

**UNIVERSITY OF HERTFORDSHIRE**

School of Computer Science

MSc Data Science and Analytics

**Final Project Report**

7COM1039 -0501-2021- Advanced Computer Science Masters Project

Masters Project 02nd Dec 2022

**Title**

Predicting Greatest Cricketer by Comparing Different Machine Learning Approaches

**Name:** Nadeem Theba – 20029831

**Supervisor:** Dr. Fouad Shat

MSc Final Project Declaration

This report is submitted in partial fulfillment of the requirement for the degree of Master of Science in Data Science and Analytics at the University of Hertfordshire (UH).

It is my work except where indicated in the report. I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged.

**Predicting Greatest Cricketer by Comparing Different Machine Learning Approaches**

Table of Contents

[List of Tables 4](#_Toc120916824)

[List of Figures 5](#_Toc120916825)

[Abstract 6](#_Toc120916826)

[1.0 Introduction 7](#_Toc120916827)

[2.0 Literature Review 9](#_Toc120916828)

[2.1 Defining Performance Metrics 10](#_Toc120916829)

[3.0 Aim and Objective 11](#_Toc120916830)

[4.0 Methodology and Implementation 12](#_Toc120916831)

[4.1 Defining Performance Metrics 12](#_Toc120916832)

[4.1.1 Batting Attributes 12](#_Toc120916833)

[4.1.2 Bowling Attributes 13](#_Toc120916834)

[4.2 Dataset Preparation 15](#_Toc120916835)

[4.3 Exploratory Data Analysis 17](#_Toc120916836)

[4.4 Data Preprocessing and Machine Learning Models 20](#_Toc120916837)

[5.0 Results and Evaluation 23](#_Toc120916838)

[6.0 Limitations 31](#_Toc120916839)

[7.0 Conclusion 31](#_Toc120916840)

[Appendix A: Batsmen 34](#_Toc120916841)

[Appendix B: Bowler 45](#_Toc120916842)

[References 54](#_Toc120916843)

# List of Tables

[Table 4.1 Player Statistics 1](#_bookmark12)8

[Table 4.2 Performance Score for Top Five Bowlers with Highest Wickets 19](#_bookmark14)

[Table 4.3 Performance Score for Top Five Batsmen (Batting Average>40 and Total Matches Played>200) 20](#_bookmark33)

[Table 5.1 Comparing Model Accuracy (Batsmen) 24](#_bookmark36)

[Table 5.2 Comparing Model Accuracy (Bowlers) 24](#_bookmark39)

[Table 5.3 Comparison of Performance Metric PCA\_batsmen 28](#_bookmark43)

[Table 5.4 Comparison of Performance Metric PCA\_bowler 28](#_bookmark47)

[Table 5.5 Accuracy Measure Metric (Batsmen) 29](#_bookmark49)

[Table 5.6 Accuracy Measure Metric (Bowlers). 30](#_bookmark52)

# List of Figures

[Figure 4. 1 Top Ten Bowlers with Highest Wickets Taken 19](#_bookmark3)

[Figure 4. 2 Comparison of Batting Averages to the Total number of matches played 20](#_bookmark5)

[Figure 4. 3 Correlation Matrix of Batsmen Data. 21](#_bookmark8)

[Figure 4. 4 Correlation Matrix of Bowlers Data 21](#_bookmark10)

[Figure 5. 1 XGBOOST Accuracy for Batsmen Data 25](#_bookmark16)

[Figure 5. 2 XGBOOST Accuracy for Bowlers Data 25](#_bookmark30)

[Figure 5. 3 AdaBoost Accuracy for Batsmen Data 26](#_bookmark34)

[Figure 5. 4 AdaBoost Accuracy for Bowlers Data 26](#_bookmark37)

[Figure 5. 5 Random Forest Accuracy for Batsmen Data 27](#_bookmark40)

[Figure 5.6 Random Forest Accuracy for Bowlers Data 27](#_bookmark41)

Predicting Greatest Cricketer by Comparing Different Machine Learning Approaches

# Abstract

The purpose of this study is to examine the performance qualities that define the calibre of a cricket player. While selecting a team for the competition, the performance data analysis of the players is crucial for making crucial business decisions. This study explores three distinct machine learning models—Random Forest, XGBOOST, and AdaBoost—in terms of their accuracy in forecasting batsmen and bowlers' performance indicators. Random Forest and XGBOOST were discovered to be the most accurate machine learning models for the prediction challenge. The purpose of this research is to provide a simple and uncomplicated strategy for discovering different important performance variables that link to a cricketer's high performance.

# Introduction

This study provides data analysis and prediction in the field of cricket by identifying important features of high-performing players and deploying various machine learning models for forecasting the performance of cricketers. The examination of batsmen and bowlers' performances will help determine the highest-performing or "greatest" cricket player of all time. Cricket is a sport where using statistics to compare team performance and predict scores can be helpful.  (James Cochran, 2017). There are various verified web sources to gain a deeper understanding of the rules and regulations of this sport (CricketRules, 2022). However, a portion of this section will provide an overview of how cricket sports operate. Each team takes turns batting and fielding in cricket, which is a sport played by two teams. Cricket can be played with just a bat, a ball, and three stumps. Each team consists of eleven players, including batsmen, bowlers, and a wicketkeeper. The objective of the batting team is to score as many runs as possible. The fielding team, on the other hand, attempts to prevent the batting team from scoring maximum runs by capturing the ball (only when it has not yet touched the ground) or preventing the ball from getting beyond the boundaries. Similarly, the fielding team's bowler endeavours to strike the ball through the wickets. In the report's "Defining Performance Metrics" section, it is detailed in greater detail how a batsman can score runs and how the fielding team can stop the opposing team in order to score the most runs. A team's bowling and batting decisions are made prior to the start of play. The leaders of both teams toss a coin to choose which team will bat or bowl first. The captain who wins the toss decides whether his side will bat or bowl first. Generally, leaders decide whether to bat or bowl first based on the characteristics of the cricket pitch. The pitch is a rectangular area of the ground measuring 22 yards in length and 10 feet in width, bounded on either side by the bowling creases and on either side by imaginary lines, one on each side of the imaginary line connecting the centres of the two middle stumps, each parallel to it and 1.52 metres from it (Lords, 2017). Typically, there are three formats for cricket: Test matches, one-day matches, and Twenty20 matches. This study's dataset will only retrieve information in ODI format. There is a vast amount of literature on predicting the outcome of matches in the IPL, T20, and world cup.

