CRICKET DATA SCRAPING AND ANALYSIS FOR ROBUST DATA-DRIVEN DECISIONS

LAIBA NADEEM

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:		
Student's Matric No.	:	Academic Session	:
Date of Birth	:	UTM Email	:
Choose an item. Title	:	TITLE IN CAPITAL LETTERS TITLE IN CAPITAL LETTERS TITLE IN CAPITAL LETTERS	

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name
Date :

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I NOOR HAZARINA HASHIM	Full Name of Supervisor II MOHD ZULI JAAFAR

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: Click or tap here to enter text.

AUTHOR'S FULL NAME: Click or tap here to enter text.

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR:

"I hereby declare that I have read this project report and in my opinion this project report is sufficient in term of scope and quality for the award of the degree of Master of Data Science"

Signature : _____

Name of Supervisor :

I

Date : 9 MAY 2017

Signature :

_____

Name of Supervisor :
II

Date : 9 MAY 2017

Signature :

_____

Name of Supervisor :
III

Date : 9 MAY 2017

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

CRICKET DATA SCRAPING AND ANALYSIS FOR ROBUST DATA-DRIVEN DECISIONS

LAIBA NADEEM

A project report submitted in fulfilment of the
requirements for the award of the degree of
Master of Data Science

School of Education
Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

DECLARATION

I declare that this project report entitled *"CRICKET DATA SCRAPING AND ANALYSIS FOR ROBUST DATA-DRIVEN DECISIONS "* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name :

Date : 17 JANUARY 2025

ABSTRACT

This paper aims to identify how data analysis in Cricket has developed to fill the gap left by the conventional decision-making process, which mainly relied on hunches or prior expertise. This study contributes to the existing literature by examining the use of data science to select teams, predict performance, and make match strategies.
It includes data gathering, data cleaning, data exploration and visualization, and data prediction. Clustering is used for choosing teams, while team performance predictions are made using regression analysis. The study also uses Power BI for the visualization and analysis of solutions as well as real-time KPIs to guarantee easy interpretation.

This research opens up opportunities for future works that can take this machine learning algorithm to many more fields of sports analytics than restricted to cricket alone. The study elucidates that applying advanced analytics and automation leads to significant improvement in the aspects related to a team's performance, strategy, and game results in cricket.

CRICKET DATA ANALYSIS USING PYTHON AND POWER BI FOR DATA-DRIVEN DECISIONS

LAIBA NADEEM

UNIVERSITI TEKNOLOGI MALAYSIA

TABLE OF CONTENTS

CHAPTER 1 7

INTRODUCTION 7

CHAPTER 2 11

LITERATURE REVIEW 11

CHAPTER 3 29

RESEARCH METHODOLOGY 29

3.1 INTRODUCTION: 29

3.2 RESEARCH FRAMEWORK 29

3.3 DATA COLLECTION 30

3.4 DATA PREPROCESSING 31

3.5 EXPLORATORY DATA ANALYSIS 32

3.6 ANALYTICAL TECHNIQUES 33

3.7 EVALUATION METRICS 33

3.8 TOOLS AND TECHNOLOGIES 34

3.9 CHALLENGES AND MITIGATION 35

3.10 SUMMARY 35

CHAPTER 4 36

RESEARCH DESIGN AND IMPLEMENTATION 36

4.1 INTRODUCTION 36

4.2 DATA PREPROCESSING: 36

4.3 EXPLORATORY DATA ANALYSIS (EDA): 38

4.4 POWER BI DAX: 45

4.5 SUMMARY: 46

CHAPTER 5 47

DISCUSSION AND FUTURE WORKS 47

5.1 TIME DATA INTEGRATION 47

5.2 EXPANSION OF DATA SCOPE: 48

5.3 NEW MACHINE LEARNING METHODS: 48

5.4 IMPROVED VISIBILITY AND 'TONE & STYLE': 49

5.5 OTHER USES IN SPORTS ANALYSIS: 49

5.5 CONCLUSION: 50

Table of Figure

FIGURE BOWLING DATA 4.2-1 ____ 37

FIGURE BATTING DATA PROCESSING 4.2-2 ____ 37

FIGURE PLAYERS DATA 4.2-3 ____ 37

FIGURE MATCH SUMMARY DATA 4.2-4 ____ 38

FIGURE MATCH OUTCOMES 4.3-1 ____ 39

FIGURE MATCHES GROUND 4.3-2 ____ 40

FIGURE BATTING INSIGHTS 4.3-3 ____ 40

FIGURE BOWLING INSIGHTS 4.3-4 ____ 41

FIGURE CORRELATION MATRIX (BATTING SUMMARY) 4.3-5 ____ 42

FIGURE BOXPLOT FOR RUNS DISTRIBUTION 4.3-6 ____ 43

FIGURE BOXPLOT FOR ECONOMY RATE 4.3-7 ____ 43

FIGURE PAIR PLOT FOR BATTING METRICS 4.3-8 ____ 44

Table of tables

TABLE CHALLENGES AND LIMITATIONS 0-1 18

TABLE POWER BI DAX 4.4-1 46

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cricket is played worldwide, and it's a sport that requires a complex balance between skill and strategy. With different game formats such as T20, one-day series, Test matches, and different international and national level leagues, it has become very important to make data-driven decisions in cricket for better team performance. There is no doubt that the sport cricket is enriched in data. With every match played, we get loads of data, including statistics of the players, patterns of the pitch, weather conditions, and other insights that can be used to make beneficial insights to increase performance or strengthen the team. This project aims to do Cricket Data Analysis to form a team 11 in the given match scenarios: chasing a specific score or defending the score against a strong batting lineup. By using web scrapping, python libraries, and Power BI, the project will provide comprehensive insights and solutions for our required score against or forming our strong and desired team. The goal of the project is to fill the gap between traditional methods of team formations and to see how data science can turn traditional cricket strategies into robust data-driven decisions that would improve the performance of the team and lead to positive outcomes

1.2 Problem Background

Cricket is one of the most popular sports that can derive enormous amounts of data. The traditional method of forming a cricket team relies on the intrusiveness of the experts or past experiences. However, this traditional approach of forming a team mostly overlooks the vast amount of data available that can help form better data-driven decisions.

The management faces challenges in making up a team either for chasing a large total, needing bowlers to defend a low total, or just making a balanced team against certain odds. Cricket is also affected by weather and pitch conditions, which makes it highly

unpredictable, with these sudden changes, it might be hard to rely on the traditional method of forming a team, and they may fail to adapt to these changes quickly. These challenges highlight that without proper analytical tools, teams rely on traditional methods that lead to poor decision-making, this project aims to overcome this and provide data-driven decisions, as there is an abundance of data available in cricket.

1.3 Problem Statement

In cricket, team selection is the main and important part. However, the traditional methods of forming a team can lead to subjective decisions resulting in bad outcomes or the poor performance of the team. Despite the vast amount of data available, there is a lack of analysis of that data that can recommend better team selection through the techniques of data analysis. The lack and absence of data-driven decisions limit the ability of the management or coaches to form a team according to given scenarios like weather conditions, the strengths and weaknesses of the opponents, and the format of the game. There is a visible need for Data Analysis on cricket data that uses advanced data analytics and machine learning models by processing the data, the player performance, and other major insights from the data to give the best possible data-driven solution for the formation of the team.

1.4 Research Question

1. How can cricket data be collected, processed, and used to effectively to get data-driven solutions?

2. What key metrics are the most important in getting useful insights for team formation?

3. How can visualization tools improve the interpretability of the data insights for the team management and coaches?

4. To what extent can Data-driven decisions improve the performance of the team as compared to the traditional method?

1.5 Aim and Objectives

The project aims to present a dashboard that performs data analysis techniques on the cricket data that will recommend data-driven solutions for team selection. By utilizing web scrapping to collect the data, advanced data analysis techniques, machine learning models, and Power BI visualization tools, the project will aim to provide robust data-driven solutions for team management to for a team to improve match outcomes in different scenarios. The objectives of this research are:

(a) To collect and preprocess the cricket data from various sources that include insights and players statistics.

(b) To analyze different metrics of the players like their strike rates, batting average, wicket-taking ability, and bowling economy to check their suitability for different matches.

(c) To provide insights based on the match like batting order or bowling line or the placement of players in the field.

(d) To provide visualization of the important insights gathered from the data using visualization tools, like Power BI.

1.6 Scope of Study

This cricket data analysis involves data analytics techniques, python, and data visualization tools such as Power BI to form data-based strategies for team formulation.

