	162	TRUE					261.2903
3	2013-08-29	AJ Finch	AUS	156	FALSE					247.6190
4	2016-09-06	GJ Maxwell	AUS	145	TRUE					223.0769
5	2024-01-17	FH Allen	NZL		FALSE	94				220.9677
6	2025-02-02	Abhishek Sharma	IND		FALSE		54			250.0000
7	2023-02-01	Shubman Gill	IND	126	TRUE					200.0000
8	2017-07-09	E Lewis	WI	125	TRUE					201.6129
9	2016-01-31	SR Watson	AUS		TRUE	86				174.6479
10	2023-11-28	RD Gaikwad	IND		TRUE					215.7895
11	2012-09-21	BB McCullum	NZL		FALSE					212.0690
12	2022-09-08	V Kohli	IND	122	TRUE	90				200.0000
13	2021-04-14	Babar Azam	PAK		FALSE					206.7797
14	2024-01-17	RG Sharma	IND		TRUE	103	69			175.3623
15	2024-02-11	GJ Maxwell	AUS	120	TRUE					218.1818
16	2024-11-15	NT Tilak Varma	IND	120	TRUE	69				255.3191
17	2015-01-11	F du Plessis	RSA		FALSE					212.5000
18	2023-12-19	PD Salt	ENG	119	FALSE	80				208.7719
19	2016-01-10	Mohammad Shahzad	AFG	118	TRUE					176.1194
20	2017-12-22	RG Sharma	IND	118	FALSE					274.4186
21	2023-03-26	I Charles	WI	118	EALSE	60	46	10		256 5213

Figure 1: Batting Data

The data is taken from ESPN Cricinfo, a reputable source containing in-depth statistics on all T20Is played and the players themselves. (ESPN, 2025a) To access the data in R, the cricketdata package made by Rob Hyndman (Hyndman et al., 2025) was used to access the Statsguru data in R as a dataframe. A screenshot of the batting dataframe is shown above in Figure 1.

The data is then filtered to only include T20Is between the test-playing full ICC members, defined in Appendix B. The data is also joined by match to determine the total impact factor of teams in their matches, both bowling and batting.

¹⁰¹ 2.2 Derivation of Impact Factor

At the core of this analysis is the novel impact factor used in comparing batting and bowling impacts. To determine the formula for batting and bowling impact, correlation to the target metric between the total impact of both teams was prioritized in addition to similarity in the distribution of the impact factors each game.

The impact factor was manually tuned to a quantity that sufficiently addressed both criteria. They are defined below in Equations 1 and 2:

¹⁰⁸ 2.2.1 Definition of Bat and Bowl Impact

Batting Impact = $(R \times 0.125) + ((SR - 130) \times 0.025) + (4s \times 0.3) + (6s \times 0.5) - \text{OutPenalty}$ (1)

where: R =Runs scored

$$SR = \text{Strike rate} = \left(\frac{\text{Runs}}{\text{Balls Faced}} \times 100\right)$$

4s = Number of fours hit

6s = Number of sixes hit

OutPenalty = Fixed penalty of 0.5 if the batter is dismissed

Example:

	A batter scores $R = 40$ runs off 28 balls, hits 3 fours, 2 sixes, and is dismissed.					
110	$SR = \frac{40}{28} \times 100 \approx 142.86$, OutPenalty = 0.5					
111	Batting Impact = $(40 \times 0.125) + ((142.86 - 130) \times 0.025) + (3 \times 0.3) + (2 \times 0.5) - 0.5$					
112	$= 5 + (12.86 \times 0.025) + 0.9 + 1.0 - 2.5 = 5 + 0.3215 + 0.9 + 1.0 - 0.5 \approx 6.72$					
	Bowling Impact = $(W \times 3.25) + (M \times 3) - ((E - 7.4) \times 0.5)$					

where: W = Number of wickets taken

M = Number of maiden overs bowled

 $E = \text{Economy rate} = \left(\frac{\text{Runs Conceded}}{\text{Overs Bowled}}\right)$

7.4 = Baseline economy rate for adjustment

 $0.5=\mbox{Penalty}$ per unit increase in economy above 7.4

(2)

Example:

113

109

A bowler takes W = 2 wickets and a E = 6.85 economy rate 3 and has no maiden overs.