The dataset is acquired from espncricinfo, an open-source cricket dataset website, in order to obtain the data of all batters and bowlers. As a result, two distinct datasets, Batsmen and Bowlers, are generated and all data processing, visualisation, and machine learning models are applied to them. Both of these datasets feature a variety of batsman and bowler performance-related metrics. Using the notion of principal component analysis, these factors are further examined to determine the performance measures for bowlers and batters (PCA).PCA is a widely used variable reduction technique for correlated data with the objective to collapse a set of correlated variables into fewer uncorrelated variables as linear combinations of the original variables (Ananda B. W. Manage, 2013). Following that, three distinct machine learning models—XGBOOST, Random Forest, and AdaBoost—are used to predict batsman and bowler performance metrics in order to select the best performing cricketer. The remainder of the paper is structured as follows. Section 3.1 of Chapter 3 describes the various parameters that influence the performance metrics of batters and bowlers. The next sections of Chapter 3 discuss data collection, preprocessing, visualisation, and machine learning model implementation. Chapter 4 examines crucial evaluation and comparison results for models. Chapters 5 and 6 give the conclusion, some ideas for further research, and the limits of this study. The project timeline implemented for this project has been illustrated in the following Gantt Chart,

# Literature Review

This study's background research was undertaken by analysing the possibilities of existing or comparable studies in the field of sports or cricket. This section includes verified background information pertinent to this study. The authors of "Unique technique for cricket match outcome prediction using XGBOOST algorithms" employ the boosting algorithm to estimate the score at the end of the next over and discuss the influence of boosting algorithms to predict the match winner depending on the present condition of the match. The model accurately predicted the outcome of a certain team's match with an accuracy of 84.4%.(NIMMI WEERADDANA, 2021). The data, like in this study, was obtained from the open-source cricket data website espncricinfo. In addition, the study compares the findings of the boosting technique against those of a number of conventional data mining algorithms, including logistic regression, nearest neighbour, and support vector machines. The XGBOOST algorithm outperforms all other algorithms and conventional data mining techniques. In addition to this work, several others have utilised the XGBOOST algorithm for prediction and data analysis in the field of cricket. Another study by Kalpdrum Passi and Niravkumar Pandey uses the number of runs scored by batters and the number of wickets taken by bowlers as indicators of their performance (Pandey, 2018). They employ conventional machine learning algorithms such as Random Forest classifier, Support Vector Machine, and Nave Bayes to solve this prediction issue. To anticipate the outcome of runs and wickets, the authors have developed four distinct performance indicators for each bowler and batter. The analysis revealed that Random Forest was the most accurate classifier, predicting runs and wickets with an accuracy of 90.74 percent. It should be emphasised that the authors have considered a large number of input factors as opposed to this study, which only considers a small number of input variables based on the performance metric formula described in later parts of this report.

After reviewing a vast amount of research, there are just a handful of cricket studies that use the AdaBoost model for predictive analysis. One such study applies Multinomial Logistic Regression, Random Forest, and AdaBoost to the analysis of cricket player performance. Significantly, the authors chose these models due to their capacity to accept a huge quantity of input data, train the decision trees and stumps, and provide reliable predictions (Sahu, 2021). The study concluded that both Random Forest and AdaBoost had a neck-to-neck prediction score, however, AdaBoost did not turn out to be the best as Random Forest achieved the highest accuracy for predicting the outcome of IPL matches. The paper sheds a light on the significance of different characteristics of a cricket player such as the number of wickets taken, the number of matches played, and so on in performing the task of team selection. It should be noted that the authors have performed a deeper data pre-processing by performing techniques such as P-value and Manual Encoding. While there is a wide range of studies related to predicting the outcome of matches in ODI and IPL, a limited number of research has been done on the predicting the highest performing cricketer or players’ performance in a match. One such research done by Arjun tan and Marius Schamschula attempts to find the greatest Barbadian cricketers by using the approach of cluster analysis. The method of cluster analysis consists of the classification of a group of objects according to certain similarities and criteria (Schamschula, 2018). While this work does not employ any machine learning techniques to address the issue, its strategy is novel and provides a mathematical knowledge of how to solve such difficulties. This research employs four fundamental cluster analysis mathematical procedures to determine the best Barbadian cricketer. Only the records of the Barbadian cricketers in the Holy Trinity of West Indian Cricket have been reviewed in this research; hence, the data is extremely limited.

This study addresses the research topic of identifying the essential qualities that make a cricketer (batsmen or bowlers) the "greatest" or a high-performing cricketer. To answer this question, one must first understand the performance characteristics and a performance metric that calculates the quality of a cricketer based on these parameters. A similar study attempted to rank IPLT20 cricketers' performances based on their strengths and weaknesses. It also compares several ranking systems to the study's own rating system. In addition, the authors identified the top ten players who had the potential to play high-quality games in the future. The authors decided to utilise the recursive feature elimination machine learning technique to evaluate the performance of each player in order to pick the features to be used to forecast their performance. Later, the research concludes with a listing of the top ten batsmen and bowlers (C. Deep Prakash, 2016). It should be emphasised that the rank performance measure presented in the study has its own limits and could be modified through external techniques. It is essential to analyse the model's accuracy ratings to see whether it overfits or underfits the data. The objective of Nadini and Sachin Kumar's study was to forecast the winning fantasy cricket team by measuring the accuracy of each model. The authors employed six distinct machine learning models, including Nearest Neighbor, Multinomial and Gaussian Nave Bayes, Support Vector Machine, Decision Trees, and Random Forest algorithm. The authors underline the need of analysing error metrics scores such as R-squared error, mean absolute error, mean squared error, mape error, etc., when processing these models. The author thinks that the 0.99 value of R2 may imply that the models used in the study are overfit (Sachin Kumar S, 2022). In addition, the authors describe how to determine whether a model is overfit when the R-squared error metric score is significantly close to 1. The study uses data from the open-source website "Cricsheet" to calculate the R2 score for the best performing model and provides information on how the study can be improved for future research.

# Aim and Objective

The purpose of this study is to examine the features of batsmen and bowlers that contribute to their high performance, with the aim of predicting these performances using various machine learning models. This study employs principal component analysis to determine the performance scores for each player based on particular attributes. Furthermore, based on the literature review, it is possible to conclude that there is a large amount of literature on machine learning approaches for predicting the outcome of a tournament match. However, statistical analysis of cricket performance criteria has only been the topic of a few of research. This study use the statistical method of Principal Component Analysis to conduct both prediction and performance analysis in an effort to fill this gap in the literature. Consequently, the following research question will be formulated for this study,

RQ: “Which characteristics contribute in predicting the next greatest cricketer?”

# Methodology and Implementation

## 4.1 Defining Performance Metrics

This section will discuss the numerous aspects that influence a cricketer's performance. In cricket, a variety of physical elements contribute to effective performance. These factors are studied and analysed by teams in order to comprehend the opponent's playing style and the most effective strategy for defeating them. However, the unpredictable nature of a cricket match is what makes it so exciting, and it is frequently impossible to predict the winning team until the last ball is bowled. As a result, there is a growing demand for sports data scientists and data analysts who can assist in making optimal judgments about squad composition and shot selection. The form of an individual and the team as a whole play a crucial role in determining the winning team. In this study, the performance characteristics of an individual player are analysed to determine the performance metrics. After examining the voluminous literature on the issue of this study, the following traits for bowlers and batsmen have been determined to be consistent in all studies measuring player performance

## 4.1.1 Batting Attributes

**Batting Average (BA):** In cricket, the batting average is an important aspect to understand the consistency of a batsman. The batting average is the total number of runs scored per dismissal of that player. The number of runs scored and number of times the player is dismissed vary for each player. As a result, the average of each batsman will be unique. Batting average of a batsman is defined as the total number of runs scored by the batsman divided by the number of times he was dismissed or the number of times that the batsman has remained not out (A). The performance of the batsmen is directly proportional to their batting average scored i.e., a higher batsmen average score signifies a higher performance of the player.

**( 1 )**

**Strike Rate (SR):** In cricket, the definition of the term ‘Strike Rate’ is different for a batsman player as compared to the bowling player. For batsman, it is defined as the number of runs the player has scored for every 100 balls faced during playtime. It can also indicate the speed of the score achieved by the batsman (Croucher, 2000). Similarly to batting average, a higher score in strike rate can indicate a high quality performance of the player or a player who has the ability to score quickly.