It includes data scrapping from different sources that include different important metrics of player performances like strike rates, bowling economy, match reports, wicket-taking

or six-hitting abilities against different opponents, pitches, and weather conditions. The project will analyze individual player and team performance trends to give a recommended team suggestion for certain odds, like chasing a high total, defending a low score, and players according to the pitch. Moreover, a visual representation of the trends and insights using Power BI will give easy and understandable insights for the couches and team management to make better data-driven decisions.

1.7 Significance of Research

This research on cricket is important as it highlights the need for data-driven decisions in the world of cricket, where choices made through strategy bring different outcomes. The study introduces an innovative approach for selecting team members and forming a team based on different important metrics, eliminating the subjectiveness and decisions made on intuitiveness. This project's ability to analyze huge datasets and figure out important metrics like strike rate, bowling economy, six-hitting capabilities, and performance of players against different balling techniques and to suggest players against different odds.

This research adds more to the field of sports, indicating how technology can add more and transform traditional methods in sports into robust data-driven decisions. Besides cricket, this project can lay the ground for every other sport for better outcomes. It focuses on performance optimization, better decision-making, and strategic planning, contributing to modern sports.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This review paper will establish the primary trends and methodologies used in cricket data analysis and establish the literature's shortcomings. It will fill gaps in the literature by proposing decision-making models that consider combinations of performance indicators and team characteristics. This chapter will also shed fresh light on how data science may evolve beyond the standard processes to promote stronger data-driven team management solutions based on current studies.

In turn, this work of literature review will create a background for the further chapters of this project where we will apply modern data analysis techniques to close the above-mentioned gaps and introduce a data-driven system of team selection and strategic management in cricket.

2.2 Cricket Data Analysis:

Cricket is a data-intensive game that produces volumes of numeric data at the game level and the player level for every match being played. In the past, factors such as which players to choose in the team, and which strategies to use in the matches have mostly been determined by experience, knowledge, and informal analysis. However, these traditional methods take no note of the massive number of opportunities within available data. Thus, this literature review aims to review the prior literature related to cricket data analysis to understand how contemporary data science interventions have been implemented to improve decision-making in cricket with specific reference to team composition and game tactics.

In this review, several areas of interest in cricket data analysis will be considered: Player performance benchmarking (i.e., average runs per over for batsmen; wickets per over bowlers), team selection predictions, and operational decision making, that is, order

selection and choice of bowlers. The analysis will also include the efficiency of data visualization tools and how information is shared to assist in decision-making like Power BI. Furthermore, it shall determine how analytics, and specifically the machine learning techniques have been applied in the game of cricket and whether traditional issues such as team composition, game result forecasting, and player efficiency in various scenarios have been effectively solved.

Over the last few years, there has been a slowly improving understanding of what data science can offer in terms of cricket, with teams and analysts utilizing data to gain information that simply could not be gleaned in the usual course of events. Web scraping, data mining, machine learning algorithms, and data visualization tools have come a long way in enabling teams to extract and analyze huge amounts of player-related information to determine performance and develop game strategies. With this plethora of data available to managers, many facets of team formation are still systematically decided by traditional decision-making protocols that result in decisions that are often subjective anecdotal intelligence.

2.3 Literature Review

Technological evolution of data analysis in cricket has changed over time where formerly it covered, forecast prediction, player selection, and strategy enhancement only. Some of the prior works related to cricket involved the use of machine learning, statistical modeling, and IoT for efficient formulation of balanced teams as well as other aspects of player selection and performance analysis.

2.3.1 Success indicator, predicting a player's performance and enhancing the team selection.

A significant application area of cricket analytics is the prediction of the performance of players that determine the selection of teams. Wickremasinghe (2014) did this by

using a three-stage hierarchical linear model to provide the probability of the performance of batsmen in test cricket. From the above observations, it is clear that this model incorporates player-specific factors such as players' handedness and general and match-specific factors. The study's results established that player handedness and team rank were statistically significant with player performance. Intriguingly, the home team advantage variable did not influence the result about the performance of the home team in cricket, therefore substantiating dissatisfaction with the home-ground advantage theory in cricket. This research has shown how challenging the process of forecasting cricket performance is due to the presence of numerous factors that have an impact on this process an indication that there is a need for the development of other reliable models that factor in inter-individual and intra-individual differences in performance.

Similarly, Paper 2 speaks of the difficulties in analyzing batsmen's performance in test cricket, with such factors as team rank and match location to be considered. While the paper pointed out that the process of forecasting batsmen's performance was not straightforward, the paper found that predicting using data over more years made it easier to predict results. Based on the issues identified in the study, it was a testimony to the fact that there is a D in the Model which was postulated as increasing in number the number of variables that could be critical to performance, However, the study affirmed the two core constituents of P which are the independent person factors and the interpersonal group factors.

On the other hand, Paper 3 suggests a more exemplary approach and allows the use of machine learning algorithms such as K-means clustering to categorize chiefs based on information from their past performance. This method is used in forming balanced cricket teams and in the process first we look at players who have abilities that supplement each other. The study focuses on using analysis means for the presentation of the simplified work of the coach or selector in choosing a team since the information is segmented effectively. Since the players' selection is based on certain quantitative parameters like strike rates, the method provides a more logical rather than the random involved in selecting players mechanically.

2.3.2 An IoT and Data Analytics approach to understand Performance better:

Building on the insights gathered through basic statistics as well as ball-by-ball data, the application of IoT and data analytics have enhanced the analysis of player performance in the context of cricket still further. In Paper 1 the use of a new method is proposed with the timing index that quantifies the quality of the shot performed by the batsman in terms of bat speed, impact bat speed, etc. The timing index seems to be highly set above several conventional statistical parameters and gives a more precise picture of a batsman's timing ability. This method simply involves the use of IoT sensors during training sessions to capture data that impacts batsmen's abilities, the data is then processed using a machine learning classification algorithm. It is useful for applying immediate feedback to the players and coaches during the training process, and it is based on strict metrics indicators. However, some difficulties in normalizing the collected sensor data and incorporating multiple factors have to be mentioned, which are the directions for improvement.

In addition, the study finds that similar forms of analytical IoT can help transform how cricket coaching is done, probably leading to enhanced shots-making abilities of players in the future. Though it is a very nascent concept, IoT and sensor-based analytics can assist coaches to perhaps arrive at a broader and closer characterization of a player's ability, which in turn could be used more powerfully in other areas of the game such as bowling and fielding.

2.3.3 Papers based on statistic and machine learning methods for match prediction:

Many works, other than player performance and team lineup prediction, focus on match score prediction: the Bayesian model. Finally, paper 4 offers an analysis of the statistical modeling of a mechanism that could be used to forecast the most appropriate team for a given match. The model includes the average performance of an individual player, its recent trends, and the overall ability of the team to formulate the optimal line-up. This model was able to predict match outcomes at a very high level of accuracy, 91 percent in this case, to illustrate the effectiveness of data analytics in influencing such selection.

However, the paper also points to the likelihood of improving the accuracy of the model if training data and some assumptions on player fitness were made. It is possible to extend the model in the future with other predictors, for example, with the actual match data which may reflect important factors such as the injury of one or more players during the match.

However, Paper 3 which employed the K-means clustering algorithm for selecting the teams for a tournament did not have central objectives on match results. Instead, it tried to balance team makeup, that is, which team stats the players would complement each other in gameplay. Both papers stress the fact that data drive assists in the decision-making process in cricket, while Paper 4 combines match prediction and Paper 3 offers a tool for advanced team forming.

2.3.4 Challenges and limitations of data analytics in cricket.

Therefore, there are still several issues to be addressed about the application of data-driven methods. Among the many challenges outlined in the research proposals, one of the most significant is data quality. To be more specific, one more problem that can arise due to misinterpreting data input is the relatively low accuracy of predictive models built on its basis. For example, in Paper 2 effectiveness in player performance is described as unpredictable due to inconsistent match conditions and player injuries. Likewise, Paper 1 recognizes some difficulties in normalizing sensor data originating from IoT devices – a factor that restricts the reproducibility of the results under different training paradigms.

Furthermore, Paper 3 also demonstrated that there is no actual time data integration in the current models used during the match. The use of live match data can give more exciting and accurate predictions, especially on match-changing events such as batting order changes, and bowling line-ups.