114

Bowling Impact = $(2 \times 3.25) + (0 * 3) - ((6.85 - 7.4) * 0.5) = 6.5 + 0.55 * 0.5 \approx 6.775$

Both batting and bowling equations consider a baseline adjustment for strike rate and economy rate based on the average score of around 149 in T20Is. (Statsguru, 2025) Both example players did very well in their respective games and had high similar impacts of around 6.75, showing the relevance of the heuristic. Afterwards, the batting and bowling impacts were grouped by their totals by game for each team. Afterwards, the differences between total impact factors are used as a heuristic to be correlated to the winner of the match (interpreted numerically as 1, 0). As can be seen in the code, the correlation value is 0.773 for 1029 T20Is, and the R^2 is 0.598, approximately 0.6 which indicates high explainability.

122 2.2.2 ANOVA and Histogram Validation of Bat/Bowl Similarity

To validate the assumption that the distributions are similar, an analysis of variance (ANOVA) is performed by modeling the batting impacts on the bowling impacts and the result is displayed in Table 1. The pvalue that is almost 0 shows that the distributions are very similar. In addition the visual validation of the histogram in Figure 2 shows how similar the total impact factors are, and that they can be used for modeling.

Term	$\mathbf{D}\mathbf{f}$	$\mathbf{Sum} \ \mathbf{Sq}$	Mean Sq	F value	$\Pr(>F)$
Total Bowling	1	19376	19376	95.95	$< 2 \times 10^{-16}$
Residuals	2068	417623	202		



Table 1: ANOVA Results

Figure 2: Impact Distributions

127

¹²⁸ 2.3 Determining Impact Factors to be used in Prediction

In this study, the performance of players in a T20 match is evaluated through a series of factors that encompass both their batting and bowling abilities. The impact of each player is adjusted for various gamespecific conditions, including venue-based adjustments and player rankings. Below, the rationale behind the decision-making process is outlined for each factor used in the calculation of player impact.

133 2.3.1 Player Ratings and Impact Calculation

Each player's performance is assessed based on their batting and bowling ratings, which are derived from historical performance data. These ratings are extracted from the player_rankings_2 dataset, which includes batting impact (BatImpactperGame) and bowling impact (BowlImpactperGame) for each player. These ratings reflect the overall contribution of a player to their team's performance in previous matches.

138 2.3.2 Ground Buffs (Venue Factors)

The performance of players is also adjusted based on the specific characteristics of the ground where the match is played. Each ground has unique attributes that can influence player performance, such as pitch conditions, weather, and altitude. The **venue_factors** dataset provides ground-specific adjustments known as "ground buffs". These buffs are applied to both batting and bowling ratings to account for the venue's impact on player performance. The following ground buffs are considered:

- Batting Buff (BattingScale2): This factor adjusts the batting impact based on ground-specific conditions such as pitch type, boundary size, and weather.
- Bowling Buff (BowlingScale2): This factor adjusts the bowling impact based on similar ground conditions that affect how bowlers perform, including pitch conditions and boundary dimensions.

¹⁴⁸ 2.3.3 Weighting by Player Ranking

The weight assigned to each player's performance is influenced by their relative ranking in the team. Batting
and bowling impact are adjusted according to player rankings:

• Batting Impact: The top 8 batsmen (top and middle order) in the playing XI receive a weighted adjustment to their batting impact. The adjustment is proportional to the player's ranking, with higher-ranked players receiving a greater weighting. The weight is calculated as a function of the player's position in the batting order, with later batsmen receiving lower weights to reflect their reduced role in the match. Bowling Impact: Similarly, the top 6 bowlers are given weighted adjustments to their bowling impact,
 based on their ranking. The adjustment is calculated in a similar manner, with higher-ranked bowlers
 receiving greater weight. The bottom bowlers are excluded from the impact calculation, as they are
 less likely to bowl in the match.

160 2.3.4 Impact Per Game Adjustments to be used for prediction

Once the initial ratings are adjusted for venue buffs and player rankings, the resulting impact values (BatImpactperGame_2 and BowlImpactperGame_2) are calculated for each player. The calculations are performed iteratively for both teams, with batting and bowling impacts adjusted according to the above factors:

• For batsmen, the weight is scaled according to their position in the batting order.

- For bowlers, a similar scaling is applied, with the most prominent bowlers receiving higher weights.
- The adjusted values for each player are stored and used to update the overall team impact.

