Purnendu Bhowmik Shuvro

Fall-25_CSE303_S2_Analysis of Cricketer's Performance in T20I using Classification

Mini Project Full

Document Details

Submission ID
trn:oid:::29306:124412085

Submission Date
Dec 13, 2025, 11:37 PM GMT+6

Download Date
Dec 14, 2025, 12:39 AM GMT+6

File Name
Fall-25_CSE303_S2_Analysis of Cricketer's Performance in T20I using Classification.docx

File Size
112.6 KB

11 Pages

2,770 Words

15,168 Characters

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

26 Not Cited or Quoted 10%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
3% Publications
7% Submitted works (Student Papers)

Match Groups

- **26** Not Cited or Quoted 10%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
3% Publications
7% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
iptek.its.ac.id		<1%
2	Student papers	
University of Western Australia on 2024-09-06		<1%
3	Internet	
www.slideshare.net		<1%
4	Student papers	
General Sir John Kotelawala Defence University on 2025-12-07		<1%
5	Internet	
link.springer.com		<1%
6	Internet	
www.naturalspublishing.com		<1%
7	Student papers	
Liverpool John Moores University on 2025-12-10		<1%
8	Student papers	
Middlesex University on 2024-10-05		<1%
9	Student papers	
Liverpool John Moores University on 2025-05-24		<1%
10	Internet	
anzdoc.com		<1%

11 Internet
www.gnauniversity.edu.in <1%

12 Student papers
Dublin Business School on 2025-08-28 <1%

13 Student papers
Higher Education Commission Pakistan on 2025-07-11 <1%

14 Student papers
University of Huddersfield on 2025-01-10 <1%

15 Student papers
University of Westminster on 2024-11-10 <1%

16 Internet
eudoxuspress.com <1%

17 Internet
www.arxiv-vanity.com <1%

18 Internet
www.livemint.com <1%

19 Student papers
General Sir John Kotelawala Defence University on 2025-09-01 <1%

20 Student papers
Napier University on 2024-08-21 <1%

21 Publication
Rucia V. November, Haiyan Cai, Mogammad Sharhidd Taliep, Clement Nyirenda, L... <1%

22 Student papers
University of Essex on 2024-09-18 <1%

23 Publication
Reinertsen, Erik. "Dichotomizing Illness from Cardiovascular and Locomotor Activ... <1%

24 Publication
Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%

25 Student papers

University of West London on 2024-04-29 <1%

EAST WEST UNIVERSITY

Research Paper

Course Name: Statistics for Data Science

Course: CSE 303, **Section:** 02

Semester: Fall 25

Project Title: Analysis of Cricketer's Performance in T20I using Classification

3

Submitted To:

Dr. Mohammad Manzurul Islam

Assistant Professor,

Department of Computer Science and Engineering

East West University

Submitted By:

Name:	ID:
Purnendu Bhowmik Shuvro	2023-1-60-085
Md. Ahsiul Karim	2022-3-60-074
Md. Junaid Rashid	2022-3-60-086

Submission Date: 09 December, 2025

Analysis of Cricketer's Performance in T20I using Classification

Purnendu Bhowmik Shuvro, Md. Ahsiul Karim, Md. Junaid Rashid

Abstract:T20I is the shortest and fastest format of cricket, where every ball is important and players must perform with high skill. This study focuses on analyzing cricketer's performance in T20 International (T20I) cricket using statistics and Machine Learning. In this work, batting and bowling data were collected from different cricket sources and cleaned to remove errors. This study uses some important measures like - strike rate, economy rate, averages and consistency were used to understand the performance of a player. Two Machine Learning algorithms - Logistic Regression and Support Vector Machine (SVM) were trained using these features to classify a player's performance and predict the match outcomes. The results showed some cricketers' performance in batting and bowling. . The algorithms performed very well, "Logistic Regression" giving 99.40% accuracy and while "SVM" performed slightly better with 99.70% accuracy. Both algorithms can effectively recognize patterns of the cricket data. But "SVM" performed far better than "Logistic Regression". The study demonstrates that combining data analysis using Matching Learning provides a clear understanding of player performance.