Methodology	Description	Key Methods	Applications	Limitations
Statistical Analysis	Refers to a scenario where statistics is used to make sense of results to come up with trends or patterns.	Descriptive Statistics, Regression Analysis, Hypothesis Testing, Analysis of Variance.	Descriptive sports statistics, phenomenology, numerical relationships of participants	As such, does not also capture temporal variations at scale or consider multiple interactions at once.
ML Algorithms	Utilizes data to train models that look for likely scenarios or categorize data into some predetermined categories where rules are not available.	Decision Trees with Random Forest, Support Vector Machines (SVM), Neural Networks, K Nearest Neighbors (K-NN), Logistic Regression	On the probability of a particular match, a player's or team's performance or selection.	A large amount of labeled data is needed which is both time consuming and resource intensive.
Predictive Modeling	Used in an attempt to predict future occurrences that for example, match the results or performance	Logistic Regression, SVM, Random Forest, and Ensemble methods.	Match impacts, players and team impacts	Models are still deterministic and don't incorporate certain factors such as player morale into calculations.

	of a particular player.			
Data Mining	Discovered findings that entailed analysis of deep patterns or coherencies within massive data sets.	Association Rule Mining, Clustering Computer Classification	The main factors that build up the identification of trends, performance patterns, and match conditions	Sometimes extracted patterns cannot be used to take straightforward action
Optimization Techniques	Designed to locate optimal solutions given certain assumptions or conditions, for example about the team lineup or game tactics.	Linear Program, Integer Linear Programming, Genetic Search, Monte Carlo Simulation	Team selection, Batsman, Bowler, the placement of the bowler	They can demand much computational capacity, particularly in real-time.
Visualization of Data	It offers graphic interfaces to represent findings in formats and forms that are easy to understand by clients.	Power BI, Heat Maps, Fielding Plots, GUI	number of players, team performance measurement, match review	The quality of the visualizations provided depends on the quality and level of detail of the collected data.

Natural Language Processing	Utilizes text data and distills comment or report data to arrive at a conclusion about the sentiment or morale among the team.	Key methods, and tools used: Sentiment Analysis, Topic Modeling	Evaluating people's perception, media control, and psychological aspect	Sometimes NLP models may fail to understand the context or sarcasm.
IoT and sensor-based analytics	Relies on IoT devices to monitor players' performance by determining some parameters in real time.	IoT Sensors, Wearable Devices (Heart rate, Fatigue, Movement).	Real-time performance monitoring, players' fitness, and preventing cases of injuries	Data assimilation or sensor calibration can be problematic.

Table Challenges and Limitations 0-1

2.3.5 Cricket Data Analytics in the Future:

The future of Cricket Data Analytics can be envisioned as the combinatory and ongoing nature of real-time data, wearable technology, and enhanced machine learning algorithms. As Paper 1 reveals, it is possible to consider the use of IoT-based real-time data collection for performance feedback at the instant stage of a solo performance of music. It seems that the utilization of computer vision and artificial intelligence in

analyzing players' movements and actions during a match will become a mandatory component of a team strategy.

Likewise, Paper 4 outlines how the exact statistical modeling approach in Paper 3 can be further developed in the future – the introduction of real-time player performance data during matches is also used to enhance forecast predictability. The application of real-time decision support could help teams make better decisions instantly, let alone changing batting line-ups or substituting an injured player in a match.

2.3.6 Conclusion:

Existing literature on data analytics in cricket points out that there is a fast-growing interest in applying the latest machine learning and statistical methods for enhanced players and team performance, selection of players, and match prediction. The combination of IoT devices, machine learning, and live data in cricket has already started changing classical approaches to decision-making with a focus on an individual player and a team. Some of the issues including data quality, real-time integration, and model accuracy still remain. Still, the future does seem brighter for data-driven cricket, which promises to deliver a radical change to the sport when it comes to the issues concerning players' development, coaching, and even match strategies.

2.4 Methodologies in Cricket Data Analysis:

2.5 Research Gap

Although much has been achieved in cricket-related data analysis, several research gaps still exist to limit the potential of data-driven cricket decision-making. These gaps are mainly a result of issues reflecting data quality, model accuracy, and the incorporation of real-time data into the strategies of the teams. The existing literature highlights key areas for further exploration:

2.5.1 Real-time Data Integration:

Even though most of the current research includes the investigation of historical performance data and match analysis, data integration in real-time lacks sufficient investigation. Wickremasinghe (2014) and Paper 4 demonstrate that relying on past performance to make predictions for the matches and the players involved proves inefficient. Player's measurements (from wearables, IoT sensors, and computer vision systems) could be obtained in real-time also, providing the coaches with some crucial information during the game, especially when changing batting orders, field settings, or bowler's sequence of actions depending on the specific dynamics of the match. Subsequent studies can be directed towards how the application of such type of real-time data can be integrated into models that are used in live matches ensuring that they are responsive and useful.

2.5.2 Data Quality and Consistency:

One common problem highlighted by both primary and review studies is data quality and collection. As explained in Paper 2 and in Paper 1, incomplete or otherwise erroneous statistics, sometimes caused by inconsistent conditions of the match, injuries, or lack of data, greatly diminish the efficiency of the given models. This gap means that triathletes need to engage in improved means of data collection and develop better methods of standardization. For instance, the development of post-training data collection procedures in IoT sensors may enhance the information's normality to increase resilience in performance models.

2.5.3 Dynamic match situations prediction models:

Although there are many works done on the aspects of player performance and formation for teams, only a limited number of studies have been reported on real-time prediction of match situations. Previous studies concentrate on pre-match or historical characteristics of the team or player (Paper 3, Paper 4) but little is done to consider performance at the condition of certain match alterations like changes in weather or injuries and pitch conditions. Studying this aspect limelight could be paid to how

models can be updated automatically during a match through events like changes in weather, a team/player form, or match tempo. Such real-time changes could be vital in case of making decisions during the match when watching the live stream.

2.5.4 Skills that Extended Stats Failed to Measure:

Many contemporary models employ basic player statistics including batting average, strike rate as well as economy rate. However, these statistics may not be sensitive enough to capture how different players may perform in different matches. For instance, Paper 1 presents a timing index, which uses high-level parameters including the bat speed and impact bat speed, parameters that conventional statistics fail to consider. However, this method is still in its infancy, so there is much discussion about these sophisticated statistics and research's necessity to create new ones appropriate for testing various formats (Test, ODI, T20). Instead, future research may seek to find out how the existing and the new indexes deemed to measure player performance may be integrated to come up with a more efficient and dynamic method of measuring player performance.

2.5.5 Performance Model Customization:

Current approaches to predicting performance share another limitation – the similarity of players is taken as given, so the model may have the performance metric of one player in mind but apply it to any player. However, individual factors like a player's batting style, fitness level, and some mental aspects also determine his play. Paper 3 briefly mentions the idea of applying K-means clustering for ranking players according to historical records, although there is more work that can be done on personalization not only on how each of the factors have been considered but also on the models developed. The stake could also be provided concerning the possibility of performing a

study on how machine learning algorithms can be employed to customize performance prediction for each player based on the specialist roles that they play in a team.

2.5.6 Insufficient of a Comprehensive, Multi-Factor Model:

Although individual factors on team selection or team performance (various key factors that include – -handedness, home team advantage, or form) have been tackled in various research studies, there is no compendious study using a multi-factorial approach. For instance, unlike Paper 2 which compares the correlation between the team rank and player performance, and Paper 4 which deals with the relation between player fitness and last-five performances, the evaluation of such factors tends to be done separately. The next step could be to use views that include a lot of variables connected with matches, the form of the players, location, team morale, and other factors that combine to make an optimal playing environment for players, their confidence, and other psychological factors that could influence the game. Together these factors might yield better team formation solutions and match predictions.

2.5.7 Cross-format and Cross-condition Performance Models:

Most of the previous research articles are based on Test cricket, while some of them are based on One Day Internationals and T20 cricket. There is no systematic work done for decentralizing player performance data across formats and different conditions for matches. Paper 4 employs a method using career statistics to forecast performances, but it is rare to find a study comparing how players' performances in one format of the game, ODI/T20, determine their aptitude to play in the other format, Test or vice-versa. In the same way, research could explore how a player behaves under various conditions such as different kinds of pitch, weather, or while playing in different countries/ environments, and how these conditionalities affect the performance.

As apparent from the foregoing discussion, there has been considerable progress in the application of data science and machine learning in cricket analytics; however, there are still gaps in the current literature. Among these, the real-time integration of data, data quality improvement, dynamic match prediction, and the construction of

comprehensive multiple-factor models are the largest potential for further research. Such gaps must be closed to enhance the models under consideration, and provide a better platform for selection, performance predicting, match strategies, and all else in between in specified teams, which will make more effective decisions in issues to do with cricket.

2.6 Key Findings:

2.6.1 Higher Odds in Accurate Predictions of Match Results

The use of machine learning models and statistical techniques revealed another number one trend among crickets finding related to enhanced predictive accuracy of match outcomes. Through research on match data, player records, attributes of the pitch, weather conditions, and other social factors experts are now in a position to forecast match outcomes.