¹⁶⁸ 2.3.5 Calculation of Total Player Impact

Finally, the adjusted batting and bowling impacts are compiled into a total impact score for each player. This score is split into batting and bowling impacts, with the first 8 players in the batting order contributing to the batting total, and the next 6 players contributing to the bowling total. The impact scores for each player from both teams are then consolidated into a total impact matrix, which represents the overall performance potential of the two teams based on individual player ratings and match conditions.

In summary, the impact factors in this methodology are derived from a combination of player rankings, ground-specific adjustments, and positional weights. These factors provide a comprehensive assessment of player performance, accounting for both individual ability and contextual factors, such as venue and team composition.

178 **2.4** Modeling

To predict T20I match outcomes based on player-level performance metrics, several classification models were implemented and evaluated using a dataset where each match was represented by the total impact scores of 28 players (14 per team). As a baseline, a simple heuristic predicted the team with the higher cumulative impact score as the winner. This naive method yielded an accuracy of **64%**, providing a benchmark for more advanced models. A logistic regression model with elastic net regularization was trained using 5-fold cross-validation to tune the hyperparameters α and λ . (Tay et al., 2023) To convert predicted probabilities into binary outcomes, optimal classification thresholds were selected by averaging the best thresholds across ROC curves in 10-fold cross-validation. (Robin et al., 2011) Repeating this process over 30 iterations, the logistic regression model achieved a maximum accuracy of **70.24%**, outperforming the heuristic approach and all other models.

To enhance explainability, the coefficient weights were extracted from the logistic model. Features associated with top-order batters and key bowlers—such as the highest run scorers, highest strike rate performers, and wicket-takers—were consistently assigned large absolute coefficients, indicating their strong influence on match outcomes. Notably, impact scores from players in the top three batting positions had the greatest positive association with winning probability, while low-performing bowlers negatively influenced predicted outcomes. In addition, Batters 7-8 and Bowlers 5-6 had large coefficients, showing that the model respected lineup depth and all-rounder 'X-factor' performances.

A random forest model with 500 trees was also tried, tuning the mtry parameter via internal 5-fold CV. This model achieved a maximum accuracy of **67.8%**. Feature importance rankings from the random forest agreed with the logistic regression: early-order batters and strike bowlers had the greatest influence, though the non-linearity of the random forest also surfaced some interactions among lower-order players.

Lastly, a linear Support Vector Machine (SVM) was evaluated using a grid of cost values $C = 2^{-5}, \ldots, 2^{5}$ and 5-fold cross-validation. (Meyer et al., 2023) The best SVM model reached an accuracy of **69.7%**, comparable to the logistic regression model.

Among all approaches, logistic regression offered the best performance while also allowing for direct interpretability through model coefficients. This suggests that linearly weighted combinations of individual player impacts can be powerful predictors of T20I match outcomes.

206 **3** Results

²⁰⁷ 3.1 Player and Venue visualizations

Now that the impact factors are clearly defined, they can be used to construct rankings of players by batting, bowling, or all-around play and venue factors. Similarly, teams can be ranked by their recent impact factors. The full visualizations of the Top 15 in Career Impact and Impact per Game for batsmen are found in Appendix A.1 containing Figures 4, 5, 6, and 7. All-rounders are defined as individuals in the top two-thirds of all players in batting and in the top 55th percentile of all players in bowling. The average venue impact was also min-max scaled to be used as a multiplier in the predictive modeling phase.

214 3.2 Model Results

The modeling results are organized below in Table 2.

Model	Mean Accuracy (%)	95% CI (%)
Heuristic (Total Impact Comparison)	64.00	[64.00, 64.00]
Logistic Regression (Elastic Net)	68.32	[66.42, 70.24]
Random Forest	66.08	[64.91, 67.80]
Support Vector Machine (Linear Kernel)	67.42	[65.53, 69.70]

Table 2: Model performance based on mean accuracy and 95% confidence intervals over 30 iterations

The following data in Table 3 is the model's output for select matches.