**( 2)**

**Total Runs (TR):** In cricket, a run is the unit of scoring and the team with the most runs (scores) win the match. The aim of the batsmen in cricket is to try to score as many runs as possible throughout their innings. To attain a single run (or score), the batsman has to strike the ball and both batsmen and his batting partner has to run towards each other’s position before the bowler outs either of them. It is possible to score runs without running the length of the pitch, if a batter succeeds to hit the ball past the boundary line. Four runs are scored if the ball bounces or rolls along the ground to past the boundary, whereas six runs are scored if the ball reaches past the boundary without touching the ground.

**Hundreds:** A hundred or also called ‘century’ is a score which is greater than or equivalent to 100 runs by a single batsman in a single innings. This performance metric will calculate the number of centuries scored by each player over the course of their career.

**Fifties:** A fifty or also called ‘half century’ is a score which is greater than or equivalent to 50 runs by a single a batsman in a single innings. This performance metric will calculate the number of half centuries scored by each player over the course of their career.

**Sixes:** As mentioned earlier, a score of six is considered when the batsman succeeds to hit the ball past the boundary line without ball touching the ground. This performance metric will help evaluate the players who scored the highest sixes throughout their career.

**Fours:** Similar to scoring a six, four can be scored when the ball hits past the boundary line by bouncing or rolling on the ground. This performance metric will help evaluate the batsmen with better skills of hitting a four in a match.

## 4.1.2 Bowling Attributes

**Wickets Taken (WT):** The primary objective of the bowling team in cricket is to gain wickets by getting the batsman out. There are numerous ways to consider a ball a wicket, including run out, bowl caught, stumped, and leg before wicket. This performance attribute will represent the total number of wickets a bowler has taken during his career.

**Economy Rate (ER):** The bowler's economy rate is the average amount of runs conceded per over bowled. The economy rate is inversely proportional to the bowler's performance quality. A bowler with a lower economy rate is seen as performing at a higher level than one with a higher economy rate. A lower score indicates that the bowler has successfully prevented the opponent from scoring runs.

**( 3)**

**Bowling Average (BA):** The bowling average represents the average number of runs conceded per wicket taken by a bowler. Similarly to the Economy Rate, a lower bowling average indicates superior performance. Consequently, this performance metric is indirectly proportional to the bowler's performance.

**( 4)**

**Strike Rate (SR):** The strike rate in bowling is the average number of balls bowled per wicket for a given bowler. This performance statistic is also inversely proportional to the bowler's performance. Therefore, a bowler with a lower Strike Rate is more likely to take wickets than a bowler with a higher Strike Rate.

**( 5)**

Principal component analysis is used in this study to evaluate and rate the performance of both bowlers and batters. Based on the research that Manage and Scariano did to successfully use principal component analysis in multivariate data analysis to rank the cricket batsmen and bowlers who played in the 2012 Indian Premier League, this study introduces two performance metrics to measure the level of performance of batsmen and bowlers (Ananda B. W. Manage, 2013),

**( 6)**

**( 7)**

The numbers multiplied with each characteristic is a result of calculation performed by Manage and Scariano to obtain the correlation matrix among each of the characteristic of batsman and bowler. It is to be noted that the smaller values indicate better performance in the PCA (Ananda B. W. Manage, 2013). However, in order to facilitate comprehension and comparison, negative indices are used. Higher scores for both PCA batsmen and PCA bowlers reflect a player's superior performance. In any sport, it is difficult to quantify athletic performance; however, Manage and Scariano have developed a straightforward method for calculating a player's performance based on certain characteristics.

## 4.2 Dataset Preparation

The data used in this study is obtained from a verified source for cricket data: www.espncricinfo.com (ESPNcricinfo, 2022). ESPNcricinfo is a premier sports website devoted only to cricket. It provides a vast data repository for the sport of cricket. The website gives cricket statistics for men and women based on innings, years, country, player classification, and ground. Users can quickly select data based on their needs and export the results to PDF. In this study, the men’s cricket dataset has been extracted from the HTML espncricinfo website using the pandas library. Later, these data are converted to an a.csv file, which the Pandas library will import and read in order to conduct various data operations. The task of data collection has been performed while keeping ethical issues in consideration. Three main ethical issues that could be potentially faced while conducting research are (Research, 2007),

* Concerns regarding deception or voluntary participation in research
* Factors relating to the breach of confidentiality of personal data
* Matters relating to particular groups of individuals whose characteristics might imply additional vulnerability.

As a result, care is taken to ensure that the data collection source's creditors are acknowledged in the study and that no personal biases or viewpoints influenced the data collection procedure. For this study, two distinct datasets (batsmen and bowlers) consisting of players' career-long performance are investigated, including as many as 27,988 cricket batsmen and 2,200 cricket bowlers. The batting statistics comprise numerous performance metrics for batsmen, such as the total number of career matches played, the total number of runs scored, the number of balls faced, the total number of hundreds, half-centuries scored, etc. Similarly, the data for bowlers includes parameters such as wickets taken, career average, total matches played, and so forth. The following list provides a breakdown of several major properties in both datasets,

1. **Player -** This column stores the names of each player along with their country name
2. **Span-** This column stores the start and end year of each player's career in cricket
3. Inns - This column stores the total number of innings played by each player over the course of his career
4. **Runs -** This column stores the total number of runs scored by the player over the course of his career in batsmen data. Similarly, it stores the total number of runs gained on their thrown balls over the course of his career in bowler’s data.
5. **HS -** This column stores the total number of high score achieved by the batsmen over the course of his career
6. **Ave** - This column stores the total number of average scored by the both batsmen and bowler over the course of his career
7. **BF -** This column stores the total number of balls faced by the batsmen over the course of his career
8. **SR -** This column stores the total number of strike rate achieved by both the batsmen and bowler over the course of his career
9. **Mat -** This column stores the total number of matches played the player over the course of his career
10. **Econ –** This column stores the total number of economy rate achieved by the bowler over the course of his career

Using Python 3.9.0 and numerous Python-supported libraries, the full dataset was analysed, inspected, and applied to machine learning methods. The data was loaded and analysed using the pandas, numpy, matplotlib, and seaborn packages. In addition, libraries like Sklearn have been utilised to import numerous machine learning algorithms for forecasting player performance indicators. After loading the dataset, data cleansing is the first stage in data preparation. This study's dataset contains as many as 2798 data values and 16 features for the batter's dataset and 2200 rows with 14 characteristics for the bowler's dataset, which reflect various information on the performance statistics of the players. Data cleaning is the process of removing faulty data values from a dataset, such as null values, improperly formatted values, misspelt values, incomplete values, and duplicate values. Data cleansing is a crucial component of the data analytics process. A dataset that is incomplete or dirty would affect the accuracy of the results; therefore, in this study, the dataset has been preserved as intact as feasible to increase the dependability of insights.