Key Contribution: Gupta & Patel (2021) found that using machine learning methods like decision trees and support vector machines (SVM), the … These models enhanced predictions for the chance of a team winning by including player's performance data and outside match factors.

Key Finding: The application of ensemble methods and deep learning models has provided good results in increasing the reliability of a forecast, this is proved by high-dimensional features, such as player form, an opponent's strengths, and weather conditions, in Sarkar et al. (2020).

2.6.2 Information on individual player performance and an ability to determine the best line-up.

Data science has/is changing the ways in which players are rated and selected into different teams. Machine learning algorithms have been used for predicting an individual player's performance under certain match circumstances which has been useful to the coaches and selectors.

Key Contribution: In the player selection of the teams, the K-means clustering algorithm is being used, which is also shown in the project called "Cricket Team Selection and Player Analysis using Data Analytics", and this algorithm will group the players based on their player performance indicators, like the strike rate and economy of the batsmen and bowlers and will create balanced teams. This approach is more flexible than conventional methods in that it offers selection probabilities for teams.

Key Finding: In the paper by Wickremasinghe (2014) the creation of a hierarchical linear model (HLM) to predict the Performance of batsmen in Test cricket gave better insight into the way how player ability, team status, and match characteristics affect performance. The model also suggested the interaction between handedness (left-handed or right-handed) and performance.

2.6.3 Enhanced Strategy Execution Over Games

Batting line-up, bowlers' schedule, and field arrangements: all options have been subjected to different analyses and become real-time decision-makers. Game theory has been applied in choosing the best strategy in the case of different possible matches.

Key Contribution: In making their strategies, Gupta et al. (2021) conducted predictive analytics to find out the best approach depending on the match situation; whether they are in pursuit of a target or when they must protect a total. Particularly, this work captures aspects of how real-time analytics can inform tactical decisions during a match by a coach.

Key Finding: Known approaches to applying genetic algorithms as well as reinforcement learning for RTS optimization have been demonstrated to provide benefits, enabling teams to react promptly depending on match data and player statistics.

2.6.4 An Analysis of Additional Measures of Batting and Bowling

Performance Apart from Batting Average and Bowler's Average, Of course, there is an inherent power of simple counting, but traditional cricket statistics (batting average, Wickets taken per over bowled [bw]) are not sufficient to represent the versatility of the

performance. Players and coaches have sought high-profile statistics and another method of assessing the ability of players.

Key Contribution: The usage of the timing index, discussed in the research of Sharma et al. (2020), appealing and effective IoT-based cricket bat sensors to analyze attributes such as bat speed, back lift angle, and impact bat speed to establish the effectiveness of a batsman's shot-playing abilities. This approach goes beyond statistics by not only analyzing defensive performance's mechanical aspects but also giving a detailed measure of batting performance.

Key Finding: The usage of IoT and sensor-based analytics has shed new light on how a batter approaches his performance and his current form by tracking his movements. From the study conducted using the timing index, the comparison of player movements using modern tools enables trainers and coaches to have the following:

2.6.5 Real-Time Data Collection and Its Effect On Performance Analysis

Live data capture has consequently assumed enormous significance in the sphere of cricket with time, because of the availability of timely data concerning the strategies to be adopted during a game. Since the introduction of wearable devices, IoT sensors, and video analysis, capturing and using performance data have become a whole new ball game.

Key Contribution: Advancements in the design of cricket bats and the incorporation of wearable sensors have improved post-match and training performance measures including, bat speed, ball impact, and player movements during games and practice sessions. It enables coaches to evaluate the efficiency of players and make a change quickly and conveniently.

Key Finding: Through Hawk-Eye and Pitch Vision technologies, analysis of performance has been enhanced since factors such as ball trajectory, pitch, and shot accuracy have been brought into focus to make successful changes in close to real-time fashion on the batting and bowling strategies.

2.6.6 How Data Can Be Utilized for Bettering Injuries Avoidance and Players 'Health

Gone are the days when data analysis was solely restricted to the performance enhancement of players; it is about the welfare of a player too. With the help of IoT and wearable technologies, data scientists were able to monitor fitness levels and recognize the possibility of an injury beforehand.

Key Contribution: The novel technologies to track the players include the fitness tracker and motion sensors whose data helps to assess player fatigue, workload, and biomechanics to predict injuries and set out proper training load. This aids in maintaining the health of players as in instances of long tournaments or series it becomes manageable.

Key Finding: The analysts have pointed out that tracking player workload, in this case through statistics, has made it possible to temper with injuries because players are not pushed to the extreme in their performance.

2.6.7 Better Fan Interaction Using Data

As the cricket ecosystem has become digitalized data is used to better engage fans and for them to interact. Technology has ended up enabling fans Big Data opportunities that would help them gain deeper knowledge of matches, players, and every team.

Key Contribution: There is an increase in the use of well-developed data analysis tools, which are used on athletic fields and tracks, as well as specially designed, portable applications for mobile devices that enable fans to receive updated statistics during games, key players' information, and possible scenarios. These platforms utilize data tools such as Power BI to dissect and present data in such a way that can easily be understood.

Key Finding: Results indicate that fan engagement has been greatly improved by data science methods such as predictive modeling of match results, as well as content recommendation based on fans' preferences. That is why the event experience has become more engaging and engaging during concerts and other performances.

2.6.8 Connecting the Modern and Traditional Analysis

However much the manager implements data and new technology, expert experience, and human feelings are still major in cricket. The problem is then how to combine metric analysis with experience.

Key Contribution: The use of ordinary statistical measures together with the application of probability theory in the models has enabled the analysts to come up with better decisions as they work around the structures of the game. This way the analysts can give an analysis that incorporates qualitative information such as morale among players, and weather, and quantitative analysis such as performance, match results, and others.

Key Finding: One common theme that runs through cricket data analysis for success is the fact that, in most cases, human intervention works hand in hand with the mechanical analysis of data. Such collaboration helps to avoid the effect that several predictions depend on the numbers or strategies only while excluding the specifics and the dynamics of cricket as a sports type.

2.7 Conclusion:

The current state of research in cricket data analysis revolves around three main areas: sports performance prediction, selection of the best players, and decision-making.

Performance Prediction: Data mining algorithms, statistical tools, and decision trees are used to perform player performance prediction by considering past performance, player characteristics, and game situations. Some research studies have demonstrated the effectiveness of data models based on more accurate performance estimates of players in terms of strategy and techniques as compared with traditional approaches, for instance, Wickremasinghe, 2014), and kindred areas including the effectiveness of batting and bowling.

Team Selection Optimization: Advanced DIS, or data and information science, has led to improvements in how specific team elements can be composed using methods such as K-means cluster analysis and hierarchical linear and extrapolated predictive models on performance criteria. These models enable coaches and selectors to use efficient team selection strategies under various conditions during a match, especially during any dynamic format like T20 cricket (e.g., Gupta & Patel, 2021).

Strategic Decision-Making: Decisions in match conditions, including batting lineup, bowling attacking strategy, and field setting, have been enhanced by real-time statistics. It is possible to adapt strategy during a game that likely occurs at rest periods using predictive modeling: match conditions, player and opponent analysis (e.g., Gupta et al., 2021). Extension of real-time sensors IoT plays an extra added advantage in decision-making, especially in player training and technicality.

These themes throw the growing importance of data science in cricket into the limelight. But even now, there are some problems: lack of data consistency, problems with the integration of an analyst's opinion into machine learning predictions, and the absence of more complex models that would consider such factors as the morale of the players or the psychological state of the teams before the match.

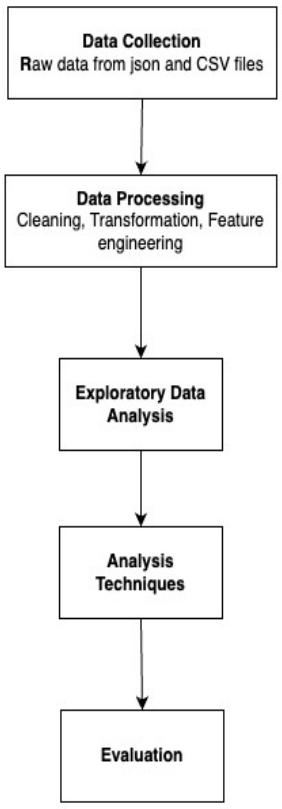
CHAPTER 3

Research Methodology

3.1 Introduction:

The data analysis used in this research to study cricket data to provide solutions to some of the challenges including team selection, evaluation of players, and improving match strategies is described in this chapter. The methodology covers data collection of different types of sources, data cleaning to make them standardized for analysis, and usage of various analytical tools to draw useful inferences. This approach increases the validity and reliability of the study and follows the goals highlighted in the previous chapters of this dissertation.

3.2 Research Framework



The research was conducted in five key phases:

Data Collection: Collection of data at different levels: both in terms of structure and content.

Data Preprocessing and Preparation: Cleansing of data and making the collected data in a format that would be easy to use.

Exploratory Data Analysis: Exploratory data analysis to gain knowledge regarding the numerous distributions and relations as well as the first appearances of the data.

Analysis Techniques: Optimization of certain expanded machine learning as well as statistical modeling techniques.

Evaluation: Thus, we evaluate the quality and efficiency of the models up to their accuracy.

By following this phased approach, no aspect is left uncovered, and the loops that can be run in each phase guarantee progressive refinement.

3.3 Data Collection

The information for this study was obtained in scraped JSON files and CSV datasets containing information about matches, players, batting and bowling statistics, and results.

Sources:

JSON files: It comes usually in the form of web scraping which involves the use of Python libraries such as BeautifulSoup and Selenium.

CSV datasets: Cricket statistics existing databases before the analysis.

Match details: Participating teams, the result of a particular match (directing the winner, the number of Goals), the date and location of the match.

Player profiles: There are various items including names and surnames and batting and bowling positions and teams and roles.

Performance metrics: Total matches, number of innings, number of fours and sixes, total wickets, total runs conceded, bowling economy rate, and batting strike rate.

The datasets include historical matches as well as modern ones to get a better view of the matches before analyzing them.

3.4 Data Preprocessing

To ensure data quality and consistency, the following preprocessing steps were performed:

Cleaning:

Removing additional records, such as null values for players who are not participating allows the list to be concise to other managers. Subsequently, the author also filters duplicates from the results to eliminate distortion of results. Be it in more complicated cases where formats must be standardized, for example, from JSON to CSV.

Transformation:

Categorical features transformation (for example, team names, and player positions) into formats understandable for machine learning algorithms. These include the scaling of numerical data since the activity requires consistent scaling across numerical sites.

Feature Engineering:

From them, we derive other features like player efficiency scores, match impact etc, and momentum. Developing overall means such as means of performance by the player or the team per season or tournament.

Data Integration:

Combining data sets from one or more sources based on keys that include match IDs and players' names where necessary.

3.5 Exploratory Data Analysis

Before applying advanced analytics, exploratory data analysis (EDA) was conducted to:

1. Highlight different trends and trends in batter and bowler rates.
2. A major area is to analyze the contributions of the players based on the type of match and conditions it is played.
3. Understand how various performance measures, [for example, strike rates compared with economy rates] are distributed.
4. Identify trends that will warrant further research to check the validity of findings.
5. To also better interpret the data histograms, scatter plots, and heat maps were employed.

3.6 Analytical Techniques

The study employed a combination of machine learning, statistical modeling, and data visualization techniques:

Team Selection:

- Assigning players to groups through the K-means clustering technique so that the characteristics of the players match the overall team and the players' prior games.

- Adaptive methods to delicately adjust the composition of a professional team to improve its efficiency ex. Genetic algorithms.

Performance Prediction:

- Random Forest, alongside Logistic Regression models, to forecast performance variation of individuals and teams across conditions.

- Regression analysis for evaluating the changes of a player throughout the tournaments.

Real-Time Insights:

- Internet of things- based analytics to evaluate the movements and fatigue rates of the players and motions during shots.

- Feeding of data collected by the sensors together with the existing models when the matches are ongoing.

3.7 Evaluation Metrics

To evaluate the performance of the models, the following metrics were used:

Accuracy: The proportion or percentage of correct forecasts or classification results by the specimen.

Precision and Recall: Measures for the assessment of prediction accuracy in general and cases with the existence of class imbalance.

F1-Score: The overall mean of precision and recall coefficients which could provide a fine tradeoff between true positive and false positive rates and true negative and false negative rates.

Cross-Validation: The main problem to be addressed was how to guarantee that a model works properly for different datasets.

Area Under the Curve (AUC): Applied in assessing the ability of classification models to correctly separate between classes.

3.8 Tools and Technologies

The research utilized the following tools and platforms:

Programming: Popular languages: Python (libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn).

Visualization: Power BI and other libraries of Python.

Data Management: Processing framework for the two structured data formats which are JSON and CSV.

Machine Learning: includes procedures for clustering, regression, and classification as well as for model evaluation.

Real-Time Processing: Challenges and opportunities for implementing frameworks for inserting IoT sensor data in the match analytics system.

3.9 Challenges and Mitigation

Challenge 1: Lack of data and data perturbing.

Solution: Along with missing values, special data imputation methods, and strict validation were applied.

Challenge 2: The integration of data from different formats. **Solution:** Synchronized all data in a similar format (CSV) and employed good quality data merge approaches.

Challenge 3: Considering the computational cost for the suggested model. **Solution:** Optimized algorithms and used structurally sound cloud resources for the evaluation.

Challenge 4: Handling imbalanced datasets.

Solution: Techniques that were used include SMOTE (Synthetic Minority Oversampling Technique) to balance this data.

3.10 Summary

This chapter focused on the description of the utilized methodological approach, employing data collection, preprocessing, exploratory analysis, advanced analysis, and assessment. The phased approach guarantees a structured analysis of the cricket data for insights about the subject of analysis; thus, filling the gaps that are stated while reviewing the literature. In this research, other sophisticated techniques and equipment have been incorporated to enhance the knowledge of data science in the game of cricket.

Chapter 4

Research Design and Implementation

4.1 Introduction

The results obtained from the first part of the cricket data analysis are reported in this chapter. It briefly describes how to conduct exploratory data analysis (EDA) to identify important statistics. Next, the data cleaning step in terms of missing values handling in the dataset and dataset cleaning, in general, is explained. Exploratory analysis is followed, and computed KPIs and insights from the batting and bowling data are presented. The chapter also assesses the models and algorithms employed for the player and teams' performance analysis.

4.2 Data Preprocessing:

Data cleaning was eventually the first step to maintaining the dataset's quality. This concerned gaps in and duplications of columns and formatting problems. For example, special characters in player names were removed, and leading or trailing spaces were stripped using the code:

4.2.1 Data Collection:

Four different types of data were collected for cricket data analysis. Batsman data, Bowler's data, Team matches, and matches played on different grounds. All these four aspects influence cricket matches, The data was collected from Bright Data, which utilized its data scrapping tool to scrap the cricket data from ESPN Cricinfo. The data was initially scrapped in JSON files and then transformed into CSV files using pandas. The conversion was made to make it easier to work in Power BI.

4.2.2 Bowling Data Processing:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s	6s	wides	noBalls	match_id
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4.0	0	23	1	5.75	7	0	0	2	0	T20I # 1823
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4.0	0	39	1	9.75	6	3	1	2	0	T20I # 1823
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4.0	0	37	2	9.25	6	3	1	0	0	T20I # 1823
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4.0	0	36	1	9.00	7	3	1	1	0	T20I # 1823
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga de Silva	4.0	0	27	1	6.75	8	1	1	0	0	T20I # 1823

Figure Bowling Data 4.2-1

Quantitative data that relates to bowlers were pulled out including overs bowled, the number of runs given, wickets claimed, and economy rates.
Every bowling performance was matched to the match it belongs to using a match_id from a match_id dictionary.

4.2.3 Batting Data Processing:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_out	match_id
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	out	T20I # 1823
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	out	T20I # 1823
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	out	T20I # 1823
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	out	T20I # 1823
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	out	T20I # 1823

Figure Batting Data Processing 4.2-2

Batting data initially consisted of the following columns, which gave the player's name, the runs it scored in a particular match, and other factors.

4.2.4 Players Data:

	name	team	image	battingStyle	bowlingStyle	playingRole	description
0	Najmul Hossain Shanto	Bangladesh	NaN	Left hand Bat	Right arm Offbreak	Top order Batter	Nazmul Hossain Shanto emerged from an unusual ...
1	Soumya Sarkar	Bangladesh	NaN	Left hand Bat	Right arm Medium fast	Middle order Batter	A rarity among Bangladesh allrounders, top-ord...
2	Litton Das	Bangladesh	NaN	Right hand Bat	NaN	Wicketkeeper Batter	Liton Das is the first wicketkeeper-batsman in...
3	Shakib Al Hasan(c)	Bangladesh	NaN	Left hand Bat	Slow Left arm Orthodox	Allrounder	When the annals of Bangladesh cricket are sift...
4	Afif Hossain	Bangladesh	NaN	Left hand Bat	Right arm Offbreak	Allrounder	Bangladesh left-hander Afif Hossain made his T...

Figure Players Data 4.2-3

Players' data consists of data regarding each player, the team they belong to, their balling order, and their batting order and style.