Match	Win Probability (%)	Actual Winner	Margin
IND vs RSA (T20 WC Final 2024)	IND: 50.46	IND	7 runs
NZ vs PAK (Unseen Match, March 26)	NZ: 55.45	NZ	8 wickets
BAN vs. WI Dec 19 2024	BAN: 50.62	BAN	80 runs
IND vs. PAK Oct 23 2022	IND: 64.5	IND	4 wickets

Table 3: Predicted Win Probabilities and Actual Outcomes for Select T20I Matches

216

215

217 4 Discussion

4.1 Analysis of Players and Venues

219 4.1.1 Player Analysis: Virat Kohli

Virat Kohli is considered by many to be the world's best batsman, scoring the third most runs of all time and having the second most total hundreds. (ESPN, 2025b) However, many have shown doubt about his performances in the shortest format of the game, especially due to concerns about his strike rate. (TOI Sports Desk, 2024)

The impact factor method considers him one of the best, if not the best T20I batsmen despite his strike rate. He is ranked the best batsman by total career batting impact, while he finds the third spot in batting impact per game. His low strike rate is not enough of a hindrance due to the constant amount of runs he amasses. The impact factor is able to find value in his ability to facilitate the innings.

228 4.1.2 Player Analysis: Ajantha Mendis

Ajantha Mendis (labeled BAW Mendis) was a spinner for Sri Lanka between 2008 and 2014. He is the only player to take 6 wickets in a T20I twice, and also won Emerging Player of the Year in 2008. Labelled a mystery spinner, when batsmen figured him out, he began to get hit all around the park. His biggest criticism was the lack of consistency, and he disappeared from Sri Lanka's XI in favor of Rangana Herath
and other spinners. (Cricbuzz, 2025)

The impact factor method has a shortcoming when considering a spinner like Mendis. Recent matches do not get weighted more since the model does not have a time factor, and high performances in a small dataset could push his impact as high as the third rank. To address it in model predictions, adding a random effect to the player impact based on the number of games played could make a difference.

The theoretical framework would be that the more matches someone plays, the less deviation their impact score would have. Comparing Mendis and Kohli, where the former played one thirds the games the latter did, Mendis's variance parameter estimate would have 3 times the variance of Kohli's parameter and thus represent his inconsistency. Player dropping ensures the model stays representative regardless of consistency because inconsistent players tend to be dropped.

243 4.1.3 Player Analysis: Glenn Maxwell

Glenn Maxwell is considered the quintessential T20 player for the combination of his golden arm and power 244 hitting. He can steal wickets from loose shots, and send loose deliveries 100 meters away. Most of his impact. 245 however comes from batting. Unorthodox shots make him hard to stop when he gets going (ESPN, 2023), 246 and his high batting impact scores and semi-decent bowling performances push him up the all-rounder list. 247 This is in contrast to someone like Wanindu Hasaranga higher on the list, who is a handy batsman lower 248 in the order, but does most of his damage with the ball in his hands. Hardik Pandya is different from both. 249 being middle-of-the-pack in bowling and batting. The impact factor method gives them high, yet roughly 250 equal impacts which correspond to a correct aggregation. 251

²⁵² 4.1.4 Venue Analysis: England vs. Subcontinent Venues

The venue analysis by impact factor is mostly accurate to real-life ground impacts. 12 out of the top 20 best batting venues are located in either India or Pakistan, reflecting the popular opinion that many pitches in the subcontinent are batting paradises.

However, the grounds in England tend to be more balanced and favor the bowlers more. In total impact relative to batting, Lord's places 79th out of 105 venues. The famous slope at Lord's makes bowling relatively favorable as well. Southampton, The Oval and Headingley are all in the 60-70 range, reflecting England's moderate preference for bowling, showing that the scaling based on impact factor is an accurate measure of ground impacts.

²⁶¹ 4.1.5 Team Rankings Based on Recent Matches

Based on the impact factor, rankings based on the total accumulated impact can be constructed for the last 6 months. The data shows that India is far and away the most in-form team, followed by Australia and Afghanistan. South Africa and Pakistan are lower on the list, especially due to recent series losses to India and New Zealand, respectively. Ireland is also placed higher since they have not played the top teams in the last 6 months outside of Zimbabwe. However, overall, the list is an accurate representation of the current T201 landscape.



Figure 3: Rankings based on performance for last 6 months

²⁶⁸ 4.2 Analysis of Models

Four different approaches were tested for predicting T20I match outcomes: a simple heuristic model, logistic regression with regularization, random forests, and support vector machines (SVM). The heuristic model, which selects the team with the highest total player impact, achieved an accuracy of approximately 64%, serving as a useful baseline. Logistic regression outperformed the others, reaching a maximum test accuracy of 70.24%, with hyperparameter tuning and threshold optimization via cross-validation. Random forests and SVMs followed closely with peak accuracies of 67.8% and 69.7%, respectively.