Introduction: Cricket is an outdoor sport using bat and ball, which is also known as a respected and gentle sport, and holds a significant status among the sports specially in team sports. It is played in the center of the sports, which is known as a pitch. This sport is contested between two teams, each team playing in batting and bowling/fielding in the two innings. There are 3 formats played in this sport both internationally and domestically, they are: Test, ODI, T20I. T20I is the smallest format in terms of time, which the sport makes dynamic and fast-paced. This format was introduced in 2003, the match was between Hampshire and Sussex in English domestic cricket, in the international match between Australia and New Zealand on

17th February 2005 in Auckland. The world cup of this format was started in 2007 in South Africa, where India became first ever champion, beating Pakistan by 5 runs in the nail-biting final [1].

In T20I format, the more impact action or performance is required from the cricketers, because it consists of 20 overs or 120 balls, which is fewer than other formats of Test and ODI. Test is played in 4 or 5 days in the international stage and ODI is a 50 overs game in 8 to 9 hours, which is 2.5 overs more than T20. Each ball is important in the T20I, the cricketer has to perform with the developed skills.
Batter, Bowler and All Rounder are needed to perform in this format with proper and accurate skills. Batter plays more risk taking shots with an aggressive mindset, which means focus on hitting more boundaries, either four or six, but there are some batter who focus on anchor role, meaning scored more running between the wickets, one, two or three, tries to play less dot balls because few balls in this format, sometimes this batter tries to hit fours or six, if the delivery is suitable to the batter. Bowler's tactics are to deliver more accurate dot balls to build on the batter so that the batter can play the wrong shots, to deliver wicket taking ball, to try to give a few extras, such as: wide or not ball to the opposition's team so that total runs can be less or check to win the match.

Nowadays, T20 is emerging as a short timed sport, which is popular among the young audiences. Most of the teams are using a data driven platform for analysing the performance among the cricketers in this format. This approach helps the efficient in this format.

Literature Review: In this section, 15 research papers are studied for this project. To determine the machine learning model that is used for analysing the cricketer's performance in T20I.

Sajib at el. 2024 [1] worked on Bangladesh men's cricket team performance in T20I with 5 cricketers. The *logistic regression* model was used for analysis in this paper. Waseem at el. 2025 [2] proposed on cricketer performance in T20I using survival analysis mainly on the batter. Zaman at el. 2025 [3] introduced analysis on the performance of All-Rounder in T20I cricket. *Cluster Analysis* was used for analysis in this paper. Waqas at el. 2025 [4] worked on a hybrid approach for team selection

in T20I cricket. *Logistic Regression* applied for the approach, which was given perfect accuracy in this paper. The accuracy was 97.5% for bowling and 95.3% for batting. Roy at el. 2025 [5] proposed a model based on K-means clustering to study Bangladesh Cricketer's T20 performance.

Jadwani at el. 2023 [6] introduced the approach on the analysis of evaluating cricketers performance in T20I. *Clustering* algorithm was used in this model. Prakash at el. 2022 [7] proposed an evaluation method and index of performance of cricketers in T20I using machine learning and deep learning. *K-means clustering* and *Random forest algorithm* were used in this method. Kasande at el. 2023 [8] worked on the automatic machine learning that selected the cricketer based on the performance. Both classification and regression were used in this model. Manage at el. 2020 [9] worked on the classification of the all rounder in the limited over cricket. Here *Logistic regression* and *Support Vector Algorithm(SVM) were implemented* in this machine learning model. *ROC curve* was used for the accuracy. November at el. 2025 [10] proposed the hybrid analytical approach to determine the main performance of the cricketers. Random forest algorithm and lasso logical regression were developed for this model. The accuracy was 85.9% from this hybrid model.