In this study, the first stage in the data cleaning procedure is the elimination of irrelevant observations. The undesirable observations may include duplicate data values, null values, etc. Consequently, both datasets are analysed for duplicate, null, improperly formatted, and misspelt data values. The dataset contained zero null values; however, a large number of rows contained special characters that were removed along with incorrect columns. A few columns of the batsman and bowler dataset contained special characters that had to be eliminated prior to further processing. In the batsman dataset, a new column titled '100 Sum 50' was added to store the sum of hundreds and half centuries scored by each player. This column was included since it would be used to determine the batsman's performance score, as discussed in the next sections. All the data values in both datasets were converted to lowercase for easier comprehension and to avoid words with irregular capitalization. Finally, the data types of each column in the batsmen and bowler data were examined, and it was discovered that nearly every feature had the erroneous data type. This is as a result of the inclusion of special characters in nearly every column of both databases. The majority of these special characters are '-', which represents empty or null values, '\*', which symbolises the batsmen not being out, and '+', which represents a number larger than a specified value in certain columns, such as 4s and 6s in the batsmen dataset. These characteristics included 4s – the total number of fours each player scored, 6s – the total number of sixes each player scored, and HS – the total number of high scores for each player. Originally, each of these features had an object datatype, which was later converted to an int64 datatype. Further investigation revealed that these three columns had special characters, which clarifies the object datatype.

## 4.3 Exploratory Data Analysis

After data cleaning, the batsmen data structure has been reduced from 2798 rows and 16 columns to 2554 rows and 16 columns. Similarly, the bowler's data structure has decreased from 2200 rows and 14 columns to 1742 rows and 13 columns. Before beginning analysis, it is crucial to eliminate any unnecessary information from the dataset. The goal of exploratory data analysis is to investigate various aspects of cricket and identify the best performers in terms of batting and bowling. For visualisation, the libraries Matplotlib and Seaborn are used for analysis. 25% (also known as the "quantile value") of players have gathered total scores lower than approximately 29 runs, 50% (also known as the "median value") of batsmen have scored total runs lower than approximately 125 runs, and 75% of players have gathered total scores lower than approximately 529 runs. It can be seen in Table 1 that just 25 percent of batsmen scored more than 29 runs in their careers, with some not even scoring a century. Sachin Tendulkar holds the record for the most overall runs scored in cricket with 18,426 runs, including 49 centuries and 96 fifties. Sachin Tendulkar retired from cricket in 2012, but his total runs remain unsurpassed. The smallest number of total matches played by a batsman is one, while the most total matches played by a player over his career is 463. In terms of a bowler's data, the number of wickets taken is a crucial performance metric. 534 is the maximum number of tickets purchased by M. Muralidaran, as observed.

**Table 4.1: Player Statistics**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Batsmen** | | | | **Bowler** | | | |
|  | **Runs** | **Average** | **Balls Faced** | **Strike Rate** | **Wkts** | **Average** | **Balls** | **Strike**  **Rate** |
| **25% percentile** | 29.0 | 8.6 | 54.0 | 0.0 | 3.0 | 27.8 | 166.0 | 34.3 |
| **50% percentile** | 125.0 | 16.5 | 194.0 | 50.0 | 11.0 | 34.8 | 490.5 | 42.5 |
| **75% percentile** | 529.7 | 26.0 | 728.5 | 65.3 | 37.0 | 45.7 | 1580.0 | 54.5 |
| **min** | 0.0 | 0.0 | 1.0 | 78.6 | 1.0 | 1.0 | 2.00 | 2.0 |
| **max** | 18426.0 | 264.0 | 21368.0 | 328.5 | 534.0 | 251.0 | 18811.0 | 234.0 |

According to the statistics, 25% of the bowlers took 29 wickets, while 50% and 75% of the bowlers scored 125 and nearly 529 wickets, respectively. In Table 3.1, the player's statistics for a handful of significant traits are given. Visualization is an essential component of Exploratory Data Analysis for gaining insight into the data and identifying patterns in the players' attributes through the use of maps and graphs.

Chart, bar chart

Description automatically generated**Figure 4. 1 Top Ten Bowlers with Highest Wickets Taken**

An important parameter which highly influence the performance of bowlers is the ability to take the number of wickets in a match. As displayed in the graph Figure 3.1, it is observed that the top ten players with the highest wickets taken have scored a very high score of the performance metric, PCA\_bowler as tabulated below,

**Table 4. 2 Performance Score for Top Five Bowlers with Highest Wickets Taken**

|  |  |  |
| --- | --- | --- |
| **Bowler** | **Total Wickets Taken** | **PCA\_bowler** |
| M Muralidaran | 534 | 193.48333 |
| Wasim Akram | 502 | 178.97661 |
| Waqar Younis | 416 | 144.90312 |
| Wpujc Vaas | 57.68 | 131.02843 |
| Shahid Afridi | 41.02 | 121.59493 |

Table 3.2 includes the names of the top five bowlers in terms of the overall number of wickets taken throughout the course of their careers, as well as their performance measure score (PCA bowler). To evaluate the performance quality of these bowlers, the list of bowlers with the highest PCA bowler score must be calculated, since a higher number of scores indicates a higher performance quality. The top five bowlers were determined prior to modelling by computing the performance metric PCA bowler range between 120 and 193. Refer to the report's Results and Evaluation section for additional information. However, based on the range, it can be concluded that the bowler's ability to take the most wickets has a significant impact on their total performance.

Chart, bubble chart

Description automatically generated**Figure 4. 2 Comparison of Batting Averages to the Total number of matches played**

The above Figure 3.2 depicts the relationship between the Batting Average (X-axis) and the total number of matches played (Y-axis). The graph represents the batting average of the players as compared to the number of matches they played. The red square is drawn to signify the players with a high-performance metric score of PCA\_batsmen. These numbers signify that a high performing batsman lies with the score of their batting average above 40 for the total number of matches played above 200. As a result, it can be suggested that playing a high number of matches does not necessarily mean that the player will have a high number of batting average. Moreover, the players with the highest number of batting average have played a significantly lower number of matches in their career. A significantly good chunk of data is concentrated between the batting average range of 0 to 20, showing that the data contains a good number of players who are inexperienced and played in very few matches in their career. The players with batting average above 40 and total number of matches played above 200 are tabulated below,

**Table 4. 3 Performance Score for Top Five Batsmen (Batting Average>40 and Total Matches Played>200)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Batsmen** | **Batting Average** | **Total number of Matches** | **PCA\_batsmen** |
| SR Tendulkar | 44.83 | 463 | 9447.42609 |
| KC Sangakkara | 41.98 | 404 | 7211.49154 |
| RT Ponting | 42.03 | 375 | 6596.52128 |
| V Kohli | 57.68 | 262 | 6935.01069 |
| SC Ganguly | 41.02 | 311 | 6360.1216 |

The performance of five batters with batting averages above 40 and a total number of matches played in excess of 200 is depicted in Table 3.3. Prior to modelling, the top five cricketers, as determined by computing the performance metric PCA batsmen range between 8000 and 11500. To find out more about this, please go to the Results and Evaluation section of the report. Consequently, based on the values of the PCA batsmen in Table, it can be concluded that these cricketers are high-quality players, which explains why just a few are located within the red box of the graph.

The correlation matrix obtained from the batsmen data, revealing correlation coefficients for various variables, is depicted in Figure 3.3. This table makes it easy to identify correlations between potential pairings of features. There is a perfect positive linear correlation between the columns Match and Innings, followed by the column pairs Match-Runs, Runs-Innings, and Balls Faced-Runs. On the basis of this information, it may be concluded that these columns are highly interdependent. For instance, the more the number of matches a player has played, the greater his chances are of raising his total on the scoreboard.

