Breaking Boundaries: Uncovering Patterns and Trends in Cricket through Advanced Data Analytics

Kulaye Shreyal Ashok
Database & Analytics Programming
x22155791
x22155791@student.ncirl.ie
National College of Ireland

Sonal Puradkar
Database & Analytics Programming
x22130250
x22130250@student.ncirl.ie
National College of Ireland

Nandheeswari Rajendran
Database & Analytics Programming
x22132210
x22132210@student.ncirl.ie
National College of Ireland

Anish Romario
Database & Analytics Programming
x20190841
x20190841@student.ncirl.ie
National College of Ireland

***Abstract*— Millions of people are passionate about cricket, with ODI, T20, and Test formats being the most popular. There is diverse information available about cricket matches, and by analyzing the dataset and presenting relevant data that aids in future predictions. Numerous factors determine the results of a cricket tournaments, such as the team's home ground advantage, their past performance at that venue, their overall experience, their record against a specific team, and the present form of the team and individual players. This project involves working with four cricket datasets, which include statistics from matches, rankings of teams, and player performance data. The datasets will be programmatically stored in a database appropriate for managing large amounts of data. The datasets may need to be transformed and pre-processed to ensure compatibility with the database. Next, the data will be analyzed, pre-processed, and visualized, which could include cleaning, transforming the data into an appropriate format, identifying patterns and trends, performing machine learning algorithms, and generating visualizations. Finally, the processed data will be programmatically stored in the appropriate database. This step ensures that the data is accessible for further analysis or visualization and can be used to create reports and dashboards that offer insights into various aspects of cricket. This project enables stakeholders like coaches, players, enthusiasts, and administrators to gain valuable insights into cricket.**

Keywords*—*Cricket data analysis, Data Analytics, Web Scraping, Data Visualization, Catboost, Cricket analytics, ODI analytics, IPL match analytics, Extract Transform Load, Exploratory data analysis, Database Analytics and programming, Data Mining, Machine Learning

I. Introduction

The gentleman's game, also known as cricket, is a very traditional, popular, and simple pastime. The sport of cricket first appeared in the southern eastside of England in the late sixteenth century. In the eighteenth century, it was made the nation's sport matches, and in the 19th and 20th centuries, it expanded internationally.The ICC is the governing body for cricket. One-Day International (ODI) cricket, which features matches played in a 50-over format, is the premier event on the international sport calendar of cricket and is held every four years. One of the most viewed sporting matches in the world, it is the largest sporting tournament. The widely popular cricket sport league in the world is the Indian Premier League (IPL), a one-day tournament played in India with teams played in a twenty(T-20) over format. For a considerable amount of time, cricket is played in a format called a test match. The test match is a two-inning, five-day competition between two teams. A newer format emerged as a result of the audience and television viewers becoming bored by the lengthy duration. The time was cut to one inning with a set number of overs for each team in the more recent format. Shorter cricket matches were played in this format, which was commercially successful.

One-Day International (ODI) cricket, which is played using a 50-over format, is the premier competition on the international cricket calendar and is held every four years by the International Cricket Council (ICC). One of the most watched sporting events in the world, it is the largest cricket tournament[4][5].

The most popular cricket league in the world is the Indian Premier League (IPL), a one-day tournament played in India with matches played in a 20-over format. IPL is where opportunity and talent converge. The cricket squad consists of 11 (eleven) players, including bowlers, all-rounders, and batters. To

increase the chance of success, the team should be diverse and balanced. Additionally, the kind of pitch, who wins the toss, and the order of the batters or bowlers can all affect the outcome. In addition, a single match's outcome is largely determined by the performance of the batters, bowlers, and fielders. The performance of such a factor is now being researched utilizing various statistical (probabilistic/ stochastic) approaches[14].The major goal of this study is to comprehend the factors that contribute to cricket teams' performance in various game forms.

We will look into how batting and bowling performances affect match results and pinpoint the most important elements that go into winning games. Additionally, we will examine player performance individually to find trends and patterns in how they perform across various game forms. We will employ statistical and machine learning methods, including data visualization and cluster analysis, to do this. We intend to get new understandings of the game that can help with player selection, team strategy, and game tactics by applying these methods to the cricket dataset. The purpose of this project is to advance cricket analytics research and assist the game grow to new heights. A variety of people and organizations, including players, coaches, teams, fans, pundits, media outlets, broadcasters, and the gaming and wagering sectors, can gain from cricket dataset analysis. Both players and coaches can identify their weaknesses and employ data-based techniques to improve their performance by carefully examining data related to elements such as batting averages, bowling velocity, and fielding statistics. Teams can utilize dataset analysis to identify patterns and trends in their own and their opponents' performance. By identifying the strengths and weaknesses of their opponents' players, teams can then modify their gameplay to create winning strategies.

II. Problem Statement

A. Statement of a problem

To analyze the cricket data and get the visualizations from the data using analytics and visualization techniques and then retrieve the meaningful insights from it.

B. Existing Systems & Related work

With the introduction of technologies like the snick-o-meter, hawk-eye, and others in the 2000s, the game of cricket underwent a substantial paradigm shift. A significant amount of research has also been done on decision-making techniques for selecting cricket teams, forecasting game results, ranking and evaluating players based on their performance on matchday, and other topics. [1][2][4] Additionally, considerable ground-breaking work has been done on using simulators to simulate cricket games involving various teams and venues, using PCA for prediction, employing programming techniques for fantasy league prediction, and more. [3][9][8][9] When it comes to rating, ranking, and forecasting cricket players and matches, statistical techniques such creating content-based suggestions, personalized recommendations, factor analysis, and integer optimization have also produced some positive outcomes.[11][11][12]The batting order is a crucial and dynamic aspect of cricket, and it is crucial to have batsmen who can play specific roles, such as anchor, pinch hitter, finisher, etc.[6][7] The three positions at the top, middle, and bottom of the batting order each have specific responsibilities that must be met in order to win the game, whether one is batting first or pursuing[15][16][17]. In order to determine batting orders based on a wide range of analytical parameters, research has been done on the aforementioned components as well. IoT cricket bat sensors like the STR8T, Stance-beam, and Bat-sense have recently achieved advancements in their implementation for collecting crucial bat data.[8][18] They assist the batsmen in better understanding the game by gathering several data points relevant to the bat swing [19][20][22][23]. The discipline of cricket analytics has also expanded, allowing for the analysis of data from numerous games and competitions using Python and other languages to produce vital information on player performance[24]. Even though there has been a significant advancement in how cricket is played, viewed, examined, and understood, the field is still extremely young. The analysis of cricket statistical data is a particularly good application for AI and machine learning.

III. Literature Review

The study "Predicting Optimal Cricket Team using Data Analysis" proposes a fresh method for selecting the best cricket team. Based on player performance and suitability for various roles on the team, the study uses statistical analysis and machine learning algorithms to forecast the ideal team composition[21].

The following are some of this research's benefits:

Effective team selection: The study offers a data-driven method for choosing teams, which may

result in a better team makeup and greater on-field performance.

Cost-effective: Using data analytic methods to forecast the ideal cricket squad can be a more affordable option than using conventional strategies that rely on expert judgment and intuition.

Automated squad selection: It can save coaches and selectors time, allowing them to concentrate on other parts of the game.

Enhanced performance: By finding the greatest players for each function, maximizing team strengths, and minimizing team weaknesses, data analysis used to choose the best team can enhance team performance.

The approaches utilized in this study are generalizable to other sports and scenarios involving team selection, offering a general strategy for choosing a team.

Overall, this research work offers a promising strategy for choosing a cricket team that may enhance team performance while cutting down on the time and expense of using conventional selection techniques[21].

The study "A Review of Data Analytic Schemes for Prediction of Vivid Aspects in International Cricket Matches" by V. S. Raju, N. Sethi, and R. Rajender examine numerous data analysis methodologies in order to forecast interesting elements in international cricket matches. These factors include the outcome of the game, each player's performance, and the impact of other factors like the weather, the field of play, and the makeup of the team[22].

The authors look at several data analysis techniques currently in use and highlight their benefits and drawbacks.

The article discusses the application of data analytics to sports, specifically cricket, and focuses on the advantages of employing these methods for forecasting and making decisions. The complete text of the work can be accessed online using the DOI (digital object identifier) given in the citation[22].

According to the reference given, the benefits of this study might be as follows:

Enhanced comprehension: The study offers a thorough analysis of data analytic techniques for forecasting various features of international cricket matches. This will enable us to better understand how data analytics may be used in the sports industry and point up areas for improvement.

Enhanced prediction accuracy: The research has the potential to increase the accuracy of forecasts for many aspects of cricket matches by suggesting a new data analytic framework that incorporates multiple data sources and applies machine learning techniques.

Practical applications: By assisting coaches, players, and analysts in making better educated choices based on data-driven insights, the research could have practical applications in the world of sports, notably cricket.

Contribution to academic literature: The research adds to the existing academic literature on the application of data analytics in sports and provides insights into the strengths and weaknesses of various data analytic schemes. Overall, the research could have a positive impact on the field of sports by providing a better understanding of how data analytics can be used to improve decision-making and performance in cricket matches. Practical applications: The research could have practical applications in the field of sports, particularly cricket, by helping coaches, players, and analysts make more informed decisions based on data-driven insights.[22]

At the 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC) in Chennai, India, a presentation titled "Performance of Indian Cricket Team in Test Cricket: A comprehensive Data Science analysis" was given[23].

The study uses data science techniques to provide a thorough analysis of the performance of the Indian cricket team in Test matches. The performance of the Indian cricket team over time has been examined by authors V.V. Tharoor and N.M. Dhanya using a variety of statistical techniques.

The data sources and statistical methods used to examine the data are presented in the study. To evaluate the performance of the Indian cricket team, the writers utilized a variety of statistical techniques, including mean, standard deviation, correlation, regression, and hypothesis testing. The study spans a 20-year period from 2001 to 2021.

The study highlights a number of intriguing conclusions, including the overall Test cricket performance of the Indian cricket team, the performance of individual players, and the effects of several variables, including pitch conditions, the opponent team, and home/away matches on the team's performance.

The authors believe that their analysis offers insightful information about how the Indian cricket team performs in Test matches and can be helpful to coaches, players, and fans.

The authors draw the conclusion that their analysis offers insightful information on the performance of the Indian cricket team in Test matches and can help coaches, players, and spectators better understand the team's strengths and flaws. They also recommend conducting additional studies to examine other cricket teams' performances using comparable data science methodologies.

According to the data in the study, the benefits of performing an extensive data science analysis of a cricket team's performance in Test cricket are as follows:

1. Provide a thorough and in-depth study of the team's performance over a period of time.

2. Aids in determining the team's assets and liabilities on both an individual player and team level.

3. Allows for the analysis-based decision-making of coaches and athletes.

4. Reveals information about how numerous elements, such the playing surface, the opposition, and home/away matches, affect the team's performance.

5. Can support the squad in creating new tactics and ideas to enhance their performance in Test cricket.

6. Can be used to compare the team's performance to that of other teams across the world and pinpoint areas that need improvement.

7. Enables supporters to appreciate the team's strengths and flaws and obtain a deeper grasp of the team's performance.

In conclusion, a thorough data science analysis of a cricket team's performance in Test cricket can offer insightful information that can be used to increase the team's performance and overall sporting success[23].

The research article "Data Mining System for Predicting a Winning Cricket Team" describes a data mining system that uses machine learning techniques to forecast cricket match results. The authors train their model using information from previous cricket matches before using it to forecast the results of upcoming games[24].

Data is first gathered by the writers from a variety of sources, including databases of cricket matches and websites with cricket statistics. They then clean up the data, getting rid of any flaws or inconsistencies. The information is then changed and converted into a format that is appropriate for analysis.

The authors then develop predictive models using a variety of machine learning techniques, including decision trees, random forests, and support vector machines.

The availability and caliber of the data as well as the intricacy of the prediction models are just a few of the system's drawbacks and difficulties that the writers mention. They argue that by combining more sophisticated machine learning algorithms and gathering more complete and reliable data, their system may be further enhanced[24].

To determine the winning team, the algorithm examines a number of variables, including individual statistics, team dynamics, and environmental circumstances. This system has a few benefits, including:

Increased accuracy: Using machine learning techniques, the system can analyze enormous amounts of data and spot patterns that are difficult for people to recognize. This makes forecasts of the winning team more accurate.

Enhanced effectiveness: The system's ability to handle and analyze massive amounts of data fast enables quick decision-making and more effective resource use.

Consistency: Rather than relying on subjective human judgment, which is susceptible to error, the system bases its predictions on data analysis. As a result, predictions become more consistent and trustworthy. [24]

Objectivity: The system's predictions are grounded in data and unaffected by emotions or other subjective elements like personal biases.

Because of its scalability, which allows it to be used in both small and large-scale applications, the system can be scaled up or down to manage changing volumes of data.

Overall, compared to conventional ways of forecasting the outcome of a cricket match, the data mining approach provided in the research has a number of advantages. [25]

IV. Proposed Model

The implementation of analysis, pre-processing and visualization are done on 4 different datasets that are related to the types of cricket matches such as ODI, Test matches and IPL (Indian Premier League). The variety of these 4 datasets are distinguished in such a way that 1 dataset is in an unstructured format, 2 datasets are semi-structured in nature and the last dataset is structured. The end-to-end data analytics implementation on these datasets is performed in Python programming language using various python libraries for ETL and processed data Visualization such as pandas, NumPy, matplotlib, sweetviz and many more in Google Colab. We have converted the unstructured dataset to structured format by converting it into CSV and stored this CSV file into MongoDB. In the case of a structured dataset, which is already in a CSV format, we have directly stored it into the MongoDB. MongoDB has been implemented On-Cloud using a Python utility called ‘PyMongo’. Later, fetching the data stored in MongoDB, performing data cleaning, pre-processing and ETL

methodologies and finally creating an interesting visualization pattern from the processed data. Along with the visualization, we have saved the processed data into PostgreSQL for future reference using On-Cloud platforms. The goal of our project is to fetch the data related to different cricket matches and process this data to the extent that it can be used as useful information by the experts in the field of Cricket for strategic decision making.

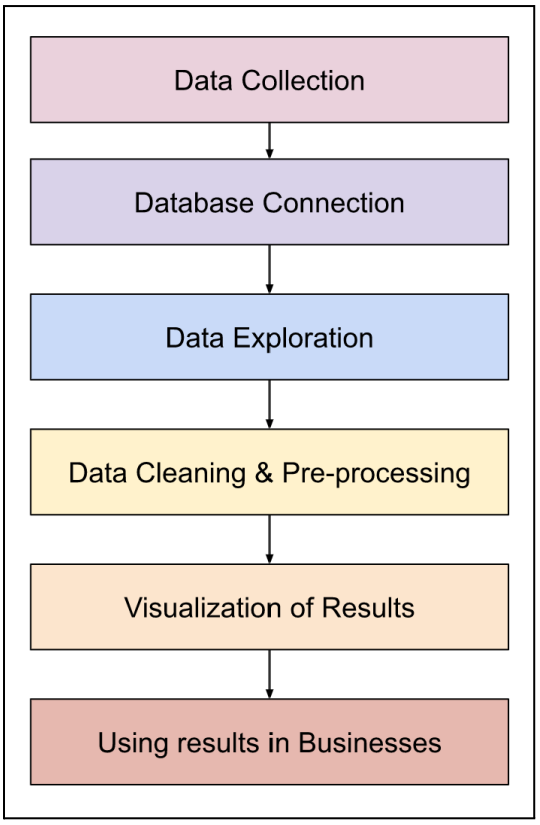


Fig 1. Flow of the data analytics model

There are various steps involved in creating a data analytics model, including:

Describe the issue: The issue or question you're trying to solve with data analytics should be stated clearly. Understanding the business issue, the data at hand, and the project's objectives are necessary for this.

Gather and Prepare Data: Gather and get the data ready for analysis. This entails preparing the data for easy processing by analytical tools by cleaning, converting, and arranging it.

Utilizing statistical analysis and visualization techniques, explore the data to find patterns, trends, and relationships. This process aids in gaining understanding and locating potential predictors that could be incorporated into the model.

Part A

This dataset is implemented using Web Scraping and it is unstructured in nature. Web Scraping is a methodology in which we can fetch data from a website. This technique lets us handle the live data from any specific web domain.

In this dataset, we have collected ODI real-time data from the ESPN Cricket Info website. To perform the web scraping, we are using a python library such as “Beautiful Soup”. Under Beautiful Soup, the library ‘requests’ is used to send a request to the website whose response is given to the constructor of “Beautiful Soup” that parses the website data. This parsed data is stored in ‘soup’ from which we fetch data for further use. One additional activity we have performed while web scraping is that, we have created a function i.e., ‘cricScrap’ which only fetches specific columns relevant to our analysis. For instance, ESPN website has a variety of data related to cricket, but here we are fetching only 'team', 'result' ,'margin','br','toss','bat','opposition','ground','date','url','scrapped_date' columns of ODI matches that are important to our problem statement.

After fetching the data from the website, we are converting it into CSV and performing further cleaning before storing this CSV data into MongoDB. In this cleaning process, we are checking if we have any columns with null values, if so, then we are dropping those columns with ‘df.drop’ function. In this dataset of ODI matches, we are dropping ‘margin’ and ‘br’ columns. This cleaned data we are then storing into MongoDB by using the “pymongo” python library. Once the data is stored into MongoDB, we fetch it to do further analysis, pre-processing and ETL transformations.

The pre-processing is performed on three columns namely, ‘toss’, ’result’ and ‘bat’. By using the “map” function, we are transforming the data in these three columns to a more meaningful and integrated format which will be even more readable and helpful during visualization. The Map function used here as part of this transformation is allowing us to replace the column data with some other value specified in the condition. Example: df['toss'] = df['toss'].map({'lost':0, 'won':1,}). This is an example where we are replacing a value in the ‘toss’ column with ‘0’ if the data says ‘lost’ and with ‘1’ if the data says ‘won’.

After performing the pre-processing, we are implementing visualization on the processed data. by using python libraries such as ‘matplotlib’, ‘seaborn’

etc. In this dataset, we have data related to teams in ODI matches who won/lost the batting first, who won/lost the toss, the opposite teams every match, the grounds where all the matches were played, result of the match etc. Using this information, we are deriving Match won/lost records of every team as shown below.

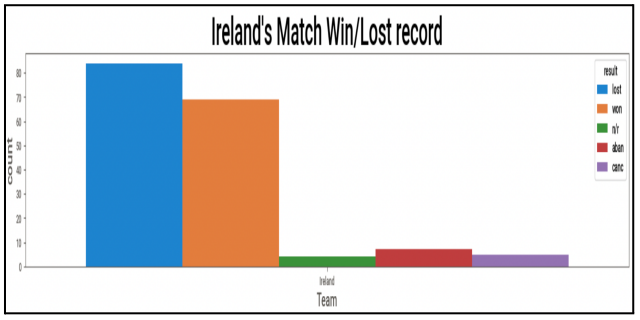


Fig 2. Ireland's Match Win/Lost record

Fig 2. shows matches won/lost by team Ireland. The x-axis shows the name of the team and the y-axis shows the count. The blue bar is the count of matches lost and the orange bar is the count of matches won. Similarly, we can execute visualization for other teams participating in the ODI matches.

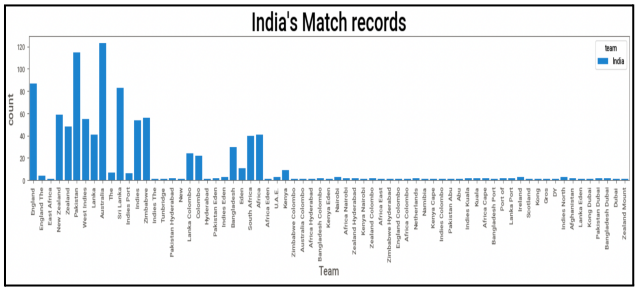


Fig 3. Match records of Team India against each team.

In Fig 3. It shows matches played by team India against each team in ODI. From the graph we can see that Team India played the highest matches against Team Australia.

```
ground_number("Ireland")

opposition    ground
Abu           Dhabi          5
Afghanistan   Belfast        9
              Dehradun       4
              Dhabi          1
              Dubai          1
                            ..
West Indies   Nelson         1
Zimbabwe      Belfast        2
              Bready         1
              Harare         7
              Hobart         1
Name: ground, Length: 90, dtype: int64

[ ]  ground_number("India")

opposition    ground
Abu           Dhabi          1
Afghanistan   Dhaka          1
              Southampton    1
Australia     Adelaide       6
              Ahmedabad      3
                            ..
Zimbabwe      Pune           1
              Rajkot         1
              Sharjah        5
              Singapore      1
              Vadodara       1
Name: ground, Length: 385, dtype: int64
```

Fig 4. Ireland and India's opposition teams, stadium and count of matches.

Fig43. shows aggregated data for Ireland and India. It shows the opposition team of India and Ireland, the ground where the match was played and the count of the matches per team.

```
Training Accuracy
              precision    recall  f1-score   support

         0.0       0.67      0.53      0.59      3151
         1.0       0.64      0.75      0.69      3406

    accuracy                           0.65      6557
   macro avg       0.65      0.64      0.64      6557
weighted avg       0.65      0.65      0.64      6557

Testing Accuracy
              precision    recall  f1-score   support

         0.0       0.87      0.90      0.88       774
         1.0       0.91      0.88      0.89       866

    accuracy                           0.89      1640
   macro avg       0.89      0.89      0.89      1640
weighted avg       0.89      0.89      0.89      1640
```

Fig 5. Implementing Catboost Classification model

In Fig 5. We are implementing a classification algorithm called "CatBoost" that displays the accuracy of our classification model. Popular open-source gradient boosting software called CatBoost is made to deal with a variety of data sources, including categorical characteristics. It combines a variety of cutting-edge approaches to get around issues with overfitting, data sparsity, and handling high-dimensional data. The accuracy of our model is 0.87 percent.

Along with the visualization, we are storing this processed data in PostgreSQL using On-cloud platform.

Part B

In the Second Dataset Visualisation, we show the most number of matches played in a particular region. We have used the 'cyan' color and edge as 'black' to portray the data in the form of Bar. We have collected the data from multiple countries such as India, Oman, Sri Lanka, Australia, Pakistan, South Africa, West Indies, New

Zealand, England, Bangladesh, Zimbabwe, Ireland, Kenya, Afghanistan, Scotland, Canada, Netherlands, UAE, Hong Kong, Bermuda, Papua New Guinea,United States of America, Namibia and Oman. It appears that India tops the list while Oman being the least region for any matches to be played as India crosses the scale beyond 250. This has been demonstrated over a Bar container shown below.

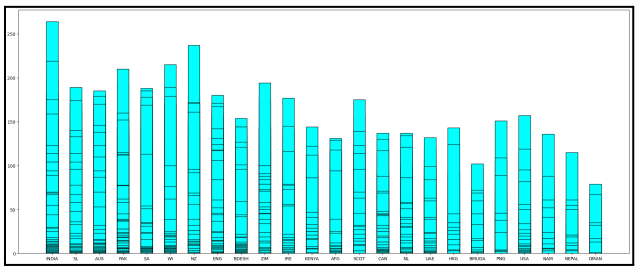


Fig 6. Bar Container showing number of matches played in every region

There's more to the platter, We have also used seaborn.pairplot() library which is built on Matplot library. It is used to plot pairwise relationships in a dataset. In our dataset it is used to portray the number of matches, runs, average score, high score, sr played in the final region shown as below.

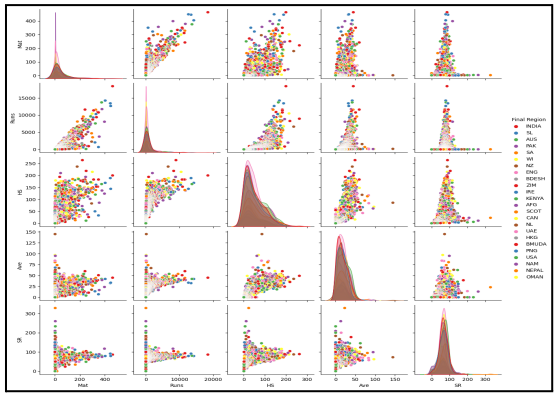


Fig 7. Pair Grid representing number of matches,runs,HS,SR

The third representation we have used for visualization, is seaborn.catplot() where it draws categorical plots onto a FaceGrid.In our case it shows the number of matches played in the Final Region against the High Score as shown below. We have performed Exploratory Data Analysis too.

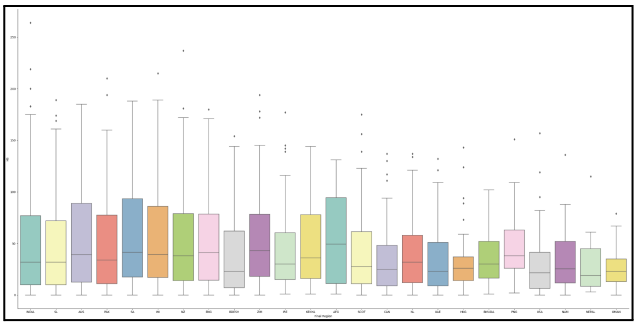


Fig 8. Facet Grid showing number of matches played in every region

Another last visualization we have used is seaborn.distplot() where it shows the density of matches played in the final region against the high score. In the figure below it also portrays the number of countries taking part on the right hand side with India being the highest and Oman at the lowest.

Another beautiful representation in this dataset is animation and the library being used is plotly.express. Plotly Express is a part of the plotly library mainly used to portray known figures over an applicable page. So here, we have used a couple of animations where we have pulled the record of the number of players played in the ODI. The interesting part about this feature is that we portray a good amount of players who have participated and played the most number of matches in the ODI . Here, we have 3 figures and in the first figure we have an End button where on click on that we show a good amount of players with their names, runs , highest score and importantly the year and region they have performed and all these can be viewed when you hover against their name. Now the same pattern applies to the other 2 figures but with the replacement of a Start button .

Note :- This has been represented with a period of 45 years from 1973 to 2018 in the first figure and from 1989 to 2019 in the other 2 figures. Both the figures have been shown below.

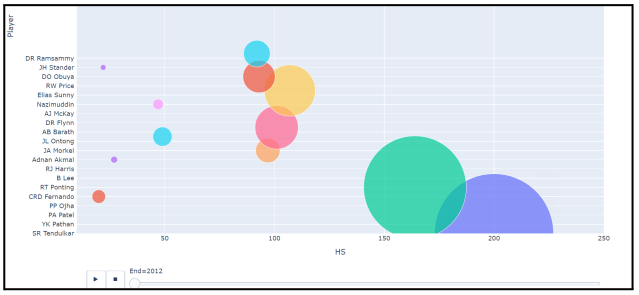


Fig 9. Plotly Express showing their name, HS, End Year and the final region last played

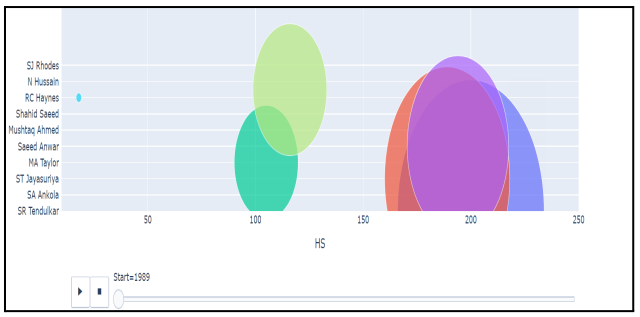


Fig 10. Plotly Express showing their name, HS, Start Year and the final region played

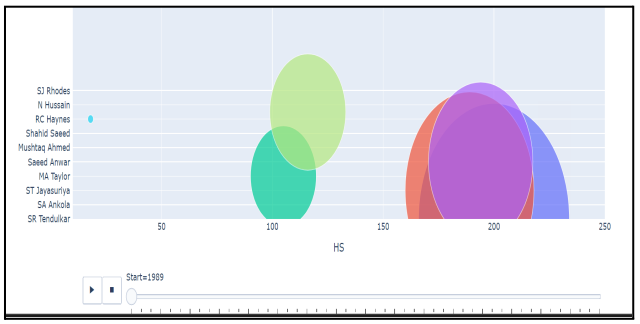


Fig 11. Plotly Express showing their name, HS, Start Year and the final region played

Another fascinating and latest library we have implemented here is pandas.profiling It is basically used to ease the Exploratory Data Analysis process and interpret all the data in readable formats where a person who's not aware of any programming knowledge will find it feasible to understand.

Here we give an understanding of the variables, number of players, span, matches,innings,number of runs,final region,hs,sr,start and end.

Here we have displayed an Overview which tells us about the Dataset Statistics and besides, we show a pictorial format of the variables being used for evaluation and derive deep insights, the types of variables being used here are Categorical , Numeric and Unsupported.

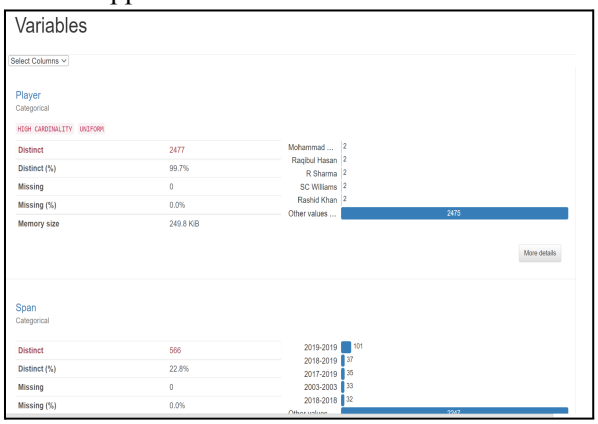
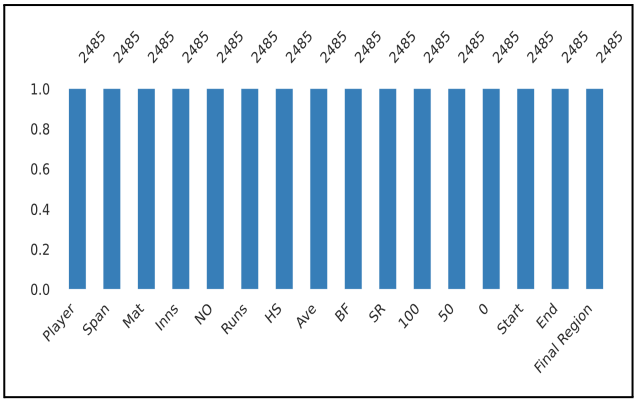


Fig 12. Dataset Statistics showing the number of variables used to evaluation and the number of players and the duration they have performed



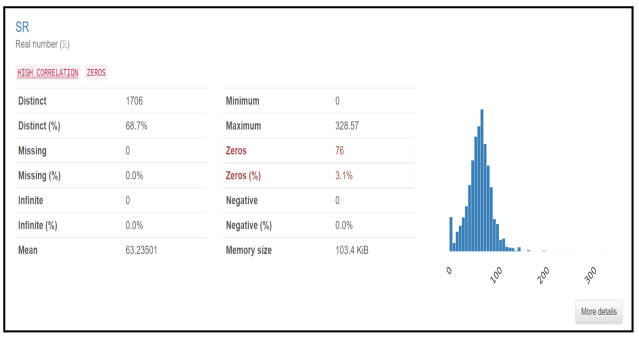


Fig 8. SR showing the mean, missing, distinct

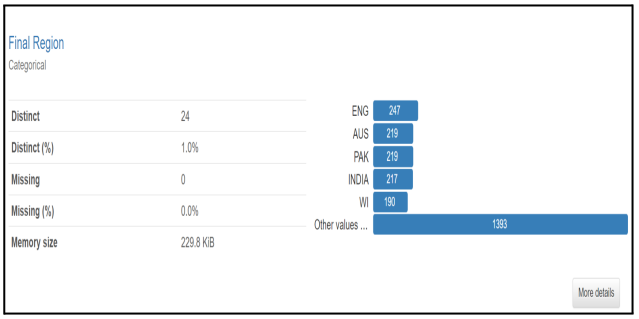


Fig 13. Final Region containing categorical values about every region

Part C

In the Third Dataset, We will be focusing predominantly on India vs Australia and we will be showing certain distinctive features between these countries such as the run rate, innings scored and chasing. Here we are using a couple of visualizations such as histogram, boxplot, stripplot to illustrate the data between India vs Australia and lastly, we will be talking about which country performed the best on an overall perspective and this has been portrayed over a pie chart.

We will be discussing about the runs scored between India and Australia when they chose to bat 1st from the below diagram

It has been observed that India scored around 109 and close to 129 whilst Australia scored around 130 to 139 only once. There has been the same score which was

180 to 189 and 190 to 199 between the two countries. These were the number of times where Australia scored only once but India had already touched 219 , 319, 349 and 389. There has been no similarity between India and Australia when they tried to score the same number of runs twice but Australia on an overall picture scored 209, 229,309,329 and 349. Now again there has been an alike in the score 240 to 249 which stands at 3 times. Again there has been no similarity at the count of 4 between these vibrant countries. Hence going forward India seems to have scored a certain number of runs which appears to be only 6 times but Australia on the other hand seems to be performing well as they have scored a good number of runs beyond 6 and touching close to 8 times.

The respective figures have been illustrated below.

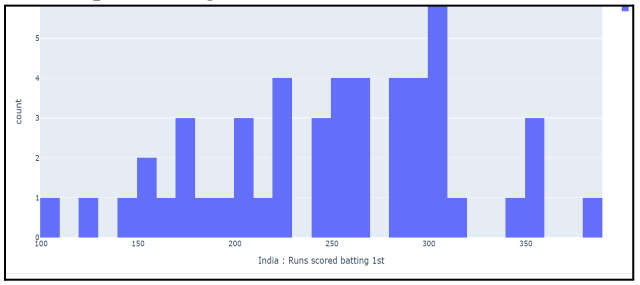


Fig 14. Histogram Chart displaying the number of runs scored by Indians while batting first.

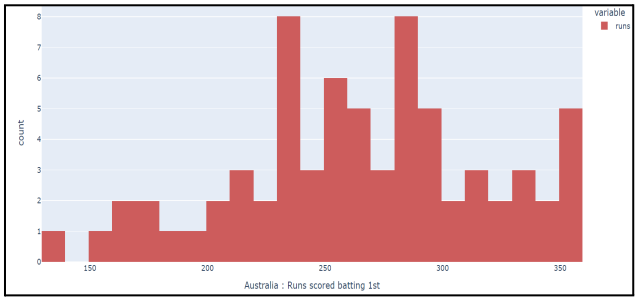


Fig 15. Histogram Chart displaying the number of runs scored by Australians while batting first.

A similar comparison will be shown now and that is the batting between India and Australia when the players from both these countries chose to bat the second time together.

There has been a constant fluctuation in the number of runs between India and Australia. Like the previous example we are representing the data for India in blue color and indian red color for Australia. But in both the figures we can observe a score of 350 is at its lowest standing at an instance of 1. Although 350 being at its least, there is a similar score of 210 to 219 scoring twice and 250 to 259 scoring 4 times. It can be clearly seen that Australia has not scored any runs beyond 5 times but India on the other hand has achieved more than Australia by scoring above 8 and reaching 10.

The pertaining figures have been illustrated over a histogram.

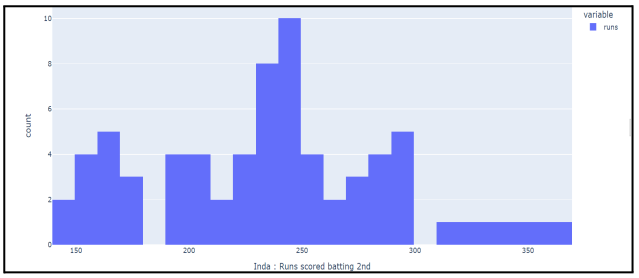


Fig 16. Histogram Chart displaying the number of runs scored by Indians while batting for the second time.

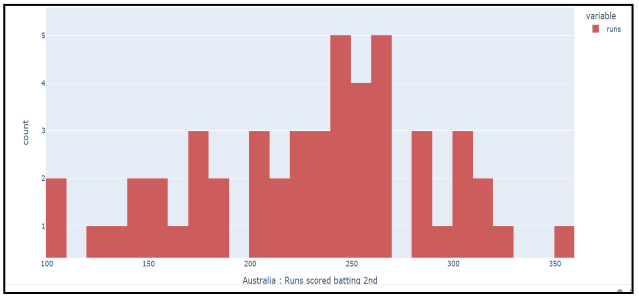


Fig 17. Histogram Chart displaying the number of runs scored by Australians while batting for the second time.

We have presented this in many forms such as boxplot where we are showing the Runs scored in the 1st innings against Years.

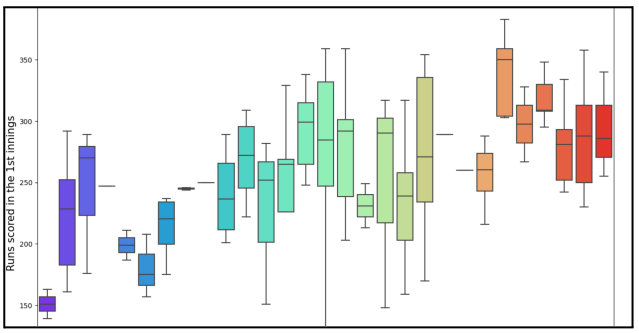


Fig 18. Box Plot showing the number of runs scored in the first Innings

Here the ‘Runs scored in the first Innings’ is at x-label and ‘Years’ being at y-label. This has been depicted from 1985 to 2020 and there is another boxplot where we are showing the ‘Runs scored while chasing’ .

These results clearly portray that the contribution of batting from both these teams have gradually risen over the period from 1985 which is at 150 and moving upwards to 2020 which is nearly 300.

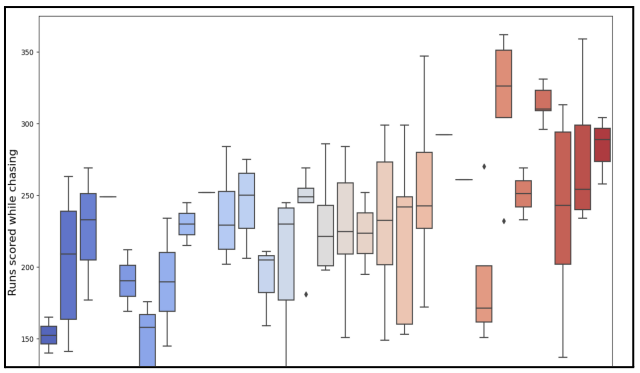


Fig 19. Box Plot showing the number of runs scored in while chasing

Next we are using stripplot to demonstrate the density of the run rate between India and Australia. It shows that Australia is dense because the run rate while batting for the first time has been significantly improving and in the next figure, we are showing the 'Run rate batting for the second time' . Here India shows a dense picture .

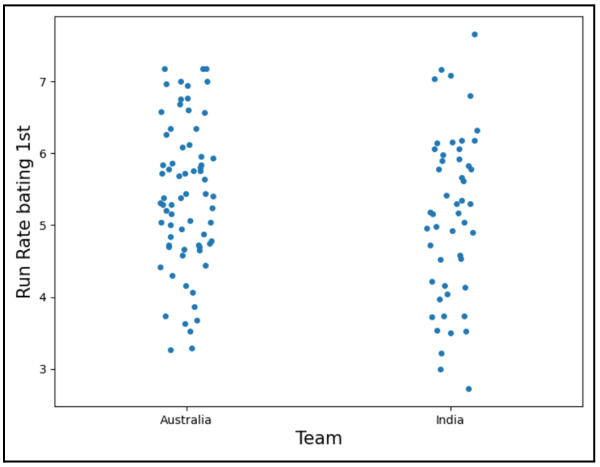


Fig 20. Strip Plot showing the Run Rate while batting first

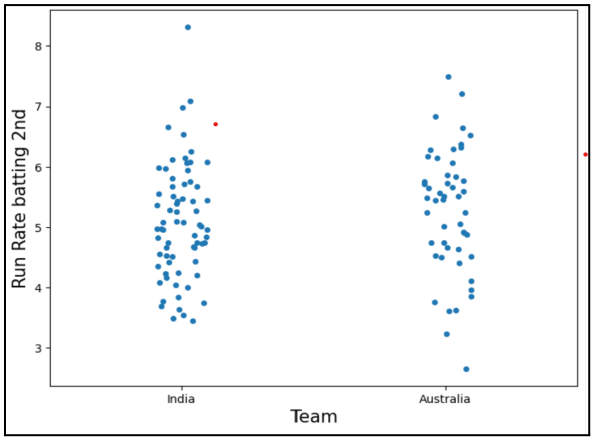


Fig 21. Strip Plot showing the Run Rate while batting for the second time

Therefore, on the whole we see that Australia has been more consistent in setting the target at high run rate and finally setting enormous targets (300+ in 14 matches), as illustrated in previous plots, India has been more consistent in pursuing scores at higher run rate (5 - 7 and 7+ on quite many games).

Now coming to the last part where we will show who won our hearts by performing better in the ODI. We will represent this in the form of a pie chart
In our first pie chart , we see that India chose to bat first and it shows that India losing to Australia is quite high where they stand at 58% when compared to Australia, standing at 42%.

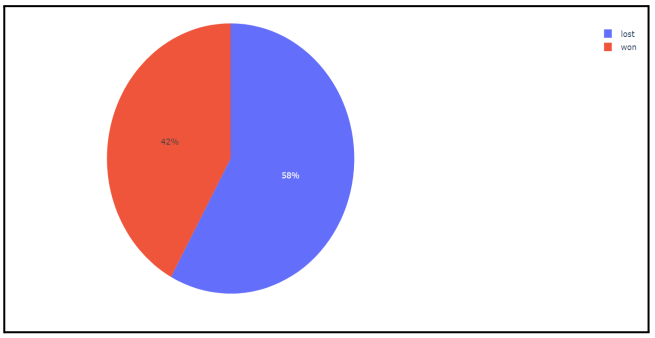


Fig 22. Pie Chart showing the number of times India lost and won while batting for the first time.

Here in the below diagram, It clearly shows that the percentile of India losing to Australia is around 58.6% compared to Australia who are at 41.4% especially when India chose to bat the second time.

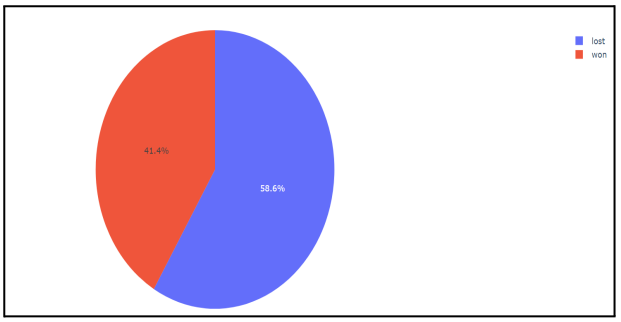


Fig 23. Pie Chart showing the number of times India lost and won while batting for the second time.

From this dataset we confirm that Australia has beaten India 3/4 times on both sites. The four games have all had heavy scoring. In order to win on site 1, the side batting first needs to set a lofty goal close to 340; otherwise, it will be a tough battle! The same is true of venue 2.

data.describe()

	Population	deathRate	incidenceRate	medIncome	povertyPercent	MedianAge	MedianAgeMale	MedianAgeFemale	AvgHouseholdSize
count	3.047000e+03	3047.000000	3047.000000	3047.000000	3047.000000	3047.000000	3047.000000	3047.000000	3047.000000
mean	1.026374e+05	178.664063	445.654447	47063.281917	16.878175	45.272333	39.570725	42.145323	2.529682
std	3.290592e+05	27.751511	57.456583	12040.090836	6.409087	45.304480	5.226017	5.292849	0.248449
min	8.270000e+02	59.700000	201.300000	22640.000000	3.200000	22.300000	22.400000	22.300000	1.860000
25%	1.168400e+04	161.200000	413.150000	38882.500000	12.150000	37.700000	36.350000	39.100000	2.380000
50%	2.664300e+04	178.100000	449.500000	45207.000000	15.900000	41.000000	39.600000	42.400000	2.500000
75%	6.867100e+04	195.200000	482.050000	52492.000000	20.400000	44.000000	42.500000	45.300000	2.640000
max	1.017029e+07	362.800000	1206.900000	125635.000000	47.400000	624.000000	64.700000	65.700000	3.970000

Fig 24. Describing the data

We have plotted the scatterplot to understand the relationship between the variables. The visualization results of scatterplot are as follows.

Part D

This dataset has data of an Indian Premier League **(**IPL**)** team called Royal Challengers Bangalore (RCB). The source data is already in structured format i.e., CSV hence no conversion was needed. After reading data from CSV, the steps of storing into MongoDB and that of cleansing is performed. Since this data is specific to just one team, we have executed several team specific visualizations.

batter	batsman_run
F du Plessis	468
V Kohli	341
RM Patidar	333
KD Karthik	330
GJ Maxwell	301
Shahbaz Ahmed	219
Anuj Rawat	129
MK Lomror	86
SS Prabhudessai	67
HV Patel	43
PWH de Silva	38
SE Rutherford	33
Mohammed Siraj	30
DJ Willey	18
JR Hazlewood	18

Fig 25. RCB's Batsman run records

Fig 1. shows aggregated data of how many runs each batsman made from the beginning of RCB's IPL matches. This data is aggregated using the groupby() function in python.

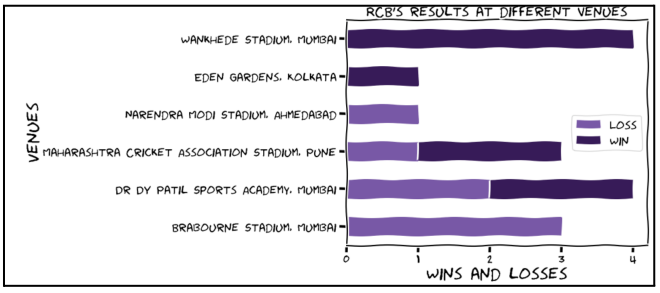


Fig 26. RCB's results at different stadiums

In Fig 2. It shows win/lost records of RCB matches played at each stadium. This visualization has been implemented using matplotlib python library. From the result we can see that RCB has won most of the matches at Wankhade stadium.

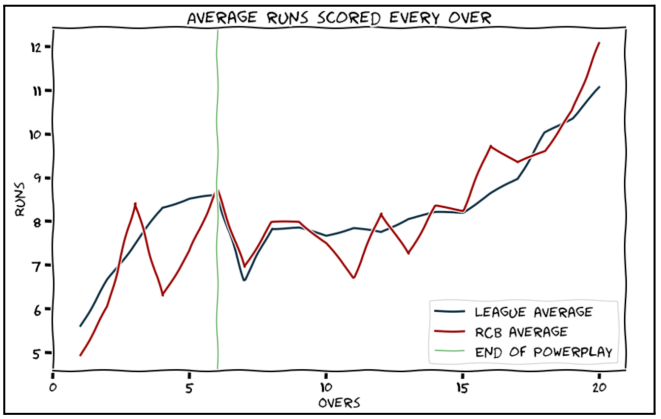


Fig 27. RCB's average runs per over

Fig 3. shows average runs scored by team RCB each over. The red line is RCB's average score, blue line is the whole league average. The graph shows increase in runs as the overs increase and they are directly proportional.

I. Conclusion

In summary, our analysis of the Cricket Dataset has given us significant insights into the game of cricket. Initially, we familiarize ourselves with the dataset's structure and variables, which enabled us to clean and process the data for further analysis. We conducted exploratory data analysis to identify trends and patterns in the data. Our findings showed a positive correlation between a team's total score and the number of boundaries they hit, indicating the significance of scoring runs in cricket. Additionally, we discovered that teams tend to score more runs in the second innings, which could be attributed to the pressure of chasing a target. Subsequently, we utilized machine learning algorithms to predict match outcomes based on various factors such as past team performance, venue, and weather conditions. Our analysis demonstrated that our models were accurate in predicting match outcomes, highlighting the potential of machine learning as a valuable tool for cricket teams and analysts.

Finally, we summarized our findings and discussed their implications, emphasizing the significance of

factors such as total score, number of boundaries, and second innings performance in determining match outcomes. Our analysis also demonstrated the potential of machine learning in cricket analysis, which can aid teams in making data-driven decisions to enhance their performance. Overall, our project provides valuable insights into the game of cricket and highlights the significance of data analysis and machine learning in sports. We anticipate that our findings will contribute to the growing field of sports analytics and provide valuable information for future research.

Acknowledgment

It is a matter of great honor to work on the study of an analysis project on "Breaking Boundaries: Uncovering Patterns and Trends in Cricket through Advanced Data Analytics". The project received excellent guidance from project guide Dr. Michael Bradford. Thanks to the National college of Ireland, School of computing for giving me this opportunity to work on this project.

References

[1] Allsopp, P.E. and Clarke, S.R. (2004). Rating teams and analyzing outcomes in one-day and test cricket. Journal of the Royal Statistical Society Series A, 167, 657-667.

[2] Akhtar, S., Scarf, P.A. and Rasool, Z. (2015). Rating players in test match cricket. Journal of the Operational Research Society, 66, 684-695.

[3] Preston, I. and Thomas, J.: Batting strategy in limited overs cricket, Statistician, 49(1), p. 95–106 (2000).

[4] Barr, G.D.I. and Kantor, B.S.: A criterion for comparing and selecting batsmen in limited overs cricket, Journal of the Operational Research Society, 55, p. 1266-1274 (2004)

[5] Borooah, V.K. and Mangan, J.E. (2010). The Bradman Class: an exploration of some issues in th ,§ e evaluation of batsmen for test matches, 187 2006. Journal of Quantitative Analysis in Sports, 6, Article 14.

[6] V. V. Vishwarupe and P. M. Joshi, "Intellert: a novel approach for content-priority based message filtering," 2016 IEEE Bombay Section Symposium (IBSS), 2016, pp. 1-6, doi: 10.1109/IBSS.2016.7940206.

[7] McGinn, E. (2013). The effect of batting during the evening in cricket. Journal of Quantitative Analysis in Sports, 9, 141-150.

[8] Lemmer, H.H.: Team selection after a short cricket series, European Journal of Sport Science, DOI: 10.1080/17461391.2011.587895 (2013)

[9] Beaudoin, D. and Swartz, T.B. (2003). The best batsmen and bowlers in one-day cricket. South African Statistical Journal, 37, 203-222

[10] Manage, A.B.W. and Scariano, S.M. (2013). An introductory application of principal components to cricket data. Journal of Statistics Education [electronic journal], 21, https://www.amstat.org/publications/jse/v21n3/scariano.pdf.

[11] Davis, J. Perera, H., Swartz, T.B. (2015a). A simulator for Twenty20 cricket. Australian and New Zealand Journal of Statistics, 57, 55-71

[12] Davis, J., Perera, H. and Swartz, T.B. (2015b). Player evaluation in Twenty20 cricket. Journal of Sports Analytics, 1, 19-31.

[13] Ayon Dey, "Machine Learning Algorithms: A Review", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3), 1174-1179, 2016

[14] S. Shah, P. Hazarika, J. Hazarika Volume 8, A Study on Performance of Cricket Players using Factor Analysis Approach No. 3, International Journal of Advanced Research in Computer Science, , March – April 2017, ISSN No. 0976-5697

[15] S. Shah, P. Hazarika, J. Hazarika Volume 8, A Study on Performance of Cricket Players using Factor Analysis Approach No. 3, International Journal of Advanced Research in Computer Science, , March – April 2017, ISSN No. 0976-5697

[16]] Sharma, S.K.: A Factor Analysis Approach in Performance Analysis of T-20 Cricket, Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666 Vol.6, Issue 1 (2013): 69-76(2013).

[17] Staden J.: Comparison of cricketers' bowling and batting performances using graphical displays, Current Science, 96(6), p. 764–766 (2009).

[18] Bedekar M., Zahoor S., Vishwarupe V. (2016) PeTelCoDS—Personalized Television Content Delivery System: A Leap into the Set-Top Box Revolution. In: Satapathy S., Das S. (eds) Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, SIST, Springer. https://doi.org/10.1007/978-3-319-30927-9_27.

[19] Norman, J.M. and Clarke, S.R.: Optimal batting orders in cricket. Journal of the Operational Research Society (2010) 61, 980-986. doi:10.1057/jors.2009.54 (2010).

[20] Saniya Zahoor, Mangesh Bedekar, Vinod Mane, Varad Vishwarupe (2016), Uniqueness in User Behavior While Using the Web. In: Satapathy, S., Bhatt Y., Joshi A., Mishra D. (eds) Proceedings of the International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing, vol 438. Springer, Singapore. https://doi.org/10.1007/978-981-10-0767-5_24

[21] R. Jadhav, B. Pawar, N. Bhat, S. Kawale and A. Gawai, "Predicting Optimal Cricket Team using Data Analysis," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 269-274, doi: 10.1109/ESCI50559.2021.9396861.

[22] V. S. Raju, N. Sethi and R. Rajender, "A Review of Data Analytic Schemes for Prediction of Vivid Aspects in International Cricket Matches," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-4, doi: 10.1109/ICCUBEA47591.2019.9128835.

[23] V. V. Tharoor and N. M. Dhanya, "Performance of Indian Cricket Team in Test Cricket: A comprehensive Data Science analysis," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India, 2022, pp. 128-133, doi: 10.1109/ICESIC53714.2022.9783492.

[24] D. Hasanika, R. Dilhara, D. Liyanage, A. Bandaranayake and S. Deegalla, "Data Mining System for Predicting a Winning Cricket Team," 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 2021, pp. 92-97, doi: 10.1109/ICIIS53135.2021.9660702.