Bowala at el. 2022 [11] introduced a model that identifies the efficient performance on the bowler. Pramanik at el. 2022 [12] proposed the performance analysis of cricketers in T20 matches using classification. *K-nearest neighbors(KNN)* implemented in the model. Bharadwaj at el. 2024 [13] worked on the analysis on the cricketer's performance in machine learning. *SVM* was used in this research. Iltaf at el. 2025 [14] proposed the analysis on the performance of pace bowlers. Wickramasinghe at el. 2023 [15] worked on the application of the bowling performance in the machine learning model.

Methodology: In this section, the methodology is presented in the process of this project.

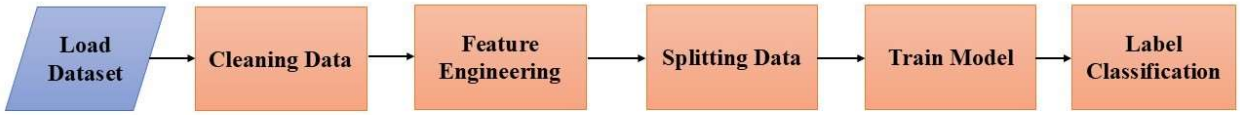

Fig: A flow chart of the project

Load Dataset: The datasets are loaded in the google colab from Kaggle. 3 csv files are used for the project to analyze the performance of cricketers. They are:

1. *players_info.csv*
2. *t20i_Batting_Card.csv*
3. *t20i_Bowling_Card.csv*

Cleaning Data: There are some null values in the dataset. Using ***dropna()*** function for removing the null values. All 3 csv files are merged into the one dataset using the ***pandas*** library from Phyton. Some column names have to change due to merging the dataframe.

Feature Engineering: Here, some cricket statistics are used for the project. Using those methods to find the cricketer's performance more accurately. For Batting,

1. *Strike rate = (Total runs / total balls faced) * 100*
2. *Boundary Percentage = (((Total fours scored * 4) + (Total sixes scored * 6)) / Total runs) * 100*
3. *Batting Average = Total runs / Total out*
4. *Consistency = Mean runs per innings / Standard Deviation per innings*

For Bowling,

1. *Economy = Total runs conceded / Total overs bowled*
2. *Bowling strike rate = Total balls bowled / Total Wickets*
3. *Dot ball percentage bowler = (Total dots bowled / Total balls bowled) * 100*
4. *Wicket taking consistency = Total wickets / Matched bowled*
5. *Hard Hitting control = Total sixes conceded / Total overs bowled*
6. *Bowling average = Total runs conceded / Total wickets*

Splitting Data: 1660 samples are used for training models. 80:20 is the splitting ratio of the data. 1328 samples are used for trains and 332 samples are used for tests.

Train model: Logical regression and SVM are implemented for training the data in the project.

Label Classification: After training both algorithms in the model, which is now ready for classification from the cricketer's performance.

Experimental Setup: This section describes the tools , dataset, pre-processing steps , analytical pipeline and model implementation used in our project

01.**Development Environment**

The project was implemented using :

a) Python (Jupyter Notebook / .ipynb environment)

b) Required Libraries

- NumPy
- Pandas
- Matplotlib / Seaborn
- Scikit-learn
- CSV file reading modules

The notebook served as the platform for running code, testing, visualization, and exporting outputs.

02 . Dataset Description

The project used cricket performance and match-related datasets extracted from publicly available cricket sources such as kaggle.

The dataset contained player statistics , match outcome labels and feature variables such as score, location , role performance ,boundaries, wickets etc.

03. Software / Hardware specification

10

a) Processor intel core i5

b) ram 8gb

c) Os : windows 11

04. Data Pre-processing

In this notebook we cleaned and prepared the date by loading the CSV files , cleaned and renamed the messy column names , dealt with incomplete and missing data, converted categorical values into numerical form and using only the relevant match and player records needed for the analysis .