Chart, treemap chart

Description automatically generatedChart, table, treemap chart

Description automatically generated**Figure 4. 3 Correlation Matrix of Batsmen Data Figure 4. 4 Correlation Matrix of Bowlers Data**

The correlation matrix produced from the bowler's data is depicted in Figure 3.4. Identifying a connection between two aspects can aid in comprehending the significance of features in relation to one another and provide a tip as to which column to remove and which to retain. The top three feature combinations with the best positive correlation for the bowler's statistics are balls-innings, balls-runs, and innings-wickets. For instance, the longer a bowler bats, the greater his odds of obtaining an opportunity to take a wicket.

## Data Preprocessing and Machine Learning Models

Before processing machine learning models, it is crucial to ensure that the data is in the correct format, as different models require the data to be in different formats. After cleaning the batting and bowling datasets, data preprocessing was performed to make the data acceptable for machine learning models in order to improve the accuracy and efficiency of the machine learning models used in this work. As part of data preprocessing, it is essential to delete redundant columns, as data frequently contains columns that are irrelevant to the analysis. Consequently, the batsmen dataset contains two columns: "Span," which stores the length of a player's cricket career, and "Mat", which stores the total number of matches played by the batsmen over his whole career. These columns do not contribute to the PCA performance indicators for batters, and maintaining them in the model could result in a less accurate score. A new column has been added to the PCA performance statistic for batsmen that totals the number of fifties and hundreds recorded by each player throughout their career. In the case of batsmen data, the objective variable for all machine learning models will be a column titled "PCA Batsmen," which computes the performance score for each player using the method described in the preceding section. Given that average and strike rate are of the float data type, the performance score recorded in "PCA Batsmen" will likewise be of the same type. The remaining columns will function as dependent variables. The "Player" column, which holds each player's name, is one of the most crucial characteristics. This column is essential for evaluating each player's performance score and identifying those with the highest scores. However, this column is composed of object-type string data. As this column will be required, it must be transformed into a machine-readable format using label encoding. As machines can only interpret numbers and not words, converting category columns to numerical columns is necessary for understanding machine learning methods. In this method, a unique integer is assigned to each label based on its alphabetical order. Consequently, each player will be assigned a positive, unique number. This technique has been implemented using Sklearn's preprocessing module, which permits LabelEncoder encoding of a categorical column. Once applied, a new column is established in each dataset to record the corresponding batsmen's allocated ids. Separating the data into dependent variables and target variables is the next stage. The goal variable will be "PCA Batsmen" for the batsmen dataset and "PCA Bowler" for the bowlers dataset. The machine learning models are deployed independently on each dataset in order to facilitate the evaluation of results for both batsmen and bowlers. Following the data cleaning process, the dependent features will consist of all leftover columns in both datasets. In addition, the "Player" column will no longer be regarded as a dependent variable, as it will be replaced by the encoded column "Player encoded." As shown below, three machine learning models are employed in this study to predict the performance metric for each player.

**XGBOOST (Extreme Gradient Boosting)**

Tree boosting is an effective and widely used machine learning technique, and XGBOOST is a scalable end-to-end tree boosting supervised machine learning algorithm that has been shown to give accurate results while using much fewer resources than existing systems to solve regression and ranking problems (Tianqi Chen, 2016). In this model, the depth of each tree has been set to 10, and it will continue to divide until it reaches the maximum depth specified for the tree. This model's learning rate is set to 0.1, which multiplies the tree values by 0.1 to slow model fitting and prevent overfitting. In XGBOOST, subsampling is performed once for each tree built. Each time a new depth level is reached in a tree, the subsample ratio of columns is set to 0.8. Because of its superior predictive ability and performance, which was achieved by improving the gradient boosting framework, this model was used in this study.

**Random Forest**

Random Forest machine learning algorithm was developed by Leo Breiman, a statistician from University of California at Berkley to improve the classification of diverse data using random sampling and attributes selection (Livingston, 2005). It is a supervised machine learning technique that constructs decision trees from several samples and calculates the mean for regression issues. This method has been widely used in the literature on the same topic as this study and is widely regarded as a superior option for sports prediction challenges. It adds additional randomization to the model when building trees to search for the best feature from a random subset of features, reducing the overfitting problem, decreasing variance, and increasing the accuracy score(Wisconsin–Madison). Before fitting the data into the model, the data will be partitioned into train and test, with X storing all dependent features and y the target feature. The data is split 70/30, with 30% in testing and 70% in training, as it is standard practise to split data in this manner to reduce errors. (Hai-Bang Ly, 2021). The random state parameter is used to divide the dataset prior to shuffling. For this model, the random state has been set to 42 to ensure that the train and test sets are identical after 42 randomizations. n\_estimators, the required number of trees before calculating the maximum average of forecasts, is an additional crucial parameter of this model. The greater the number of trees, the better the performance; nevertheless, it may take longer to process. The n estimators are set to 1000 for processing the random forest model in this investigation.

**AdaBoost (Adaptive Boosting)**

Freund and Schapire introduced Adaboost algorithm in 1995 to improve the predicting ability of learning system (TU Chengsheng, 2017). It is among the earliest boosting techniques that may be applied to classification and regression applications. This model's learning rate is set at 0.2, which determines the loss function used to calculate the weight of the base models. This setting is offered to reduce each classifier's contribution. This model's exponential loss is convex and grows exponentially for negative values, making it more susceptible to outliers. Compared to random forest and the XGBoost method, this model employs only 50 trees, as opposed to thousands in random forest and the XGBoost algorithm. This model was chosen to predict the performance metrics of cricket players due to the ability of the adaboost algorithm to improve the prediction power by converting a number of weak learners into strong learners, where a weak learner represents a classifier with a weak correlation to the true classification.

# Results and Evaluation

This study evaluates three models based on their respective accuracy scores and accuracy measure scores, such as R2 score, MAE, MSE, and RMSE. Table 4.1 displays the accuracy score derived from processing each model to predict the performance metric score for batters (PCA batsmen), whereas Table 4.2 displays the accuracy score derived from processing each model to predict the target variable (PCA bowler) for the bowlers data. The performance statistic for batters and bowlers was defined in the section titled "Defining Performance Metrics." Evidently, the XGBOOST model and Random Forest model provide consistently superior accuracy than the other two models. Similar research reaches the conclusion that boosting algorithms such as XGBOOST outperform other widely used machine learning models in forecasting the outcome of cricket matches (NIMMI WEERADDANA, 2021). In this study, different sizes of training and test sets are used to find the best combination that gives the most accuracy with the results tabulated below,

**Table 5.1 Comparing Model Accuracy (Batsmen)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning Model** | **Accuracy for predicting PCA\_batsmen (%)** | | |
| **Train (90%) Test (10%)** | **Train (70%) Test (30%)** | **Train (60%) Test (40%)** |
| XGBOOST | 99.96 | 99.96 | 99.08 |
| Random Forest | 99.92 | 99.92 | 99.92 |
| AdaBoost | 98.87 | 99.01 | 97.89 |

**Table 5.2 Comparing Model Accuracy (Bowlers)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning Model** | **Accuracy for predicting PCA\_batsmen (%)** | | |
| **Train (90%) Test (10%)** | **Train (70%) Test (30%)** | **Train (60%) Test (40%)** |
| XGBOOST | 99.39 | 99.69 | 98.35 |
| Random Forest | 99.98 | 99.74 | 99.74 |
| AdaBoost | 94.90 | 91.94 | 92.13 |

