PREDICTING PEAK PERFORMANCE OF A CRICKET PLAYER USING MACHINE LEARNING AND DATA ANALYTICS

by

Akif Azam
19101165
Tarif Ashraf
19101195
Ahmad Al Asad
19101196
Kazi Nishat Anwar
19101209
Ilhum Zia Chowdhury
19101214

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
September 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Akif

Akif Azam
19101165

Tarif

Tarif Ashraf
19101195

Ahmad Al Asad
19101196

Kazi Nishat Anwar
19101209

Ilhum Zia Chowdhury
19101214

Approval

The thesis/project titled "Predicting Peak Performance of a Cricket Player using Machine Learning and Data Analytics" submitted by

1. Akif Azam(19101165)
2. Tarif Ashraf(19101195)
3. Ahmad Al Asad(19101196)
4. Kazi Nishat Anwar(19101209)
5. Ilhum Zia Chowdhury(19101214)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September 25, 2022.

Examining Committee:

Supervisor:
(Member)

Tanvir Rahman

Tanvir Rahman
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In the modern era, the game of cricket has evolved into a batting friendly lexicon more than ever. However, bowlers adapting to every suitable condition can also change the dynamics of the game. Prior studies were carried out, mostly focusing on team combinations and batting analytics but did not highlight the batter and bowler's potential. This paper seeks to understand the conundrum behind this impactful performance by determining how much control a player has over the circumstances and generating the "Effective Runs" and "Effective Wickets," two new measures we propose.We first gathered the fundamental cricket data from open source datasets. However, variables like the pitch, weather, and control were not readily available for all matches. As a result, we compiled our corpus data by analyzing ball-by-ball commentary of the match summaries that led us to determine the control of the shots played by the batter as well as deliveries that were in control by the bowler. Our dataset comprised seven renowned international cricketers. For batters we prepared the dataset, encoded, scaled, and split the dataset to train and test Machine Learning Algorithms and predict the impact the player will have on the game. Multiple Linear Regression and Random Forest give the best predictions accuracy of 90.16% and 87.12%, respectively. On the other hand, for bowlers, we upscaled the wickets taken by the bowler and set a threshold accordingly. Given that the threshold was met, we concluded that the effective wickets taken by the bowler were impactful with regards to the overall match performance. Machine Learning classifiers were trained to predict this impact of a bowler. The best individual accuracy result was provided by Logistic regression for the Spinners at 73.21% and SVM Classifier for the Seamers at 79.17%. However, the overall best average precision for both types of players was observed at 78.75% by Logistic Regression.

Keywords: Corpus Dataset; Machine Learning; Cricket; ODI; Commentary Analysis; Prediction; Classification; Regression

Dedication

We dedicate this research to our loving family.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, we would like to convey our gratitude to late Mr. Hossain Arif, Assistant Professor, CSE, Brac University, for his constant guidance, which has considerably aided us in moving forward with our research. We express our heartfelt condolences on our professor's untimely passing.

And finally, to our supervisor Mr. Tanvir Rahman sir for his kind support and advice in our work. He helped us whenever we needed help.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

BF	Balls Faced
DNB	Did Not Bat
FN	False Negative
FP	False Positive
kNN	k-Nearest Neighbors
MLR	Multiple Linear Regression
ODI	One Day International
rbf	Radial Basis Function
SR	Strike Rate
SVC	Support Vector Machine Classifier
SVM	Support Vector Machine
SVR	Support Vector Machine Regressor
$T20$	Twenty Twenty
$TDNB$	Team Did Not Bat
TN	True Negative
TP	True Positive

Chapter 1

Introduction

1.1 Game of Cricket

1.1.1 History

Amid the seventeenth century, two groups of 11 men gathered in Sussex for a competitive duel [1]. This duel later transitioned to be a major source of entertainment throughout various age groups across Britain. All that was needed for this contest was a set of “stumps, a ball and a wooden bat” in the center of a circular ground. The game’s objective was to prevent the attacking team from hitting the stumps alongside making sure the defendants hit the ball as far away as possible. This was a major foreshadowing of the game renowned today as “Cricket” worldwide.

This game of cricket, over time, has gone through fundamental changes. The game once played for pleasure has now emerged as an international enigma. One of the key parts of the game is the number of points scored, which is called “runs” in cricketing grammar. This challenge is to be completed over a set number of balls “bowled,” where every six balls bowled are referred to as an “over.” Additionally, a designated number of overs form an “Inning” that varies in different formats of the game played.

1.1.2 Game Mechanism

Cricket is played between two teams and is divided into two Innings. In each inning, one of the teams “bat” – defending the stumps – while the other team “field” – attacking the stumps with the ball. The batting team sends out two “batters” at a time; one of them is the “striker,” facing the “bowler,” while the other stands on the non-striking end. Each over is bowled by a bowler. There is a “wicketkeeper” from the fielding team behind the striker’s stumps, and the rest of the players are “fielders.”

The batting team attempts to score as many runs as possible. Several ways to score runs are: hitting the ball away and running between the wickets, each counts as a run; hitting the ball far enough to cross the boundary. Going over the boundary results in six runs, whereas going across results in four for the batting team. Furthermore, if the batting team runs between the wickets despite hitting the ball, extras known as “byes” are added to the team total. In addition, extras are also conceded in rare cases by the bowling side whenever the bowler makes a mistake, for example, no-ball and wide ball.

On the other hand, the challenge for the fielding team is to constrict the runs to a minimum. While in the process, the goal of the bowlers is to eliminate the playing batter by taking his “wicket,” which can be accomplished in several ways: hitting the stumps; catching the ball after a direct hit from the striker’s bat – “caught out”; hitting the striker’s leg when it is in front of the stumps – “leg before wicket” (LBW). Hitting the stumps has its own set of variations: the bowler directly bowls and hit the stump – “bowled out”; the wicketkeeper breaking the stumps with the ball when the striker is out of his crease – “stumped out”; any of the fielders breaking the stumps when the batters are running between the wickets – “run out.” A rare way of getting out is “hit wicket” when the batter himself hits the stumps mistakenly.

1.1.3 Playing Formats

The game is now played in 3 major formats, “Test Cricket,” “One Day International (ODI),” and “Twenty- Twenty (T20)”. Test cricket is the epitome of the sport transpiring over five days. Here the teams generally bat and bowl for two innings. Each day is transcribed into three sessions which comprise 30 overs. The only way to win in this format is for the fielding team to take all the batting team’s ten wickets; otherwise, the game is drawn as the batting side’s innings is not complete. Thus, the fielding team has to take more wickets while giving away fewer runs. Another popular format is ODI which comprises two innings each of 50 overs. For the batting team, the goal is to score as many runs as possible since the number of overs is limited. Last but not least, the most entertaining and exciting format that transformed cricket’s dynamics – the T20 format. In this electrifying format, each team plays 20 overs, similar to the two innings rule of ODI cricket.

1.2 Motivation

In recent times, cricket has spread worldwide within different age groups. The level of interest in the game is at its summit. Currently, 106 countries are now involved with cricket. Within this, 12 countries are full members, while the remaining 94 are associates [2]. Moreover, the game’s popularity is continually rising each year, indicating the evident globalization of the sport. Furthermore, according to Dave Richardson, ICC chief executive, in [3], claims have been made about cricket being featured in the 2028 Olympic Games in Los Angeles since 1900. Quoted from [3], Greg Barclay, the ICC chair, “We have more than a billion fans globally, and almost 90 per cent of them want to see cricket at the Olympics…whilst there are also 30 million cricket fans in the USA.”

As the sport continues to be globalized, there is a massive pool of possible players, at a professional level, worldwide. The biggest stage of the game is the world cup. This is a 50-over format of the game. According to [4], The World Cup event is one of the most-watched sporting events in the world. Here, the players are constantly striving to give their best. The players are constantly striving to perform better at every game level. This format is so challenging for the players because of the number of balls played, added by several factors. To start with, there are a minimum of 8-12 international teams that a player needs to play against. Out of these, the high-ranked teams' bowlers pose a big threat to the opposition.

One of the essential aspects of the game is the "pitch," being of several pitch types. Firstly, "green pitches" or "damp pitches" favor the bowlers. These pitches comprise uneven surfaces, including a thin layer of grass or has moisture, resulting in the ball speeding up and skid through once it hits the pitch. Therefore, this pitch gives the fast bowlers a competitive edge over batters, as the ball will swing and bounce more than usual. Batters scoring on this type of pitch will have to prove their skills, playing proper shots with control following their basic cricket grammar. Secondly comes "dead pitches" or "flat tracks," continuously rolled, a more batter-friendly variant. Batters are prone to score more in flat tracks because there is a lack of bounce and turn for the bowlers to take advantage of. The ball does not grip the surface, making this condition unfavorable for bowlers. Challenging as it may, skillful bowlers will still get wickets and put pressure on the opposition, proving their importance to the team and highlighting their performance. Lastly comes "dry pitches" that are often likely to be dusty, consisting of a soft surface, which lets the ball grip and turn alongside uneven bounce, making it the best condition for spinners to perform. Subsequently, the batter has a moderately hard time adjusting to such conditions; however, the batter scoring the most runs will be the performer of the game.

In addition, another condition that allows us to define a player's impact depends on the weather condition. Weather factors such as wind conditions, humidity, the surrounding temperature, etcetera play a vital role in determining how a player performs. For example, windy conditions are ideal for bowlers to make the ball swing, hence a bowler-friendly environment. However, due to the moisture content, humid conditions reduce the cohesion and make the pitch weak. This results in a greater struggle for batters. Temperature varies in different parts of the world. Adapting to different temperatures is a part of the player's growth cycle. Subcontinental players prefer a warmer condition, while the opposite is applicable for those outside the subcontinent.

A batter has to have the adequate skill to go against such line-ups and score as much as possible . Additionally, ODIs take place in various venues throughout the year. Different venues prepare different kinds of pitches for matches to take place. A batter has to be versatile enough to be able to play on all green, dry, and flat pitches. These pitches might or might not favor the batter, thus testing a batter's ability. Given that this is a format of 300 balls, a batter has to organize his innings and be calculative in selecting which shots to play against the variation of deliveries he is facing. Moreover, facing such deliveries, the batter also has to be aware of the strike rate, the number of times he's hitting the ball with the middle part of the bat, and choosing between which ball to play or leave. A batter deserves to be acclimated whenever they perform sublimely in adverse conditions, thus having an "Impact" on the game's outcome. The impact, in this case, reflects not only the runs scored but also the amount of "Control" they have had. This control is a better representation of the impact of the performance. Most of the past works are used to predict player performance [5] or identify the best team [6]. In contrast, our research seeks to assess a player's influence while considering his control in all sorts of situations. This paper aims to decipher the dilemma behind this impactive performance through our research on how much control a player has over the situation and using different Machine Learning algorithms and compare their results to generate "Effective Runs." To summarize, our proposal is unique in that it considers the control of batters in measuring the player's impact on the overall match.

Many studies have been performed to determine a batter's effectiveness across all formats; however, very little research has been conducted on a bowler. Our goal is to create a corpus dataset and using those data it helped us to create our measure. The pitch and weather significantly impact a bowler's effectiveness in a particular game. Thus, considering all the factors, we have developed a new measure that can depict a bowler's impact more comprehensively: the control for a bowler depending on the runs conceded. Using this control, we are upscaling the wicket score and, upon finding this, are categorizing these values. Given our threshold, we can predict whether the bowler was impactful or not. This will eventually help the team management and analysts to find the best-suited bowler. Our research aims to analyze the datasets, identify a bowler's influence, and assess several classification methods' confusion matrix, accuracy, and precision.

The impact performance is decided by how much "control" a cricketer has over the course of the match being a batter or a bowler. This will be the key to selecting a player in the team combination. Furthermore, this impactive performance directly correlates to the game's outcome and determines the direction the game will progress in. All of this brings us to our scheme, which, employing machine learning and data analytics, will benefit in predicting when a player reaches his best in his career. This proposed scheme is not confined to the selectors only but also the entire team management committee, for instance, coaches, franchise owners, sponsors, etcetera.

Therefore, the dilemma remains:
How can we propose a better scheme that can measure this impactive performance of a particular cricketer?

1.3 Aims and Objectives

We aim to decipher the dilemma behind this impactive performance through our research on how much control a player has on the situation by using Machine Learning algorithms such as Multiple Linear Regression (MLR), Polynomial Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, k-Nearest Neighbors, and Logistic Regression and compare their results to generate a metric that incorporates "Effective Runs" and "Effective Wickets."

The main motive of our research is to develop a better-performing model that can be used to predict the player's impact on the game. Individual analysis of batters and bowlers is done using different features like pitch type, weather, opposition, temperature, and the playing style (control) in these situations. Machine Learning is the key to determining a player's impact. Distinct models and data sets will be used to analyze players of different categories. A new parameter called "Effective Runs" for batter and "Effective Wickets" for bowlers to find the influence of a player on the game's result, for which the following objectives are to accomplished:

1. To provide relevant data to the player selection committee to choose a batter and a bowler who will be preferable in a specific pitch, weather, and ground. These data were collected by analyzing the post commentary of each match.

2. We used machine learning algorithms which imposed a better performing meter to assess a batter and bowler's performance or control. For batter, the control was calculated by adding the number of deliveries that were middled or left alone divided by the total number of balls faced. Similarly, for bowlers this control was based on the runs conceded per delivery to determine the fraction of deliveries controlled by the bowler.

3. We classified the batter's and bowler's impact after considering many features that have a positive correlation with impact. This impact was based on the Runs Scored for the batter and Wickets Taken for the bowler and the Control measurement.

Chapter 2

Literature Review

In this current world of entertainment, irrespective of the sport in context, classification, and ranking of players has become an important factor for many and has thus brought to life a vast area of interest for many researchers.

This chapter of the paper aims at proving the importance of research in the field of cricket through reviewing previous relevant works in the field.

S. Akhtar, P. Scarf, and Z. Rasool [7] developed a new player rating system. The authors used Multinomial Logistic Regression on 104 test match data. To measure each player's total impact, they modelled match result probabilities using data from each session's batting, bowling, and fielding performances. The greatest player in a game, a series, or a year might then be determined using the proposed contribution technique that rates players over time.

Jhanwar, M.G., Pudi, V. in [8] incorporated used game conditions and the team's strengths to measure each player's contribution. Focusing on a total of 786 matches of the top 9 ODI-playing teams (2017), they proposed the "Work Index," that shows how much work a team still has to complete in order to reach its goal. This assessed each player's performance, allowing for consistent comparisons of players inside and across positions. The authors obtained an accuracy of 86.80% when they predicted the player of the match award for 51 ODI matches from 2006 to 2016. This was done to further validate their methodology.

H. Saikia and D. Bhattacharjee [9], using a Multilayer Perceptron (MLP) Neural Network, predicted the performances of batters by analyzing their performance from the first three seasons of IPL. The authors further calculated the actual performances of the batters from the fourth season and got an accuracy of 66.67% for their model. This model could help the selectors decide which batters to buy for their team.

A. Kaluarachchi and S. V. Aparna [10] developed a software tool CricAI by analyzing how factors like home game advantage, day/night effect, winning the toss, and batting first affect the outcome of the match using Bayesian classifiers in Machine learning. This tool has applications in increasing the chances of victory through simple tweaks in certain factors in the game.

T. B. Swartz in [11] used statistical analysis, utilizing data from 427 International ODI matches from the 1990s, to conclude that there is no competitive advantage in winning the coin toss and also the log-odds of the probability of winning increases by around 50% when playing on one's home-field.