05.Exploratory Analysis

In EDA , we see the data from different angles to understand how the features behave . The included ,

a) Checking frequency distributions to see how values are spread

b) Visualizing performance trends over time

c) Inspecting correlations between different variables

d) Studying how each feature affects player or match outcomes

These insights helped us to choose the important features .

06. Model selection and implementation

We used supervised machine learning methods to predict cricket match outcomes . Historical match data helped the models learn patterns so they could make predictions on new , unseen matches . To keep things simple , interpretable and efficient , we chose two algorithms one is logistic regression and support vector machine (SVM)

Logistic regression:

Logistic regression was used as our starting model . It predicts the chance of winning or losing based on different match-related features .

Support Vector Machine (SVM):

SVM was used to handle more complex relationships in the data . Cricket performance data often isn't perfectly linear , so SVM helps capture deeper patterns using kernel functions.

We have followed some evaluation process to make sure the model produces the correct result. We divided the data into a training set and a testing set . Then we run each model several times . This helped us check whether the accuracy stayed stable instead of giving inconsistent and random results . We also used plots and tables to compare them and understand their behavior more clearly .

Result and Discussion : In this section we are going to present a comprehensive analysis of T20I batting and bowling performances using aggregated career statistics obtained from the processed dataset .

1. Batting Performance Analysis

Leading run scorers : In the analysis, Virat Kohli is the highest run scorer in T20I cricket , scoring 4037 runs with average above 51 . Other leading batter Rohit sharma - 3900+ runs , Babar Azam 3485+ runs .

2. Boundary Distribution

The analysis of four and sixes indicates top batter Babar Azam , Rohit sharma , Kohli and warner dominate the fours category .

3. Bowling Performance Analysis

The results show that Tim Southee has taken the highest wickets in T20I cricket . Other major wicket -takers are Shakib Al Hasan , Rashid Khan , Lasith Malinga , Ish Sodhi .One of the most important findings is that Rashid Khan is the most economical bowler among the top wicket - takers , maintaining an economy below 7 and taking over 130 wickets .

Model Training Overview :

In this project two machine learning models one is logistic regression and one is support vector Machine (SVM) were trained to predict match outcomes .

1. Logistic Regression Performance

Logistic Regression performed very strongly, and achieved . Accuracy is 0.9940 and Precision is 0.99. This indicates the model made almost no incorrect prediction and handled both classes consistently .

2. Support Vector Machine (SVM) :

The SVM performed better . Accuracy is 0.9970 and Precision is 1.0 Both models performed well and SVM produced the best overall result ans showed perfect balance across all metrics .

Conclusion :

In this project we can understand how cricketers perform in T20I matches using statistics and machine learning . After combining batting and bowling data , cleaning it , and creating meaningful performance feature , we are able to analyze players more clearly and make models that give correct results

In our findings we see top batters like Virat kohli , Rohit Sharma and Babar Azam continue to dominate with consistent scoring . Bowlers such as Tim Southee, Shakib AL Hasan and Rashid khan show outstanding performance in wicket-taking The machine learning models performed well. Logistic Regression gave an accuracy of 99.40% and SVM performed better at 99.70% . It means that machine learning can recognize patterns in cricket data very efficiently .

Overall, this study shows how data and machine learning can help us understand T20 cricket in a deeper way. With more detailed data in the future—such as ball-by-ball information, match conditions, or player form—this approach can become even

more powerful and help coaches, analysts, and teams make smarter decisions in the game.

