Cricket Data Analytics using Python

A
PROJECT

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
In
INTERNET OF THINGS

By

Pranjal Dubey (0901CE211094)

Under the guidance of

Dr. Praveen Bansal
Assistant Professor & Coordinator
Centre for Internet of Things

CENTRE FOR INTERNET OF THINGS
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE,
GWALIOR (M.P.) – 474005

JUL-DEC 2023

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this project entitled Cricket Data Analytics Using Python which is being submitted in the partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Internet of Things to Centre for Internet of Things is an authentic record of my own work carried out under the guidance of Dr. Praveen Bansal . Assistant Professor & Coordinator. Centre for Internet of Things.

The matter presented in this project has not been submitted elsewhere by us for the award of any other degree/diploma.

Pranjal Dubey

Name & Signature Student

Date: 22/11/23

Place: Gwalior

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief.

Minor Project Coordinator

22/11/23

Dr Yashwant Sawle
Assistant Professor,
Centre for Internet of Things
MITS, Gwalior

Guided/Forwarded by

22/11/2023

Dr Praveen Bansal
Assistant Professor &
Coordinator of Centre for Internet of Things, MITS, Gwalior

ACKNOWLEDGEMENTS

Madhav Institute of Technology & Science, Gwalior, who has been supportive of my project and career goals and who worked actively to provide me with valuable academic time to pursue these goals. I would also like to extend my regards to **Dr. Praveen Bansal** who has supported me throughout this project with his knowledge and guidance. Each of the members of the Centre for Internet of Things provided me with extensive personal and professional guidance and taught us a huge deal about both scientific research and life in general. This work would not have happened without the financial and academic support of the Madhav Institute of Technology & Science, Gwalior.

Executive Summary

With millions of admirers worldwide, cricket is one of the most popular sports in the world. In addition, there is a lot of data on individual performance, team tactics, and match results, making the game intricate. By gaining insights from this data through data analytics, teams can improve their playing lineups, strategies, and training regimens. In order to help spectators better understand the game and forecast future results, it can also be utilised to spot trends and patterns in the gameplay. The goal of this project is to analyse cricket data using Power BI and Python. Python is an effective programming language that works well with jobs involving data analysis. Using Power BI, interactive dashboards and reports can be made for data visualisation purposes. In general, this project will show off Python and Power BI's capabilities for data analytics related to cricket. Many stakeholders, including cricket teams, cricket fans, media and broadcasters, and cricket sponsors, will benefit from the project's outputs.

ABBREVIATIONS

BI	Business Intelligence
ICC	International Cricket Council
CSV	Comma Separated Values
JSON	JavaScript Object Notation
HTML	Hypertext Markup language
DAX	Data Analysis Expressions
KPI	Key Performance Indicator
PY	Python

Table of Contents

ACKNOWLEDGEMENTS **3**
Executive Summary **4**
ABBREVIATIONS **5**
LIST OF FIGURES **7**
CHAPTER - 1 ..... **8**
INTRODUCTION **8**
1.1 overview 8
1.2 Data Analytics Definition 9
1.2.1 Types of Data Analytics 9
1.3 OBJECTIVES 10
CHAPTER - 2 LITERATURE REVIEW **11**
2.1 Early Cricket Data Analytics 11
2.2 Modern Cricket Data Analytics 12
CHAPTER – 3 ..... **14**
METHODOLOGY **14**
3.1 Data collection using web scraping 14
3.2 Data cleaning and transformation in Python Pandas 15
3.3 Data Transformation in Power Query. 16
3.4 Data modelling and building parameters using DAX 17
3.5 Build dashboard in Power BI 18
3.6 Collect insights from the dashboard and select the final 11 players 19
CHAPTER – 4 ..... **20**
RESULTS & DISCUSSIONS **20**
4.1 Player Analysis and Final 11 20
CHAPTER – 5 ..... **23**
CONLUSION **23**
REFERENCES **24**

LIST OF FIGURES

S NO.	FIGURE NAME	PAGE NO.
1.	Cricket Data Analytics using Python and Power BI	8
2.	Cricket Sabermetrics	11
3.	Web Scraping using Bright Data	14
4.	Data Cleaning and Transformation	15
5.	Data Transformation in Power Query	16
6.	Data modelling and building parameters using DAX	17
7.	Build Dashboard in Power BI	18
8.	Data Insights	19
9.	Anchors	20
10.	Fast Bowlers	21
11.	All Rounders	21
12.	Finishers	22
13.	Best 11	22
14.	Openers/ power hitters	22

CHAPTER - 1

INTRODUCTION

1.1 overview

Each cricket match generates a large amount of data, making the sport complex. You can learn a lot about player performance, team tactics, and match results from this data. The process of gathering, preparing, and evaluating data in order to derive valuable insights is known as data analytics. Two effective tools that can be used for cricket data analytics are Python and Power BI. Python is a programming language that works well for analysing and manipulating data. A business intelligence tool called Power BI allows users to build interactive reports and dashboards. Cricket teams, players, and coaches can all use the insights from our analysis to perform better. Teams can utilise our data, for instance, to determine which players are performing well, choose the optimal starting lineup for every game, and create winning team plans. With the help of our analysis, players can pinpoint their areas of weakness and monitor their development over time.

Fig 1.1 Cricket Data Analytics using Python and Power BI

With the use of data analytics, this project seeks to improve knowledge of cricket as a sport. We will conduct a thorough analysis of cricket data using Power BI and Python, which will include:

- Web scraping: To obtain cricket data from multiple websites, including ESPN Cricinfo, CricketArchive, and ICC Cricket, we will employ Python.

- Data cleaning and transformation: The scraped data will be cleaned and converted into a format that is appropriate for Power BI using Pandas, a Python package for data manipulation and analysis.
- Data visualisation: To produce interactive dashboards and reports that meaningfully and perceptively display the cricket data, we will employ Power BI.

1.2 Data analytics definition

Data analytics, similar, to data mining involves the process of analysing types of data both structured and unstructured. By employing methods, procedures, algorithms and systems it aims to extract knowledge and insights. The collection, transformation and analysis of data are carried out in order to identify patterns, trends and associations. The application of data analytics can lead to cost reduction, streamlined processes and informed decision making. Numerous sectors such as healthcare, finance, retail and manufacturing leverage data analytics for purposes. For example healthcare organisations utilize data analytics to reduce costs while improving outcomes and identifying research opportunities. Financial institutions employ it for risk assessment purposes like fraud detection as for developing innovative products and services. Retailers use data analytics to target marketing campaigns while optimizing inventory levels based on insights, into customer behaviour. Manufacturing companies rely on data analytics to enhance product quality by reducing waste and increasing productivity.

1.2.1 Types of Data Analytics

These four categories of data analysis are frequently employed in conjunction to gain a comprehension of data. For an instance a company may utilize analysis to grasp its performance diagnostic analysis to recognize the factors impacting its performance predictive analysis to anticipate future performance and prescriptive analysis to determine the optimal actions needed to achieve its objectives.

Types-

- **Descriptive analytics:** The goal of this kind of analytics is to describe historical events. To summarise data, it makes use of descriptive statistics like means, medians, and modes.
- **Diagnostic analytics:** This kind of analytics goes beyond just reporting events; it also explains why they occurred. It finds relationships between variables by using methods like regression analysis and correlation.

- **Predictive analytics:** This kind of analytics forecasts future events based on past data. To create predictive models, it makes use of methods like time series analysis and machine learning.
- **Prescriptive analytics:** This kind of analytics suggests courses of action that can be followed to accomplish desired results. It determines the optimal course of action by applying techniques like simulation and optimisation.

1.3 OBJECTIVES

The use of Data Analytics in Cricket aims to achieve the following primary objectives:

1. Identify trends and patterns in player performance over time.
2. Clean and transform the scraped data into a format that is suitable for Power BI analysis
3. Create interactive dashboards and reports using Power BI to visualize the cricket data in a meaningful and insightful way.
4. A Power BI dashboard that provides interactive visualizations of cricket data.
5. The ability to use cricket data analytics to improve player performance, team strategies, and match outcomes.

CHAPTER - 2

LITERATURE REVIEW

A lot of data is produced during a cricket match. Using this data, teams' tactics, player performance, and match results can all be better understood. Gathering, sanitising, and evaluating data in order to derive valuable insights is the process known as data analytics. For data analytics on cricket, two effective tools are Python and Power BI. Python is an excellent programming language for working with and analysing data. Interactive dashboards and reports can be made with Power BI, a business intelligence tool.

2.1 Early Cricket Data Analytics

When it came to evaluating player performance, analysts in the early days of cricket data analytics relied on fundamental statistical metrics like batting average, bowling average, and fielding average. These metrics were helpful in giving a broad picture of a player's performance, but they lacked subtlety and were unable to shed light on more intricate elements like bowling and batting technique or the effects of various environmental factors on player performance. Sabermetrics, a novel approach to data analytics in cricket, first appeared in the late 1970s. Bill James, who employed sophisticated statistical analysis to refute conventional wisdom regarding baseball performance, popularised sabermetrics. Cricket analysts quickly adopted sabermetrics, and new statistical metrics like expected batting average, slugging percentage, and on-base percentage were used to evaluate player performance. Cricket data analytics has involved the use of sabermetrics, which is still in widespread use today.

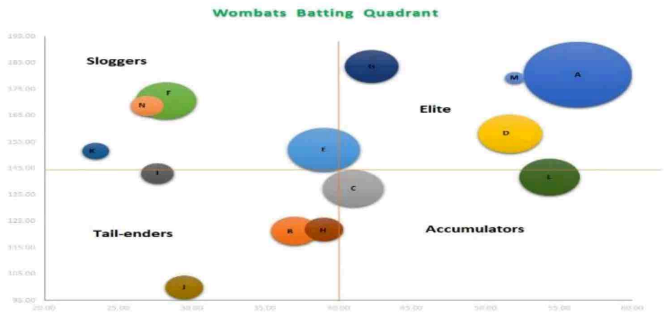

Fig 2.1 Cricket Sabermetrics

In the same way that data analytics has been used for centuries to analyse and enhance player and team performance, cricket is a centuries-old sport. However, as new technologies and approaches have emerged to enable more complex and precise analysis, the approaches used for cricket data analytics have changed significantly over time.

2.2 Modern Cricket Data Analytics

Naglah, A., et al. (2023) Teams, coaches, and players are using the quickly expanding field of modern cricket data analytics to boost output. Machine learning, artificial intelligence, and big data analytics are just a few of the methods used in modern cricket data analytics. Player performance analysis is one of the main areas of focus for contemporary cricket data analytics. Baseball, bowling, and fielding are just a few of the sports in which player performance can be monitored with the help of contemporary data analytics tools. After that, this data can be utilised to pinpoint problem areas and create specialised training curricula.

Alftab, M., et al. (2023) Player performance analysis is one of the main areas that modern cricket data analytics is focused on. Player performance in batting, bowling, and fielding can all be monitored with the help of contemporary data analytics tools for cricket. Afterwards, using this data, customised training plans and areas in need of improvement can be created. Modern cricket data analytics is also being used to enhance the fan experience, in addition to player performance analysis and strategy development. Broadcasters are utilising data analytics, for instance, to offer more in-depth and educational commentary. Additionally, data analytics is being used to create new fan-interactive features like player comparisons and real-time match predictions.

Nimmagadda, V. S. R., et al. (2019) With new tools and techniques appearing frequently, the field of modern cricket data analytics is one that is fast developing. The emergence of big data has been one of the biggest new developments in recent years. big data describes extensive and intricate datasets that are challenging to analyse through conventional means. Nonetheless, big data analytics may be utilised to glean insightful information from cricket data with the aid of strong computers and complex algorithms. Using machine learning to data analytics for cricket is another significant advancement in this field. Through the use of machine learning, computers can become more intelligent without explicit programming by learning from data and making predictions. Several different cricket data analytics applications have been developed using machine learning.

Gupta, A., Sharma, S., & Kumar, V. (2021) To create fresh approaches and methods, contemporary data analytics in cricket is also being employed. To determine the ideal batting and bowling strategies for various situations, for instance, teams can employ data analytics. Assembling strategies to take advantage of an opponent's weaknesses can also be done by teams using data analytics. The application of modern data analytics in cricket extends beyond

strategy development and player performance analysis to enhance the fan experience. To give more in-depth and educational commentary, for instance, broadcasters are utilising data analytics. New interactive features for fans, like in-the-moment match predictions and player comparisons, are being created using data analytics.

CHAPTER – 3

METHODOLOGY

3.1 Data collection using web scraping

Data extraction from websites is known as web scraping. It can be used for projects involving data science, Market analysis etc. A multitude of tools and methods are available for extracting data from websites. A popular method is to use a web scraping library like Scrapy or Beautiful Soup. These libraries offer functions for web page data extraction and HTML parsing. We can store the data from the website in a database, CSV file, or JSON file, among other formats. Data analysis and other uses are then possible with this information.

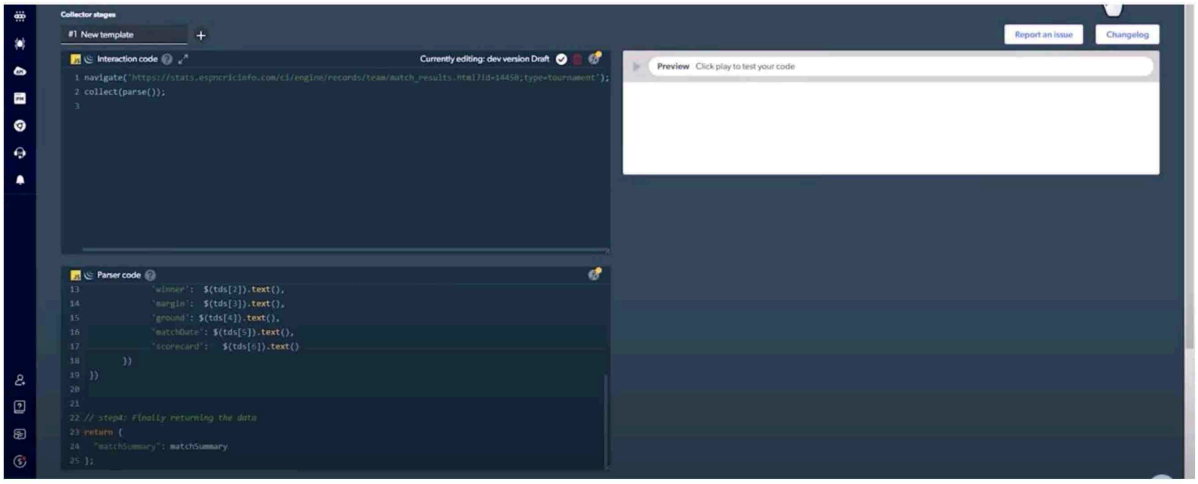

Fig 3.1 Web Scraping using Bright Data

3.2 Data cleaning and transformation in Python Pandas

```python
#import necessary libraries
import pandas as pd
import json
```

(1) Process Match Results

```python
with open('t20_json_files/t20_wc_match_results.json') as f:
    data = json.load(f)

df_match = pd.DataFrame(data[0]['matchSummary'])
df_match.head()
```

```python
df_match.shape
```

```
(45, 7)
```

Use scorecard as a match id to link with other tables

```python
df_match.rename({'scorecard': 'match_id'}, axis = 1, inplace = True)
df_match.head()
```

Create a match ids dictionary that maps team names to a unique match id. This will be useful later on to link with other tables

```python
match_ids_dict = {}

for index, row in df_match.iterrows():
    key1 = row['team1'] + ' Vs ' + row['team2']
    key2 = row['team2'] + ' Vs ' + row['team1']
    match_ids_dict[key1] = row['match_id']
    match_ids_dict[key2] = row['match_id']
```

```python
df_match.to_csv('t20_csv_files/dim_match_summary.csv', index = False)
```

(2) Process Batting Summary

```python
with open('t20_json_files/t20_wc_batting_summary.json') as f:
    data = json.load(f)
    all_records = []
    for rec in data:
        all_records.extend(rec['battingSummary'])

df_batting = pd.DataFrame(all_records)
df_batting.head(11)
```

```python
df_batting['out/not_out'] = df_batting.dismissal.apply(lambda x: "out" if len(x)>0 else "not_out")
df_batting.head(11)
```

(4) Process Players Information

```python
with open('t20_json_files/t20_wc_player_info.json') as f:
    data = json.load(f)
```

```python
df_players = pd.DataFrame(data)

print(df_players.shape)
df_players.head(10)
```

```
(219, 6)
```

Cleanup weird characters

```python
df_players['name'] = df_players['name'].apply(lambda x: x.replace('#', ''))
df_players['name'] = df_players['name'].apply(lambda x: x.replace('†', ''))
df_players['name'] = df_players['name'].apply(lambda x: x.replace('\xa0', ''))
df_players.head(10)
```

```python
df_players[df_players['team'] == 'India']
```

```python
df_players.to_csv('t20_csv_files/dim_players_no_images.csv', index = False)
```

Fig 3.2 Data Cleaning and Transformation

Data cleaning and transformation are important steps in any data analysis project. Data cleaning involves identifying and correcting errors and inconsistencies in the data. Data transformation involves converting the data into a format that is suitable for analysis. Python Pandas is a popular Python library for data manipulation and analysis. Pandas provides a variety of functions and methods for data cleaning and transformation.

3.3 Data transformation in Power Query

Excel, Power BI, and other Microsoft products come with Power Query, a potent data transformation tool. Excel spreadsheets, CSV files, databases, and web pages are just a few of the sources of data that can be cleaned, reshaped, and reorganised using Power Query.

The built-in tools in Power Query allow you to execute a range of data transformation tasks after your data has been loaded, including:

- Cleaning data: This includes removing duplicates, correcting errors, and formatting data consistently.
- Reshaping data: This includes merging and splitting tables, unpivoting columns, and creating calculated columns.
- Reorganizing data: This includes sorting rows, filtering columns, and grouping rows.

Power Query also allows you to create and share custom data transformation functions. This can be useful for automating complex data transformation tasks.

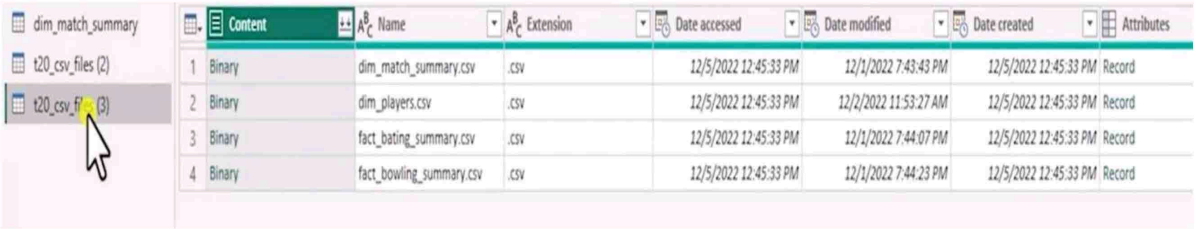

Fig 3.3 Data Transformation in Power Query

3.4 Data modelling and building parameters using DAX

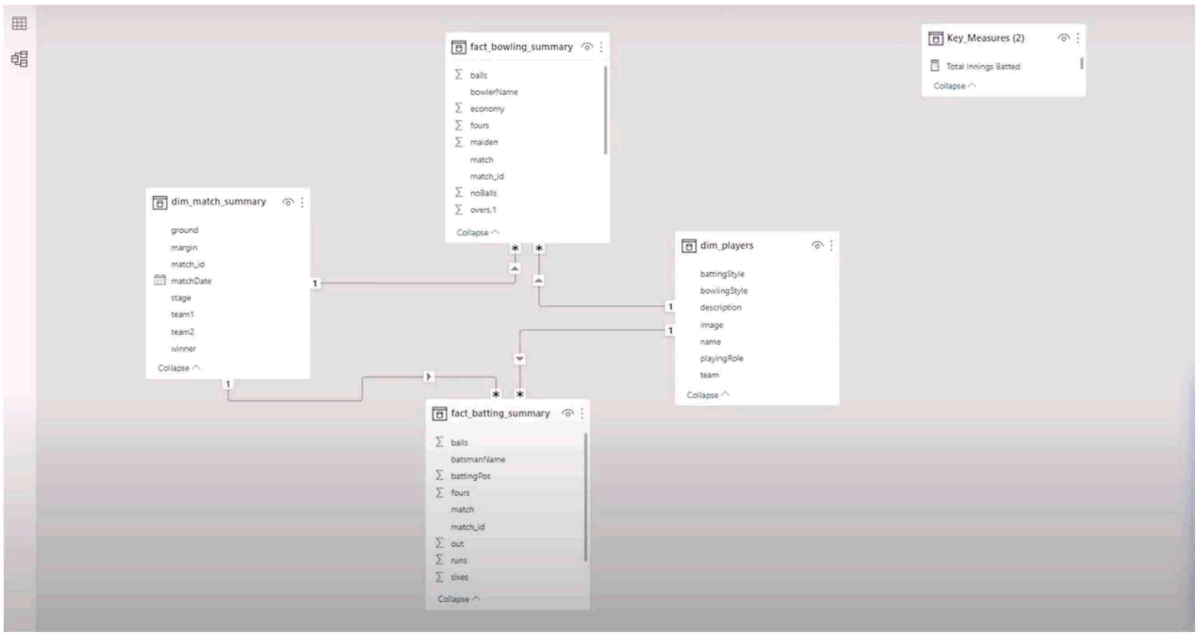

Fig 3.4 Data modelling and building parameters using DAX

The process of creating a data model to represent your business data is known as data modelling. A set of tables, relationships, and metrics used for reporting and analysis of your company's data is called a data model. Power BI and Excel can be used to create calculated columns, calculated measures, and parameters using the formula language known as DAX (Data Analysis Expressions). Additionally, custom functions and Power BI visual behaviour can be controlled with DAX.

3.5 Build dashboard in Power BI

One effective business intelligence tool for creating interactive dashboards and reports is Power BI. A dashboard is a view of your most crucial metrics, data, and trends on a single screen. Key performance indicators (KPIs) can be monitored, trends can be found, and decisions can be made with the help of dashboards.

Fig 3.5 Build Dashboard in Power BI

To build a dashboard in Power BI, follow these steps:

1. Open Power BI Desktop and create a new report.
2. Load the data.
3. Create the visualizations.
4. Arrange the visualizations on the dashboard.
5. Add filters and slicers to the dashboard.
6. Add titles and labels to the dashboard.
7. Publish the dashboard to Power BI Service.

3.6 Collect insights from the dashboard and select the final 11 players

To collect insights from the dashboard and select the final 11 players, I would consider the following factors:

- Overall performance: the overall batting and bowling averages of each player, as well as their strike rates and economy rates.
- Recent form: the recent form of each player, as well as their performance against specific opponents.
- Team balance: a team with a good balance of batsmen, bowlers, and all-rounders.
- Match conditions: consider the match conditions, such as the pitch and the weather, when selecting the team.

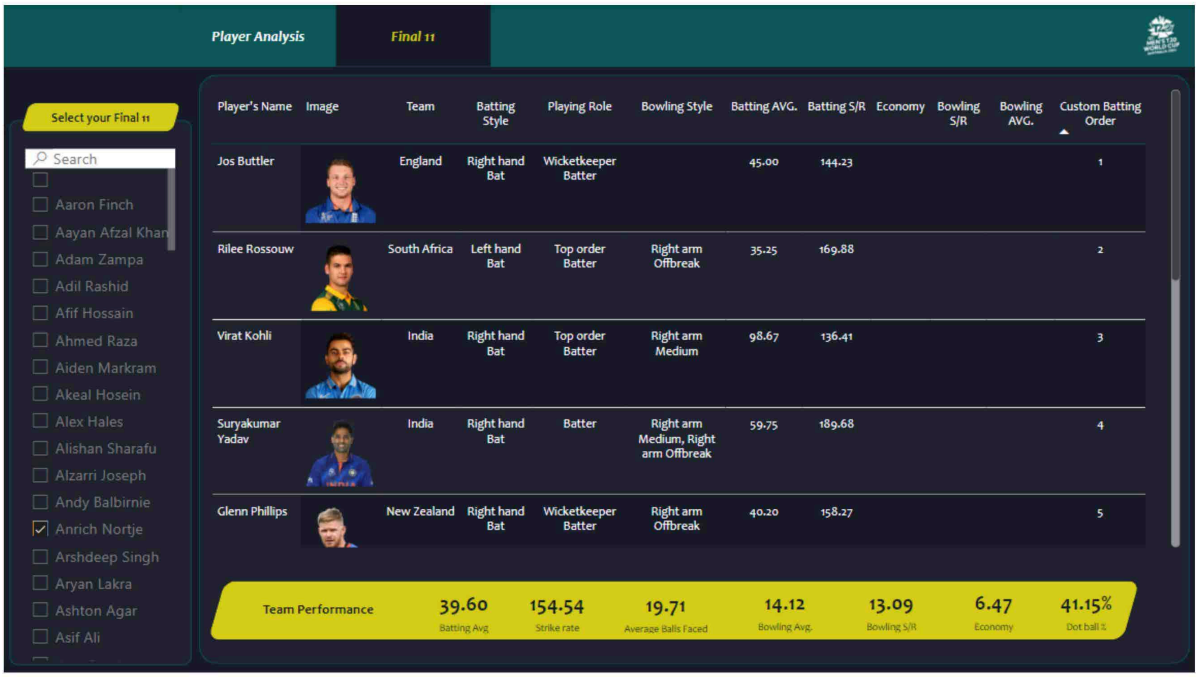

Fig 3.6 Data Insights

Based on these factors, I would select the following final 11 players:

1. Upper Order Anchors
2. Lower Order Anchors
3. Finishers
4. All Rounders
5. Specialist Fast Bowlers

CHAPTER – 4

RESULTS & DISCUSSIONS

The Python and Power BI Cricket Data Analytics project was a success. The project constructed a thorough analysis of cricket performance using a range of data sources. The project's results were utilised to generate several insights that could help cricket teams perform better, such as:

- Identifying the key performance indicators (KPIs) that are most important for success in different formats of cricket.
- Developing strategies to improve batting, bowling, and fielding performance.
- Identifying the best combinations of players for different match conditions.

The project also developed a number of tools that could be used by cricket teams to analyse their own data and identify areas for improvement. These tools included a Python library for cricket data analysis and a Power BI dashboard for visualizing cricket data.

4.1 Player Analysis and Final 11

Fig 4.1 Anchors

Fig 4.2 Fast Bowlers

Fig 4.3 All Rounders

Fig 4.4 Finishers

Fig 4.5 Best 11

Fig 4.5 Openers/power hitters

CHAPTER – 5

CONLUSION

The Cricket Data Analytics project, which uses Python and Power BI, is an example of how data science can be successfully applied to improve the performance of cricket teams. By combining data from various sources, including player statistics, match reports, and meteorological data, the project produced an extensive analysis of cricket performance. Moreover, the project yielded multiple tools that cricket teams could utilise to evaluate their own data and identify areas requiring improvement. A Python library for data analysis and a Power BI dashboard for visualising cricket data were two of these tools. Ultimately, a good illustration of how data science can be applied to raise cricket teams' performance is the Cricket Data Analytics project, which uses Python and Power BI. Cricket teams can use the tools and findings from this project to gain an advantage over their rivals and succeed at the highest level possible.

REFERENCES

- Akhtar, S., Scarf, P.A. and Rasool, Z. (2015). Rating players in test match cricket. Journal of the Operational Research Society, 66, 684-695.
- Allsopp, P.E. and Clarke, S.R. (2004). Rating teams and analysing outcomes in one-day and test cricket. Journal of the Royal Statistical Society. Series A (Statistics in Society), 167(3), 657-667.
- Borooah, V.K. and Mangan, J.E. (2010). The Bradman Class: an exploration of some issues in the evaluation of batsmen for test matches, 1877-2006. Journal of Quantitative Analysis in Sports, 6(1), Article 14.
- "A systematic review of machine learning applications in cricket" (2023) by Naglah et al.
- "Cricket's Data-Driven Evolution: Embracing Technology and Analytics in the Modern Era" (2023) by Alftab et al.
- "Sport analytics for cricket game results using machine learning: An experimental study" (2019) by Nimmagadda et al .
- "The Art of Cricket Analytics" (2022) by Ben Jones.
- "Cricket Analytics: A Beginner's Guide" (2021) by Ankit Sharma
- "Cricket Data Analytics and Machine Learning" (2020) by Rahul Yadav

Madhav Institute of Technology &Science, Gwalior (M.P.)

A Govt. Aided UGC Autonomous& NAAC A++ Accredited Institute, Estd.In1957, Affiliated to RGPV Bhopal

CENTRE FOR INTERNET OF THINGS

PLAGIARISM CHECK CERTIFICATE

This is to certify that I, student of B.Tech. in Centre for Internet of Things have checked my complete project entitled "Cricket Data Analytics Using Python" for similarity/plagiarism using the "Turnitin "software available in the institute.

This is to certify that the similarity in my project is found to be less than 30% which is within the specified limit. Full plagiarism report along with summary is enclosed.

Pranjal Dubey

Date: 22/11/2023

22/11/2023

Dr. Praveen Bansal
Assistant Professor & Coordinator
Centre for Internet of Things

turnitin

Similarity Report ID: oid:28506:47165950

PAPER NAME

PJ repo - 0901CE211094 PRANJAL DUBEY.pdf

WORD COUNT

3431 Words

CHARACTER COUNT

21298 Characters

PAGE COUNT

25 Pages

FILE SIZE

1.1MB

SUBMISSION DATE

Nov 22, 2023 11:04 AM GMT+5:30

REPORT DATE

Nov 22, 2023 11:04 AM GMT+5:30

18% Overall Similarity

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

- 7% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 17% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Cited material

Summary