Logistic regression was selected for final prediction due to its strong performance and interpretability. The model provides probabilistic outputs, allowing uncertainty to be quantified. For example, in the 2024 T20 World Cup Final between India and South Africa, the model assigned India a win probability of 50.46%, reflecting an essentially even matchup which can be seen in the small result margin of 7 runs. Similarly, for the March 26 match between New Zealand and Pakistan, the model predicted a New Zealand win probability
of 55.45%, indicating a slight edge.

These results suggest that logistic regression provides both competitive performance and actionable probabilities for match outcome predictions. Moreover, its explainability makes it suitable for understanding how individual player features contribute to team-level predictions.

284 5 Conclusions

This paper has been maximizing the value of a simple model in T20I prediction. More complex models can 285 consider strategy with the toss before matches, as many pitches get better to bat on because of the dew in 286 the night, meaning there is a significant advantage in batting second. (Krishnaswamy, 2022) In addition, 287 recent player form can change the parameter passed into the logistic regression. For example, Virat Kohli 288 was in a bad run of form during the T20 World Cup, only crossing 10 runs twice in the 7 matches before 289 the final. Specific bowler-batsman combinations are not considered either, such as the Indian top order's 290 issues with left arm pace bowlers. (Iyer, 2023) Overall, however, this paper's methods have had an accuracy 291 of 70% in predictions and an explainable impact factor that can be used to quantify team performance. This 292 framework can streamline cricket analysis and objectively pit players and teams against each other through 293 the power of data. 294

295 6 Acknowledgements

²⁹⁶ I would like to thank the Wharton Sports Analytics and Business Initiative for the opportunity to share my
²⁹⁷ work, and especially Professor Adi Wyner for organizing the Wharton High School Data Science Competition,
²⁹⁸ which sparked my interest in sports analytics. I am also grateful to the broader cricket analytics community,
²⁹⁹ particularly ESPN Cricinfo and the open source contributors to the R Project. Special thanks to Dr. Rob
³⁰⁰ J Hyndman, the creator of the cricketdata package, for enabling data-driven sports research. I would also
³⁰¹ like to thank my grandfather, Giridhara Govinda Rao for introducing me to the game of cricket and inspiring
³⁰² my love for it.

303 References

- Cricbuzz. (2025). Ajantha mendis. Cricbuzz. Retrieved April 4, 2025, from https://www.cricbuzz.com/
 profiles/1390/ajantha-mendis#:~:text=Ajantha%20Mendis%20won%20the%20Emerging, has%
 20accomplished%20that%20feat%20twice.
- ESPN. (2023, November). Afghanistan vs australia 07 11 2023. *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/series/icc-cricket-world-cup-2023-24-1367856/afghanistan-vsaustralia-39th-match-1384430/live-cricket-score
- ³¹⁰ ESPN. (2025a). Espncricinfo. ESPNcricinfo. Retrieved April 4, 2025, from https://www.espncricinfo.com/
- ESPN. (2025b). Virat kohli records. *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo. com/cricketers/virat-kohli-253802/tests-odi-t20-records
- Hyndman, R., Gray, C., Gupta, S., Hyndman, T., Rafique, H., & Tran, J. (2025). Cricketdata: International
 cricket data [R package version 0.3.0, https://github.com/robjhyndman/cricketdata]. https://pkg.
 robjhyndman.com/cricketdata/
- Iyer, R. (2023, September). Why do indian batsmen struggle against left-arm pacers? rohit sharma and
 co.'s record against left-armers detailed. *Sportingnews.com*. Retrieved April 4, 2025, from https:
 //www.sportingnews.com/in/cricket/news/why-do-indian-batsmen-struggle-against-left-arm pacers-rohit-sharma-and-cos-record-against-left-armers-detailed/aomdbaj3kqbxdyimaxsefxrm
- Krishnaswamy, K. (2022, April). The (un)dew advantage: What choice do teams batting first have? *ESPN- cricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/story/ipl-2022-pbks-vs-kkr the-undew-advantage-what-choice-do-teams-batting-first-have-1308698
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien* [R package version 1.7-14]. https://CRAN.R-project.org/package=e1071
- R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical
 Computing. Vienna, Austria. https://www.R-project.org/
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12, 77.
- ³³⁰ Statsguru. (2025). Total match results: T20 internationals. *Cricinfo*. Retrieved April 4, 2025, from https:
- //stats.espncricinfo.com/ci/engine/stats/index.html?class=3;orderby=matches;orderbyad=reverse;
 template=results;type=aggregate
- Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1), 1–31. https://doi.org/10.18637/jss.v106.i01