S. R. Iyer and R. Sharda [12] used neural networks to forecast each cricketer's future performance based on data from their previous performances. Cricketers were divided into performer, moderate, and failure. The authors progressively trained and evaluated their neural network models by collecting data on players from 1985 to the 2006-2007 season and dividing it into four sets of data. These models forecasted the cricketer's performance in the near future. Recommended cricketers for the 2007 World Cup were identified using the ratings obtained and heuristic methods.

M. Shetty et al.; in [13] used a model for predicting player performance and finding the best all rounder player while focusing on factors like pitch type, weather, ground, opposition, and several extra features. ODI data for several Indian cricketers were used to train and test on Logistic Regression, Support Vector Machine (SVM) Classifier, Decision Tree, and Random Forest where Random Forest resulted in the best outcome. The authors' model received 76%, 67%, and 95% accuracy for batters, bowlers, and all-rounders respectively, which was then used to select the best combination for the Indian Cricket team.

P. Somaskandhan et al.; [14] aimed to identify the set of qualities that have a significant influence on the outcome of a game. Employing statistical analysis and different machine learning algorithms while minimizing the use of domain knowledge, SVM gave the best accuracy, which was then used to examine possible combinations of different features to find the set with the highest accuracy. The result, with an accuracy of 81%, was the set of attributes: high individual wickets, number of bowled deliveries, number of the thirties, total wickets, wickets in the power play, runs in death overs, dots in middle overs, number of fours and singles in middle overs.

M. Bailey and S. R. Clarke [15], in an effort o predict the results of ODI cricket matches, a number of factors were developed that could each individually account for statistically significant percentages of the variation related to the anticipated run totals and match results. The match outcome was predicted using a Multiple Linear Regression model with data from 2200 ODI matches played before January 2005. Prediction variables were numerically weighted based on statistical significance.

M. Khan and R. Shah [16] used Data Mining techniques to find the parameters that play a vital role in forecasting the outcome of an ODI cricket match and to measure the accuracy of the prediction. They investigated the statistical relevance of variables such as home field advantage, winning the toss, game strategy, match type, competing team, venue familiarity, and season the match is played in. The authors employed Logistic Regression on previously played match data to determine which factors contribute to prediction. For model training and prediction analysis, SVM and Nave Bayes Classifier were utilized. Comparative analysis was done from the various sets of models represented using graphical representation and confusion matrices. A bidding scenario is also taken into account to clarify the decisions that may be made after the model has been developed. The effect of this option on the model's cost and payback is also investigated.

Shah, P. [17] introduced a new measure for an individual performance called "Quality," describing the importance of the opposition faced in measuring a player's performance. For example, taking into account the runs scored by the batter the bowler is bowling against, or the number of wickets taken by the bowler the batter is facing, the author proposed a potential method at identifying good batters or bowlers based on player-vs-player information.

M. K. Mahbub, M. A. M. Miah, S. M. S. Islam, S. Sorna, S. Hossain, and M. Biswas [18] researched to determine the starting 11 for the Bangladesh (ODI) cricket squad. Their primary objectives were to identify potential cricket team members who would be effective, evaluate each player's strengths, and rate the individuals. They created their own scoring systems for bowlers and batters, respectively. And the players are chosen for the team if they pass a particular mark. With the Support Vector Machine, they could predict the team with 94% accuracy for the batter and 93% for the bowler.

In 2018, A. I. Anik, S. Yeaser, A. G. M. I. Hossain, and A. Chakrabarty [5] researched choose the best players using machine learning, which was based on past playing records. From these players, the winning team combination was then determined. They used feature selection algorithms to find out the attributes that related to the output feature. Then machine learning models such as Linear Regression, and Support Vector Machine was used to predict the runs scored by a batter and runs given by a bowler. Moreover, they have also deployed a Neural Network in the bowler dataset to find the performance comparison.

In 2019, N. Rodrigues, N. Sequeira, S. Rodrigues, and V. Shrivastava [19] offered a technique for choosing players that considers their performance against specific opposition. The model will use regression to predict both the batting and bowling measures. These metrics can be incorporated into the player selecting procedure. The dataset used to train the model is a player's prior performance against a specific opponent. The model takes into account the opposition and the match venue. Based on the input fields, a rank-wise list of all the batters and bowlers is produced, which the selectors can use to choose the squad according to the desired combination.

V. V. Tharoor and N. Dhanya [20] from India used Exploratory Data analysis on the performance of the Indian Cricket team. Several data visualization techniques were used to compare and contrast the statistical data for Batter, Bowler, Captaincy, and National record. The study concluded an increase in the performance of the Indian cricket team over time by taking into account the overall team performance, win-loss ratio, successful captaincy and more. The Random Forest Classifier had the best accuracy relative to the other classifier models tested to demonstrate the effect of the number of overs on the match result.

E. Mundhe, I. Jain, and S. Shah [21] created a web application to do predictive analysis on a live T-20 match in order to forecast the result and the winner of the match before it starts. The Multivariate Polynomial Regression technique has a 67.3% accuracy rate, which means that 6 out of 10 times the actual score came within the projected score range. The system was used to predict the runs scored for a live match at the end of the 20th over. The accuracy of the Random Forest Classifier algorithm in predicting the outcome of the game using past data is 55%.

S. Priya, A. K. Gupta, A. Dwivedi, and A. Prabhakar [22] aimed to compare the analysis of several machine learning algorithms for predicting the winning team. Numerous supervised classification algorithms are implemented, including Logistic Regression, Random Forest Regression, k-Nearest Neighbor and Support Vector Machine, Naive Bayes, and Decision Tree to achieve this goal. They discovered that the Random Forest Classifier gave them the model's most outstanding accuracy (74%).

D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, S. Vasudevan, V. Veeramani Kannan, and S. Sagubar Sadiq [23] predicted the result of an ongoing IPL match. The prediction was implemented using various machine learning algorithms such as Gaussian Naive Bayes,Support Vector Machine, k-Nearest Neighbor and Random Forest. Various models were used at different point in time of the match with necessary feature selection. Different teams had varying accuracy with a certain model.

Lastly, in a study by M. A. Pramanik, M. M. Hasan Suzan, A. A. Biswas, M. Z. Rahman, and A. Kalaiarasi [24] several non-ensemble and ensemble classifiers were used to analyze their performance in predicting match outcomes in Bangladesh Premier League (BPL) T20 matches. They predicted match results in two ways: using only pre-match features which gave highest accuracy of 64.58% from kNN and also by forecasting match outcomes based on all historical data, including post-match features which gave an accuracy of 93.39% from Gradient Boosting Algorithm.

From the above discussion, we see that other than traditional ways of determining the performance of players, like how many runs have been scored by the batter or how many wickets have been taken by the bowler, there is a vast pool of features and attributes that play a crucial role in defining the performance of a player. External conditions such as pitch type, weather, and opposition plays their role in the game outcome. Nevertheless, there has been little research on how the control of the batter in bowler-friendly conditions or the control of a bowler in a batter-friendly condition affects the performance and outcome of the game. Hence, the approach in predicting when a player will reach his best performance should not be restricted to a small domain alone, but exploring more variables should be the next stage at any improvement in this area of study.

Chapter 3

Methodology

Our primary purpose is to predict the impactive performance of a cricket player in certain conditions. Our first step is to collect data. This includes collecting performance data of different world-class players over their careers. This data corresponds to figures for their performance in the ODI format. Our data would include all aforementioned conditions affecting a player's performance as features. We generated two separate corpus dataset for bowlers and batters. Then we will use Machine Learning Algorithms to predict the impact of a batter and a bowler. We train the data and then, test using actual data, and finally compare the algorithms.

3.1 Workplan

Our initial phase starts with comprising readily available data from open source databases. Then we manually generated more data by analyzing the commentary for individual games. We integrated a threshold to the variables to place weights on the values, while we derived some new variables from the previously collected data. For bowlers, this new measure was then generated with a threshold set to categorize a player being impactful or not impactful. For batters, we generated a measure that can represent the impact of a batter. Classification Models are infused upon bowlers while regression models are applied on batters, which were then used to predict this impact. Figure 3.1 provides a high-level view of the model design.

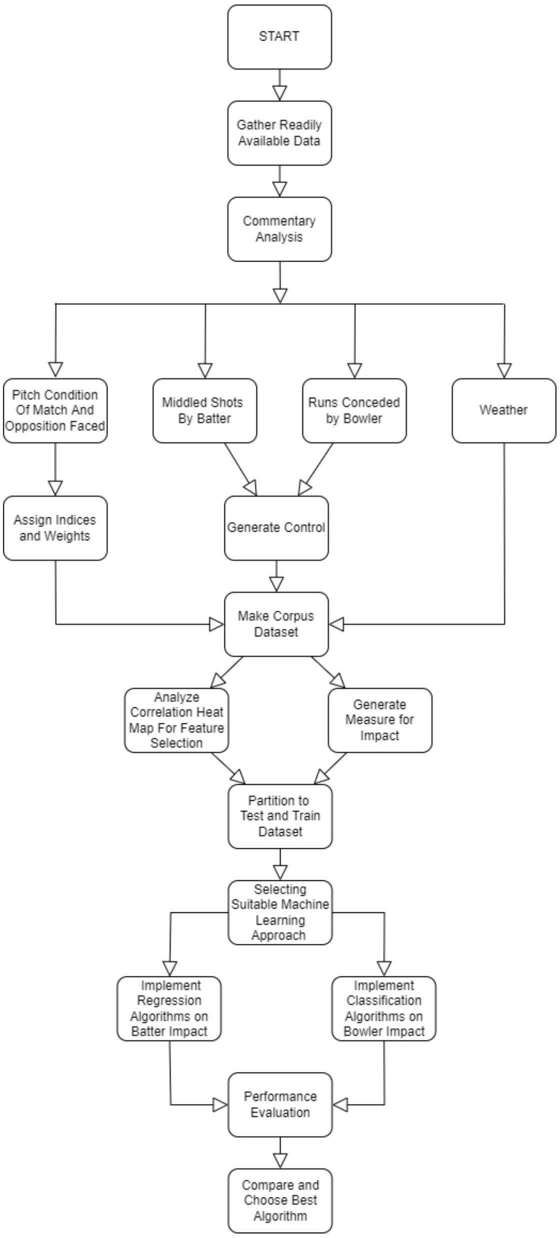

START
Gather Readily Available Data
Commentary Analysis
Pitch Condition Of Match And Opposition Faced
Middled Shots By Batter
Runs Conceded by Bowler
Weather
Assign Indices and Weights
Generate Control
Make Corpus Dataset
Analyze Correlation Heat Map For Feature Selection
Generate Measure for Impact
Partition to Test and Train Dataset
Selecting Suitable Machine Learning Approach
Implement Regression Algorithms on Batter Impact
Implement Classification Algorithms on Bowler Impact
Performance Evaluation
Compare and Choose Best Algorithm

Figure 3.1: Proposed workplan for the research.

The conditions mentioned earlier, such as opposition, pitch type, and many others, will be major features in the collected data. Data collection leads us to the imperative task of Data Preprocessing: it is important to format and scale all related data and make it feasible. Next comes generating a correlation heatmap. According to the data, this will help us determine which features affect the player's performance. We will devise an Impact Formula that will act as our dependent variable using these features.

Our next stage includes splitting the data into test and train sets. First, the train data is used to train our Machine Learning Algorithms that we plan to use. Then we use our trained models to predict the impact of a player. This will then be validated with actual test data on the player.

The performance of each model in predicting a player's effect will then be compared and evaluated in order to decide which is the best accurate classification model for bowlers and regression model for batters.

Chapter 4

Dataset

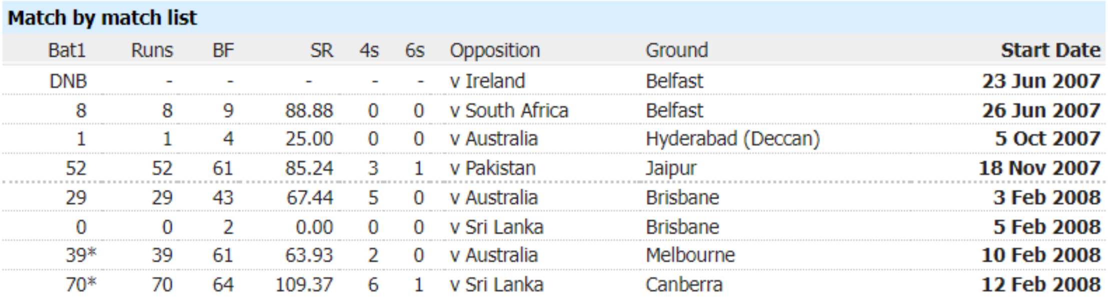
4.1 Data Collection

There is a lot of research being done in cricket because it is such a global and popular sport, and it goes beyond just selecting the best players for a squad [6] to predict the performance of cricket players [5]. Data collection in this field can be both easy and challenging, depending on the form of research. General statistics such as Runs Scored or Balls Faced are available on websites such as Espncricinfo. On the other hand, more in-depth variables such as Control, Pitch reports, etc. of many matches are not available to the general public unless it was a significant factor in raising the hype during the actual match event. Our research area combines both the first and the latter cases. We did our preliminary data analysis by collecting data for the Indian cricket batter Rohit G Sharma over his entire ODI career from June 23, 2007, till February 11, 2022, a total of 230 cricket matches. Then we further strengthened our models and deductions using two more batters. The Australian batter David A Warner had played a total of 128 ODI matches in his career from January 18, 2009 till November 29, 2020. Lastly we introduced the New Zealand prodigy, Kane Williamson who played a total of 151 ODI matches in his career from 10 August, 2010 till 13 March, 2020. For bowlers, we chose 2 fast bowlers and 2 spinners based on their number of matches played and their performance in diverse conditions. We chose the Australian fast bowler Mitchell A Starc who played a total of 99 ODI matches in his career from 22 October, 2010 till 26 July, 2021. Next comes a promising fast bowler from New Zealand, Trent Boult, who has risen through the ranks by showing his tremendous bowling prowess. He played a total of 93 ODI matches in his career from 11 July, 2012 till 26 March, 2021. For spinners, we included the Pakistani spin wizard, Saeed Ajmal, who played a total of 113 ODI matches in his career from 2 July, 2008 till 19 April, 2015. Lastly, we included the veteran spinner from India, Ravichandran Ashwin, who played a total of 113 ODI matches in his career from 5 June, 2010 till 21 June, 2022. This brings the total number of matches with data gathered for both bowlers and batters to 927. Some of the variables were generated, while others were manually determined and recorded.

The first phase of data collection for both batter and bowler was taken from Espncricinfo which has its own database, Statsguru, which contains all the general stats for a cricketer. Furthermore, from individual player's commentary analysis we have included new features and created further formulas to predict the impact.

4.1.1 Batter

For batters,we entered queries to show match by match batting data in all venues - home, away, or neutral - over individual batters' ODI career. Here [25], we collected the data with the following variables: "Bat1," "Runs," "BF," "SR," "4s," "6s," "Opposition," "Ground" and "Start Date." Bat1 contained values like "DNB," "TDNB," "52" and "30*" where DNB means the player did not bat in that game, TDNB means the team did not bat, 52 means the player got dismissed after scoring 52 runs, and 70* means the player scored 70 runs and was not dismissed by the end of the match. Runs contain the runs scored in that match, and BF is the number of deliveries faced by the batter. SR is their strike rate, 4s and 6s are the numbers of 4s and 6s scored, respectively. Opposition is the team the player played against, and the Ground is the venue played on, with the Start Date being the date of the game.