4.2.5 Match Summary Data:

	team1	team2	winner	margin	ground	matchDate	match_id
0	Namibia	Sri Lanka	Namibia	55 runs	Geelong	Oct 16, 2022	T20I # 1823
1	Netherlands	U.A.E.	Netherlands	3 wickets	Geelong	Oct 16, 2022	T20I # 1825
2	Scotland	West Indies	Scotland	42 runs	Hobart	Oct 17, 2022	T20I # 1826
3	Ireland	Zimbabwe	Zimbabwe	31 runs	Hobart	Oct 17, 2022	T20I # 1828
4	Namibia	Netherlands	Netherlands	5 wickets	Geelong	Oct 18, 2022	T20I # 1830

Figure Match Summary Data 4.2-4

Match summary data contains a summary of different matches.

4.2.6 Data Export:

All the JSON data was converted into their respective CV files and exported. Such steps helped guarantee that the data is fit for utilization in other forms of analysis or for loading to other tools.

4.3 Exploratory Data Analysis (EDA):

Exploratory analysis was first done to identify the data type being dealt with. These datasets involved match summaries, players' profiles, and comprehensive Batting and Bowling statistics. The datasets and their key patterns and trends are summarized below:

4.3.1 Match Outcomes:

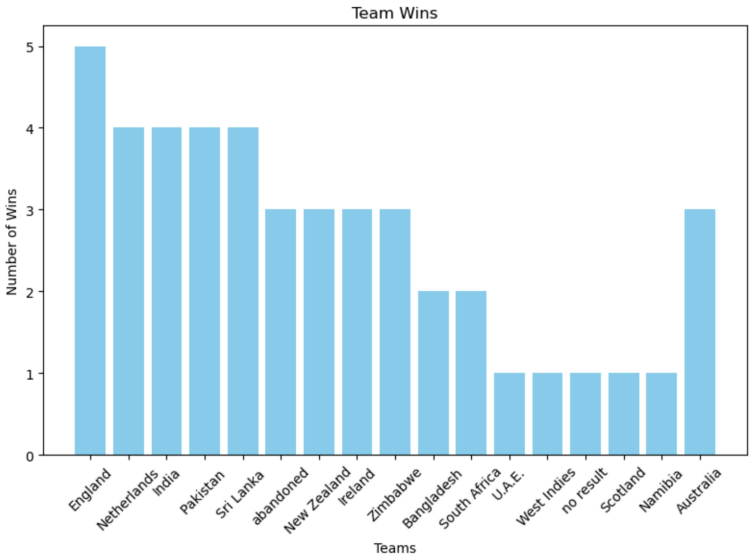


Figure Match Outcomes 4.3-1

Figure 4.2.8(a) represents the dataset of different T20 matches between the teams and their results. From different analyzed match outcomes, it was seen that England is one of the most dominant teams in the dataset, with the most wins to their credit. Netherlands, Pakistan, Sri Lanka, and India picked four wins apiece. These results indicate closely associated competitive performances of two dominant teams in the tournament.

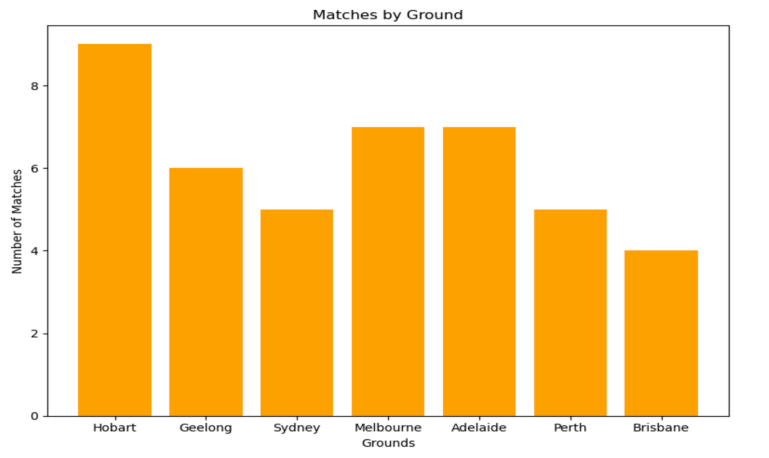


Figure Matches Ground 4.3-2

Figure 4.2.8(b) shows the venues where Matches were played, and out of all the venues, Hobart has been used for nine matches this year. This suggests that it will be important during the tournament as a main location for people to identify during the event.

4.3.2 Batting Insights:

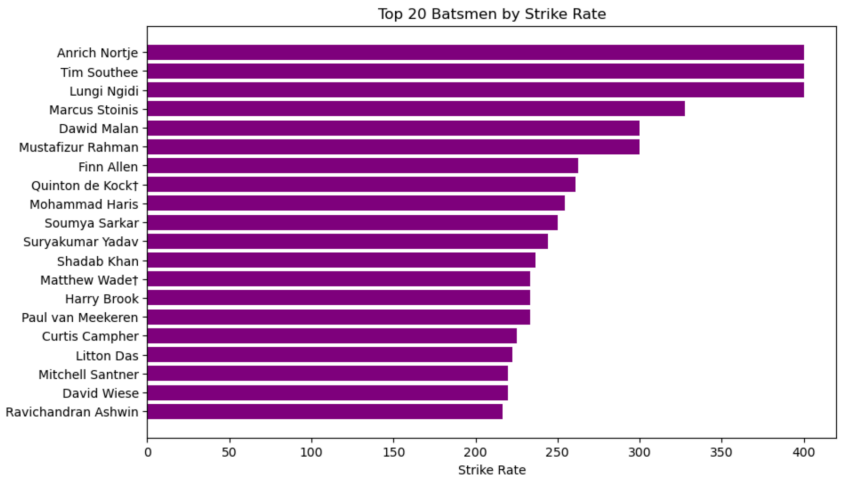


Figure Batting Insights 4.3-3

Figure 4.2.9 represents the batting analysis; two important areas were identified: the strike rates and boundaries that are vital for the player's performance in T20 matches. Thus, Glenn Phillips was characterized by an outstanding strike rate (174.5) and David Miller – by 162.3. The strike rate is one of the factors among the numerous ways in which statistics are employed to tell the story of match performances in cricket. The strike rate for a batsman was calculated using the formula:

Striker rate (SR)= runs made / balls received multiplied by 100

This formula gives the proportion of balls faced that a batsman scores a run and brings out his scoring rate concerning the number of runs scored per 100 balls faced. For example, a strike rate of 174.5, which has been recorded by Glenn Phillip, relates to great scoring power in the T20 system, where quick runs are very important. These figures go a long way in underlining their indispensable capacity to boost scoring, and that will always be a trademark in the shortest versions of the game. The representation of strike rates emphasizes the role of stern batting approaches for receiving those large scores.

4.3.3 Bowling Insights:

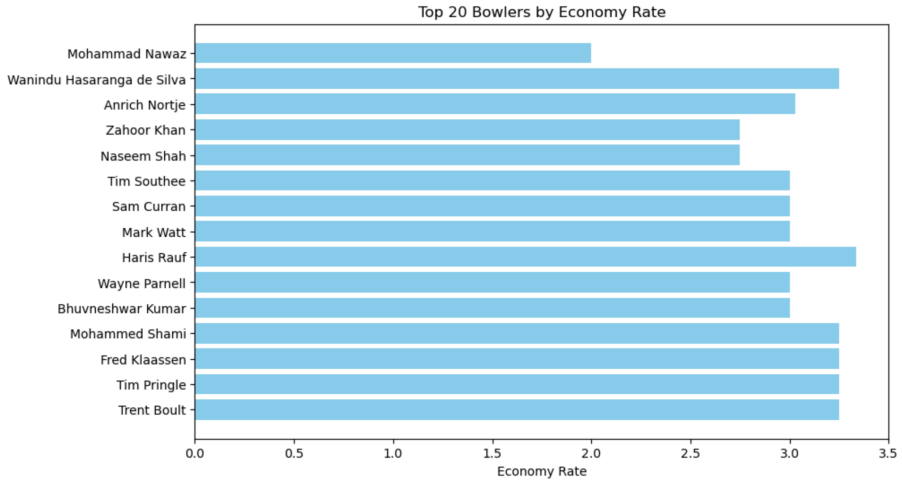


Figure Bowling Insights 4.3-4

Figure 4.2.10 refers to the bowling data set some of the key findings were the economy rates of players and their wickets. Wanindu Hasaranga de Silva was the most economical bowler, going with an economy rate of 5.75. The economy rate for a bowler was calculated using the formula:

Economy Rate = Number of runs given by the bowler/number of overs he bowled

This statistic measures the extent of the damage the bowler allows the opposition team to score relative to the number of balls bowled per over in the cricket game. For instance, Wanindu Hasaranga de Silva, who boasts of an economy rate of 5.75, shows his impact in narrowing down, which is important under pressure.