TOI Sports Desk. (2024, April). 'usually he isn't this slow': Former india cricketers express concerns about 335 virat kohli's strike rate. The Times of India. Retrieved April 4, 2025, from https://timesofindia. 336 indiatimes.com/sports/cricket/ipl/top-stories/usually-he-isnt-this-slow-former-india-cricketers-337 express-concerns-about-virat-kohlis-strike-rate/articleshow/109615198.cms 338 Vistro, D., Rasheed, F., & David, L. (2019). The cricket winner prediction with application of machine 339 learning and data analytics. International Journal of Scientific & Technology Research. https:// 340 www.kclas.ac.in/wp-content/uploads/2021/01/The-Cricket-Winner-Prediction-With-Application341 Of-Machine-Learning-And-Data-Analytics-.pdf 342

15

343 Appendices

³⁴⁴ A Player and Venue Visualizations, Code Used for Analysis and

³⁴⁵ Figure Generation





Figure 4: Total Career Bat/Bowl Impact Top 15



Figure 5: Bat/Bowl Impact Top 15 per game



Figure 6: All Rounder by game/career impact



Figure 7: Best Venues for Batting/Bowling

347 A.2 Code

- $_{348}$ $\,$ All code used to process data, generate figures, & tune parameters/validate model at
- $_{\rm 349}~{\rm https://github.com/ArchithSharma/CricketPredictions.}$

³⁵⁰ B Glossary of Cricket Terms Used

351 Batting Terms

352 Runs Total number of runs scored by a batsman.

Strike Rate The average number of runs scored per 100 balls faced. A higher strike rate indicates more
 aggressive scoring.

Boundary A scoring shot that results in four (4) or six (6) runs.

Fours & Sixes A four occurs when the ball crosses the boundary after touching the ground; a six is when it crosses without touching the ground.

Batting Impact Score A custom metric combining runs, strike rate, and boundaries to evaluate batting
 effectiveness.

Top Order The group of batsmen who come out to bat first and score most of the runs, typically the first
 three or four guys in.

Middle Order The group of batsmen that are typically all-rounders and power hitters to try and get quick
 runs near the end.

364 Bowling Terms

- ³⁶⁵ Wickets The number of batsmen a bowler dismisses.
- ³⁶⁶ Maiden Overs Overs in which no runs are conceded.

³⁶⁷ Economy Rate Runs conceded per over bowled. A lower economy indicates more efficient bowling.

³⁶⁸ Spinner A bowler who bowls slow but gets the ball to turn, deceiving batsmen.

³⁶⁹ **Pacer** A bowler who bowls quick to try and beat batsmen with variations and leave them helpless.

- ³⁷⁰ Bowling Attack The group of bowlers in a team and their combined effectiveness.
- Bowling Impact Score A weighted metric assessing bowling performance using wickets, maidens, and economy rate.

373 Match Context

- **T20I** Twenty20 International a short-format cricket match where each team bowls a maximum of 20 overs.
- Venue Factor A numerical measure of how batting-friendly or bowling-friendly a stadium is, based on its
 average scoring.
- ³⁷⁷ Playing XI The eleven players selected to participate in a match.

378 Strategy Terms

- 379 All-Rounder A player who contributes significantly with both bat and ball.
- Toss The coin toss where the captain who wins decides whether to bat or bowl first.
- ³⁸¹ Run rate Runs per over (6 balls).
- ³⁸² **Pitch** Surface that the bowlers bowl to the batsmen on, can heavily influence team decisions by favoring
- batting or a certain kind of bowling.

³⁸⁴ Test playing nations list

Afghanistan, Australia, Bangladesh, England, India, Ireland, New Zealand, Pakistan, South Africa, Sri
Lanka, West Indies, Zimbabwe.