In all three-train test split cases, it can be noticed that Random Forest and XGBOOST attain approximately identical accuracy scores. The difference between the prediction accuracy ratings of the two models in the datasets is between 0.20 and 0.30. When processing both datasets with a 70/30 train-to-test ratio, it is possible to conclude that a respectable accuracy score is reached. If the data preprocessing and data cleansing stages are omitted, these scores will plummet significantly. Consequently, if considerable data pre-processing and cleaning have been conducted, the model will not need many adjustments to achieve a high accuracy score. In addition, all models, with the exception of Random Forest, detect a decrease in accuracy when the train is set to 60% and the test is set to 40%, as observed in the data for batsmen. This is because the models contain a small amount of training data for predicting the target variable and a huge amount of testing data. In forecasting the performance measure for bowler (PCA bowler), only the XGBOOST model experienced a 1% drop in performance accuracy for the bowler dataset, but the Random Forest and AdaBoost models maintained the same score. All three models are trained to predict the PCA batsman and PCA bowler performance measure scores for batters and bowlers, respectively. In all datasets, Figures 4.1 and 4.2 demonstrate the comparison between the actual score and the score predicted by the XGBOOST model.

Chart

Description automatically generated**Figure 5. 1 XGBOOST Accuracy for Batsmen Data**

**Figure 5. 2 XGBOOST Accuracy for Bowlers Data**

**Chart, line chart

Description automatically generated**

Figure 4.1 illustrates a comparison of the actual value and predicted value for the target variable "PCA batsmen" by the XGBOOST machine learning model, whereas Figure 4.2 demonstrates a comparison of the actual and predicted values for "PCA bowlers" by the same model. The 70% training to 30% test data ratio for model processing has produced both of these diagrams, yielding an accuracy of 99.96% for predicting PCA batsmen and 99.69% for predicting PCA bowlers. Almost all of the projected values are identical to the actual value of the target variable, confirming the graph's accuracy. In Figure 4.2, the gap between actual and anticipated values for the target variable is discernible between -100 and -150. These figures are then saved in a separate dataframe in order to compare the actual value to the model's prediction. As the value of both performance metrics grows, performance quality also improves. Several of the actual values for PCA batsmen are 49.70, 0.00, and 74.07, and the XGBOOST model predicts values that are comparable to these: 49.26, 0.04, and 73.49. Similarly, the real values for the performance metric PCA bowler are 56.33, 99.34, and 52.58, as opposed to the Random Forest model's projected values of 55.88, 99.10, and 51.59. The similarity between the numbers of batters and bowlers' statistics was readily apparent and explains the 99% accuracy of both models.

Chart

Description automatically generated with medium confidence**Figure 5. 3 AdaBoost Accuracy for Batsmen Data**

Chart, line chart

Description automatically generated**Figure 5. 4 AdaBoost Accuracy for Bowlers Data**

The accuracy scores obtained by applying the AdaBoost machine learning model to the batsmen and bowlers datasets are shown in Figures 4.3 and 4.4, respectively. In contrast to the 99% accuracy rates of Random Forest and XGBOOST machine learning models, this model's accuracy score for the bowler's dataset was 92%, as opposed to 99% for the batsmen's dataset. Figure 4.4 indicates a minor decline in the projected values for PCA bowler compared to the actual values, which are substantially lower and were observed at the graph's peak. In the batsmen dataset, AdaBoost predicted values such as 51.11, 113.94, and 76.07 for actual values such as 49.70, 126.67, and 74.07. Similarly, the projected values for the bowler's dataset are 72.71, 77.38, and 53.71, whereas the actual values are 72.37, 76.83, and 45.46. Compared to the XGBOOST and Random Forest models, AdaBoost has forecast the correct values, with a number of predictions being less than or greater than the actual values.

**Figure 5. 5 Random Forest Accuracy for Figure 5.6 Random Forest Accuracy for Bowlers**

Graphical user interface

Description automatically generated with low confidence**Batsmen Data Data**

Diagram, histogram

Description automatically generated

Figures 4.5 and 4.6 depict a visual comparison of real to predicted values for PCA batsmen and PCA bowlers, respectively, using the Random Forest machine learning model. The graph in red displays the actual values, whereas the graph in blue reflects the algorithm's anticipated values. For the values between -100 and 0, a decrease in PCA bowler prediction is observed. Random Forest has achieved greater accuracy in predicting both the PCA batsmen and PCA bowlers' performance metrics than AdaBoost, which is a boosting technique similar to XGBOOST. Similar to the XGBOOST model, the Random Forest model predicts values that are relatively close to the actual values, with a few forecasts slightly higher or lower than the actual values. For the top two models, Random Forest and XGBOOST, a comprehensive examination of each feature's effect on the model has been conducted. According to an evaluation, for the batsmen dataset, the columns representing total runs and total fours scored by each batsman had the highest feature relevance for the Random Forest machine learning model. The columns representing the total number of wickets taken by each bowler and the bowler's strike rate are the most essential features contributing to the Random Forest model in the bowler's dataset, followed by the bowling average of each player. Similarly, the XGBOOST model ranked innings and runs as the most important features for the batsmen dataset, followed by strike rate and average. Bowling average and wickets were recorded by the XGBOOST model. considered the most significant features for model processing. This investigation has assisted in determining which characteristics contribute most to the forecasting ability of both machine learning models. The identification of high-performing or "greatest" cricketers is an essential aspect of this study. The factors that contribute to a cricketer's high level of performance have been explored in earlier sections of the report. To evaluate high-performing cricketers, the PCA algorithm for batsmen and bowlers considers distinct characteristics. Following Table 4.3 and Table 4.4 are comparisons of the top five players with the highest PCA batsmen and PCA bowler scores, respectively, prior to modelling by each of the three machine learning models employed in this study.

**Table 5.3 Comparison of Performance Metric PCA\_batsmen**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Top five performing batsmen** | | | | |
|  | | | | |
|  | ***Before Modeling*** | | | | |
| Name | SR Tendulkar | KC Sangakkara | ST Jayasuriya | RT Ponting | Dpmd Jayawardene |
| PCA\_Batsmen | 9447.42 | 7211.49 | 6956.52 | 6935.01 | 6360.12 |
|  |  | | | | |
|  | ***After Modeling*** | | | | |
| **XGBOOST** | | | | |
| Name | SR Tendulkar | KC Sangakkara | ST Jayasuriya | RT Ponting | Dpmd Jayawardene |
| PCA\_Batsmen | 8490.46 | 7109.82 | 6807.53 | 6775.93 | 6375.45 |
|  | **Random Forest** | | | | |
| Name | SR Tendulkar | KC Sangakkara | ST Jayasuriya | RT Ponting | Dpmd Jayawardene |
| PCA\_Batsmen | 8490.64 | 7109.82 | 6807.53 | 6775.93 | 6375.45 |
|  | **AdaBoost** | | | | |
| Name | SR Tendulkar | KC Sangakkara | ST Jayasuriya | RT Ponting | Dpmd Jayawardene |
| PCA\_Batsmen | 8489.54 | 7085.36 | 6801.49 | 6770.89 | 6370.44 |