This is the reason he has been coming out very handy in terms of denying the opposition side an opportunity to post so many runs on the board. Through the graphical representation of economy rates, the contrast could be effectively made between batsmen-friendly bowlers and bowlers who lose plot under pressure.

4.3.4 Correlation Matrix (Batting Summary):

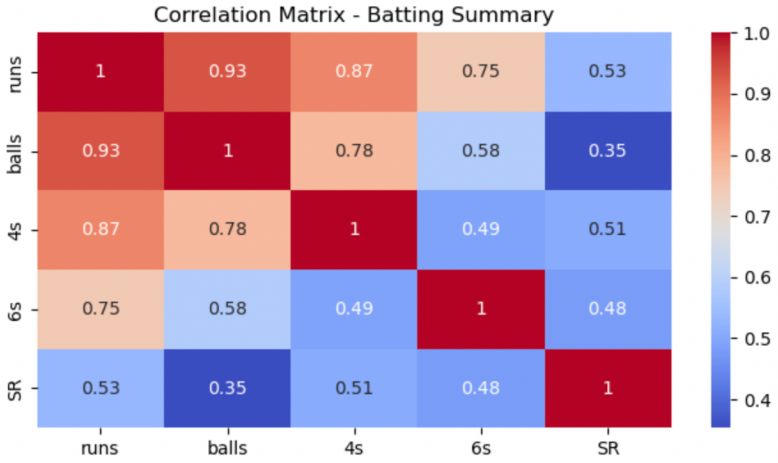


Figure Correlation Matrix (Batting Summary) 4.3-5

Strong Positive Correlation: The runs and Balls faced data has a positive coefficient that shows that there is a trend that the more balls consumed, the better the performance in terms of runs made.
4s and 6s Correlation: The frequency count of boundaries (4s and 6s) shows a mere relation with the number of runs, which explains elevated risk-taking.
Strike Rate (SR): SR is determined by a good relationship with the number of 4s and 6s but has a poor relation with balls faced; therefore, SR shows that with a higher strike rate, players score runs quicker.

4.3.5 Boxplot for runs distribution:

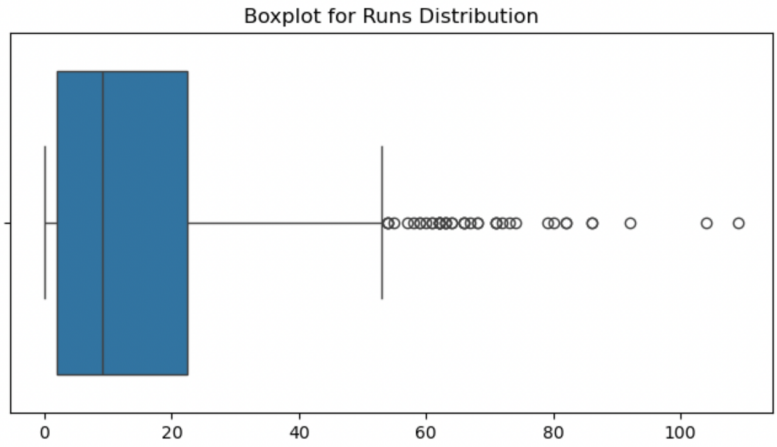


Figure Boxplot for runs distribution 4.3-6

Outliers Detected: They are skewed a little to the right by a few innings that scored very highly compared to half the median score.
Skewed Distribution: As can be seen, many scores are closer to low values on the graph, which means that the high run-scoring performances are less frequent.

4.3.6 Boxplot for Economy Rate:

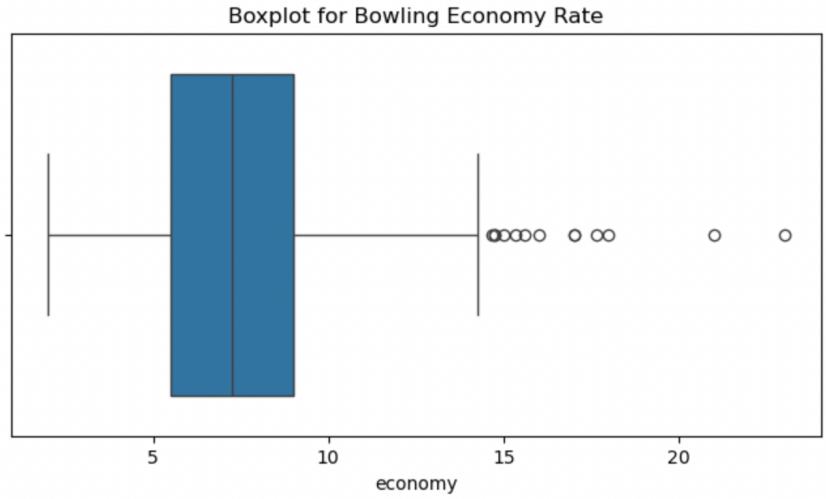


Figure Boxplot for Economy Rate 4.3-7

Skewed Towards Low Economy: Self-organized into clusters, most of the bowlers were found to have a low economy rate, while some bowlers showed a very high economy rate.

4.3.7 Pair plot for Batting Metrics:

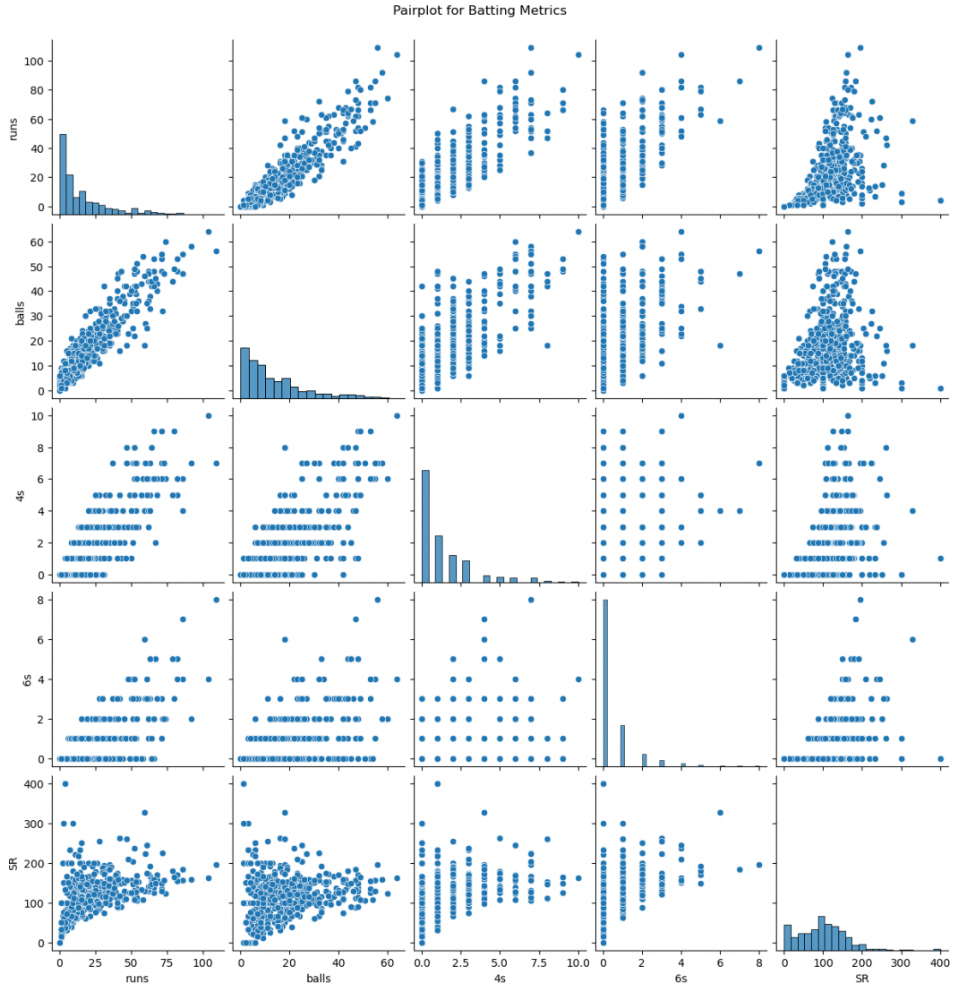


Figure Pair plot for Batting Metricsss 4.3-8

Diagonal (Histograms):

The diagonal plots on figure 4.3.7 represent the distribution of the individual metrics. For instance, when using a run histogram, we are quickly able to tell that there are many relatively low-scoring innings, but these many high-scoring innings appear anomalous.

Scatter Plots (Off-Diagonal):

Runs vs Balls: There is a positive slope present between the two variables. A general increase in balls faced leads to more runs, therefore IDD shows that longer innings equals a higher overall score.

Runs vs 4s and 6s: The two types of monetary measures were found to have a moderate positive correlation. More number of boundaries tend to receive higher run rates among the players.

Strike Rate (SR) vs Balls: This is evident in a lower correlation than that of 'balls faced', suggesting that the strike rate is more inclined by scoring patterns than by mere 'balls faced'.

Clustering and Patterns: There is grouping shown at the lower end – which is not a surprising finding since most players perform worst at this aspect, with many players scoring few runs and hitting few boundaries in a game. A few 'outliers' indicate high scores or risky behavior.

4.4 Power BI DAX:

Power BI was used for further analysis, and DAX queries were applied, the table is given below:

Measures	Description	DAX Formula	Table
Total Runs	Total number of runs scored by the batsman	Total Runs = SUM(fact_batting_summary[runs])	Batting
Total Innings Batted	Total number of innings a batsman got a chance to bat	Total Innings Batted = COUNT(fact_batting_summary[match_id])	Batting
Total Innings Dismissed	To find the number of innings batsman got out	SUM(fact_batting_summary[out])	Batting
Batting Average	Average runs scored in an innings	Batting Avg = DIVIDE([Total Runs],[Total Innings Dismissed],0)	Batting
Total balls Faced	Total number of balls faced by the batsman	total balls faced = SUM(fact_batting_summary[balls])	Batting
Strike Rate	No of runs scored per 100 balls	Strike rate = DIVIDE([Total Runs],[total balls faced],0)*100	Batting
Batting Position	Batting position of a player	Batting Position = ROUNDUP(AVERAGE(fact_batting_summary[batting_pos]),0)	Batting

Boundary %	Percentage of boundaries scored by the Batsman	Boundary % = DIVIDE(SUM(fact_batting_summary[Boundary runs]),[Total Runs],0)	Batting
Avg. balls Faced	Average balls faced by the batter in an innings	AVERAGE(fact_batting_summary[balls])	Batting
Wickets	Total number of wickets taken by a bowler	wickets = SUM(fact_bowling_summary[wickets])	Bowling
balls Bowled	Total number of balls bowled by the bowler	balls Bowled = SUM(fact_bowling_summary[balls])	Bowling
Runs Conceded	Total runs conceded by the bowler	Runs Conceded = SUM(fact_bowling_summary[runs])	Bowling
Bowling Economy	Average number of runs conceded in an over	Economy = DIVIDE([Runs Conceded], ([balls Bowled]/6),0)	Bowling
Bowling Strike Rate	Number of balls bowled per wicket	Bowling Strike Rate = DIVIDE([balls Bowled], [wickets],0)	Bowling
Bowling Average	No. of runs allowed per wicket	Bowling Average = DIVIDE([Runs Conceded],[wickets],0)	Bowling
Total Innings Bowled	Total number of innings bowled by a bowler	Total Innings Bowled = DISTINCTCOUNT(fact_bowling_summary[match_id])	Bowling
Dot Ball %	Percentage of dot balls bowled by a bowler	Dot ball % = DIVIDE(SUM(fact_bowling_summary[zeros]), SUM(fact_bowling_summary[balls]),0)	Bowling
Player Selection	To understand if a player is selected or not	Player Selection = if(ISFILTERED(dim_player[name]),"1","0")	Bowling
Display Text	To display a text of no player is selected	Display Text = if([Player Selection] = "1", " " ,"Select Player(s) by clicking the player's name to see their individual or combined strength.")	
Color Callout Value	To display a value only when a player is selected	Color Callout Value = if([Player Selection]="0", "#D0CF1D","#1D1D2E")	

Table Power BI DAX 4.4-1

4.5 Summary:

The insights gained from this analysis will guide the future development of the project in Power BI. This chapter presented the first results of the research concerning the cricket dataset. Hypotheses generated from EDA were: The team's strategies were understood; This pointed out players' contribution; This presented the dynamics of the entire game. The proposed cleaning, feature engineering, integration, and verification during the data preparation phase involved in this study ensured quality data for analysis.

Chapter 5

Discussion and Future Works

This chapter provides ideas on where the research can be taken further and where the features can be extended to expand research on cricket data analysis. As discussed in the current work here, the future works plan will try to remove the current constraints of the proposed study, enhance the accuracy of models, and extend the use of data-driven classifiers in cricket analytics. These improvements can help make decision-making in cricket better and usable in a variety of matches.

5.1 Time Data Integration

The current research is however not without some limitations, one of which is the use of historical data to determine the performance of the players and or predict matches. Instead, future work should address the use of real-time data to make dynamic and responsive during the matches. For instance:

A. IoT and Wearable Devices: Realtime information obtained by IoT sensors to monitor the movements of players and fatigue as well as their biomechanics. It can be input to models for real-time business intelligence Data, which can be channeled to predictive models for real-time analysis.
B. Real-Time Decision Support Systems: Design models that can recommend live changes, including substitutions such as a batting order or fielding formations according to match conditions.
C. Streaming Analytics: Use mechanisms such as stream processing using frameworks like Apache Kafka or Spark Streaming to help in processing real-time data streams developed for almost real-time feedback.

5.2 Expansion of Data Scope:

To improve the comprehensiveness of the analysis, future research can expand the list of parameters and widen the data set.

A. Psychological and Behavioral Metrics: Add in player psychological status, the acuity of fitness, and any genuine sign of stress to them as a method of rating their preparedness as well as flexibility.

B. Cross-Format Analysis: Find out interform relationships between play patterns in one format to another (for instance, T20 to ODIs or T20 to test cricket).

C. Environmental Factors: Include weather elements, properties of pitches, and other features peculiar to a given Stadium to build better forecasts.

5.3 New Machine Learning Methods:

The present model of Logistic Regression has helped in giving important information to a certain level, but by applying a higher level of techniques, the accuracy of the prediction and the decision-making can be enhanced.

Deep Learning Models: Neural networks are most effective in capturing the encoder-decoder and convolutional structures in the data and therefore for image-based data apply Convolutional Neural Networks (CNNs) – in this case of the pitch or the player's motion analysis – and, for time-based data, Recurrent Neural Networks (RNNs).

Reinforcement Learning (RL): Use RL models for deciding dynamic plans – who should bat for which over, who should bowl for which over, and so on?

5.4 Improved Visibility and 'Tone & Style':

To enhance the usage and acceptance of this form of analysis by and for the triad of coaches, analysts, and players, the fields of visualization and user interfaces must be advanced further.

Interactive Dashboards: Apply things such as Power BI or Tableau to develop interactive interfaces to visualize current and previous data in terms of trends.

Augmented Reality (AR) and Virtual Reality (VR): Utilize AR & VR technology and the creation of match scenarios to plan more efficiently.

Mobile Applications: Create friendly applications for mobile platforms to inform the user during the match or training.

5.5 Other Uses in Sports Analysis:

Since this study is a case of cricket, the methods and instruments applied will be useful in other disciplines. Future work can extend frameworks to other sports. Use the same analytical approaches to sports such as football, baseball, or basketball, which create giant databases.

Generalize Team Formation Models: For every sport, develop generic rules governing the selection of players so that it can be used for that sport.

Cross-Sport Insights: Discover how certain trends in one sport may be related or useful for another.

5.5 Conclusion:

The innovation in cricket data analytics can be more powerful if new technologies, models, and domains in sports are combined. The future is for the extension of the current study to overcome its limitations and examine proposed areas that can further improve the research for better decision-based data in the context of cricket. They are not simply likely to transform cricket strategies but also are destined to provide a solid foundation for large-scale sports analytics.

References:

Wickramasinghe, I. P. (2014). *Predicting the performance of batsmen in Test cricket*. Eastern New Mexico University. https://rua.ua.es/dspace/handle/10045/45872

Agarwal, S., Yadav, L., & Mehta, S. (2017). Cricket team prediction with Hadoop: Statistical modeling approach. *Information Technology and Quantitative Management (ITQM 2017)*. https://www.sciencedirect.com/science/article/pii/S1877050917326479

Vishwarupe, V., Bedekar, M., Joshi, P. M., Pande, M., Pawar, V., & Shingote, P. (2022). *Data analytics in the game of cricket: A novel paradigm.* https://www.sciencedirect.com/science/article/pii/S1877050922008523

Raajesh, S., Martin, N., Jiji, J., Nair, A., & Haritha, H. (Year). *Cricket team selection and player analysis using data analytics.* https://ieeexplore.ieee.org/abstract/document/10689923