References:

1. Sajib, A. H., Limon, S. H., Naser, A. I., & Saha, G. (2023). Analysing factors impacting Bangladesh men's T20 cricket performance. *Scientific Journal of Sport and Performance*, *3*(1), 79–94. https://doi.org/10.55860/ENNL4841

2. Muhammad Waseem, Sulaiman Khan, Zakir Ali, Roohullah, & Anam Murtaza Shah. (2025). *Comparative Analysis of Players in T-20 International Cricket Using Survival Analysis*: https://doi.org/10.55966/assaj.2025.4.1.0110. `, *4*(01), 2021–2030. Retrieved from https://assajournal.com/index.php/36/article/view/691

3. Qamruz Zaman, Muhammad Ibrahim Khattak, Sidra Nawaz, Mohib Ullah Khan, & Ghazala Sahib. (2025). *A Cluster Analysis of the Performance of Allrounders in T-20 International Cricket.* Dialogue Social Science Review (DSSR), 3(1), 520–544. Retrieved from https://dialoguessr.com/index.php/2/article/view/193

4. Muhammad Waqas, Qamruz Zaman, Imran ullah, Fozia Mahsood, & Asma Shahnaz. (2025). A Hybrid Approach to T-20 Cricket Team Selection: Combining Probabilistic and Machine Learning Techniques. *Dialogue Social Science Review (DSSR)*, *3*(1), 978–996. Retrieved from https://dialoguessr.com/index.php/2/article/view/233

5. T. J. Roy, M. A. Mahmood, A. Mohanta, D. Roy, J. T. Jyoti and P. K. Ghosh, "*A Machine Learning Approach to Analyze the Performance of Bangladesh Cricket in T20*," 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2022, pp. 129-134, doi: 10.1109/ICISET54810.2022.9775839.

6. Jadwani, Yash, James Denholm-Price, and Gordon Hunter. "*A machine learning-based approach to analyse player performance in T20 Cricket Internationals*." 10th MathSport International Conference Proceedings 2023.

7. Deep Prakash, Sanjay Verma, *A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket*, Decision Analytics Journal, Volume 2, 2022, 100025, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2022.100025. (https://www.sciencedirect.com/science/article/pii/S2772662222000029)

8. Kasande, Manasi, and Sunita Jahirabadkar. "*Automated Player Selection for a Cricket Team using Machine Learning.*" (2023).

9. Manage, A. B. W., Kafle, R. C., & Wijekularathna, D. K. (2020). *Classification of all-rounders in limited over cricket - a machine learning approach*. Journal of Sports Analytics, *6*(4), 295-306. https://doi.org/10.3233/JSA-200467 (Original work published 2020)

10. November, R.V.; Cai, H.; Taliep, M.S.; Nyirenda, C.; Leach, L.L. *Identification of Key Performance Indicators for T20—A Novel Hybrid Analytical Approach.* Appl. Sci. 2025, 15, 6483. https://doi.org/10.3390/app15126483

11. Bowala, S. M. B., Manage, A. B. W., & Scariano, S. M. (2021). *Modeling T20I cricket bowling effectiveness: A quantile regression approach with a Bayesian extension.* Journal of Sports Analytics, *7*(3), 197-221. https://doi.org/10.3233/JSA-200556 (Original work published 2021)

12. M. A. Pramanik, M. M. Hasan Suzan, A. A. Biswas, M. Z. Rahman and A. Kalaiarasi, "*Performance Analysis of Classification Algorithms for Outcome Prediction of T20 Cricket Tournament Matches,*" 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 01-07, doi: 10.1109/ICCCI54379.2022.9740867.

13. Bharadwaj, Falak, et al. "*Player Performance Predictive Analysis in Cricket Using Machine Learning.*" Revue d'Intelligence Artificielle 38.2 (2024).

14. Iltaf, Ali, et al. "*Predicting International Success of Pace Bowlers in T20 Cricket.*" MathSport International 2025 (2025).

15. Wickramasinghe, L., A. Leblanc, and S. Muthukumarana. "*Semi-parametric Bayesian estimation of sparse multinomial probabilities with an application to the modeling of bowling performance in T20I cricket.*" Annals of Biostatistics and Biometric Applications 5.1 (2023): 1-13.