Match by match list

Bat1	Runs	BF	SR	4s	6s	Opposition	Ground	Start Date
DNB	-	-	-	-	-	v Ireland	Belfast	**23 Jun 2007**
8	8	9	88.88	0	0	v South Africa	Belfast	**26 Jun 2007**
1	1	4	25.00	0	0	v Australia	Hyderabad (Deccan)	**5 Oct 2007**
52	52	61	85.24	3	1	v Pakistan	Jaipur	**18 Nov 2007**
29	29	43	67.44	5	0	v Australia	Brisbane	**3 Feb 2008**
0	0	2	0.00	0	0	v Sri Lanka	Brisbane	**5 Feb 2008**
39*	39	61	63.93	2	0	v Australia	Melbourne	**10 Feb 2008**
70*	70	64	109.37	6	1	v Sri Lanka	Canberra	**12 Feb 2008**

Figure 4.1: Statsguru database for batters from stats.espncricinfo.com.

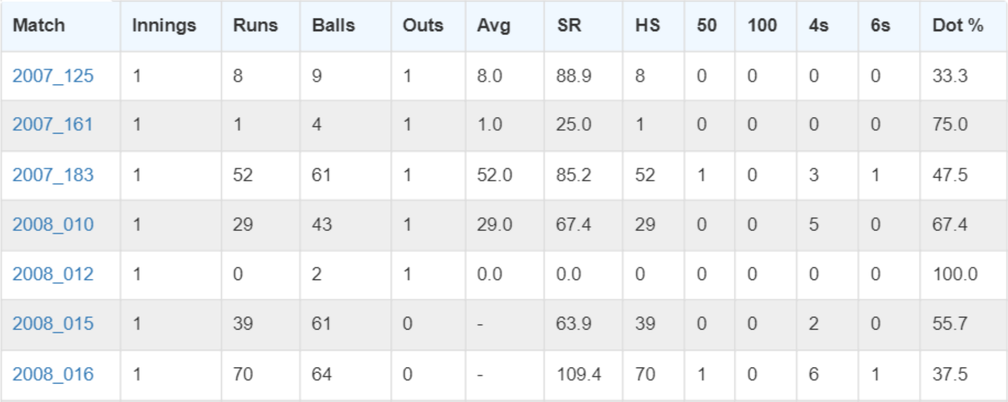
After this, we used the collected variables to generate some more variables for our model. We calculated the "Others" scored, which contains running between the wickets like singles, doubles, etc. We used the 4s, 6s, and Runs data to calculate "Others."

$$Others = Runs - (4s * 4) - (6s * 6) \tag{4.1}$$

This Others data was then used to calculate the "Running between the Wickets fraction."

Next, we collected more data that was not readily available in any database. We looked at the results, and "Win/Loss," "Team Runs," and "In at Position number" data were collected individually for each of the 230 matches we are working with. Using the data for Ground we previously collected, we checked each of the venues and categorized them into three: Home, Away, and Neutral. Home is when the venue is in the same country as the player's team, in essence, India; Away is in the country of the opposition team, and Neutral is when neither is the case.

We then used another source, Cricmetric [26], for data collection. This has another database that contains an additional statistic called the Dot Ball percentage. We search for Sharma's stats grouped by matches for his ODI career. The number of deliveries faced does not always result in scoring runs. Dot ball percentage depicts the percentage of deliveries that the batter faced without scoring any runs. From this percentage and the number of balls faced, we can calculate the number of Dot



Match	Innings	Runs	Balls	Outs	Avg	SR	HS	50	100	4s	6s	Dot %
2007_125	1	8	9	1	8.0	88.9	8	0	0	0	0	33.3
2007_161	1	1	4	1	1.0	25.0	1	0	0	0	0	75.0
2007_183	1	52	61	1	52.0	85.2	52	1	0	3	1	47.5
2008_010	1	29	43	1	29.0	67.4	29	0	0	5	0	67.4
2008_012	1	0	2	1	0.0	0.0	0	0	0	0	0	100.0
2008_015	1	39	61	0	-	63.9	39	0	0	2	0	55.7
2008_016	1	70	64	0	-	109.4	70	1	0	6	1	37.5

Figure 4.2: Cricmetric database from www.cricmetric.com.

Balls. Using this variable, we can now determine the Scoring Shot: Balls Faced - Dot Balls, the number of deliveries for which the batter scores runs. We created another variable, Scoring Rate, determined by dividing the Runs scored by the Scoring Shot. There is a difference between the Strike Rate and the Scoring Rate. Strike Rate is the Runs scored per Balls Faced, whereas Scoring Rate is the Runs scored per Scoring Shots, excluding the Dot Balls in the game.

The next statistic column was about the strike rate, calculated by dividing the number of runs scored by the number of balls faced. As the data included the strike rate as a percentage, for our model implementation, we converted them into ratios of 1.

$$Strike\ Rate = \frac{Runs\ Scored}{Balls\ Faced} \tag{4.2}$$

In addition, we deduced other areas of data needed for the model implementation that we have previously discussed. These criteria include the Dot Ball and its percentage, Scoring Shot, Middled, Left Alone and finally, the control percentage, which has been explained previously. The data from cricmetric contained the Dot Ball% only. This is used to determine the Dot Balls using the following equation:

$$Dot\ Balls = Dot\ Ball\% \ * \ \frac{Balls\ Faced}{100} \tag{4.3}$$

We also measured the scoring rate, which is implemented by the equation below:

$$Scoring\ Rate = \frac{Runs\ Scored}{Scoring\ Shot} \tag{4.4}$$

Where,

$$Scoring\ Shot = Balls\ Faced\ \ -\ \ Dot\ Balls \tag{4.5}$$

We have introduced another column known as "Others," which is the other runs scored without boundaries. Another column is running between the wickets, which is implemented by the equation below:

$$Running\ Between\ the\ Wickets\ \ = \frac{Others}{RunsScored} \tag{4.6}$$

In the next column, we have introduced the result and labeled "Win" if a team wins the match and "Loss" if a team loses the match. The term "No Result/Draw" is also labeled, and it means the match is either cancelled due to weather conditions as a "Storm" or "Bad Light," whereas the term "Draw" emphasizes if a match is tied. Moving on, The next two columns are basically "Team Run", the run scored by the team in that particular innings, and "In@Pos" which is defined as the batting position the batter bats in.

4.1.2 Bowler

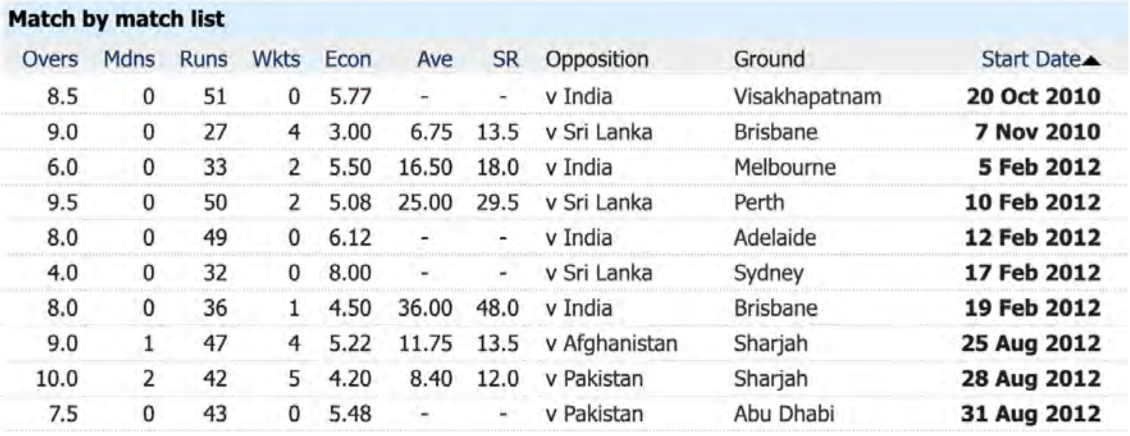
For bowlers, we divided our preliminary data collection strategy differently for the different types of bowlers. We chose 2 fast bowlers and 2 spinners based on their number of matches played and their performance in diverse conditions. We chose Mitchell A Starc and Trent A Boult as fast bowlers while Saeed Ajmal and Ravichandran Ashwin were our spinners and collected their entire ODI career statistics - a total of around 400 matches. The database Statsguru, by Espncricinfo [25], provided the bowling records of "Overs," "Maidens," "Runs," "Wickets," "Economy," "Average," "Strike Rate," "Opposition," "Ground," and "Start Date." These data were additionally complemented by collecting some extra data which were not readily available, such as - "Win/Loss," "0s," "1s," "2s," "3s." Some matches directly mentioned the number of dot balls, "0s," in that match, but for other cases the individual runs conceded per delivery were collected from the commentary. The aforementioned variables were then used to derive some more important features. "Overs" contained the number of overs bowled, but this was broken down to calculate the number of balls bowled in that match. "Ground" contained the venue data and hinted at the data for "Home/Away." The columns in the preliminary dataset are - "Overs," "Maidens," "Runs," "Wickets," "Economy," "Average," "Strike Rate," "Opposition," "Ground," "Home/Away," "Start Date," "Total Balls Bowled," "Win/Loss," "0s," "1s," "2s," "3s," etc.

The main criterias to judge the effectiveness of a bowler in a particular match are number of wickets, bowler strike-rate, economy and average. The average number of deliveries bowled every wicket is the definition of a bowler's bowling strike rate. A bowler's ability to swiftly take wickets increases with a decreasing strike rate. The economy rate is the average number of runs conceded every bowled over. In most circumstances, the lower economy rate indicates the bowler is performing better. The quantity of runs a bowler has given per wicket taken is known as their bowling average. The bowler is doing better when their bowling average is lower.

$$Strike\ Rate = \frac{Number\ of\ deliveries\ bowled}{Total\ Wickets\ Taken} \tag{4.7}$$

$$Economy = \frac{Runs\ Conceded}{Total\ Overs\ Bowled} \tag{4.8}$$

$$Average = \frac{Runs\ Conceded}{Total\ Wickets\ Taken} \tag{4.9}$$



Match by match list

Overs	Mdns	Runs	Wkts	Econ	Ave	SR	Opposition	Ground	Start Date▲
8.5	0	51	0	5.77	-	-	v India	Visakhapatnam	**20 Oct 2010**
9.0	0	27	4	3.00	6.75	13.5	v Sri Lanka	Brisbane	**7 Nov 2010**
6.0	0	33	2	5.50	16.50	18.0	v India	Melbourne	**5 Feb 2012**
9.5	0	50	2	5.08	25.00	29.5	v Sri Lanka	Perth	**10 Feb 2012**
8.0	0	49	0	6.12	-	-	v India	Adelaide	**12 Feb 2012**
4.0	0	32	0	8.00	-	-	v Sri Lanka	Sydney	**17 Feb 2012**
8.0	0	36	1	4.50	36.00	48.0	v India	Brisbane	**19 Feb 2012**
9.0	1	47	4	5.22	11.75	13.5	v Afghanistan	Sharjah	**25 Aug 2012**
10.0	2	42	5	4.20	8.40	12.0	v Pakistan	Sharjah	**28 Aug 2012**
7.5	0	43	0	5.48	-	-	v Pakistan	Abu Dhabi	**31 Aug 2012**

Figure 4.3: Statsguru database for bowlers from stats.espncricinfo.com.

Figure 4.3 depicts the numbers generated by one of the players we researched on Mitchell Starc, from a period of 13 matches. Generally, a bowler is considered to be in rhythm when his economy is around 6.00, indicating he gave away a run every ball to the opposition. The lower the economy, the better the performance. From the figure above, it can be observed that Starc had an overall economy of 3.00 on November 7th, only offering 3 runs per over he bowled, the best figures he produced in that time frame. Furthermore, the lower a bowler's strike rate, the more frequently he or she fulfills the primary purpose of bowling, which is to take wickets. It is seen that on August 28th, Starc had a strike rate of 12.0, which showed he got a wicket every 12th ball. On the other hand, some of the rows are blank in which he didn't manage to bag any wickets. Furthermore, the average of a bowler goes on to showcase his class. On November 7th Starc only let the opposition score 6.75 runs per wicket he took whereas his worst performance was when he leaked 36.00 runs per wicket.

4.2 Corpus Data

For data collection, we opted to choose derived data collection, where we used the readily available raw data. We had to convert some of the existing data points from various data sources to create new data for data collection. This derived data provided new insights as we combined it with other information, which helped us reach a definitive conclusion. For some of the data which were not readily available, we had to read the commentary, other tabular data, or the match results and analysis to deduce new data points and formulas.

The next stage of data collection required us to make our corpus data. We proposed a quantifiable way to weigh each "Opposition" played against. A major factor in the performance of a player is the type of pitch they play on. This performance varies with the playstyle of the player. Similarly, "Weather" is a game-changer where both the batter and the bowler may benefit if they can plan their game properly. Our work focused hugely on the "Control" of the player. For a batter, this control was generated from how individual deliveries from the bowlers turned out. How well he middled the ball and whether he left unplayable balls that would threaten his wicket. A bowler's capability is judged by their performance against the top-rated ODI teams. Similarly, for a bowler, the control was generated on the basis of how less expensive he was throughout the match. If the bowler bowled a dot ball, he didn't concede any run hence an entire delivery was considered as control, for 1 run conceded we recorded half of the delivery was in control and a quarter for 2 runs. Any delivery conceding 3 or more runs were resulted as not in control at all.

4.2.1 Opposition

We accumulated the ICC ODI team ranks for each opposition team. Next, we assessed their threat levels towards the players. We accumulated the ICC ODI team ranks for each opposition team for each match. Next, we assigned an index to each rank, designed based on the threat level teams set forth for batters. However, the team rankings fluctuate every so often and it is challenging to differentiate the weight between consecutive ranks like 2nd and 3rd or 7th and 8th etc. To reduce this disparity, the team ranks were divided into groups of 3. The top 3 ranked teams pose an equal threat, for which we assigned them a weight of "5". Similarly, "4" is given to teams ranked from 4 to 6, "3" given to teams ranked 7-9, "2" for teams ranked 10-12, and "1" for teams ranked 13 and below.

4.2.2 Pitch

"Form is temporary, class is permanent" is an open secret in the game of cricket. A batter in good form will manage to prevail amidst a lineup of bowlers and will score runs despite the conditions he is playing in, showcasing his class and consistency. On the other hand, even if the conditions are dire and the pitch doesn't provide much advantage, a bowler in form will flourish against a lineup of batters, with the change of variations and line and length of the deliveries he bowls, changing up the tempo and thinking one step ahead every time to outsmart the batter. One of the biggest factors that correlate with the consistency of both the batter and bowler is the pitch that they face the opposition on. Over time there have been many debates among experts discussing the differences in pitches in different countries and how it impacts the game.

The pitch of a cricket game is the main strip where the game is played and stretches 22 yards. The type of pitch determines the type of game it will be, so batters and bowlers have to prepare themselves accordingly. We have organized our data based on 3 main features that uphold critical characteristics of the game. According to the article [27] we have categorized pitches into three: A green top, a dusty or dry pitch, and a dead or flat track.

Collecting the data for all pitches was not straightforward. In fact, generalized assumptions were necessary at times. Espncricinfo was the base for the data accumulation. The commentary alongside toss review, team lineups, and weather report is the primary source for pitch details. The “pitch report” section often directly referred to the condition of the pitch where the commentators discussed it before the start of play. At times the information we were seeking was not at the beginning. Thus, we resorted to the aftermath of the match, where the captains of each side made remarks on the overall dynamics of the game during the post-match presentation, often commenting about the pitch in their dialect.

However, there were cases where the pitch was not mentioned throughout the match’s overall summary. These situations were dealt with with some strict assumptions that corresponded with the nature of home and away pitches and how it played out generally throughout the years. Additionally, we also analyzed the batmen and bowling performances to back our assumptions. For instance, if spinners took the bulk of the wickets in a given match, it can be inferred that the pitch was dry and gripping, which resulted in a turning wicket. Further presumptions were made based on performances of pace bowlers, top-order and lower-order batters.

The performances of batters and bowlers vary considerably with different pitches. A flat track, for example, is the most suitable for the batter , while it is always the hardest for a bowler to perform in, irrespective of their bowling style. In contrast, a green top is a bowler-friendly pitch and is the most challenging for a batter to perform. In contrast, seamers or fast bowlers perform better in green pitches, whereas spinners perform better in dry pitches. A batter showing high performance in a green top indicates that the batter has more control over the deliveries he faces. A dry pitch tests the batter as well as gives them an equal opportunity to perform. Hence, keeping these as a basis, we encoded the pitch types for each type of bowler and batter in the Table 4.1.

Table 4.1: Proposed Pitch Indices.

Pitch Type	***Batter***	***Bowler (Seamer)***	***Bowler (Spinner)***
Green Top	2	1.5	1
Dusty or Dry	1.5	1	1.5
Dead or Flat	1	2	2

The primary sources for the pitch details were the commentary and post-match presentation. In some situations with no details, assumptions were made regarding the nature of the pitch based on the general location and venue. For example, with a few exceptions, subcontinent pitches are mostly flat tracks that don't offer any pace or bounce. The Wankhade stadium in Mumbai is known as a “bowler's grave” because of the lack of movement for bowlers that offers no swing or spin. Adding to that the stadium is too small to even save runs while fielding, proving it to be a child's play for any good batter to score continuous boundaries.

4.2.3 Weather

Weather is one of the most crucial aspects which can favor either a batter or a bowler. The weather, the toss, and other small factors all have a significant impact on how a cricket match turns out. When selecting a choice after winning the toss, captains frequently take the weather into account. The batter can score more runs if there is good weather and clear skies. On the other hand, the bowling side can benefit from a cloudy or windy atmosphere.

Weather conditions have been derived into four categories: “Clear,” “Sunny,” “Windy,” and “Overcast.” The adaptability of a player in any of these categories will define their performance. The weather information was collected from the pre-match analysis commentary where the initial weather report was considered. An important aspect to note is that conditions might change overtime as the day progresses. We took into account the weather that was right at the beginning of the first ball. Although there had been some ambiguity, the data was kept mostly consistent in regards to some keywords, such as: “hot and humid” for sunny, “breezy” for windy, and “cloudy and dark skies” for overcast. First of all, the batters are always relieved when the sky is clear and blue since the ball doesn't swing as much under these conditions. Even while a clear sky has little bearing on seam bowling, if the sun is pounding down strongly on the field, it often quickly evaporates all of the pitch's moisture, turning it into a batter's paradise. Because of this, clear days tend to see more runs scored than days that are cloudy and gloomy.

Even though green pitches are rare in modern cricket, if the weather is warm, the grass will also dry up pretty rapidly, helping the batters. However, the ball swings and glides in the air under cloudy circumstances, which is why bowlers prefer to bowl in them, especially fast bowlers. However, as opposed to clear sky, these circumstances favor quick bowlers. Again, it benefits the bowlers if there is rain or if the humidity is high. The pitches also have a lot of moisture in them when it rains, which fast bowlers may take advantage of. In addition, in such circumstances, the pitch takes a while to dry up, giving the bowlers enough chances to attack the batters. While the wind assists the spinners in turning the ball in the air, a rainy pitch is the worst nightmare for any spinner. The moisture from the pitch handicaps the spinners' ability to grip the ball in any way, which results in loss of control in regards to line and length when bowling. Generally, a flat pitch is disadvantageous for a spinner, but a flat track with cracks on it is always a sight that encourages them as they can exploit the pitch for additional turn and movement.

4.2.4 Control

One of the most prominent factors determining the player's peak performance is "control". Control basically indicates how a batter confidently played along with the innings with finesse and graceful timing. It shows the quality and skill of analyzing the balls faced. A batter is in control of a delivery when they can sway the outcome of that delivery the way they want. This is mostly when the batter strikes the ball with the middle part of their bat, which we are calling "middled" deliveries, or when the batter intensionally refuses to strike a delivery, "left alone" deliveries. The basic formula for determining the control of a batter, then, is as shown in Equation 4.10 [28].

$$Control_{\text{Batter}} = \frac{Middled \ + \ Left \ Alone}{Balls \ Faced} \tag{4.10}$$

Espncricifno uses this to generate the control of batters when they do a noteworthy performance in a match. This is where we faced some difficulties. Espncricinfo only provides the control statistics for a player and makes it publicly available when that player is either man of the match or has a noteworthy performance. Out of all the 230 matches, we had 53 available control statistics. So we had to find the control for the remaining matches ourselves by doing delivery by delivery commentary analysis. So initially, we did the commentary analysis on 29 of the 53 available matches from the dataset of Rohit Sharma and compared our calculated control with the actual control available at Espncricinfo. The comparison is expressed in a graph shown in Figure 4.4.

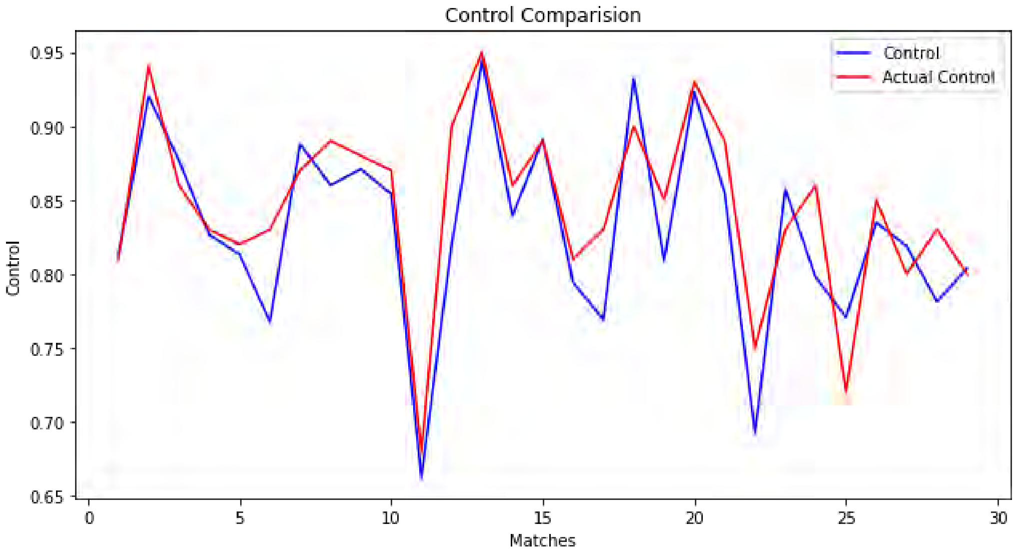


Figure 4.4: Comparison of Actual Control and Calculated Control.

From this figure, we can differentiate the Actual Control and the Control we calculated by our commentary analysis. We used this to calculate the uncertainty in our commentary analysis and deduce that our calculated control has 3.24% uncertainty with the actual control. thus proving our commentary analysis to be of high accuracy with low uncertainty. This low uncertainty depicts that our measurements are more precise.

We performed a commentary analysis on various batter's innings to determine the calculated control. In order to find the control, our commentary analysis included tallying all the middled and left alone shots and excluding the shots which were edged and not connected properly with the bat. For this analysis, we mainly focused on basic cricketing shots which require the batter to connect the shots properly. Some commonly used adjectives and the words "middled" and "left alone" were used in some cases. For left alone, some common keywords used were: "left alone," "ducked," "stepped away," "moved away," " no shot offered," "watches into keeper's gloves," "brings his bat down on it," "let go" and "shoulders arm." For middled we searched for some keywords such as: "middled," "defended," "nudged," "clipped," "swayed," "drives firmly and straight," "nibbled," and "controlled" and adjectives such as: "clobbered," "wonderful," "magnificent," "amazing," "smashed it," "clips it," "timed-to perfection," etc. which described the shots. We had to use our intuition to analyze how the shot was played in some cases. The shots which were excluded were mostly edged shots, even if the batter would hit the ball for "4" or "6". The types of edges are top edge, inside edge, outside edge, and some keywords used are: "poor timing," " leading edge," "not in control," etc. Other than edged shots, there were also missed shots where the batter attempted to strike but failed. Keywords for this category included "missed," "tried to steer it", "failed to make contact," "beaten," etc.

Similarly, in order to proclaim our measure to predict a player's impact, we had to go through a number of approaches for a bowler's control. It was considerably harder to generate bowler control than batter control. Initially ,we took the data and formula from Espncricinfo. It was required to identify all deliveries during which the batter has control before subtracting the total control from 1.This would give the bowler's control.Here [11], a bowler's control is defined as the deliveries that are not in control of the batter. Being a batter-friendly game, this measure of control hugely benefits the batters and does not highlight a bowler to their best. Using this measure,the bowler control never surpassed 0.4, even when they performed admirably well.

Next, we tried to read the commentary for bowler's and tried to find in which delivery the bowler is in control, just as we did for the batter. However, due to lack of information in the commentary section we also couldn't proceed with this method as well. For example, at which length the ball landed, whether the ball swung or not was not given for all deliveries. After reading the commentary, it was quite difficult to draw any conclusions about whether the bowler had control.

So, we put forward our own measure at generating a control, based on the runs conceded in each delivery. From the match summary we found the number of 4s and 6s conceded by the bowler, but for the number of 1s, 2s, 3s and 0s (dots) we had to read from the commentary and generate a tally count for all the matches.

$$Control_{Bowler} = \frac{0s \ + \ 1s * 0.5 \ + \ 2s * 0.25}{Total\ Balls\ Bowled} \tag{4.11}$$

We considered a dot ball with 0 runs conceded as an entire delivery in control, for 1 run conceded we say half of the delivery was in control and a quarter for 2 runs. Any delivery conceding 3 or more runs were resulted as not in control at all. Using the aforementioned approach, we determined the “Control” for each of the 400 matches and updated the previous dataset. Then we combined the data for the two seamers into one dataset and the two spinners into another separate dataset.

4.3 Impact

Our novelty includes creating new features which basically defines the impact for both batters and bowlers. For bowlers we introduce a new scale to determine the effective wickets taken by a bowler while considering the control with each delivery and then we set a threshold to classify the bowlers as “impactful” or “not impactful.” We implemented Machine Learning Classification Algorithms to predict this “Impact.” For batters we have introduced a new parameter of a player’s performance that incorporates both runs and control of a batter, eventually leading to the “Effective Runs” scored, which we are stating as “Impact” thus implementing Machine Learning Regression Algorithms to predict it.

4.3.1 Effective Runs

Table 4.2: The Impact Formulae in a Progressive Way.

First Formula	SR * $e^{(2*Control)} * Pitch\,Index * Opposition\,Index$
Second Formula	SR * $e^{(2*Control)}$
Final Formula	Runs Scored * $e^{Control}$

The impact formula was reached through a series of experiments. Our initial approach was to use some commonly used variables in cricket terms and generate a heatmap. The variables included “Strike Rate,” “Runs,” “Control,” “Opposition Ranking,” “Opposition Index,” “Pitch Index.” From analyzing the heatmaps in Figure 4.5, we progressively reached a formula. Table 4.2 shows the formulae that we came upon step by step.

Our first formula included the variables: Strike Rate, Control, Pitch Index, and Opposition Index. Then we excluded the Pitch and Opposition Indices and decided to keep them as independent variables. Our heatmaps show a weak correlation - close to 0 - between Opposition Ranking and Opposition Index with Strike Rate and Runs. The Pitch Index shows a weak negative correlation of around 0.2 or below. The exponential was used to highlight the significance of the control on the impact. Finally, the strike rate was replaced with runs because the strike rate can vary even if the runs scored are relatively low. We initially decided to double the control exponent to increase the significance of the control. But then we decided to remove the 2 from the formula since we are multiplying it with the Runs. We do not want the impactive Runs to be too high. The final formula we reached is:

$$Effective\ Runs\ =\ Runs\ Scored\ *\ e^{\text{Control}} \tag{4.12}$$

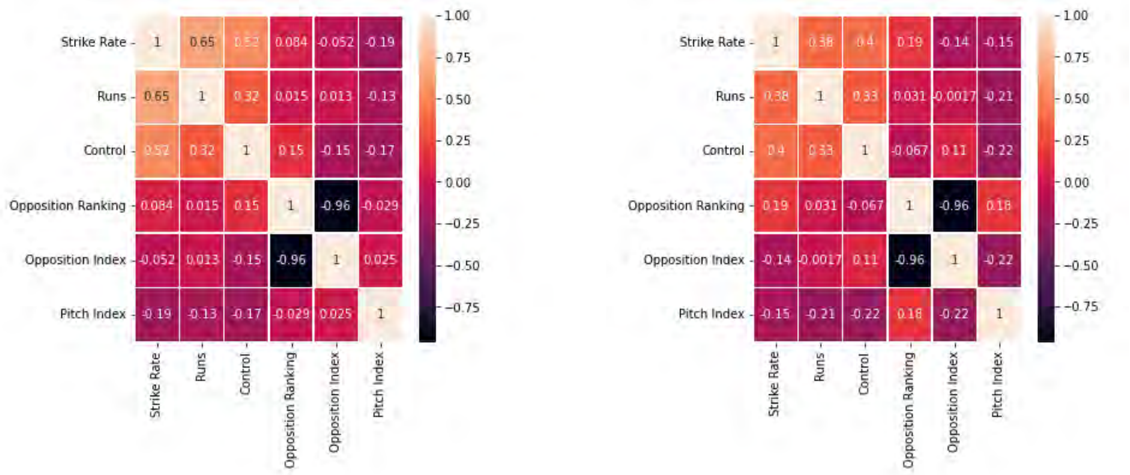


Figure 4.5: Heatmap of Important Variables over Rohit Sharma's and David Warner's career.

To explain, if someone has more Left Alones, then their runs will be less, concurrently less impact. If someone's Runs and Control are high, their impact will be high. When someone scores 70 Runs and has a control of 80% or 0.8, then according to our formula, the Impact is 155.8. Hence, even though they scored only 70 Runs, their high control leaves an impact of 156 runs on the match.

4.3.2 Effective Wickets

A significant part of a bowler's score in a cricket match is the number of wickets taken. To assess the impact of a bowler, it is crucial to take into account the number of wickets they take. But there are instances when despite taking 1 or no wickets, they give a highly economical performance where they concede low runs in more deliveries. It is challenging to consider such games, hence we opted to upscale the number of wickets taken with respect to our previously generated control exponentially. This way, we can make a formulation of "Effective Wickets" and produce a scale for impact by taking into account both the runs conceded and wickets taken as shown in Equation 4.13.

$$Effective\ Wickets \ = \ Wickets\ Taken \ + \ e^{\text{Control}} \tag{4.13}$$

This "Effective Wickets" was then rounded down to their nearest integers, since the number of wickets cannot be a continuous variable. This is when we set up a threshold for our scale. If a bowler takes 1 wicket, but due to their high control in the deliveries, the effective wickets taken turn out to be more than 2, then we are categorizing it as the bowler being impactful in the game. Essentially, the closer the effective wickets taken is to the original wickets taken, the lower the impact. To make it feasible, we set a threshold for our scale as follows - effective wickets taken less than 3 are "not impactful," and effective wickets taken equal to or greater than 3 are "impactful." This adds another column "Impact" in our dataset where 0 represents "not impactful" and 1 represents "impactful."

4.4 Pre-processing

We first imported the necessary libraries required to implement our machine learning algorithms. The main libraries include “Numpy,” “Pandas,” “Scikit learn,” “Matplotlib,” and last but not the least “Seaborn.” Numpy is used to implement multidimensional arrays and matrices. Scikit learn is used to implement the machine learning algorithms and the necessary steps required to implement the result. Matplotlib is used to implement the graphical plotting and heatmaps required to visualize the result. Seaborn is used to implement correlation between the features, and a graph can be used to visualize the correlation between the features. Pandas provide a simple data frame option that is used for data manipulation.

For bowlers, non-numerical features include “Opposition,” “Ground,” “Home/Away,” “Pitch,” “Weather.” Same features are included for batters as well except with an additional inclusion of “Out/NotOut.” The rest of the numerical features include, for batters; “Bat1,” “Runs,” “Balls Faced,” “Strike Rate,” “4s,” “6s,” “Opposition Rank,” “Opposition Index,” “Dot Ball %,” “Dot Ball,” “Scoring Shot,” “Middled,” “Left Alone,” “Control,” “Scoring Rate,” “Others,” “Running btw %,” “Team Run,” “In @ Pos#,” “Pitch Index,” “Impact.” For bowlers; “Overs,” “Maidens,” “Runs,” “Wickets,” “Economy,” “Average,” “Strike Rate,” “Opposition Rank,” “Opposition Index,” “Balls Whole,” “Balls Decimal,” “Total Balls Bowled,” “0s,” “1s,” “2s,” “3s,” “Control,” “Pitch Index” and “Impact.”

Data with missing values were first removed from consideration. The missing values were mostly from matches where the bowlers did not bowl, batters did not bat or the match got canceled for unforeseen reasons. So, discarding these data was sensible.

As with many non-numerical features, we have used the term encoding to implement in our model. Firstly we used label encoding in “Out/NotOut” to determine whether the batter was Out or Not Out. Secondly, we have imposed one-hot encoding on three specific columns. The three specific columns are “Home/Away,” which specifies the venues, “Pitch,” which says the pitch type, and “Win/Loss,” providing the outcome of the match.

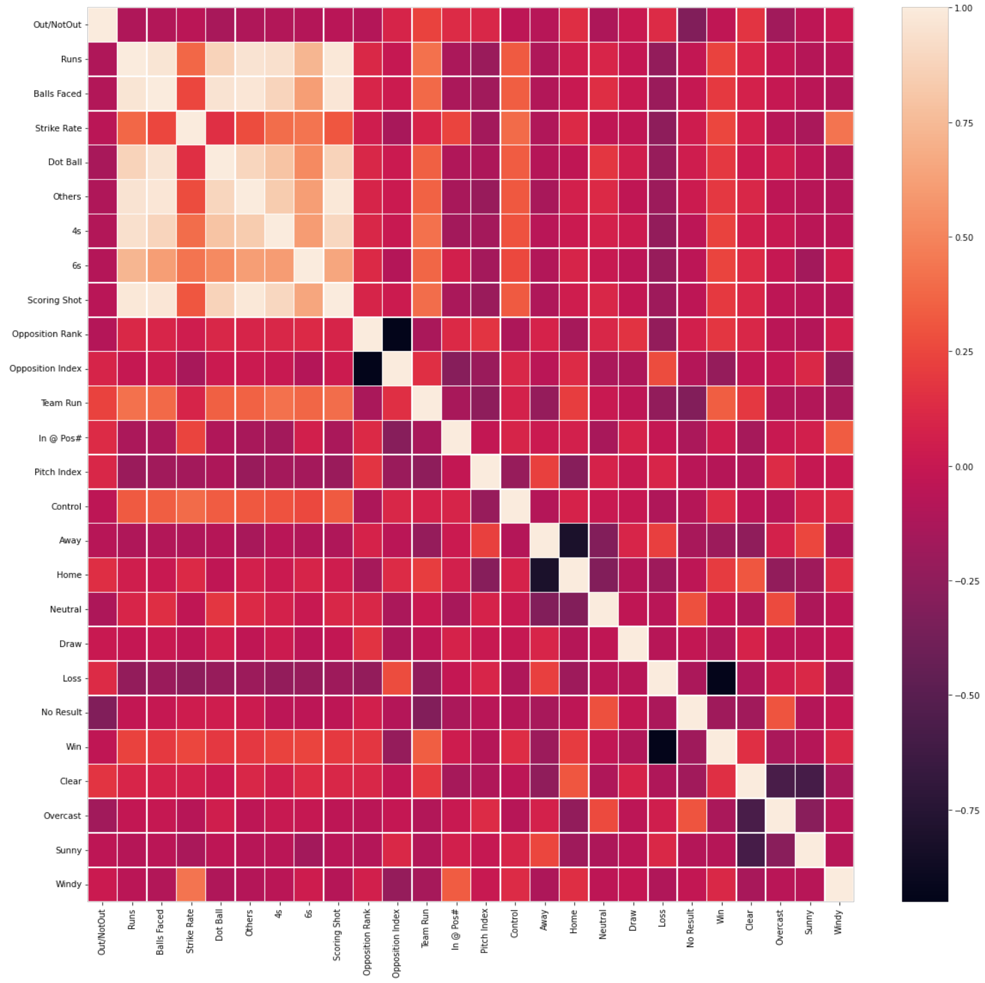


Figure 4.6: Correlation heatmap of variables over David A Warner's dataset.

Several variables were eliminated to reduce the dimension. We generated a correlation heatmap for each of the bowler types as well as the batters as shown in Figures 4.6 and 4.7. For bowlers variables like Wickets, Runs, Overs, Economy, etc., are directly connected with the formula for Effective Wickets. Other variables such as Strike Rate and Average had strong correlations with Economy. For batters, variables such as Runs, Strike Rate, 4s, 6s are connected with the formula for Effective runs. When determining the opposition index, two columns were taken into account - the opponent team, and the opposition ranking - which were eventually eliminated as features.

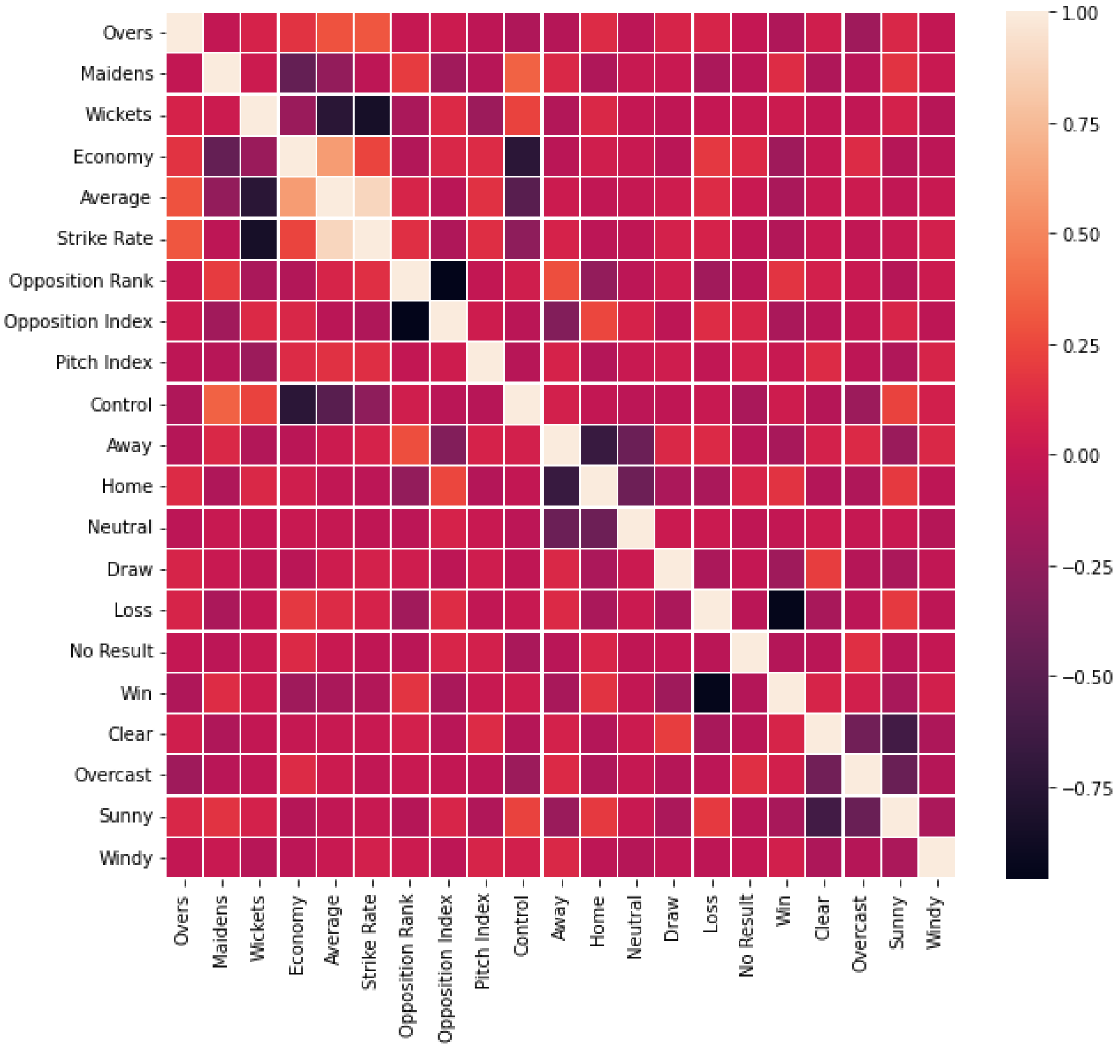
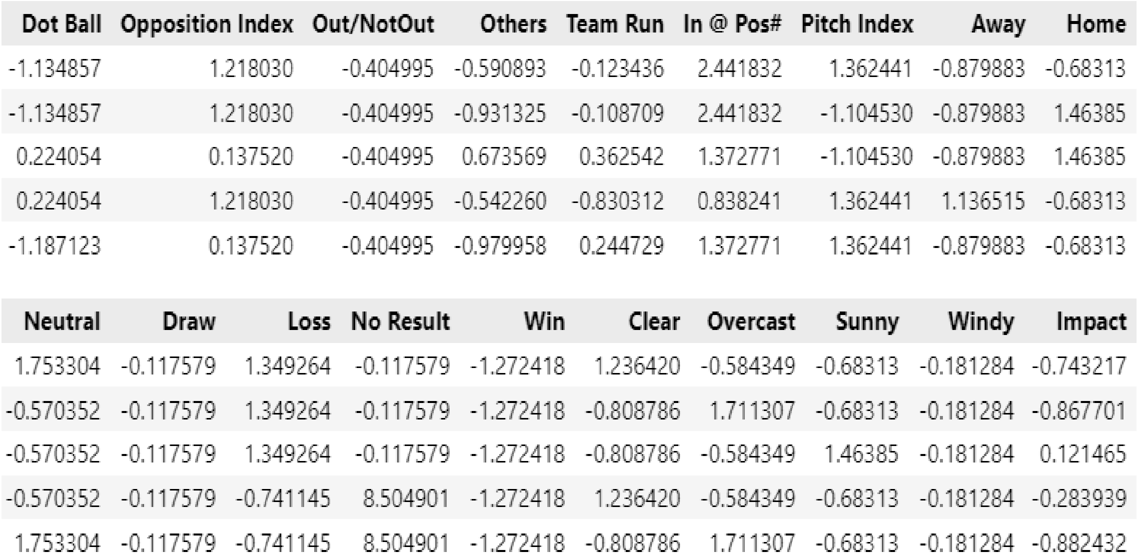


Figure 4.7: Correlation heatmap of variables over Spinners dataset.

As most of the values are greater than 1, we need to scale the data. This is done because if a variable contains relatively low values in the range of ones and in contrast another variable includes data in the range of hundreds, the algorithm can misjudge the data and give more importance to the variable with higher values. This is why all the data need to be scaled to a specific range so that the number of digits does not affect our model.

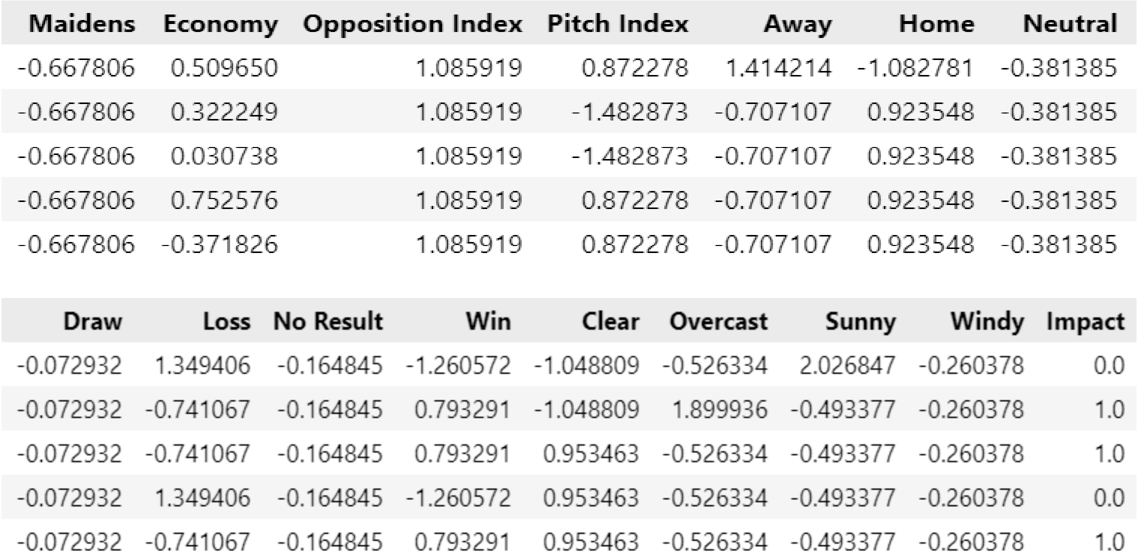


Dot Ball	Opposition Index	Out/NotOut	Others	Team Run	In @ Pos#	Pitch Index	Away	Home
-1.134857	1.218030	-0.404995	-0.590893	-0.123436	2.441832	1.362441	-0.879883	-0.68313
-1.134857	1.218030	-0.404995	-0.931325	-0.108709	2.441832	-1.104530	-0.879883	1.46385
0.224054	0.137520	-0.404995	0.673569	0.362542	1.372771	-1.104530	-0.879883	1.46385
0.224054	1.218030	-0.404995	-0.542260	-0.830312	0.838241	1.362441	1.136515	-0.68313
-1.187123	0.137520	-0.404995	-0.979958	0.244729	1.372771	1.362441	-0.879883	-0.68313

Neutral	Draw	Loss	No Result	Win	Clear	Overcast	Sunny	Windy	Impact
1.753304	-0.117579	1.349264	-0.117579	-1.272418	1.236420	-0.584349	-0.68313	-0.181284	-0.743217
-0.570352	-0.117579	1.349264	-0.117579	-1.272418	-0.808786	1.711307	-0.68313	-0.181284	-0.867701
-0.570352	-0.117579	1.349264	-0.117579	-1.272418	-0.808786	-0.584349	1.46385	-0.181284	0.121465
-0.570352	-0.117579	-0.741145	8.504901	-1.272418	1.236420	-0.584349	-0.68313	-0.181284	-0.283939
1.753304	-0.117579	-0.741145	8.504901	-1.272418	-0.808786	1.711307	-0.68313	-0.181284	-0.882432

Figure 4.8: Part of the Batter's Pre-processed Dataset.

For batters, the specific columns included as features are - "Out/NotOut," "Opposition Index," "Home/Away," "Dot Ball," "Others," "Win/Loss," "Team Run," "In at Position number," "Pitch Index," "Weather," etc. The label was "Impact."



Maidens	Economy	Opposition Index	Pitch Index	Away	Home	Neutral
-0.667806	0.509650	1.085919	0.872278	1.414214	-1.082781	-0.381385
-0.667806	0.322249	1.085919	-1.482873	-0.707107	0.923548	-0.381385
-0.667806	0.030738	1.085919	-1.482873	-0.707107	0.923548	-0.381385
-0.667806	0.752576	1.085919	0.872278	-0.707107	0.923548	-0.381385
-0.667806	-0.371826	1.085919	0.872278	-0.707107	0.923548	-0.381385

Draw	Loss	No Result	Win	Clear	Overcast	Sunny	Windy	Impact
-0.072932	1.349406	-0.164845	-1.260572	-1.048809	-0.526334	2.026847	-0.260378	0.0
-0.072932	-0.741067	-0.164845	0.793291	-1.048809	1.899936	-0.493377	-0.260378	1.0
-0.072932	-0.741067	-0.164845	0.793291	0.953463	-0.526334	-0.493377	-0.260378	1.0
-0.072932	1.349406	-0.164845	-1.260572	0.953463	-0.526334	-0.493377	-0.260378	0.0
-0.072932	-0.741067	-0.164845	0.793291	0.953463	-0.526334	-0.493377	-0.260378	1.0

Figure 4.9: Part of the Bowler's Pre-processed Dataset.

For bowlers, the specific columns included as features are - "Maidens," "Economy," "Opposition Index," "Home/Away", "Win/Loss," "Pitch Index", "Weather," etc. The label was "Impact. Figures show part of the preprocessed dataset.

Chapter 5

Implementation

Our base of implementation was the python programming language. To devise our formulae, we used the seaborn library to emphasize the correlation between our arbitrary formula and the features in our dataset. We are using all the other features as independent variables, and our aim is to use these variables to predict values for the Impact formula that we previously devised.

Since, in most cases, for a batter, the number of Runs scored is generalized as their performance in a game, along with their control over their performance, it is collectively used in the Impact formula as a dependent variable. Hence, for this, our approach should be to use regression to check the credibility of our data. On the contrary, we are appointing a threshold to the impact which we configured for a bowler. This threshold aided us in classifying a bowler as being 'impactful' or 'not impactful.' All the other features act as the independent variables and help predict the class a specific bowler falls in. Since we are answering a yes or no question about the bowler being impactful, this is a classification problem.

The Supervised Machine Learning algorithm can be Regression or Classification Algorithms. In Regression algorithms, we the output for continuous variables are predicted, but to predict the categorical variables, we need Classification algorithms.

The Regression algorithms implemented were Multiple Linear Regression, Polynomial Regression, Support Vector Machine Regression, Decision Tree Regression, and Random Forest Regression. On the contrary, k-Nearest Neighbors, Logistic Regression, and Support Vector Machine Classifier were used as Classification algorithms.

5.1 Regression

Regression is a supervised learning technique to find the relationship among more than one variable. Regression is primarily used for predicting a continuous independent value and how it is related to the dependent variables. Regression models can be divided into two parts simple and multiple. In a simple regression model, only two variables are involved and one feature, whereas, in multiple, more than 2 variables and features can be found among the variables. These can be further subdivided into linear and non-linear regressions. There are many types of regression, such as multiple linear, polynomial, support vector, decision tree, random forest etc.

We will train these models and test them to predict values and find accuracy. This accuracy is determined using the R^2(R-squared) metric.

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ variance} \tag{5.1}$$

R^2 is a goodness-of-fit metric that shows how much of the variance in the dependent variable is explained by the independent variables. The previously mentioned models use R^2 to determine their accuracies by checking the variance between the predicted and actual values. Generally, the higher the R^2, the better the regression model matches the data.

5.1.1 Multiple Linear Regression

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_{\mathrm{i1}} + ... + \beta_{\mathrm{p}} x_{\mathrm{ip}} \tag{5.2}$$

Multiple linear regression is a statistical technique for predicting a variable's outcome based on the values of two or more variables. The dependent variable is the one we want to predict using the independent variables. Linear regression attempts to determine a straight line associated with the variables. When the dependent variable has a linear relationship with more than one independent variable, then it is known as multiple linear regression.

5.1.2 Polynomial Regression

$$y = \beta_0 x^0 + \beta_1 x^1 + ... + \beta_{\mathrm{k}} x^{\mathrm{k}} + \epsilon \tag{5.3}$$

Polynomial Regression is another regression approach that represents the connection between a dependent and independent variable using an n^{th} degree polynomial. In machine learning, it's also known as the special case of Multiple Linear Regression because we turn the Multiple Linear regression equation into Polynomial Regression by adding certain polynomial terms. The dataset used in Polynomial regression for training is non-linear.

5.1.3 Support Vector Machine Regression

Support Vector Machine is a robust approach that maximizes a model's predicted accuracy without overfitting the training data. SVM is particularly well adapted to analyze data with many prediction fields. SVM works by mapping data to a high-dimensional feature space, and a hyperplane is drawn between the data to categorize them.

SVR is a sophisticated method that lets us determine how error-tolerant we are, both through an acceptable error margin and by setting our tolerance of slipping beyond that acceptable error rate. SVR uses a kernel that can be 'polynomial,' 'linear', 'rbf,' etc., where rbf stands for radial basis function. We chose our kernel to be radial based on previous observations from MLR.

5.1.4 Decision Tree Regression

The Decision Tree is a supervised learning approach that may be used to solve classification and regression issues. It is named a decision tree because, like a tree, it begins with the root node and then branches out to form a tree-like structure. It divides a dataset into smaller and smaller subsets while also developing a decision tree for each of the subsets. The result is a tree containing leaf nodes and decision nodes. Internal nodes represent dataset properties, branches represent decision rules, and each leaf node reflects the conclusion.

5.1.5 Random Forest Regression

Random Forest Regression is another supervised learning approach for regression that ensembles multiple decision trees. The ensemble learning method combines predictions from several machine learning algorithms to produce a more accurate forecast than a single model. Random Forest Regression is powerful and precise. Instead of relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. The more trees in the forest, the higher the accuracy and the lower the risk of overfitting.

5.2 Classification

Classification is a process of categorization, which is the act of recognizing, differentiating, and understanding objects. The Classification method is a Supervised Learning approach that uses training data to identify the category of new data from the testing data. A classifier learns from a given dataset or observations and then classifies additional observations into one of many classes or groupings. For example, Yes or No, 1 or 0, Impactful or Not Impactful, and so on. Classes can also be referred to as targets/labels or categories. In a classification algorithm, a discrete output function (y) is mapped to the input variable (x). There are many types of Classification algorithms such as Logistic regression, K-Nearest Neighbors, and Support Vector Machine. We will train these algorithms using our training dataset and evaluate their performances.

5.2.1 K-Nearest Neighbors

The k-Nearest Neighbors (kNN) algorithm is another supervised machine learning algorithm used for classification and regression. It is simple to put into action and comprehend. kNN works by measuring the distances between a new data point and all of the existing data, then picking the number of data points closest to the new data and chooses the most frequent label in classification or averaging the labels in regression. It does not learn the training data provided beforehand and only fits the data when there is a new set of feature data which requires classification.

5.2.2 Logistic Regression

For classification and predictive modelling, logistic regression is frequently employed. Based on a collection of independent variables, logistic regression calculates the likelihood of an event, such as voting or not voting. As opposed to linear regression, logistic regression attempts to draw an S-curve in the graphical representation of the dataset. This helps separate the dataset into binary subsets that answer a yes or no question. Because the outcome is a probability, the dependent variable ranges from 0 to 1.

5.2.3 Support Vector Machine Classification

Support Vector Machine can be used for classification and regression. However, primarily, it is used for Classification problems in Machine Learning. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors. The hyperplane separates subsets of the dataset and this is used to classify new data with respect to which side of the hyperplane the data falls in. SVM is particularly well adapted to analyze data with many prediction fields.

SVC, similar to SVR, can be used with different types of kernels like 'linear,' 'polynomial,' 'rbf,' etc. For classifying the impact of a bowler, we tuned our model with the kernel rbf.

Chapter 6

Results

6.1 Batter

Our proposed measure considers both the Runs Scored and the Control over the Balls Faced by the batter. Upon comparing the predicted values with the actual test dataset, the accuracy of each player's individually trained models is showcased in Table 6.1.

Table 6.1: Accuracy of Regression Algorithms

Regression Model	***Sharma (%)***	***Warner (%)***	***Williamson (%)***
Multiple Linear	89.14	91.53	89.80
Random Forest	85.06	91.70	84.60
Support Vector	79.88	80.21	73.13
Polynomial	76.07	70.01	56.34
Decision Tree	74.72	90.00	73.39

The Polynomial Regression model gave the weakest prediction. Polynomial Features from scikit-learn library was used which ensued an accuracy of 76.07% for Sharma, 70.01% for Warner, and 56.34% for Williamson. For Support Vector Regression, the kernel was set as 'rbf,' giving accuracies of 79.88%, 80.21%, and 73.13% distributed amongst the players.

Decision Tree Regressor was used which gave individual accuracies of 74.72%, 90.00%, and 73.39%. Ensembling 10 decision trees for the Random Forest Regressor gave overall improvements for each of the datasets: 85.06%, 91.70%, and 84.60%. Figures 6.1-6.3 show how the predictions made differ from the actual dataset values.

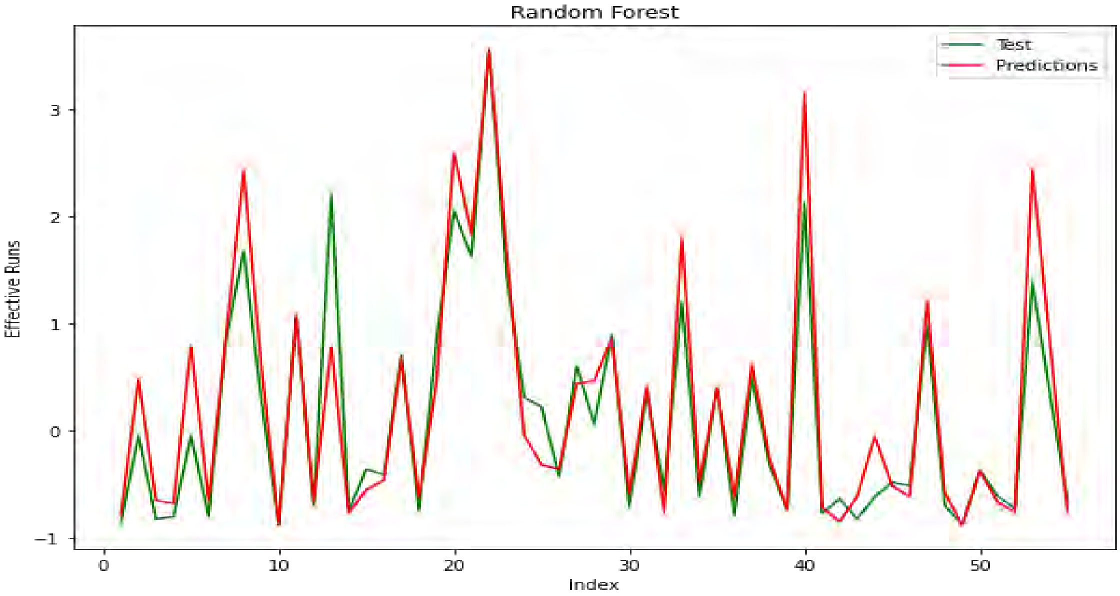


Figure 6.1: Comparison of Actual and Predicted Effective Runs by Sharma using Random Forest Regression.

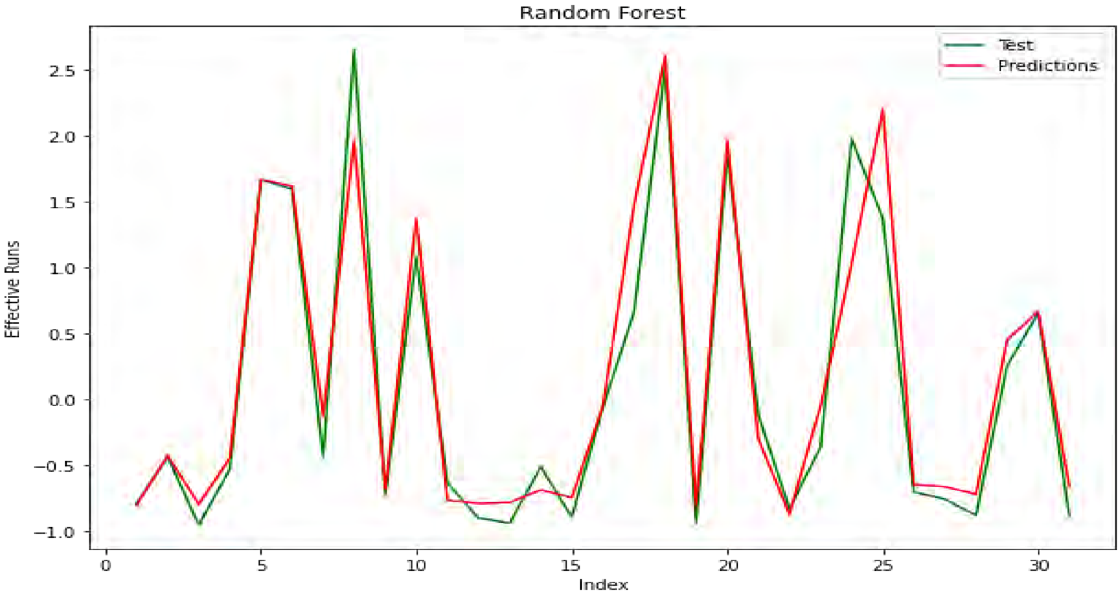


Figure 6.2: Comparison of Actual and Predicted Effective Runs by Warner using Random Forest Regression.

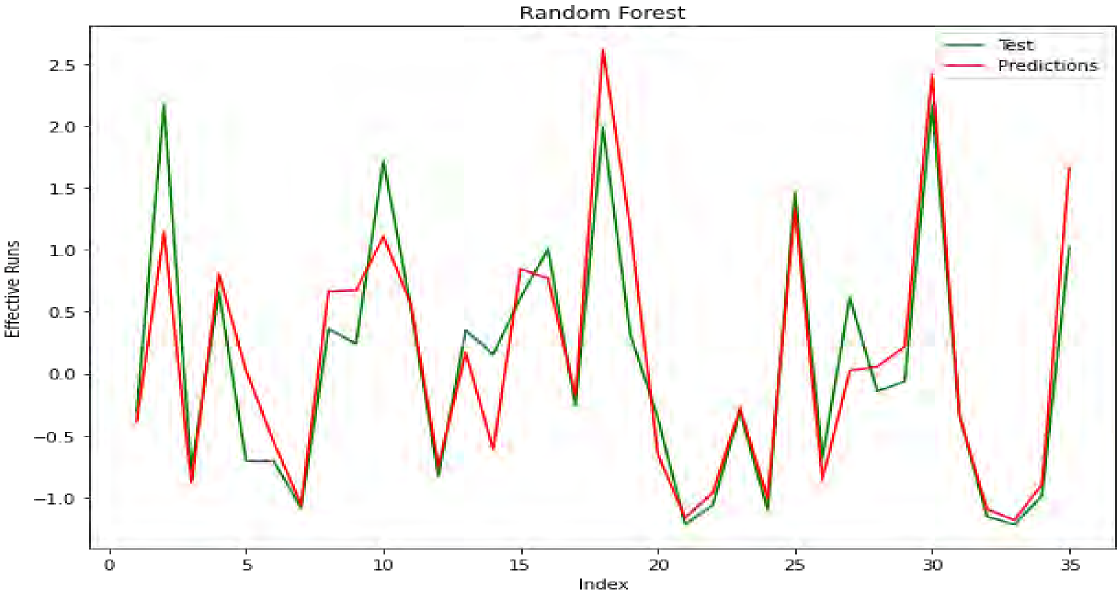


Figure 6.3: Comparison of Actual and Predicted Effective Runs by Williamson using Random Forest Regression.

However, the best outcome was reached using Multiple Linear Regression. The testing dataset for Sharma resulted in 89.14% accuracy, while Warner and Williamson gave 91.53% and 89.80% respectively. Graphical visualization of the predicted and actual values of the testing dataset is shown in Figures 6.4-6.6.

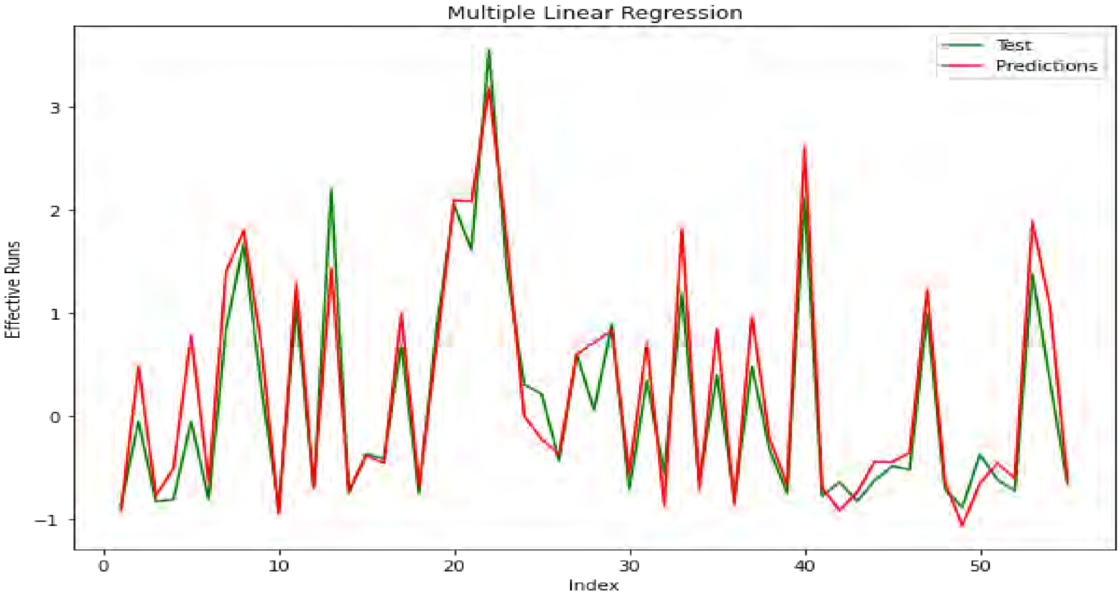


Figure 6.4: Comparison of Actual and Predicted Effective Runs by Sharma using Multiple Linear Regression.

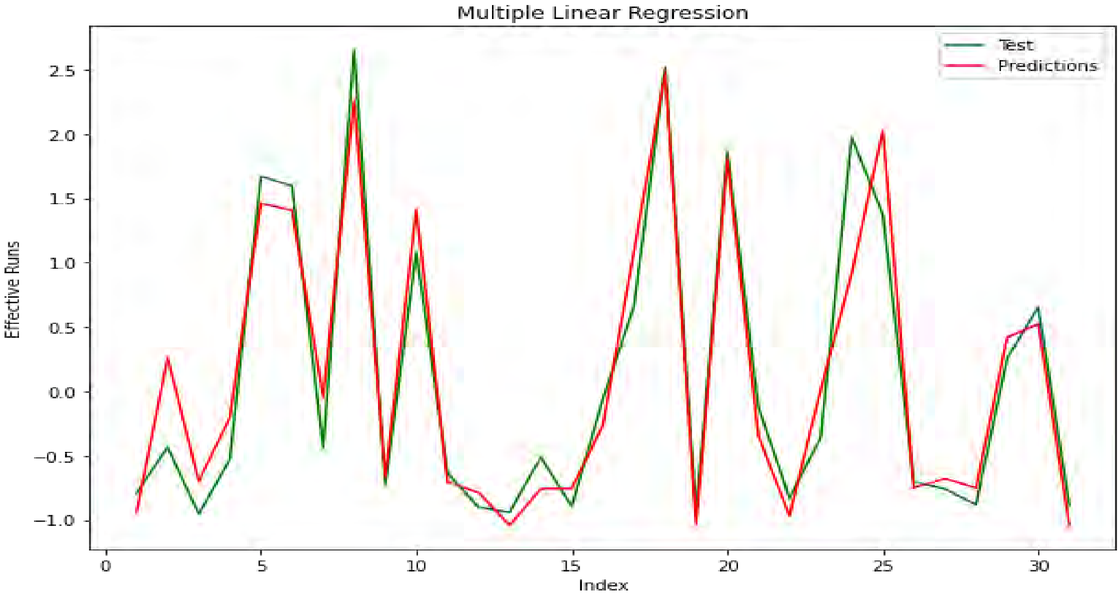


Figure 6.5: Comparison of Actual and Predicted Effective Runs by Warner using Multiple Linear Regression.

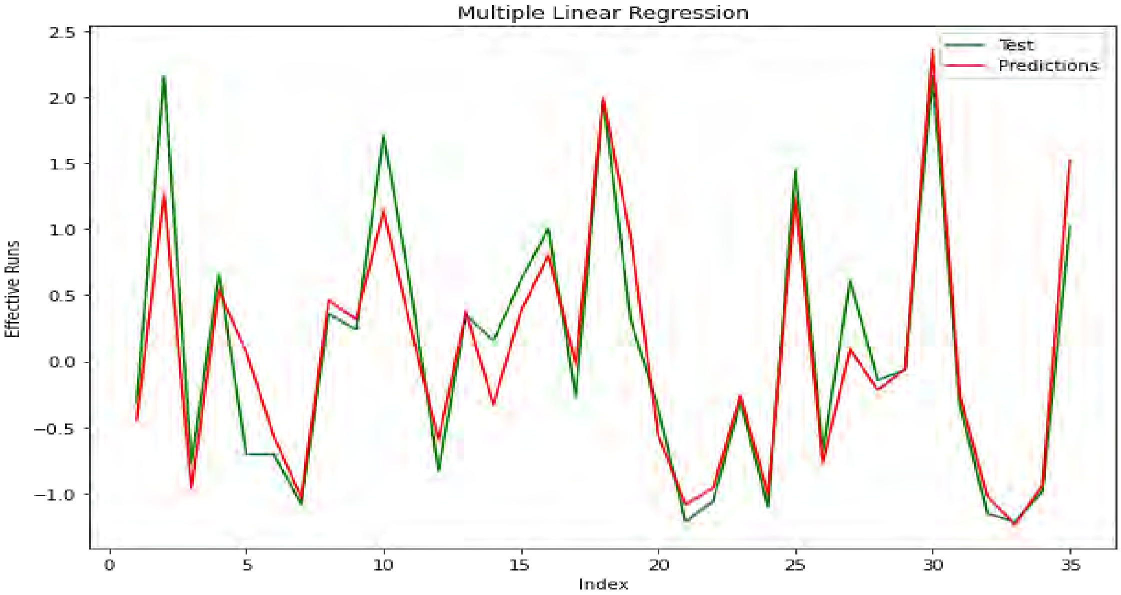


Figure 6.6: Comparison of Actual and Predicted Effective Runs by Williamson using Multiple Linear Regression.

6.2 Bowler

Machine Learning Classification Algorithms were trained to predict whether a bowler is impactful in a game or not impactful. A confusion matrix like in Figure 6.7 for each of the algorithms were used to evaluate their performaces. The accuracy of the algorithms is a fraction of the total correct predictions over the entire dataset, whereas the precision is a fraction of the total correct positive predictions over the total correct positions.

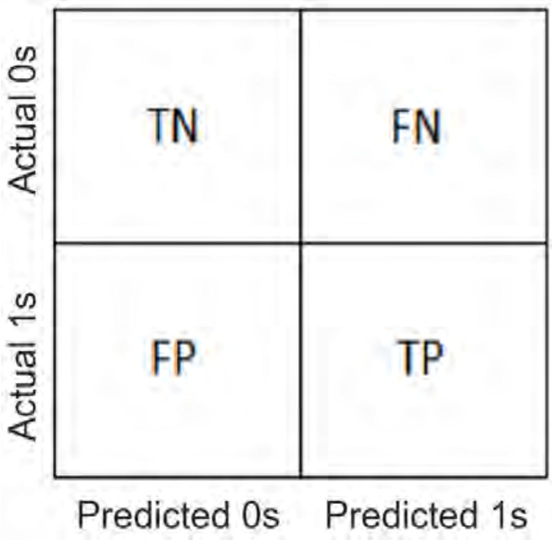


Figure 6.7: Confusion Matrix.

$$Accuracy = \frac{True\ Positive\ +\ True\ Negative}{Positive\ +\ Negative} \tag{6.1}$$

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive} \tag{6.2}$$

The accuracy of the predictions from each of the datasets is tabulated in Table 6.2 indicating the performance of the models.

Table 6.2: Accuracy of Classification Algorithms

Classification Model	***Seamers (%)***	***Spinners (%)***
k-Nearest Neighbors	77.08	64.29
Logistic Regression	75.00	73.21
Support Vector	79.17	69.64

The k-Nearest Neighbors Classifier from the neighbors module of the scikit-learn library gave an accuracy of 77.08% with the Seamers dataset, while this dropped to 64.29% for the Spinners. While tuning the model, the number of neighbors used was 5 and the 'minkowski' metric using the Euclidean distance further calibrated the model. On average, a precision score of 76.02% was observed.

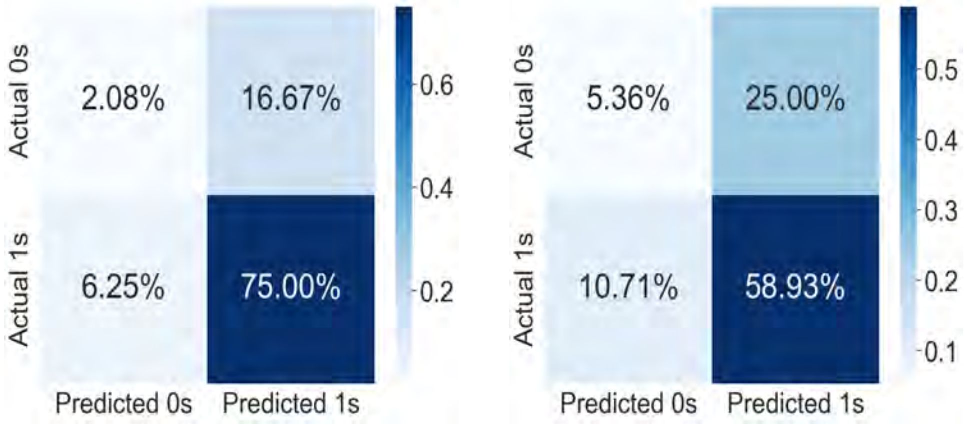


Figure 6.8: Confusion matrix of kNN Classifier over Seamers and Spinners dataset.

Logistics Regression analysis, from the linear model module of the same scikit-learn library, was performed next. The accuracies on each of the datasets were relatively close for this trained model as 75.00% and 73.21% accuracies were acquired for Seamers and Spinners respectively. The average precision score here was 78.75%.

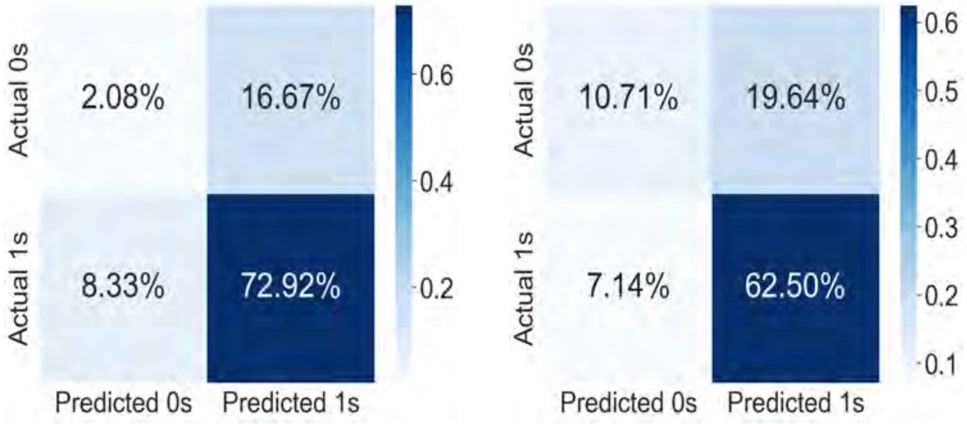


Figure 6.9: Confusion matrix of Logistic Regression over Seamers and Spinners dataset.

The last algorithm that we implemented was Support Vector Machine Classification, SVC, from the svm module. The kernel 'rbf' was used to tune this model. Support Vector Classification showed the best accuracy of 79.17% for Seamers with 69.64% for the Spinners. The average precision in this model was 76.72%.

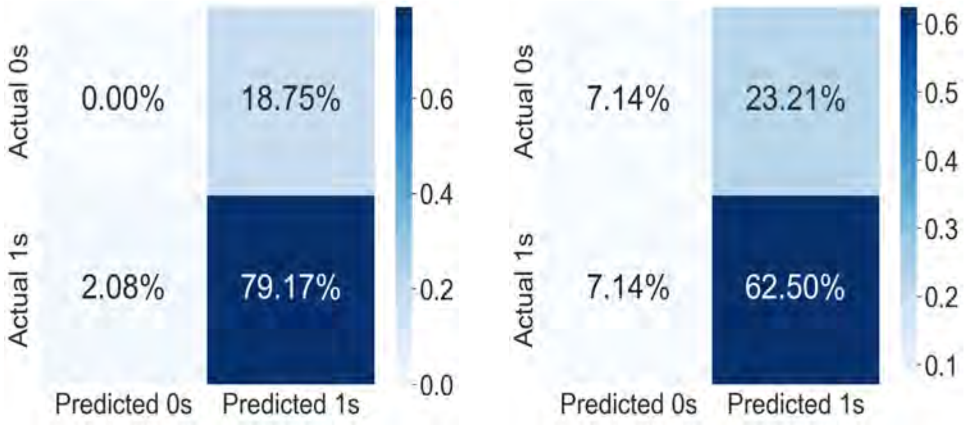


Figure 6.10: Confusion matrix of SVM Classifier over Seamers and Spinners dataset.

6.3 Model Comparison

6.3.1 Batter

The accuracy of the regression models for individual datasets has already been explained above. Upon averaging the performance of each of the models, the average accuracies resulting are visualized in the barchart in Figure 6.11.

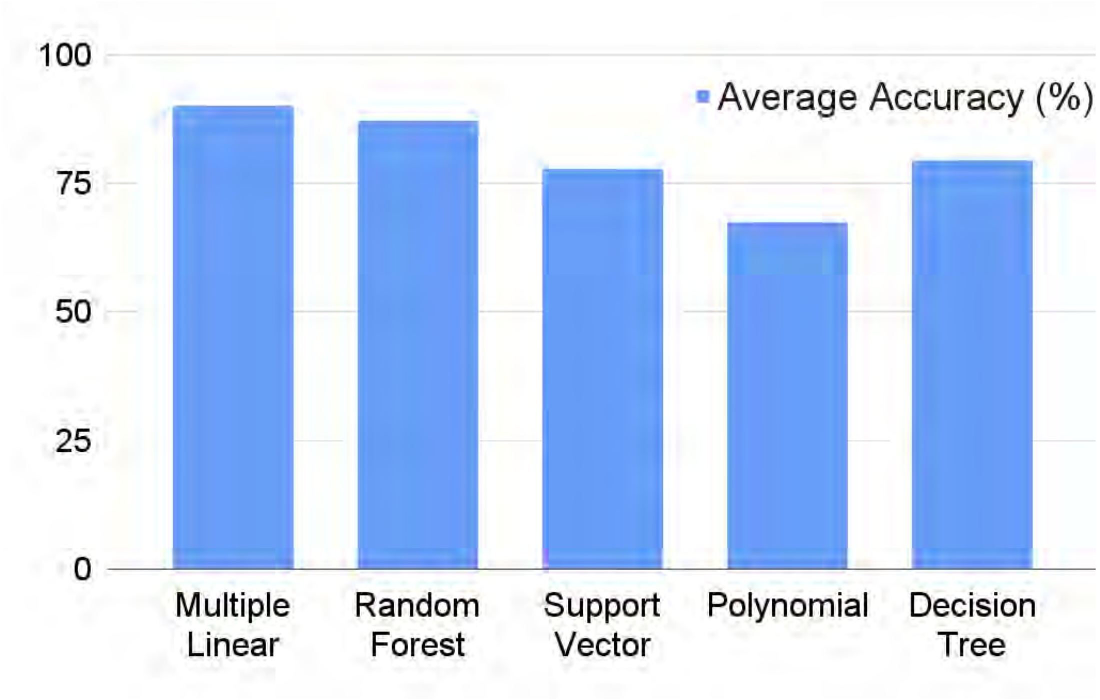


Figure 6.11: Comparison of Average Accuracies of different Regression Models.

The above barchart shows a comparison of the individual models and how well they perform in predicting our proposed Impact of a Batter. The accuracy of the Polynomial Regression averaged 67.47%. Similarly, Support Vector Machine was at 77.74%. Decision Tree Regressor gives an average of 79.37% which improved to 87.12% when 10 decision trees were used for the Random Forest Regressor. The best overall accuracy for each of the batters and average accuracy was observed with Multiple Linear Regression. Hence, using our proposed method for weighing the opposition and pitch for the batters, Multiple Linear Regression will be the best model for predicting the Effective Runs of a batter. This Effective Runs can hence be used as a measure of the batter's impact in a game.

6.3.2 Bowler

Taking both wickets and control from runs conceded into consideration, the accuracy of each of the classification models for Spinners and Seamers are visualized in Figure 6.12.

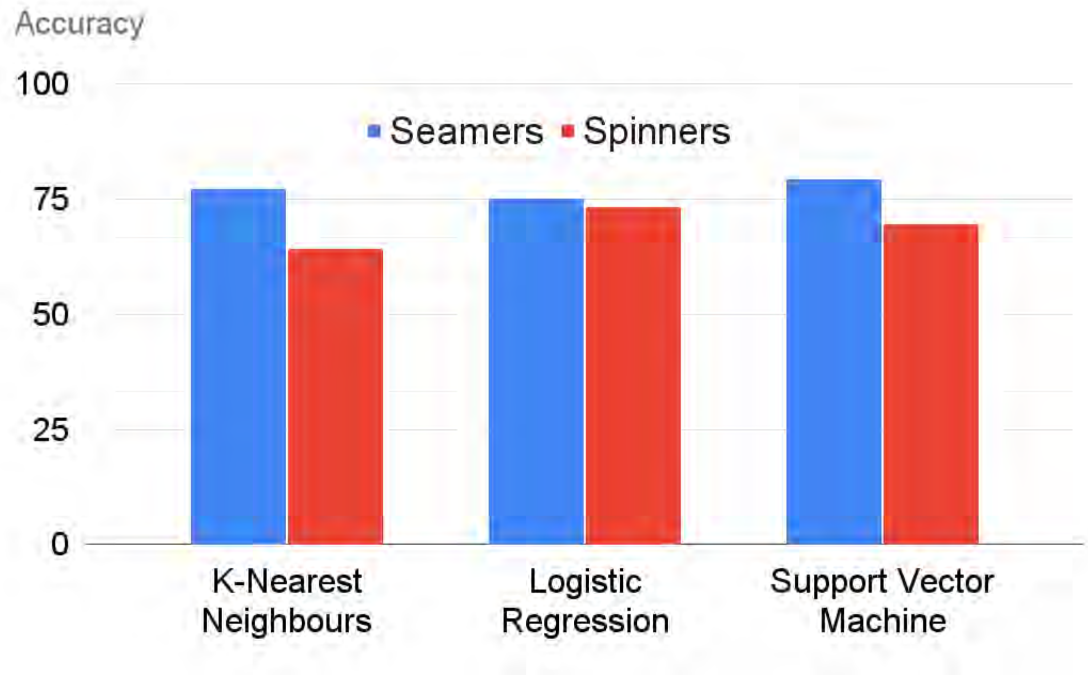


Figure 6.12: Comparison of Accuracies of different Classification Models over Seamers and Spinners Dataset.

As represented in the barcharts above, the distribution of the accuracies for the models is greater for the k-Nearest Neighbors Classifier and Support Vector Machine Classifier. But the accuracies for Logistic Regression for both the Seamers and Spinners dataset had closer values of 75.00% and 73.21%. Since these are classification algorithms, upon observing the average precision of each of the models, Logistic Regression gave the highest precision of 78.75%, while k-Nearest Neighbors and Support Vector Machine gave precisions to 76.02% and 76.72% respectively. Hence, the model with the best performance for predicting our categorized Impact from the proposed Effective Wickets of a bowler was elected to be Logistic Regression. This trained model can predict the likelihood of a bowler being impactful or not impactful in a match of ODI cricket with a precision of 78.75% as shown in Figure 6.13.

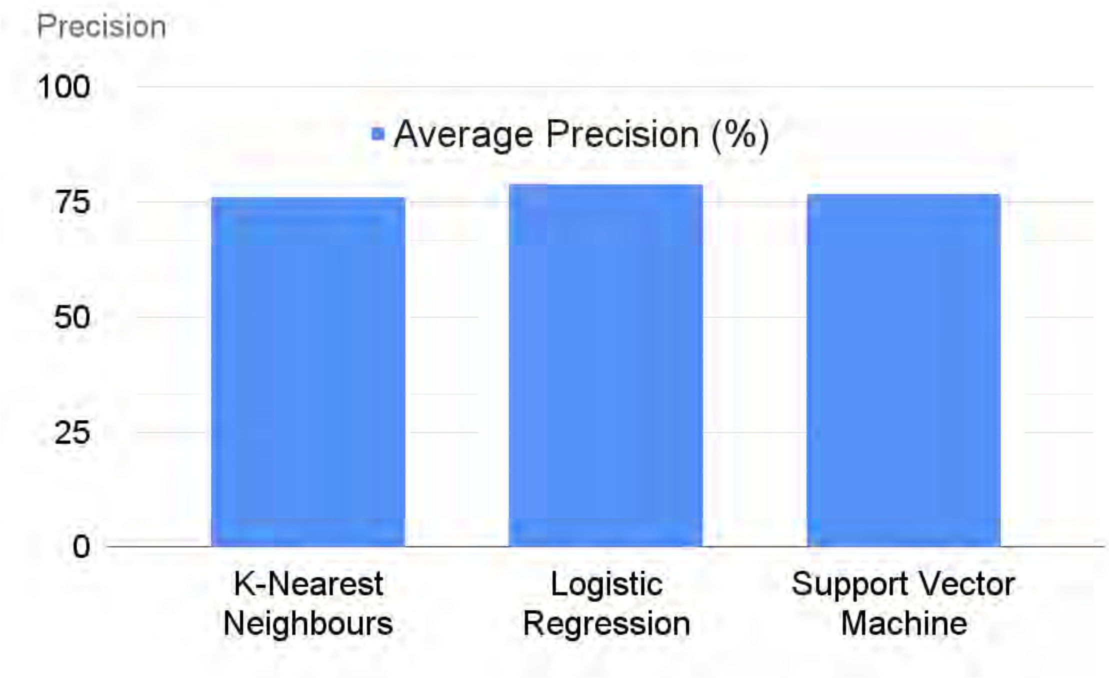


Figure 6.13: Comparison of Average Precision of different Classification Models.

This impactive performance can give the team selectors and even the players themselves an idea of how their gameplay can deviate the direction a match takes and make the sport even more interesting.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

In this era of computers, cricket has successfully settled its area of research. As the game continues to develop, the passion for the game has illuminated the youth and has given exponential rise to potentially excellent cricketers. This, inevitably, is proving to be a challenge for team management, coaches, and sponsors to accurately verdict and pick players to invest their time and money. Discovering the effectiveness of the control of a batter or a bowler significantly contributes to the outcome of a cricket match. Hence, the need for better-performing strategies to evaluate and rank players is more than ever. This leads to the demand for predictive models for future talent who are still invisible in this vast pool of players. On one end, the control of a batter was the main focus of making a new measure to determine how their performance can change the flow of a game. Features like pitch, weather, opposition, and other extra factors were used along with Machine Learning models to predict the Impact of a batter in an ODI match. Multiple Linear Regression gave the best results with an accuracy of 90.16%. This new measure, "Effective Runs," can be used to determine the impactive performance of a batter.On the other hand, cricket being predominantly a batter-oriented game can easily be dictated by a bowler in rhythm throughout the match. Hence, the demand for effectiveness of bowlers having a good control metric proves to be vital in deducing the outcome of the match. For our work, in regards to the bowler's performance, several features were incorporated. K-Nearest Neighbors provided with the best accuracy of 81.82% for seamers whie Logistic Regression showed the best accuracy for spinners being 76.09%. Overall, this research attempts to provide an improved and more developed model that utilizes past data and machine learning models to predict the impact a player will create by the role they play for their respective teams.

7.2 Future Scope

Our Future agenda is to tune our dataset further and propose new models that may even give better accuracy. In the future, our motive is to implement Artificial Neural Network (ANN) for our models. We are planning to make new measures for batters in Test and T20 cricket as well; additionally, our motive is to create new measures for bowlers and all rounders.

Bibliography

[1] R. Alston, A. Longmore, and M. K. Williams, *Cricket.* [Online]. Available: https://www.britannica.com/sports/cricket-sport.

[2] *Cricket playing nations, international cricket teams list: 2022*, Jul. 2022. [Online]. Available: http://www.cricfooty.com/international-cricket-team-icc-members-list/.

[3] *Can cricket become a truly global sport?* Sep. 2021. [Online]. Available: https://www.thecricketpaper.com/news/383065/can-cricket-become-a-truly-global-sport/.

[4] *Cricket world cup.* [Online]. Available: https://www.britannica.com/sports/Cricket-World-Cup.

[5] A. I. Anik, S. Yeaser, A. G. M. I. Hossain, and A. Chakrabarty, "Player's performance prediction in odi cricket using machine learning algorithms," in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 2018, pp. 500–505. DOI: 10.1109/CEEICT.2018.8628118.

[6] S. Chand, H. K. Singh, and T. Ray, "Team selection using multi-/many-objective optimization with integer linear programming," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1–8. DOI: 10.1109/CEC.2018.8477945.

[7] S. Akhtar, P. Scarf, and Z. Rasool, "Rating players in test match cricket," *The Journal of the Operational Research Society*, vol. 66, no. 4, pp. 684–695, 2015. [Online]. Available: http://www.jstor.org/stable/24505316.

[8] M. G. Jhanwar and V. Pudi, "Quantitative assessment of player performance and winner prediction in odi cricket," 2017.

[9] H. Saikia and D. Bhattacharjee, "An application of multilayer perceptron neural network to predict the performance of batsmen in indian premier league," *International Journal of Research in Science and Technology*, vol. 1, no. 1, pp. 6–15, 2014.

[10] A. C. Kaluarachchi and S. V. Aparna, "Cricai: A classification based tool to predict the outcome in odi cricket," *2010 Fifth International Conference on Information and Automation for Sustainability*, pp. 250–255, 2010. DOI: 10.1109/ICIAFS.2010.5715668.

[11] T. B. Swartz, "Winning the coin toss and the home team advantage in one-day international cricket matches," 2004.

[12] S. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5510–5522, 2009. DOI: 10.1016/j.eswa.2008.06.088.

[13] M. Shetty, S. Rane, C. Pandita, and S. Salvi, "Machine learning-based selection of optimal sports team based on the players performance," Jun. 2020, pp. 1267–1272. DOI: 10.1109/ICCES48766.2020.9137891.

[14] P. Somaskandhan, G. Wijesinghe, L. B. Wijegunawardana, A. Bandaranayake, and S. Deegalla, "Identifying the optimal set of attributes that impose high impact on the end results of a cricket match using machine learning," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2017, pp. 1–6. DOI: 10.1109/ICIINFS.2017.8300399.

[15] M. Bailey and S. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress," *Journal of sports science & medicine*, vol. 5, pp. 480–7, Dec. 2006.

[16] M. Khan and R. Shah, "Role of external factors on outcome of a one day international cricket (odi) match and predictive analysis," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, pp. 192–7, 6 Jun. 2015. DOI: 10.17148/IJARCCE.2015.4642.

[17] P. Shah, "New performance measure in cricket," *IOSR Journal of Sports and Physical Education*, vol. 04, pp. 28–30, May 2017. DOI: 10.9790/6737-04032830.

[18] M. K. Mahbub, M. A. M. Miah, S. M. S. Islam, S. Sorna, S. Hossain, and M. Biswas, "Best eleven forecast for bangladesh cricket team with machine learning techniques," in *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2021, pp. 1–6. DOI: 10.1109/ICEEICT53905.2021.9667862.

[19] N. Rodrigues, N. Sequeira, S. Rodrigues, and V. Shrivastava, "Cricket squad analysis using multiple random forest regression," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 2019, pp. 104–108. DOI: 10.1109/ICAIT47043.2019.8987367.

[20] V. V. Tharoor and N. Dhanya, "Performance of indian cricket team in test cricket: A comprehensive data science analysis," in *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, 2022, pp. 128–133. DOI: 10.1109/ICESIC53714.2022.9783492.

[21] E. Mundhe, I. Jain, and S. Shah, "Live cricket score prediction web application using machine learning," in *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2021, pp. 1–6. DOI: 10.1109/SMARTGENCON51891.2021.9645855.

[22] S. Priya, A. K. Gupta, A. Dwivedi, and A. Prabhakar, "Analysis and winning prediction in t20 cricket using machine learning," in *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2022, pp. 1–4. DOI: 10.1109/ICAECT54875.2022.9807929.

[23] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, S. Vasudevan, V. Veeramani Kannan, and S. Sagubar Sadiq, “Moneyball - data mining on cricket dataset,” in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1–5. DOI: 10.1109/ICCIDS.2019.8862065.

[24] M. A. Pramanik, M. M. Hasan Suzan, A. A. Biswas, M. Z. Rahman, and A. Kalaiarasi, “Performance analysis of classification algorithms for outcome prediction of t20 cricket tournament matches,” in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022, pp. 01–07. DOI: 10.1109/ICCCI54379.2022.9740867.

[25] *Statsguru | searchable cricket statistics database.* [Online]. Available: https://stats.espncricinfo.com/ci/engine/stats/index.html.

[26] *Player statistics.* [Online]. Available: http://www.cricmetric.com/playerstats.py.

[27] A. Kumar and J. Sahni, *All you need to know about cricket pitches: Preparation, different tracks, soil composition and characteristics: Cricket news - times of india.* [Online]. Available: https://timesofindia.indiatimes.com/sports/cricket/england-in-india/all-you-need-to-know-about-cricket-pitches-preparation-different-tracks-soil-composition-and-characteristics/articleshow/81289893.cms.

[28] K. Date, *Mining control statistics*, Jun. 2014. [Online]. Available: https://www.espncricinfo.com/story/kartikeya-date-mining-control-statistics-750925.