**Table 5.4 Comparison of Performance Metric PCA\_bowler**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Top five performing bowlers** | | | | |
|  | | | | |
|  | ***Before Modeling*** | | | | |
| Name | M Muralidaran | Wasim Akram | Waqar Younis | Wpujc Vaas | B Lee |
| PCA\_Batsmen | 193.48 | 178.97 | 144.90 | 131.02 | 130.37 |
|  |  | | | | |
|  | ***After Modeling*** | | | | |
| **XGBOOST** | | | | |
| Name | M Muralidaran | Wasim Akram | Waqar Younis | GD McGarth | B Lee |
| PCA\_Batsmen | 180.06 | 174.54 | 140.67 | 135.60 | 134.14 |
|  | **Random Forest** | | | | |
| Name | M Muralidaran | Wasim Akram | Waqar Younis | GD McGarth | B Lee |
| PCA\_Batsmen | 180.06 | 174.54 | 140.67 | 135.60 | 134.14 |
|  | **AdaBoost** | | | | |
| Name | M Muralidaran | Wasim Akram | Waqar Younis | GD McGarth | B Lee |
| PCA\_Batsmen | 175.80 | 173.63 | 140.00 | 135.23 | 131.11 |

To explain further, it must be understood that the performance metric score for batsmen and bowlers is based on only six and four player attributes, as defined in the section of this report titled "Defining Performance Metrics." The score of these six criteria that comprise the performance measure PCA batsmen has not yet been broken for these five players.Since the end of S. R. Tendulkar's career, no one has been able to surpass his total number of runs. Given that the total number of runs is an important factor in predicting PCA batsmen, S R Tendulkar will have the most PCA batsmen scores across all three models.The same logic can be applied to the remaining top five batsmen's characteristic values.A few crucial findings in the bowler's dataset reveal a similar tendency. First, both XGBOOST and Random Forest anticipate B. Lee's performance metric to be greater than its actual value. In addition, machine learning models indicate that GD McGarth is one of the highest-performing bowlers relative to Wpujc Vaas based on the actual performance metric score. Several characteristics of the predictive analysis are highlighted below:

* The dataset includes information on all players, from those who began their careers in the 1980s through those who began their careers in 2020. As a result, the models and performance metrics for PCA batsmen and PCA bowlers have forecast the five best batters and bowlers in the history of cricket.
* The purpose of this study is to determine the qualities that make a player "excellent" or a high performer. It may be hypothesised that the correlation of features such as total runs, batting average, total fours and sixes, etc., defined in the performance metrics formula for batters, is significantly correlated with the batsmen's performance quality. Similarly,
* The three machine learning models used in this study are designed to predict the score for the performance metrics PCA batsmen and PCA bowlers.These performance measures are directly proportional to the quality of the player's performance. In Tables 4.3 and 4.4, the highest PCA batsmen and PCA bowler scores are displayed. Upon comparison with the actual values, it was discovered that the models achieved scores that were extremely similar to the actual values of both performance criteria.

In addition, four distinct types of accuracy measure scores (R2 score, mean absolute error, mean squared error, and root mean square error) are computed for each of the three models. These scores are given in Tables 4.5 and 4.6, which compare all accuracy measurement scores for each model with a training-to-test ratio of 70% to 30%.

**Table 5.5 Accuracy Measure Metric (Batsmen)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy Measure Metric** | **XGBOOST Algorithm** | **Random Forest Algorithm** | **AdaBoost Algorithm** |
| R2 score | 0.99 | 0.99 | 0.99 |
| MAE | 4.87 | 6.79 | 7.61 |
| MSE | 160.63 | 340.16 | 444.68 |
| RMSE-Train | 24.89 | 24.89 | 55.84 |
| RMSE-Test | 18.44 | 18.44 | 21.08 |

**Table 5.6 Accuracy Measure Metric (Bowlers)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy Measure Metric** | **XGBOOST Algorithm** | **Random Forest Algorithm** | **AdaBoost Algorithm** |
| R2 score | 0.99 | 0.99 | 0.99 |
| MAE | 1.23 | 1.15 | 1.81 |
| MSE | 3.84 | 3.23 | 8.50 |
| RMSE-Train | 1.03 | 1.07 | 1.90 |
| RMSE-Test | 1.96 | 1.79 | 2.91 |

R2 Score, sometimes referred to as R-square score, is a regression score function that represents how much variance of a dependent variable is explained by an independent variable in a regression model. It is frequently employed to assess the performance of machine learning models. Typically, an R2 score above 0.6 is regarded as a respectable score, with 1.0 being the highest. The R2 for each model utilised in this investigation is same. Moreover, the R-squared score is identical for the two datasets. The interpretation of an R2 score of 0.99 is that the model can account for 99% of the variability of the dependent output attribute, leaving 1% unexplained. Consequently, a greater value of R2 is preferred, as it suggests superior performance. The R2 score is expressed mathematically as follows,

**( 8 )**

where, is the sum of squared regression andefers to the total sum of squares (University).

Mean absolute error is the average absolute difference between observed and anticipated values. Given that the actual average for the target variable is 409.07 and 67.58 for PCA batsmen and PCA bowler, respectively, this average error between the predictions and actuals in the dataset for all models can be regarded as a good value. The mean absolute error can be formally defined as follows,

MAE = []

**( 9 )**

where, represents the total number of data points and represents the difference between prediction and true value (Cort J. Willmott, 2005).

Mean Squared Error, sometimes referred to as Mean Squared Deviation, quantifies the amount of error in statistical models by analysing the average squared difference between observed and predicted values. Mean Squared Error scores for the bowler's dataset for Random Forest, XGBOOST, and AdaBoost are in the range of 1. Mean Squared Error scores for the batsmen's dataset for all models in this study range from 150 to 450. Due to the fact that MSE is not scaled, these scores may be large due to the large values included in the dataset. Mathematically, it can be expressed as follows,

MSE = [

**( 10 )**

where, represents the total number of observations, signifies the ith observed value and represents the corresponding predicted value (Murphy, 1998).

RMSE, or Root Mean Squared Error, is the standard deviation of anticipated value errors. In other words, it indicates the degree of data concentration around the line of best fit. In this study, RMSE has been determined for each Batsmen and Bowler dataframe's train and test datasets. As the train and test RMSE scores are fairly close, with a difference of less than 30 for the batsmen dataset and less than 3 for the bowlers dataset, it can be concluded that the model is not overfit. Weijie Wang and Yanmin Lu conducted an important study that demonstrates how rounding off score values affects the RMSE and MAE scores by introducing a new method of rounding off that can improve prediction accuracy scores (Lu, 2018). It can be mathematically expressed as follows,

RMSE =

**( 11 )**

where, represents the total number of data points, signifies the actual observations and represents the estimated observations.

# Limitations

This study intends to address the research question "Which characteristics contribute to predicting the next greatest cricketer?" by determining the characteristics that contribute the most to the performance score of batsmen and bowlers." The availability of additional characteristics in the data, such as weather, pitch conditions, ground, etc., could affect the identification of the most influential aspects of each player's performance. The gameplay of cricket has changed over time, which may have an effect on a cricketer's performance attributes. Over the years, the number of sixes and foures in ODI cricket has increased significantly. Because of the limitations of the features revealed in the data used for this study, it may be suggested that incorporating additional factors could significantly increase the likelihood of detecting accurate characteristics affecting cricketers' performance.

# Conclusion

This study seeks to offer a straightforward yet easy method for predicting cricket players with superior performance. This study is not confined to players' statistics for a specific innings, such as an IPL or ODI; instead, the dataset incorporates information on many attributes of each player throughout their career. Two datasets are created, one for identifying features of batsmen and the other for identifying characteristics of bowlers in the game of cricket. On the basis of these two categories of players, a number of performance metrics are explained, which are then employed in the performance metric formula to evaluate each player's performance based on a particular attribute. The proposed performance measurements and machine learning models are straightforward and may be applied directly to the types of data seen in cricket. After identifying the performance of each player with performance measures, the dataset is processed by three distinct machine learning models, which are subsequently compared based on their accuracy scores and performance evaluation metric scores, such as R-square, MAE, MSE, and RMSE for each model. On the basis of this comparison, it is determined that the XGBOOST and Random Forest algorithms outperform the AdaBoost algorithm. Similar studies can be conducted for comparable cricket analysis, including T20, ODI, and IPL matches. By analysing the numerous qualities of cricketers, this study addresses the research question, "Which characteristics contribute to predicting the next great cricketer?" Based on the performance metric formula for bowlers and batters, the following traits play a significant part in a player's high-quality performance:

1. Total number of runs scored by batsmen and bowlers during the game.
2. The number of averages achieved by the batsmen and bowler
3. The number of strike rates achieved by the batsmen and bowler
4. During the game, a single bowler took the most wickets.
5. Total number of half centuries and centuries scored by the batsmen
6. Total number of fours and sixes scored by the batsmen
7. The bowler's economy rate throughout the game.

Furthermore, XGBOOST and Random Forest forecasted bowler and batsman performance measures with the highest accuracy. According to these models, the characteristics that contribute the most to the prediction of players' performance metrics are the ability to hit a six and a batsman's high total runs and innings played. Similar to XGBOOST and Random Forest machine learning models, features such as strike rate, bowling average, and the capacity to take the most wickets influence an accurate forecast of a high-quality bowler. Finally, the lists of performance evaluation metrics for all three models are shown, where a unique pattern of identical R-square values may be detected. In all three models, the R-square performance evaluation metric achieved a score of 0.99. The R-square score near 1 indicates that the model is able to accurately predict the test data values. In some instances, it may also indicate that the model is overfit.

Certain steps could be taken during advanced research to determine whether the models are overfit or underfit. The remaining metrics are compared and evaluated based on their accuracy scores for each model. Using optimization approaches such as hyperparameter tweaking or exhaustive search, the models used in this study can be further analysed in future studies. Such optimization techniques permit the minimization of the cost function, which represents the difference between the true value of the estimated parameter and the value projected by the model. Quantifying performance in sports is a difficult and crucial undertaking, especially in competitive sports where team selection is crucial to winning or losing. Introducing three additional types of players—all-rounders, fielders, and wicket keepers—allows for the unquestionable completion of a comprehensive performance analysis. All-rounders are players who are proficient at both batting and bowling, whereas fielders are responsible for catching, collecting, and returning the ball after it has been struck by the batsman. Fielders are both hitters and bowlers, and their study can assist in identifying the performance characteristics of a superior fielder. The wicketkeeper is an important member of the fielding team because he stands behind the batsmen and has the ability to knock the ball to the stumps if the batsman is out of the boundary at the right time.The inclusion of these categories can enable a more in-depth study of all players, facilitating optimal squad selection decisions. Predictions about the best performing cricket players may be made based on the performance indicators used in this study.Furthermore, efforts can be made to improve the AdaBoost algorithm's accuracy in predicting bowler and batsman performance in cricket.

The importance of evaluating players for team selection in cricket competitions necessitates extensive data collection and statistical analysis. The prediction of high-quality performance cricketers from this study could be tested in a real-world setting. This would include calculating the performance of young players in 2016–2017 (or any X-year period) and forecasting the top five performing cricketers using the machine learning models employed in this study. Later, the accuracy of these models might be validated by identifying the top five forecasted cricketers in the ICC's top ten ranking board for the next five years, from 2018 to 2022. The ICC, commonly known as the International Cricket Council, is the global governing body of cricket, where each year the performances of international cricket players are ranked using a points system (ICC, 2022).

# Appendix A: Batsmen



Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

# Appendix B: Bowler

Graphical user interface, text, application, email

Description automatically generated

Text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application

Description automatically generated Text

Description automatically generated Graphical user interface, text, application

Description automatically generated Text

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

# References

Ahmad, H. a. (2017). Prediction of Rising Stars in the Game of Cricket . 4104-4124.

Ananda B. W. Manage, S. M. (2013). An Introductory Application of Principal Components to Cricket. *Journal of Statistics Education*, 22.

C. Deep Prakash, C. P. (2016). A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers. *International Journal of Computer Applications*.

Cort J. Willmott, K. M. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 79-82.

CricketRules. (2022). *Cricket Rules*. Retrieved from Cricket Rules | The Rules of the Greatest Game of all: https://www.cricket-rules.com/

Croucher, J. (2000). Player ratings in one-day cricket. *Fifth Australian Conference on Mathematics and Computers in Sport*, 95-106.

ESPNcricinfo. (2022). *Statistics*. Retrieved from ESPNcricinfo: https://stats.espncricinfo.com/ci/engine/records/index.html

Hai-Bang Ly, L. S.-A. (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. 15.

ICC. (2022, October 01). *Team Rankings explained*. Retrieved from ICC: https://www.icc-cricket.com/rankings/about#:~:text=The%20MRF%20Tyres%20ICC%20Player,of%200%20to%201000%20points

James Cochran, J. B. (2017). *The Oxford Anthology of Statistics in Sports.*

Livingston, F. (2005). Implementation of Breiman’s Random Forest Machine Learning Algorithm. *ECE591Q Machine Learning Journal Paper*.

Lords. (2017). *The Pitch*. Retrieved from Lords: https://www.lords.org/mcc/the-laws-of-cricket/the-pitch

Lu, W. W. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*.

Murphy, A. H. (1998). Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. 2417-2424.

NIMMI WEERADDANA, ,. S. (2021). UNIQUE APPROACH FOR CRICKET MATCH OUTCOME PREDICTION USING XGBOOST ALGORITHMS. *Journal of Theoretical and Applied Information Technology* .

Pandey, K. P. (2018). INCREASED PREDICTION ACCURACY IN THE GAME OF CRICKET USING MACHINE LEARNING. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*.

Research, W. a. (2007). Research Ethics in the Real World. In M. J. Tony Long, *Research Ethics in the Real World.* Elsevier Health Sciences.

Sachin Kumar S, P. H. (2022). DATA SCIENCE APPROACH TO PREDICT THE WINNING FANTASY CRICKET TEAM—DREAM 11 FANTASY SPORTS.

Sahu, A. (2021). Predictive Analysis of Cricket. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*.

Schamschula, A. T. (2018). Cluster Analysis Applied to the Greatest Barbadian. *International Journal of Mathematical Education*, 87-93.

Tianqi Chen, C. G. (2016). XGBoost: A Scalable Tree Boosting System. 785–794.

TU Chengsheng, L. H. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*.

University, N. (n.d.). *Coefficient of Determination, R-squared.* Retrieved from https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

Wisconsin–Madison, U. o. (n.d.). *Random Forests.* Retrieved from https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf