Quantitative Assessment of Player Performance and Winner Prediction in ODI Cricket

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science by Research
in
Computer Science and Engineering

by

Madan Gopal Jhanwar
201202018
madangopal.jhanwar@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, INDIA
JULY 2017

To my Father,

for his unconditional love and support

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **“Quantitative Assessment of Player Performance and Winner Prediction in ODI Cricket”** by Madan Gopal Jhanwar, has been carried out under my supervision and is not submitted elsewhere for a degree.

25/7/17
Date

VikramPudi
Adviser: Dr. Vikram Pudi

Acknowledgements

IIIT-Hyderabad has been a family to me for the last 5 years. During this journey, I have met several people each of whom have taught me a new thing every day. As I submit my thesis today, I feel indebted to a lot of people around me without whom my stay at IIIT would not have been such a great memory.

First of all, I would like to thank my guide, Dr. Vikram Pudi, who has been an excellent mentor throughout my research journey. He not only gave me freedom to explore my passions but also supported me when I showed interest in cricket analytics. His guidance not only helped me in my research but also made me a better person.

I don't have enough words to thank my father, mother and my family, who have shown immense faith and trust in me. They have given me continuous encouragement throughout my years of study.

I couldn't have accomplished it without the constant support from many of my friends. Saksham, thanks for motivating me in the early days of my research and helping me in getting started. Arjit, I have always seen a great friend in you. Thanks for going beyond the limits to always help me. Our discussions on cricket have given me many insights into the game. Remya, thanks for being around and always helping and motivating me. Nirat, thank you for putting in all the time and efforts. I have learnt the value of friendship and sharing from you guys.

Arihant, Abhishek, Junaid, Shipra, Somya, Kalpit, Mohit, Shubham, Sarfaraz, Nazrul, Sanatan, Rishab and Shashank, a special thanks to you all. Thank you for motivating me all along and making the college life a lively one.

I would also like to thank all my lab mates – Jaspreet, Tejas, Vamshi, Saket and Soumyajit, for the wonderful interactions and fun work sessions.

Last, but not the least, thanks to IIIT community for giving me an inspiring environment and loads of opportunities to grow.

Abstract

The statistics of professional sports, including players and teams, provide numerous opportunities for research. Cricket is one of the most popular team sports, with billions of fans all over the world. In this thesis, we address two problems related to the One Day International (ODI) format of the game. First, we propose a novel method to predict the winner of ODI cricket matches using a team-composition based approach at the start of the match. Second, we present a method to quantitatively assess the performances of individual players in a match of ODI cricket which incorporates the game situations under which the players performed. The player performances are further used to predict the player of the match award.

Players are the fundamental unit of a team. Players of one team work against the players of the opponent team in order to win a match. The strengths and abilities of the players of a team play a key role in deciding the outcome of a match. However, a team changes its composition depending on the match conditions, venue, and opponent team, etc. Therefore, we propose a novel dynamic approach which takes into account the varying strengths of the individual players and reflects the changes in player combinations over time. Our work suggests that the relative team strength between the competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual players' batting and bowling performances, forming the basis of our approach. We use career statistics as well as the recent performances of a player to model him. Using the relative strength of one team versus the other, along with two player-independent features, namely, the toss outcome and the venue of the match, we evaluate multiple supervised machine learning algorithms to predict the winner of the match. We show that, for our approach, the k-Nearest Neighbor (kNN) algorithm yields better results as compared to other classifiers.

Players have multiple roles in a game of cricket, predominantly as batsmen and bowlers. Over the generations, statistics such as batting and bowling averages, and strike and economy rates have been used to judge the performance of individual players. These measures, however, do not take into consideration the context of the game in which a player performed across the course of a match. Further, these types of statistics are incapable of comparing the performance of players across different roles. Therefore, we present an approach to quantitatively assess the performances of individual players in a single match of ODI cricket. We have developed a new measure, called the *Work Index*, which represents the amount of work that is yet to be done by a team to achieve its target. Our approach incorporates game situations and the team strengths to measure the player contributions. This not only helps us in evaluating the individual performances, but also enables us to compare players within and across various roles on a common scale. Using the player performances in a match, we predict the player of the match award for the ODI matches played between 2006 and 2016. We have achieved an accuracy of 86.80% for the top-3 positions in predicting the player of the match award, which is superior to previous works and other baseline models. This further proved the validity of our approach.

Contents

Chapter Page

List of Figures

Figure Page

List of Tables

Chapter 1

Introduction

Statistical modeling has been used in sports over decades and has contributed significantly to success on the field. Cricket is one of the most popular team sports in the world, second only to soccer [1]. Various natural factors affecting the game, enormous media coverage, and a huge betting market provide strong incentives to model the game from various perspectives. For instance, Duckworth and Lewis proposed a solution, called D/L method [2], to reset targets in rain interrupted matches which was adopted by the International Cricket Council (ICC) in 1998. However, the complex rules governing the game, the ability of players and their performances on a given day, and various other natural parameters play an integral role in shaping the course of a cricket match. This presents significant challenges in modeling the game.

1.1 History of Cricket

The origin of the game of cricket is estimated to be around mid-sixteenth century in England. However, it was already 1844 by the time international matches became commonplace. Towards the end of the 18th century, cricket became the national sport of England. With the expansion of the English empire, cricket also reached the masses. The first Test match, the earliest form of cricket, was played in 1877 when the English team toured Australia. The following year, the Australians toured England for the first time and the success of this tour ensured a popular demand for similar ventures in future.

For the organization and governance of cricket's major international tournaments, the ICC was founded in 1909 by representatives from England, Australia and South Africa. A Test match is played between two teams for 5 days. Owing to the long duration of the game and inconclusiveness of the end result, newer formats of the game were developed by ICC. These new formats also served to make cricket a faster and more exciting game. An ODI is a form of limited overs cricket, played between two teams with international status, in which each team faces a fixed number of overs. In the early days of ODI cricket, the number of overs per side was 60, with some matches having 40, 45 or 55 overs per side. However, today, the number of overs per side has been uniformly fixed at 50 overs. The first ODI was played in 1971 between Australia and England at the Melbourne Cricket Ground. Thereafter, ODIs gained a huge popularity and a total of five hundred and ninety five ODI cricket matches were played by the end of 1990. As of now, around five thousand ODI cricket matches have been played.

In order to further shorten a game of cricket and bring the timespan of a cricket match closer to the timespan of other popular team sports, Twenty20 (T20) cricket was introduced in 2005 at the international level. T20 cricket matches are similar to ODIs but restricted to a maximum of only 20 overs per side. T20 is the fastest form of cricket, however, ODIs still remains the most popular format of the game.

Today, cricket is one of the most followed team games in the world, with 106 member states, and 10 full members, namely, Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. The major ODI cricketing event is the ODI Cricket World Cup and is organized by ICC once in every four years. As of 2017, Australia has won 5 ODI cricket world cups, followed by India and West Indies with 2 wins each.

1.2 The Game of Cricket

Cricket is a game which has evolved over time. Today, international cricket is played in two major formats – the limited overs cricket and the non-limited overs cricket. A non-limited overs cricket match could last up to several days. For instance, the popular international Test match format of non-limited cricket matches lasts up to 5 days, where both the teams could potentially bat and bowl twice. On the other hand, limited overs cricket matches are made to start and finish on the same day. There are two popular formats of limited overs international cricket matches, i.e., ODIs and T20s. An ODI game generally lasts for 8 hours where each team has a maximum quota of 50 overs, and on average, a T20 international match lasts for 3 hours and is limited to a maximum quota of only 20 overs per side.

The game of cricket is played on a round or oval-shaped grassy field known as cricket ground. The borderline of the ground is known as boundary and the central part of the ground is known as pitch. The pitch is a rectangular 22 yards long clay strip with stumps at each end. The stumps are the three vertical posts that support the bails – wooden crosspieces. The pitch is usually about 60 meters from one boundary square of the pitch. Each player from the fielding team takes a location on the ground to field. One player bowls from one end of the pitch to a batsman from the batting team. One player always takes position as a wicket keeper (behind the wicket of the batsman). The remaining nine players take different positions. The team captain is responsible for assigning fielding positions to the players. Figure 1.1 shows a cricket ground with common fielding positions, and Figure 1.2 shows a cricket pitch with related dimensions and terminologies.

Of the three formats of cricket, the ODIs have gained a huge popularity and are considered the highest standard of limited overs competition. The Marylebone Cricket Club is the framer of the Laws of Cricket, the rules governing play of the game. Major rules amongst them are listed below –

- A game of ODI cricket is played between two teams of international status.

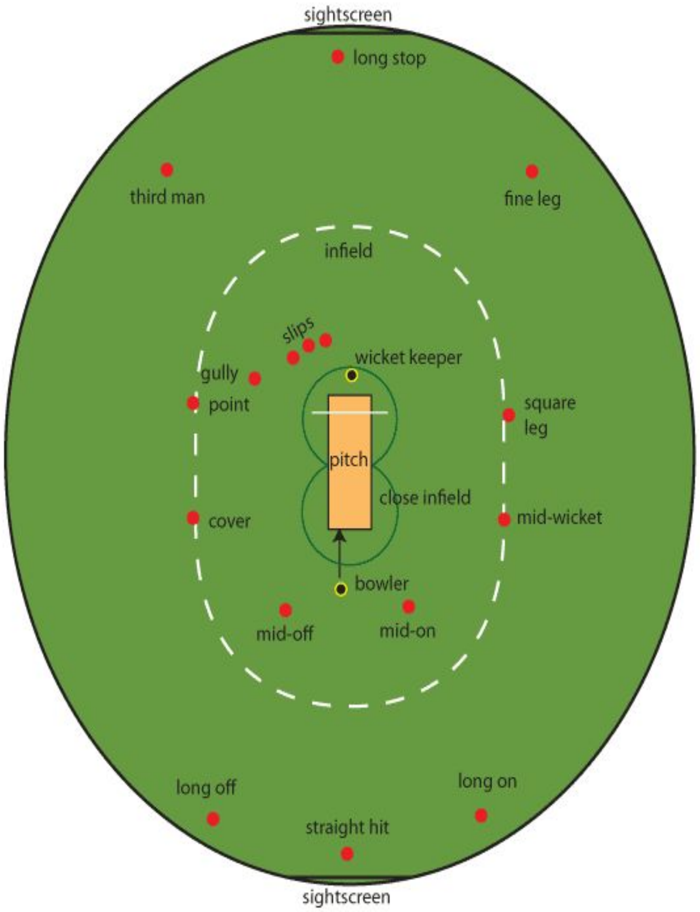

Figure 1.1: A cricket ground and some of the common fielding positions. *Source: Cricket-Australia*

- Each team has a combination of batsmen and bowlers, making up 11 players in total.

- Batsmen are the players who are specially skilled in hitting the ball to score runs and saving their wicket, whereas, bowlers are the players who are skilled in taking the opponent's wickets and restricting them from scoring runs.

- Runs can be scored in the form of 1s, 2s, 3s, 4s and 6s, whereas a wicket refers to a batsman getting out. At a given point of time, two players from the batting team play in partnership to score as many runs as possible, and therefore, a team can lose a maximum of 10 wickets.

- The game starts with a coin toss and the captain of the side winning the toss chooses to either bat or bowl first. The choice is made depending upon several factors such as weather, pitch type, venue, previous statistics, etc.

- The team batting first sets the target score in a single *innings*, where the *innings* lasts until the batting side loses all the 10 wickets or the batting side's quota of 50 *overs* is completed.

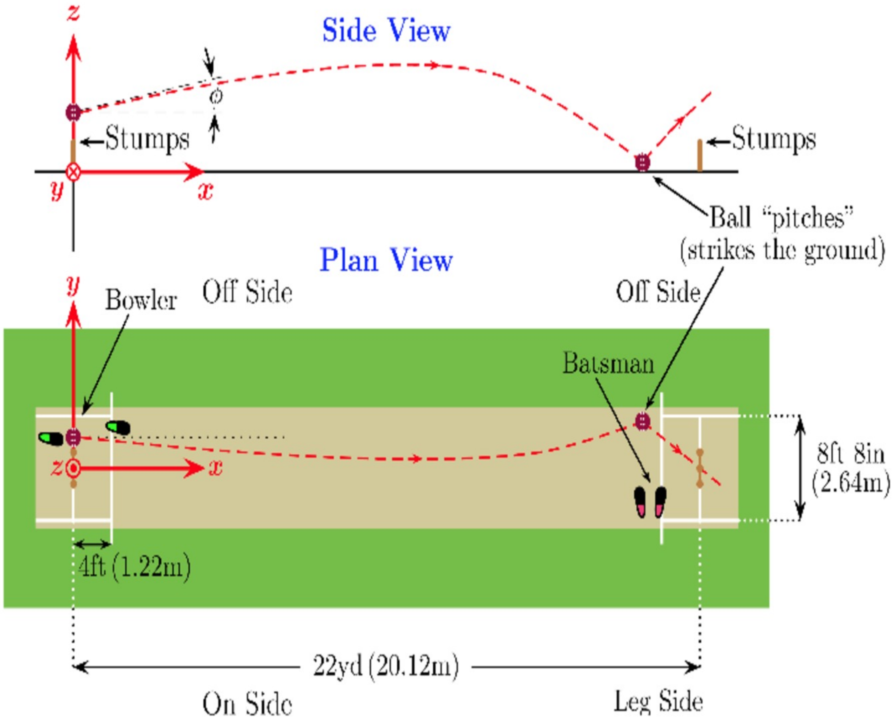

Figure 1.2: Dimensions of a cricket pitch and the related terminologies. *Source: Australia-Cricket-News*

- An *over* is defined as a set of six deliveries bowled by the bowlers. Each bowler can bowl a maximum of 10 overs. However, no bowler can bowl two consecutive overs.

- The team batting second tries to score more runs than the target score in order to win the match. Similarly, the side bowling second tries to take all the 10 wickets of the batting team or make them exhaust their overs before they reach the target score in order to win.

- If both the teams score equal number of runs at the end of their innings, then the match is declared a tie. And if the match is abandoned due to any reason such as bad light or rain, it is declared a draw.

A team generally plays with 6-7 specialized batsmen and remaining specialized bowlers, with one of the batsmen also being specialized in wicket-keeping skills. Some players are specialized in batting as well as bowling and are called all-rounders. All the players play the role of fielders too during their bowling innings. The detailed statistics of player performances for every match are logged. The number of runs scored by a player, number of balls faced, number of balls bowled, number of wickets taken and number of runs conceded, being the most popular statistics among them. At the end of a cricket match, one of the players, who played the most significant role, is awarded the player of the match award for his performance. Table 1.1 tabulates some of the common terminologies related to cricket and player statistics, and their notations that would be used constantly throughout this thesis.

Table 1.1: Notations

Notation	**Description**
$matchesPlayed$	*#ODI cricket matches played by a player*
$batInngs$	*#Matches in which the player batted*
$batAverage$	*#Runs scored divided by the #times the player got out*
$batStrikeRate$	*Average #runs scored per 100 balls faced by the player*
$bowlInngs$	*#Matches in which the player bowled*
$bowlEconomy$	*Average #runs conceded by the player per over bowled*
$bowlStrikeRate$	*Average #balls bowled per wicket taken by the player*
$Target$	*Total #runs to be scored by the batting team*
$runsRemaining$	*#Runs remaining to be scored by the batting team to get to the target score*
$ballsBowled$	*Total #balls bowled by the bowling team in the given match*
$initRunRate$	*Average #runs required per over by the batting team at the start of the innings*
$currRunRate$	*Average #runs scored per over by the batting team till the current stage of the innings*
$reqRunRate$	*Average #runs required per over by the batting team at the current stage of the match*

1.3 Aim and Contributions

In this thesis, we shed light on two of the most important issues related to limited overs cricket. First, we propose an approach to predict the winner of an ODI cricket match by considering the varying strengths of individual players forming the two teams. Second, we present a model that helps us to quantitatively assess the individual performances of players at the end of an ODI cricket match.

Over the years, several factors, such as venue of the match, toss outcome, previous head-to-head encounters between the two teams, match type (day/day-night), etc., have been studied in order to predict the outcome of an ODI cricket match. However, players are the fundamental unit of a team. Players of one team work against the players of the opponent team in order to win a match. Therefore, the strengths and abilities of the players of a team play a key role in deciding the outcome of a match. Our work suggests that the relative team strength between the two competing teams forms a distinctive feature for predicting the winner. Modeling the strength of a team boils down to modeling individual batting or bowling performances – which is the fundamental idea behind our approach. We use career statistics as well as the recent performances of a player to model him. Using the relative strength of one team versus the other, along with two player-independent features, namely, the toss outcome and the venue of the match, we evaluate multiple supervised machine learning algorithms to predict the winner of the match.

On the other hand, assessing the actual performances of the players at the end of a match is a critical task. It helps in segregating the players who are contributing to the team from the ones who are failing to deliver on the ground, which further helps in balanced team selections in future. However, evaluating the performances of players is not a straight-forward task. In a game of cricket, players have multiple roles, and different players perform under different game scenarios across the course of a match. Hence, combining and comparing the batting and bowling performances of a player, on a common scale, is a challenging task and often becomes a subjective decision.

Therefore, in the second half of this thesis, we propose a methodology to quantitatively assess the performances of individual players in a single game of ODI cricket match. We introduce a new measure, called the *Work Index*, which represents the amount of work yet to be done by a team to reach their expected target score. Work Index incorporates several important aspects of a game, including the current stage of the match, the progress so far as compared to the initial estimations, the two competing teams' strengths, etc. We measure the Work Index for both the batting as well as bowling teams, namely, the *Batting Work Index* and the *Bowling Work Index*. The former denotes the amount of work to be done by the batting team to reach the target, while the latter represents the amount of work to be done by the bowling team to restrict the batting team from reaching the target. Using these two work indices, we calculate a utility score for each player which represents his batting and bowling contributions towards achieving the team's overall goal.

Apart from estimating the player contributions in a match, assessing player performances has several other applications. Akin to many other sports, the Player-of-the-Match title is awarded to the player who played the most significant role in the match. Today, in cricket, it is chosen by the match committee and the commentators which makes it a subjective decision. Therefore, we propose a methodology to determine the player of the match using the player utility scores calculated by our approach. Moreover, comparing and ranking the players over time, for their varying roles, has been of great interest in the cricketing realm. We, further, demonstrate the adaptability of the player utility scores to help find the best batsmen, bowlers and all-rounders of all time.

The major contributions of this thesis are listed below –

- We propose methods to model batsmen, bowlers and teams, using various career statistics and recent performances of the players.
- To predict the winner of ODI cricket matches, we propose a novel dynamic approach which takes into account the varying strengths of the individual players and reflects the changes in player combinations over time.
- We propose a real-time measure, named *Work Index*, which represents the amount of work yet to be done by a team to achieve their target in an ODI cricket match.
- We introduce a new method to quantitatively assess the player performances in an ODI cricket match. Our method incorporates the game situation at all stages and enables us to compare batsmen and bowler performances on a common ground.
- We propose a method to select the Player of the Match award in ODI cricket matches, and demonstrate its quantitative superiority over other models.
- Moreover, our method can be used to find the best batsmen, bowlers and all-rounders in a given time frame, over multiple matches.

1.4 Thesis Workflow

The rest of the thesis is organized as follows – Chapter 2 gives a brief overview on related work. Studies related to all the major aspects of cricket have been discussed, followed by detailed discussions of the works pertaining to winner predictions and assessing player performances.

Chapter 3 describes our model for predicting the winner of ODI cricket matches using team composition based approach. Mechanics of the various algorithms used by our model have been explained and experimental validity of the same have been demonstrated.

Chapter 4 presents our model to quantitatively assess player performances in an ODI cricket match. The motivation behind using various statistics has been discussed. The superiority of our approach over the others has been experimentally demonstrated.

Chapter 5 concludes the thesis.

Chapter 2

Related Work

Statistical modeling has been used in sports over decades and has contributed significantly to success on the field. Popular team sports, such as Football, Basketball and Baseball, have always been characterized by a high degree of analytics. Michael Lewis' entertaining story about the use of data analysis in baseball in his book, *Moneyball: The Art of Winning an Unfair Game, 2004*, is arguably the most visible account of sports analytics. Moneyball tells the story of a manager who led his team to success despite their low budget by using computer based analytics to draft players. Although Moneyball isn't the earliest example of analytics in baseball, it sure was the catalyst for introducing the broader sports community to the potential benefits of quantitative analysis.

Cricket is one of the most popular team sports in the world, second only to soccer, with 2-3 billion fans all over the world [1]. Although the game of cricket is a relatively new and upcoming research area, it has been a statistician's delight. Here, in this chapter, we start with a brief review of the developments in the field of cricket analytics, and then we discuss the works done specific to winner predictions in cricket. Lastly, we give an overview of the studies related to assessing player performances, player ratings and rankings.

2.1 Brief History of Cricket Analytics

In literature, Duckworth and Lewis proposed a solution, called D/L method [2], to reset targets in rain interrupted matches. It is designed so that neither team benefits or suffers from the shortening of the game and so is totally fair to both. It is easy to apply, requiring nothing more than a single table of numbers and a pocket calculator, and is capable of dealing with any number of interruptions at any stage of either or both innings. The method is based on a simple model involving a two-factor relationship giving the number of runs which can be scored on average in the remainder of an innings as a function of the number of overs remaining and the number of wickets fallen. It is shown how the relationship enables the target score in an interrupted match to be recalculated to reflect the relative run scoring resources available to the two teams, that is overs and wickets in combination. This remains to be one of the most pioneering works in cricket history. The method was adopted by the ICC in 1998 for resetting targets in international cricket matches and still is in place. However, after a few years, the authors themselves proposed an amendment to the existing method, known as the D/L Professional Edition [3], to cope up with high-scoring matches when the basic model's assumptions begin to break down. Later, [4] proposed an alternative method, the VJD system, using the concepts of normal (PAR) and target scores. They constructed easy-to-use tables for employing the method using regression equations obtained from a detailed statistical analysis of a data set of closely fought matches. The VJD system provides more appropriate targets in the few situations where the D/L method seems to fail. And therefore, this method was adopted by the Indian Cricket League (ICL 2007-2009). Following that, there have been many other alternatives proposed in literature. For instance, [5, 6] proposed methods to revise the target such that the probabilities of each team winning the match, as calculated before and after the interruption, are preserved. I. G. McHale and M. Asif [7] provided a modified D/L method where they improved the functional form for the model describing methods for runs to be scored in an innings. They further suggested that it is reasonable to use a single method for both the ODI and T20 formats of the game.

Good team selection is vital for success in all sports. Therefore, the problems of optimal team selection and an optimal batting order have also been of great interest for the researchers. T. B. Swartz, P. S. Gill, D. Beaudoin, et al. [8] proposed a method for optimal or nearly optimal batting orders in one-day cricket by conducting a search over the space of permutations of batting orders where simulated annealing is used to explore the space. They use simulation to obtain the objective function, i.e., the mean number of runs per innings, as it is not available. The simulation component generates runs ball by ball during an innings taking into account the state of the match and estimated characteristics of individual batsmen. They further applied their methods to the national team of India based on their performance in one-day international cricket matches. J. M. Norman and S. R. Clarke [9], on the other hand, proposed a simplified model using dynamic programming. They showed that in all forms of cricket, significant increases in expected score result if captains allow a variable batting order and base their decision on the state of the game, rather than using a set batting order. Some [10, 11, 12] used integer programming model to select an optimal team for the ODIs and T20s. Recently, B. S, S. RP, Abhijeet, and R. S [13] proposed a methodology for objective evaluation of players for team selection. Their approach involves evaluating a player across multiple dimensions viz. batting and bowling based on role in the team, context, opponents etc. They considered all possible team level metrics that affect the outcome (win or loss) of the match, translate them to individual player metrics, develop player evaluation utility and use it for team selection. Their model was successfully able to predict the team selection with an accuracy of 83%.

There have been several other interesting works in cricket literature. B. M. De Silva and T. B. Swartz [14] used the statistics from ODIs played in 1990s to make to conclusions – contrary to widespread opinion, winning the coin toss at the outset of a match provides no competitive advantage, and the advantage of playing on one's home field increases the log-odds of the probability of winning by 0.5. H. H. Lemmer [15] proposed a method to measure strangling, a dramatic form of choking in cricket. In limited overs cricket, the team batting first sets a target for the team batting second. The latter team may win the match, draw the match or lose it by not reaching the target. Various scenarios of losing are possible; e.g. simply not reaching the target, or some form of choking after being in a strong position. Authors considered the situation where a team batting second in a limited overs cricket match needs to score only a few more runs with many balls still to be bowled, but is bowled out. This phenomenon is called ‘strangling’ because the bowling team succeeded in bowling their opponents, who were in a strong batting position, all out - they have strangled them. They proposed a criterion to measure the severity of strangling. The measure is based on the strength of the batting team just before this disaster struck.

2.2 Literature on Winner Prediction

Akin to all the other sports, winning is the ultimate goal in a game of cricket. Predicting the outcome of cricket matches in all the three major formats of the game have been of great interest in cricket literature. S. Akhtar and P. Scarf [16] forecasts match outcome in test cricket, session by session, while the match is in progress. Match outcome probabilities at the start of each session are fore-casted using a sequence of multinomial logistic regression models. These probabilities can assist a team captain or management in considering a certain aggressive or defensive batting strategy for the coming sessions. The authors further investigated how the outcome probabilities (of a win, draw, or loss) and co-variate effects vary session by session. The co-variates fall into two categories, prematch effects (strengths of teams, ground effect, home field advantage, outcome of the toss) and in-play effects (score or lead, overs-used, overs-remaining, run-rate, and wicket resources used). Their results indicate that the lead has a small effect on the match outcome early on but is dominant later. Prematch team strengths, ground effect and home field advantage are important predictors of a win early on. And wicket resources used remains important throughout a match. F. Munir, M. K. Hasan, S. Ahmed, S. Md Quraish [17] predicts the outcome of a T20 cricket match while the match is in progress. Although forecasting a T20 cricket match is a challenging problem as the momentum of the game often changes very drastically, the authors take into account the previous data of matches played between the two teams and used decision tree algorithm to predict the outcome of the match.

Similarly, predicting the outcome of ODI cricket matches have also been studied widely. M. Bailey and S. R. Clarke [18] used past data to create a range of variables that could independently explain statistically significant proportions of variation associated with the predicted run totals and match outcomes. Such variables include home ground advantage, past performances, match experience, performance at the specific venue, performance against the specific opposition, experience at the specific venue and current form. Using a multiple linear regression model, they numerically weighted the prediction variables according to statistical significance and used them to predict the match outcome. With the use of the Duckworth-Lewis method to determine resources remaining, at the end of each completed over, they updated the predicted run total of the batting team to provide a more accurate prediction of the match outcome. V. V. Sankaranarayanan, J. Sattar, and L. V. Lakshmanan [19] built a prediction system that takes in historical match data as well as the instantaneous state of a match, and predicts future match events culminating in a victory or loss. They modeled the game using a subset of match parameters, using a combination of linear regression and nearest-neighbor clustering algorithms. By using a weighted combination of both historical and instantaneous features, they simulate and predict game progression before and during a match.

While some of the works aim to predict winner while the game is in progress, others proposed models to predict the winner at the start of a match. M. Khan and R. Shah [20] identified the factors which play a key role in predicting the outcome of an ODI cricket match and also determined the accuracy of the prediction made using the technique of data mining. In the analysis, statistical significance for various variables which could explain the outcome of an ODI cricket match are explored. Home field advantage, winning the toss, toss outcome (batting first or fielding first), match type (day or day/night), competing teams, venue familiarity and season in which the match is played are the key features studied for the research. For purposes of model-building, the authors adopted three algorithms: Logistic Regression, Support Vector Machine and Naive Bayes. Logistic regression is applied to data already obtained from previously played matches to identify which features individually or in combination with other features play a role in the prediction. SVM and Naive Bayes classifiers are then used for model training and predictive analysis. Graphical representations and confusion matrices are used to represent the comparative analysis. A. Kaluarachchi and A. S. Varde [21] studied several interesting factors including home game advantage, day/night effect, winning the toss and batting first. Further, they used artificial intelligence techniques, more specifically Bayesian classifiers in machine learning, to predict how these factors affect the outcome of an ODI cricket match. Based on the emerged results, they developed a software tool called *CricAI*. The tool outputs the probability of victory in an ODI cricket match using input factors such as home game advantage available at the beginning of the match.

2.3 Literature on Player Modeling

Quantitative assessment and classification of players in a team has been of great interest amongst researchers irrespective of the sport. Some [22, 23, 24, 25, 26] discuss the various multi-criteria decision making models of player evaluation for multiplayer sports, namely, Baseball, Basketball, Cricket and Football.

In cricket, ranking and comparing batsmen and bowlers have been studied widely across multiple formats of the games. S. Akhtar, P. Scarf, and Z. Rasool [27] developed a new player rating system for test cricket. They used multinomial logistic regression to model match outcome probabilities session by session. They further used these probabilities to measure the overall contribution of players to the match outcome based on their individual batting, bowling and fielding contributions during each session. The proposed method of contribution has the potential for rating players over time and for determining the best player in a match, a series or a calendar year. They used the results from 104 matches (2010-2012) to illustrate the method. D. Beaudoin and T. B. Swartz [28] proposed a new statistic for assessing the performance of batsmen and bowlers in ODI cricket. The statistic is the ratio of runs scored to resources consumed where resources are defined according to the Duckworth-Lewis method of resetting targets [2]. A standard error has been provided to help determine real differences in performance. Various comparisons have also been made with traditional measures of performance when applied to data obtained from one-day international cricket matches.

The use of neural networks has been very popular in literature. H. Saikia and D. Bhattacharjee [29] proposed a model to predict the performances of batsmen who entered the Indian Premier League (IPL), a franchise based T20 cricket tournament, in its fourth season only, based on analyzing the performance of batsmen in the first three seasons of IPL through multi-layer perceptron (MLP) neural network. They further calculated the actual performances of these batsmen to test the external validity of the neural network model. The model was found to be 66.67 percent accurate. This proposed prediction could help the franchises to decide which batsmen they should target to buy for their team and who should not be considered at all. S. R. Iyer and R. Sharda [30] employed neural networks to predict each cricketer's performance in the future based upon their past performances. They classified cricketers into three categories – performer, moderate and failure. Authors collected data on cumulative player performance from 1985 onwards until the 2006-2007 season. The neural network models were progressively trained and tested using four sets of data. The trained neural network models were then applied to generate a forecast of the cricketers near term performance. Based on the ratings generated and by applying heuristic rules the model recommends cricketers to be included in the World Cup 2007. They further evaluated the actual performances of the cricketers in the World Cup to validate the applicability of neural networks. The results show that the neural networks can indeed provide valuable decision support in a team selection process.

Several studies include graphical representations to compare players. A. Kimber [31] proposed a graphical method to compares bowlers. In cricket, bowlers are majorly compared on the basis of bowling average, economy rate and strike rate. Therefore, they presented a simple graphical display for making simultaneous comparisons on the basis of these three summary measures. The method is illustrated with some examples from Test Match and other First-Class cricket. P. J. Bracewell and K. Ruggiero [32] developed a control chart specifically for monitoring batting performances in cricket. Issues associated with the need to retain extreme values and limited sampling opportunities were overcome by using a parametric approach to obtain theoretical quartiles from the distribution of individual batting scores. A mixed distribution, named the Ducks 'n' Runs distribution, has been proposed. A beta distribution models zero scores (ducks) and a geometric distribution describes the distribution of non zero scores (runs). This suitability of this probability distribution model has been demonstrated using data from New Zealand first class batsmen in domestic three day cricket over a four year period. Changes in the process (player performance) have been detected using rules adapted from the supplementary rules for Shewhart control charts. P. J. Van Staden et al. [33] proposed a simple way of graphically comparing the bowling and batting performances of cricketers. They illustrated the demonstration using records from the IPL. The graphs are applicable to any format of cricket and can furthermore be used to identify different types of players, for example, offensive batsmen, bowling all-rounders, etc.

G. Barr and B. Kantor [34] proposed that in the one-day game, it is clearly not good enough for a batsman to achieve a high batting average with a low strike rate. Runs scored slowly, even without the loss of wickets, will generally result in defeat rather than victory in the one-day game. Assessing batting performance in the one-day game, therefore, requires the application of at least a two-dimensional measurement approach because of the time dimension imposed on limited overs cricket. Therefore, they proposed a new graphical representation with *Strike rate* on one axis and the *Probability of getting out* on the other, akin to the riskreturn framework used in portfolio analysis, to obtain useful, direct and comparative insights into batting performance, particularly in the context of the one-day game. Within this two-dimensional framework they developed a selection criterion for batsmen, which combines the average and the strike rate (Equation 2.1). Similarly, they proposed a selection criterion for bowlers which combines the bowling economy and strike rates (Equation 2.2). As an example of the application, they applied this criterion to the batting performances of the 2003 World Cup. They went on to demonstrate the strong and consistent performances of the Australian and Indian batsmen as well as provide a ranking of batting prowess for the top 20 run scorers in the tournament.

$$batScore = batStrikeRate * batAverage \tag{2.1}$$

$$bowlScore = bowlStrikeRate * bowlEconomy \tag{2.2}$$

where $batStrikeRate$, $batAverage$, $bowlStrikeRate$ and $bowlEconomy$ are described in Table 1.1.

2.4 Literature on Assessing Player Performances

Although a variety of studies have been done in past on modeling players, quantifying the performances of individual players in a single game of cricket for both the bowlers and batsmen simultaneously has not been studied in great depth. M. I. Johnston, S. R. Clarke, D. H. Noble, et al. [35] used dynamic programming formulation to develop a method of calculating the contribution, in runs, made by each player to the team's score in a game of one-day cricket. A. Lewis [36] uses Duckworth/Lewis methodology to create alternative measures of player performances in a game of cricket. These measures take into account the stages of innings when runs are scored or conceded and wickets are taken or lost.

P. Shah and M. Shah [37] developed a new measure called the *Pressure Index* (Equation 2.3), which reflects the pressure under which a team is playing or a batsman is batting. They consider runs scored, wickets taken, balls bowled, etc. for calculating the pressure index.

$$PI = CI \times 100 + [Wk.Wt/180) \times T \times (Br/B) \times (Rr/T) \tag{2.3}$$

,where PI refers to the Pressure Index, CI is the ratio of the current required run rate to the initial required run rate, Br is the number of balls remaining in the innings and Rr is the number of runs remaining to be scored. However, the defined pressure index is valid only for the second innings of a match, as it needs a target score to calculate the required run rates. Further, no motivation was given for their formula and the factor $Wk.Wt$ is not even defined in the paper. Later, D. Bhattacharjee and H. H. Lemmer [38] found out that in many cases, the value of PI decreased when a wicket went down and the number of runs scored was below the required run rate, which is totally unrealistic.

D. Bhattacharjee and H. H. Lemmer [38] went on to propose an alternative definition of the Pressure Index. It quantifies the pressure on the teams batting or bowling in limited overs cricket matches. They use D/L resources, as proposed in F. C. Duckworth and A. J. Lewis [2], ratio of the wickets lost and the current as well as the initial required run rates to quantify the pressure on a team.

Equation 2.4 represents the pressure on the batting team. Further, they used the batting pressure index to assess the individual batting contributions of a player. $Ave(PI)$ denotes the average pressure of the entire innings, and PI_i denotes the pressure level at the end of the i^{th} ball of the second innings of the match. If R_i denotes the number of runs scored by a player on the i^{th} ball, then the pressure-sensitive updated score of the player for those runs is given by Equation 2.5. Therefore, sum of all the adjusted scores of a batsman divided by the number of balls he faced renders his total batting contribution of the player.

$$batPI = (\frac{CRRR}{IRRR}) \times \frac{1}{2}\big[exp(RU/100) + exp(\sum w_i/11)\big] \tag{2.4}$$

, where $CRRR$ is the current required run rate, $IRRR$ is the initial required run rate, RU is the D/L resources used ([39]), and $\sum w_i$ refers to the number of wicket got out.

$$R_i^* = R_i \times (\frac{PI_i - 1}{Avg(PI)}) \tag{2.5}$$

Similarly, Equation 2.6 represents the pressure on the bowling team. Further, they used the bowling pressure index to assess the individual bowling contributions of a player. The difference between the bowling pressures at the end and start of an over bowled by the bowler indicates the increase in pressure on the batting team created while the bowler bowled the over. Therefore, average increase in pressure created while a bowler bowled all of his overs represents his bowling contributions in the match.

$$bowlPI = (\frac{IRRR}{CRRR}) \times \frac{1}{2}\big[exp(RU/100) + (11 - \sum w_i)/11\big] \tag{2.6}$$

However, the method could be used to quantify the pressure on a team only for the second innings of a match, where the batting team has a fixed target to chase. Also, their approach takes into account the ratio of the wickets fell down instead of incorporating the varying strengths of individual players. This is a very critical factor because teams do not play with a fixed number of specialized batsmen. Losing 6 wickets has a different impact on a team playing with 6 specialized batsmen as compared to a team playing with 7 specialized batsmen. Furthermore, no quantitative method, in any form, of validating the approach has been discussed.

Chapter 3

Winner Prediction: A Team Composition Based Approach

With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. In this chapter, we embark on predicting the outcome of an ODI cricket match using a supervised learning approach from a team-composition perspective at the start of the match. Our work suggests that the relative team strength between the competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual player's batting and bowling performances, forming the basis of our approach. We use career statistics as well as the recent performances of a player to model him. Player-independent factors have also been considered in order to predict the outcome of a match. Finally, we show that the kNN algorithm, for our approach, yields better results as compared to other classifiers.

3.1 Motivation

The work done in literature so far has already considered most of the factors affecting the game and produced excellent results. This has given us significant insights into the game of cricket, however, a very critical aspect that the team composition changes over time has not been studied yet.

A team is comprised of 11 players, and these 11 players are replaced over time. A team changes its composition depending on the match conditions, venue, and opponent team, etc. There could be various other reasons for the same, such as a player getting injured, or getting dropped from the team for his poor performance, or taking retirement from the sport itself. The average number of player changes per match for each team and the total number of distinct players who have played for each team in our dataset is shown in Figure 3.1 and Figure 3.2 respectively.

Therefore, relying completely on historical team data is not only insufficient, but also fallacious since it does not portray the current competence of a team. Taking such obsolete factors into account might lead us to incorrect conclusions. We see this problem as a dynamic one, where we need to build a model which can reflect the changes that take place over time, thereby self-adjusting itself over every match played. Thus, having a prediction model, which takes into account the players playing in a given match, is pivotal and stays robust over time.

We built a model to predict the winner of a given ODI match based on this intuition and have achieved state-of-the-art results to the best of our knowledge.

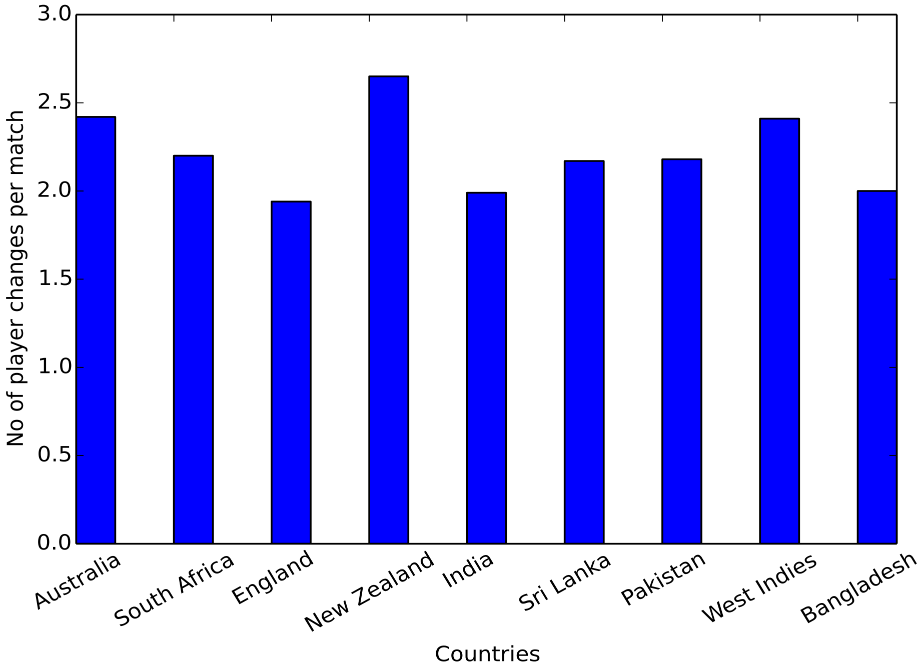

Figure 3.1: The average number of player changes per match for all the teams in the years 2010-2014. We can see that on average, a team changes around 2 players per match. Thus, about a fifth of the team changes per match

3.2 Dataset

The basic details of each match including the two competing teams, the venue, the winner and the margin of victory, along with the participating players for each match have been scrapped from the cricinfo website [40]. The career statistics of each player at the end of every international match he has played, has also been scrapped from the cricinfo website. Table 3.1 shows some sample entries for the career statistics of players at the end of a match. Notice that the career statistics of *RT Ponting* has updated from one match to another match.

For this study, we focus on all the ODI matches played between 2010 and 2014. The major data pre-processing steps taken include:

- We have restricted our study to only top 9 ODI-playing teams, namely, Australia, India, England, South Africa, Sri Lanka, Pakistan, New Zealand, Bangladesh and West Indies.

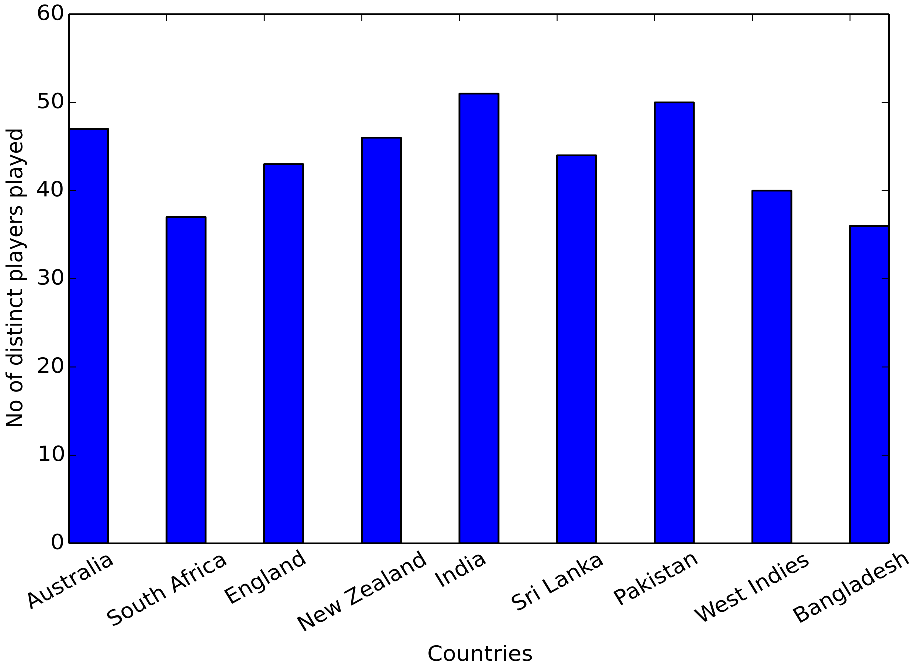

Figure 3.2: The total number of distinct players played for each team in the years 2010-2014. The high number of distinct players (35-50) in a team of 11, shows that the mentioned factors (Section 3.1) are prominent

Table 3.1: Feature Table

Match_ID	Player_Name	Country	#Matches	#Runs	#Wickets	...
65662	RT Ponting	Australia	215	7640	3	...
65664	RT Ponting	Australia	216	7669	3	...
209067	HH Gibbs	South Africa	169	5507	0	...

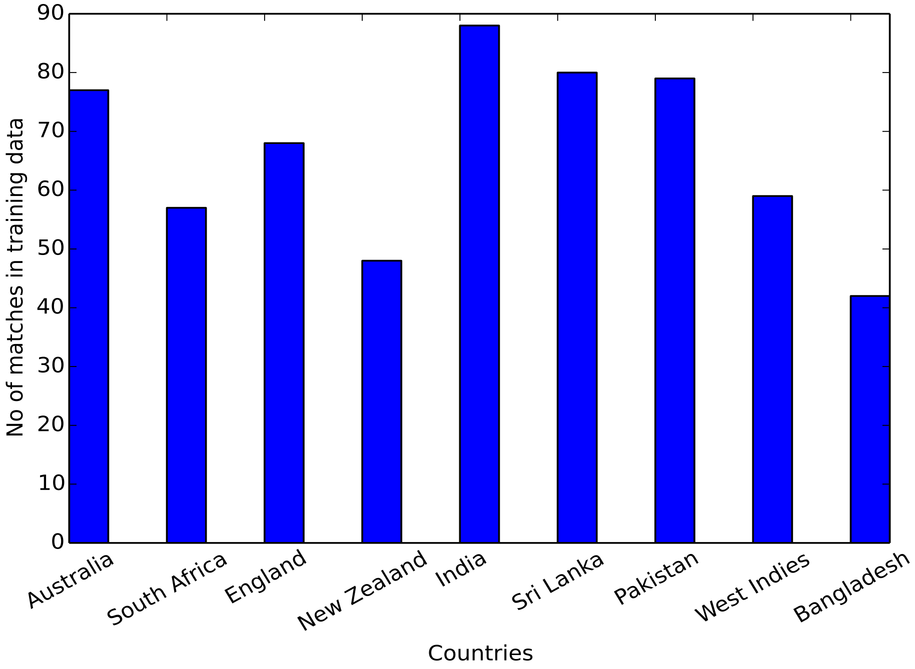

Figure 3.3: Total number of training samples for each team

- Since the impact of nature on the game cannot be foreseen, a total of 109 matches which were either interrupted by rain or ended up in a draw/tie, have been removed from the dataset.

- Venue of all the matches have been generalized to the *Country* where the match is played.

After pre-processing, there are a total of 366 matches in our dataset. We divided the dataset into two parts, namely, the test data and the training data. As we want to predict the outcome of future matches using the past data, the training dataset contains all the matches played in the years 2010 to 2013, and the test dataset contains all the matches played in the year 2014. There are a total of 299 matches in training dataset and 67 matches in test dataset.

The total number of training and testing samples for each team is shown in Figure 3.3 and Figure 3.4 respectively.

3.3 Problem Formulation

Table 3.2 defines some notations related to players and teams which we use consistently in the rest of the chapter. For a match m played between two teams A and B, the problem statement can be formalized as below.

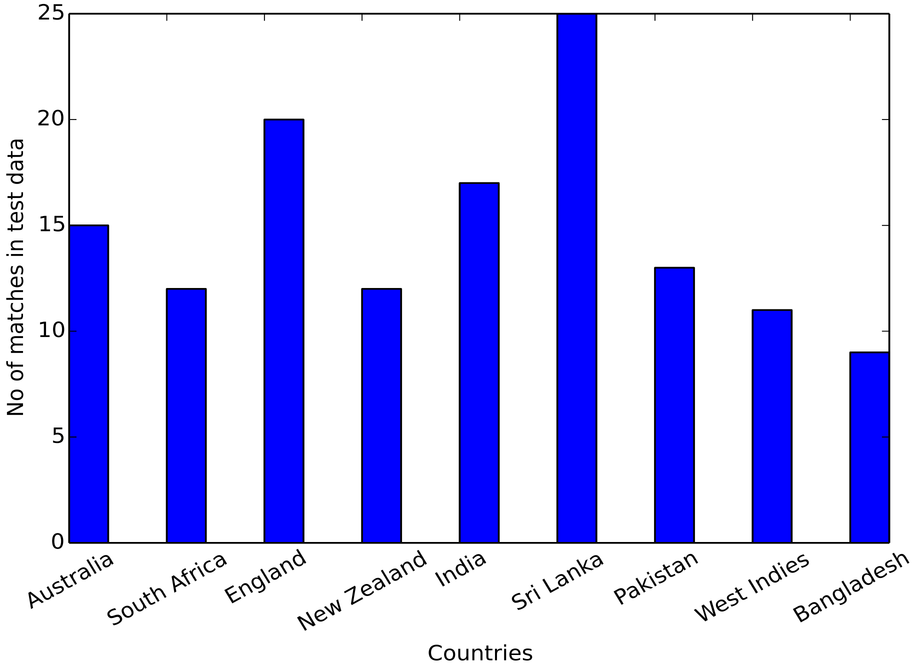

Figure 3.4: Total number of test samples for each team

Table 3.2: Notations.

Symbol	Description
$P(T,m)$	set of all players in team T playing in match m
$\phi(p, m)$	set of career statistics of player p in match m
$\phi(p)$	set of career statistics of player p
ϕ	set of career statistics
$T(m)$	the outcome of the toss for match m
$V(m)$	the venue (Country) where match m is played
$W(m)$	the winner of match m

Given:

- *Career statistics of each player* $p \in \{P(A, m) \cup P(B, m)\}$: $\phi(p, m)$
- *The Outcome of the toss:* $T(m)$
- *The Venue of the match:* $V(m)$

To Predict:

- *The Winner of the match:* $W(m)$

3.4 Methodology

A team is made up of a combination of batsmen and bowlers. Due to various reasons such as injuries, poor performances, venue, etc., the combination changes from match to match. In such a dynamic environment, considering only generalized features such as the venue of the match, the past head-to-head encounters of the two teams, winning ratios, toss decision, etc., is insufficient. Therefore, modeling the players for each team is of pivotal importance as it strongly accounts for the overall team strength. In order to predict the winner of a given match, we start with modeling the 22 players playing the match into batsmen and bowlers. We model the two teams by inspecting the individual player's career statistics and his active participation in recent matches to render the relative dominance one team has over the other. Taking base features into account, like toss decision and venue of the match along with the relative team strength, we adopt supervised learning algorithms to predict the winner of the match. The overall work-flow of our method is shown in Figure 3.5.

3.4.1 Modeling Batsmen

The batting ability of a player has a significant contribution in shaping the outcome of a match. A team usually comprises of a set of 6-7 specialist batsmen out of 11 players. The batsmen form the back bone of a team, helping it in posting a high score or chasing down a competitive total.

Matches Played, *Batting Innings*, *Batting Average*, *Number of Centuries* and *Number of Fifties* are the major features contributing to the career statistics of a batsman. As mentioned in Table 3.2, ϕ represents the set of career statistics of a player. Further, let $\phi_{Matches\_Played}$ represent the number of matches played by the player, and so on.

To model a batsman's capabilities, we have two types of datasets to get necessary insights about a player's characteristics. First, we examine his career performances to determine his potential as a contender. Second, we consider his recent match scores to analyze his prevailing *form*, where *form* of a batsman determines his contribution to the team in recent matches, which also reflects his confidence levels. Therefore, we define two scores for a batsman as follows:

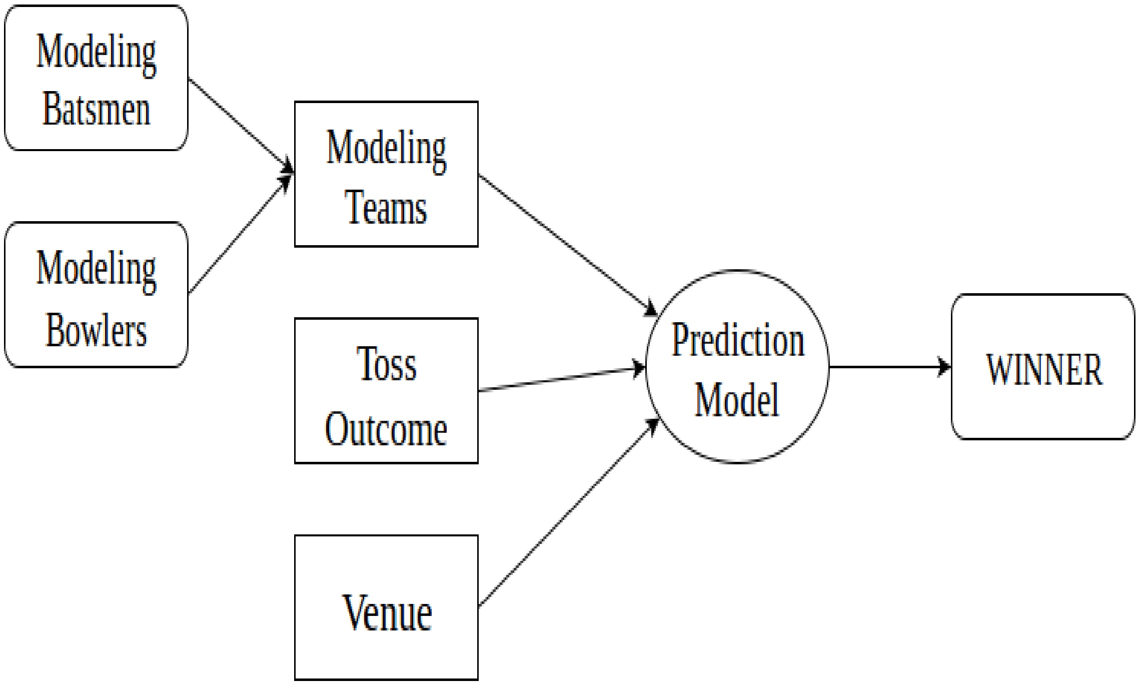

Figure 3.5: Figure depicts the overall work-flow of our approach. We model the players into batsmen and bowlers using their career and recent performance statistics, which in-turn helps in modeling teams. Using toss outcome and venue details, along with the relative team strength as the features, we use supervised models to predict the winner of the match.

- *Career_Score*: Statistics are recorded for each player during a match, and aggregated over his career. The evaluation of the performance of players in cricket using career statistics of a batsman gives us perceivable insights into his potential and his feats over time. The career statistics tell us about the longevity of a player's career, which is directly related to his potential. This means that the longer a player is retained in a team, the better is his ability to demonstrate his potential. In cricket, some of the aspects which are recorded for a batsman are the number of runs scored, batting average, number of centuries, etc.

- *Recent_Score*: In case of a batsman, it is well-known that his performance highly depends upon his recent *form*. We thus take into account the performance statistics from his recently completed matches, which directly contributes towards his current *form*. Therefore, we not only consider the career statistics, but also the player's ongoing confidence and temperament, as it has a significant impact, and notably contributes to modeling the batsman.

3.4.1.1 The Algorithm

The pseudo code of the algorithm to model batsmen for a given match is given in Algorithm 1, and is described below.

The pseudo code of the algorithm to model the batsmen for a given match is given in Algorithm 1. Lines 2-6 calculate a player's *Career_Score* using his overall career statistics. Variable u (line 3) is the ratio of the number of matches in which the batsman batted to the total number of matches he played. It captures whether the player is a full-time specialist batsman or not. Higher values of u indicate that the player often bats at the top of the batting order and hence he gets to bat in almost every match. On the other hand, lower values of u tell us that the player bats lower down the batting order and his chances of batting in the next match is also comparatively low. Variable ϕ_{Career_Score} (line 6) takes all the career statistics into account, and therefore signifies the *Career_Score* of the batsman. Similarly, lines 7-8 calculate the *Recent_Score* of a batsman. Variable M (line 7) holds the recent matches played by the player. Variable ϕ_{Recent_Score} (line 8) captures the *Recent_Score* of a batsman, which is the average number of runs scored by the player in his recent games. Since the $Career\_Score$ and the $Recent\_Score$ of players have different ranges, we have normalized them (lines 11-12) to lie in a common range of [0,1]. Finally, variable $\phi_{Batsman\_Score}$ (line 13) stores the $Batsman\_Score$ of a player which is a combination of his $Career\_Score$ and $Recent\_Score$.

Notice that the algorithm uses 7 variables namely $\{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7\}$, which represent the weights assigned to different features in modeling the batsmen. These values would be discussed in the experiment section 3.5.1.

3.4.2 Modeling Bowlers

Even though cricket is called a batsman's game, one cannot undermine the importance of specialist bowlers in a team. A team usually comprises of a set of 4-5 specialist bowlers out of 11 players. A good

Algorithm 1 Modeling Batsmen

```
Input: Players $p \in \{P(A, m) \cup P(B, m)\}$, Career Statistics of player $p$: $\phi(p)$
Output: $Batsmen\_Score$ of all the players: $\phi_{Batsman\_Score}$
for all players $p \in \{P(A, m) \cup P(B, m)\}$ do
    $\phi \leftarrow \phi(p)$
    $u \leftarrow \sqrt{\frac{\phi_{Bat\_Inngs}}{\phi_{Matches\_Played}}}$
    $v \leftarrow \mu_1 * \phi_{Num\_Centuries} + \mu_2 * \phi_{Num\_Fifties}$
    $w \leftarrow \mu_3 * v + \mu_4 * \phi_{Bat\_Avg}$
    $\phi_{Career\_Score} \leftarrow u * w$
    $M \leftarrow$ Last $\mu_5$ matches played by $p$
    $\phi_{Recent\_Score} \leftarrow mean(M^p_{Runs})$
end for
for all players $p \in \{P(A, m) \cup P(B, m)\}$ do
    $\phi_{Career\_Score} \leftarrow \frac{\phi_{Career\_Score}}{max(\phi_{Career\_Score})}$
    $\phi_{Recent\_Score} \leftarrow \frac{\phi_{Recent\_Score}}{max(\phi_{Recent\_Score})}$
    $\phi_{Batsmen\_Score} = \mu_6 * \phi_{Career\_Score} + \mu_7 * \phi_{Recent\_Score}$
end for
```

bowling unit can impact a game by restricting the opponent team to a lower score and hence directly affects the outcome of the game. To model a bowler, we are examining his career performances to estimate his potential for the next match. The past performances of a bowler and his career records give us a good insight into his capabilities.

Matches Played, *Bowling Innings*, *Wickets Taken*, *Five Wicket Hauls*, *Bowling Average* and *Bowling Economy* are the major features contributing to the career statistics of a bowler.

3.4.2.1 The Algorithm

The pseudo code of the algorithm to model bowlers for a given match is given in Algorithm 2, and is described below.

Algorithm 2 Modeling Bowlers

```
Input: Players $p \in \{P(A,m) \cup P(B,m)\}$,
       Career Statistics of player $p$: $\phi(p)$
Output: $Bowler\_Score$ of all the players: $\phi_{Bowler\_Score}$
1: for all players $p \in \{P(A,m) \cup P(B,m)\}$ do
2:     $\phi \leftarrow \phi(p)$
3:     $u \leftarrow \sqrt{\frac{\phi_{Bowl\_Inngs}}{\phi_{Matches\_Played}}}$
4:     $v \leftarrow \mu_1 * \phi_{Five\_Wkt\_Hauls} + \mu_2 * \phi_{Wkts\_Taken}$
5:     $w \leftarrow \phi_{Bowl\_Avg} * \phi_{Bowl\_Eco}$
6:     $\phi_{Bowler\_Score} \leftarrow \frac{u*v}{w}$
7: end for
```

Variable u (line 3, Algorithm 2) is the ratio of number of matches in which the bowler bowled to the total number of matches he played. It captures whether the player is a full-time specialist bowler or not. Higher values of u indicate that the player often bowls top in the bowling order and hence he gets to bowl in almost every match. On the other hand, lower values of u tell us that the player is a part-time bowler who doesn't bowl in every match he plays and his chances of bowling in the next match is also comparatively low.

Variables v and w (lines 4-5) consider other statistically significant features of a bowler such as bowling average, bowling economy, the total number of wickets taken, etc.

Finally, variable ϕ_{Bowler_Score} (line 6) takes everything into account, and therefore signifies the $Bowler\_Score$ of the player.

Notice that the Algorithm 2 uses 2 variables namely $\{\mu_1, \mu_2\}$, which represent the weights assigned to different features in modeling the bowlers. These values would be discussed in the experiment Section 3.5.2.

Also, notice that unlike batsmen, we haven't considered the recent performances of a bowler in calculating his *Bowler_Score*. This is due to the lack of data, as we do not have match-wise individual performances of every bowler.

3.4.3 Modeling Teams

The batsmen and the bowlers are the fundamental units of a team. Therefore, using the modeled batsmen and bowlers, we intend to define an overall score of a team with respect to the other. We define the batting score of a team as the summation of the batting scores of all its players. Similarly, the bowling score of a team is defined as the summation of the bowling scores of all its players. We have directly used the scores of all the players in the team score, as the variable u in the Algorithms 1 and 2 already takes care of the weighted contribution of individual players to the team score. Our algorithm to find the relative strength between two teams, A and B, competing against one another in a match m is shown in Algorithm 3. Since the *Batsman Scores* and the *Bowler Scores* have different ranges, we first normalize them to lie in the same range of [0,1] (lines 1-4). Lines 5-8 of the Algorithm calculate the batting and bowling scores of both the teams. Variable $S(A/B)$ (line 9) captures the relative strength of team A against team B. The algorithm follows the fundamental aspect of the game strategy where the batsmen of one team work against the bowlers of the other team and vice-versa.

Algorithm 3 Relative Strength between Two Teams

```
Input: Players $p \in \{P(A,m) \cup P(B,m)\}$,
    Batsman_Score: $\phi^{p}_{Batsman\_Score}$, Bowler_Score: $\phi^{p}_{Bowler\_Score}$
Output: Strength of Team $A$ against Team $B$: $S_{A/B}$
1: for all players $p \in \{P(A,m) \cup P(B,m)\}$ do
2:     $\phi_{Batsman\_Score} \leftarrow \frac{\phi_{Batsman\_Score}}{max(\phi_{Batsman\_Score})}$
3:     $\phi_{Bowler\_Score} \leftarrow \frac{\phi_{Bowler\_Score}}{max(\phi_{Bowler\_Score})}$
4: end for
5: $Bat\_Strength_A \leftarrow \left( \sum_{p \in P(A,m)} \phi^{p}_{Batsman\_Score} \right)$
6: $Bowl\_Strength_A \leftarrow \left( \sum_{p \in P(A,m)} \phi^{p}_{Bowler\_Score} \right)$
7: $Bat\_Strength_B \leftarrow \left( \sum_{p \in P(B,m)} \phi^{p}_{Batsman\_Score} \right)$
8: $Bowl\_Strength_B \leftarrow \left( \sum_{p \in P(B,m)} \phi^{p}_{Bowler\_Score} \right)$
9: $S_{A/B} = \frac{Bat\_Strength_A}{Bowl\_Strength_B} - \frac{Bat\_Strength_B}{Bowl\_Strength_A}$
```

3.4.4 Feature Construction

The choice of right features plays a key role in the success of a prediction model. For the problem at hand, which is predicting the winner of an ODI cricket match, we choose two other important features along with the relative strength of one team against the other. The first one is the venue of the match, and the second is the outcome of the toss. The venue of the match is important because of the *Home Team Advantage*. This basically means that the team playing at their home grounds has an advantage over the visiting team. This advantage is directly attributed to the psychological support that the home team gets from the audience in the ground, to the familiarity of the ground and environment, etc. The second feature is the outcome of the toss. It is has been observed in various matches that winning the toss also plays a major role in deciding the outcome of a match. The toss is directly associated with the nature of the pitch and the environment. For instance, a green pitch supports pace bowlers, so winning the toss and opting to bowl first could give the team an upper hand over the opponent team. Similarly, in humid conditions it becomes difficult for the bowlers to control the wet ball, so batting first is an optimal decision in that case.

Therefore, every match played between team A and team B in our dataset has three features: *Toss*, *Venue*, and $Stength_{A/B}$. $Stength_{A/B}$ and *Venue* are numeric features, whereas *Toss* is a binary feature. The value of *Toss* is 1 if team A is batting first, or 0 otherwise. The value of *Venue* is 1 if the match is being played at a home ground of team A, 0 if it is played at a home ground of Team B, and 2 otherwise. The value of $Stength_{A/B}$ is the relative strength of team A against team B which is calculated as described in Section 3.4.3. The target variable *Winner* defines the winner of a match. It is a binary variable. The value of *Winner* is 1 if the winner of the match is team A, and 0 if the winner is team B.

Notice that out of the two competing teams, any one of them could be considered as team A and all the feature values and the target value would update accordingly.

3.5 Experiments and Results

To assign the weights to various features in the Algorithms 1 and 2, we have used the 5-match ODI series played between India and Sri Lanka in July, 2012. A series of consecutive matches was deliberately chosen to study the impact of the recent scores of a batsman on his upcoming performances. The estimated scores of the players are compared against their actual performances. After exhaustive experimentation, the final weights are chosen such that the top 6 performing batsmen and bowlers (in terms of runs scored and wickets taken, respectively) from both the teams match with the top 6 batsmen and bowlers estimated by our algorithms. The final values of all the parameters defined in Algorithm 1, to model batsmen, are shown in the Table 3.3. Similarly, the final values of all the parameters defined in Algorithm 2, to model bowlers, are shown in Table 3.4.

Table 3.3: Values of parameters used in Algorithm 1

Parameter	**Value**
μ_1	20
μ_2	5
μ_3	0.3
μ_4	0.7
μ_5	4
μ_6	0.35
μ_7	0.65

In this section, we demonstrate the effectiveness of our approach experimentally. We show results for each of the major algorithms discussed in the previous sections.

Table 3.4: Values of parameters used in Algorithm 2

Parameter	**Value**
μ_1	10
μ_2	1

3.5.1 Modeling Batsmen

The number of runs scored by a batsman in a match depends upon various factors including the number of overs remaining, need of the hour to score quickly or to defend his wicket, the score posted by the opponent team, the position at which a batsman gets to bat, and many more. This makes it really difficult to judge the contribution of a batsman based only on numbers, given that we do not have required data to represent the exact scenario of a match. Despite this, we can make fair assumptions that a batsman scoring higher number of runs has added more value to the match, whereas the opposite might not be true. Therefore, we aim to use data to rank the players in the order of the estimated contribution they can make to guide their team towards victory with their batting skills.

We demonstrate how our model learns to update the $Batsman\_Score$ over matches and captures the dynamic nature of the game using the recent scores of the batsmen. We present the results of our model for the first 3 matches out of a 5-matches ODI series played between India and Sri Lanka in July, 2012.

Table 3.5: *Batsman Score* Table for First Match

Batsman	**Team**	**Batsman_Score**	**Runs Scored**
V Kohli	India	1.00	106
KC Sangakkara	Sri Lanka	0.78	133
DPMD Jayawardene	Sri Lanka	0.59	12
TM Dilshan	Sri Lanka	0.54	6
G Gambhir	India	0.53	3
V Sehwag	India	0.44	96

Table 3.6: *Batsman Score* Table for Second Match

Batsman	**Team**	**Batsman_Score**	**Runs Scored**
V Kohli	India	1.00	1
KC Sangakkara	Sri Lanka	0.89	NP
DPMD Jayawardene	Sri Lanka	0.60	NP
V Sehwag	India	0.57	15
TM Dilshan	Sri Lanka	0.53	50
G Gambhir	India	0.46	65

3.5.1.1 First Match

For the first match, the top 6 batsmen based on their estimated $Batsman\_Score$ along with the country and the actual number of runs they went on to score in that match are given in Table 3.5.

The top 2 batsmen, namely *V Kohli* and *KC Sangakaara* went on to score a century each, contributing significantly to their teams, and our model was successfully able to estimate their potential beforehand. Whereas, there are certain players like *TM Dilshan* and *G Gambhir* who were estimated to perform well based on their career records as well as recent performances but they failed to perform in that match. Also there are some players like *SK Raina* who was not ranked among the top 6 batsmen, but still went on to score 50 runs for his team.

3.5.1.2 Second Match

For the second match, the top 6 batsmen based on their estimated $Batsman\_Score$ are given in Table 3.6.

Table 3.7: *Batsman Score* Table for Third Match

Batsman	**Team**	**Batsman_Score**	**Runs Scored**
KC Sangakkara	Sri Lanka	1.00	73
V Kohli	India	0.89	38
DPMD Jayawardene	Sri Lanka	0.64	65
V Sehwag	India	0.63	3
TM Dilshan	Sri Lanka	0.45	4
WU Tharanga	Sri Lanka	0.44	8

Player *V Kohli* failed to score any runs despite being at the top of the list. *KC Sangakkara* and *DPMD Jayawardene* did not get to bat at all as the total number of runs scored by *India* were very less in the first innings and the first three batsmen of *Sri Lanka* won the match for them. This shows some of the unforeseeable challenges faced while modeling the game of cricket. On the other hand, *G Gambhir* and *TM Dilshan* scored a fifty each as per their estimated ranks. Notice that, owing to the high number of runs scored in the previous match, *V Sehwag* has climbed up the ladder from 6^{th} position to 4^{th} position in the table and his $Batsman\_Score$ has also improved from 0.44 to 0.57, thus showing the dynamic updates made by our algorithm over every match.

3.5.1.3 Third Match

For the third match, the top 6 batsmen based on their estimated $Batsman\_Score$ are given in Table 3.7.

The top 3 batsman have scored good number of runs with *KC Sangakkara* and *DPMD Jayawardene* getting a fifty each. Notice that the player *WU Tharanga* has gained himself a position in the top 6 batsmen because of his good performances in the last couple of matches. Also, *V Kohli* has lost the top position to *KC Sangakkara* because of his poor performance in the previous match. This shows that our model is able to capture the *form* of a batsman and is able to update the estimated $Batsman\_Scores$ in sync with it.

Given that there are several challenges associated with modeling a batsman such as varying target scores posted by the opponent team, number of overs remaining, the need to score quickly or to defend the wicket, etc., our model to rank batsmen is performing well. It is able to capture the *form* of a batsman and using it to make on-the-fly updates to the rank table over every match. While there are some missed cases, most of the times we are rightly able to estimate the $Batsman\_Score$ of all the players.

Table 3.8: *Bowler Score* Table for First Match

Bowler	**Team**	**Bowler_Score**	**Wickets Taken**	**Bowling Economy**
MG Johnson	Australia	1.00	0	5.60
DE Bollinger	Australia	0.81	2	6.40
M Morkel	South Africa	0.66	4	2.20
LL Tsotsobe	South Africa	0.65	2	4.00
JH Kallis	South Africa	0.63	0	6.00
DW Styen	South Africa	0.58	2	5.70

3.5.2 Modeling Bowlers

Bowlers in cricket matches play a crucial role in deciding the outcome of a match. Similar to batsmen, judging the contribution of a bowler in a match using numbers has many loopholes. The contribution of a bowler is not limited to taking wickets, but also in restricting the rate at which the opponent is scoring runs. On many occasions, the job of a bowler is to make one batsman play more balls than the other and the number of runs conceded is insignificant at those times. Therefore, we aim to use data to rank the players in the order of the estimated contribution they can make to guide their team towards victory with their bowling skills.

Here, we demonstrate how our algorithm is successfully able to filter out the best bowlers among all the players and rank them based on their skills. We show the results of our algorithm for 2 sample ODI cricket matches from our dataset in the following sections.

3.5.2.1 First Match

The first match we are considering, was played between Australia and South Africa in October, 2011. The top 6 bowlers based on their estimated $Bowler\_Score$ along with the country they play for, the number of wickets they took in that match, and their bowling economy rate are given in Table 3.8.

4 out of top 6 bowlers did perform well. *M Morkel* proved to be the best bowler for *South Africa* as estimated by our model. There are 4 *South Arfrican* bowlers in the top 6 bowlers, and *South Africa* did win the match with the help of a top class bowling performance from its bowlers. Contrary to what our model estimated, *MG Johnson* failed to perform according to his potential. Notice that all of these players are amongst the specialist bowlers in their respective teams and our algorithm is able to filter them out from all the 22 players playing the match.

Table 3.9: *Bowler Score* Table for Second Match

Bowler	**Team**	**Bowler_Score**	**Wickets Taken**	**Bowling Economy**
KD Mills	New Zealand	1.00	1	3.60
JM Anderson	England	0.86	3	3.44
GP Swann	England	0.68	1	3.30
TG Southee	New Zealand	0.54	3	3.70
TT Bresnan	England	0.40	0	3.66
MJ McClenaghan	New Zealand	0.34	2	4.90

Table 3.10: Modeling Teams Results

$Team_A$	$Team_B$	$Date$	$S_{A/B}$
Australia	South Africa	Oct 23, 2011	0.032
England	Sri Lanka	Jul 1, 2011	-0.228
Australia	Bangladesh	Apr 13, 2011	0.025

3.5.2.2 Second Match

The second match we are considering, was played between England and New Zealand in May, 2013. The top 6 bowlers based on their estimated $Bowler\_Score$ are shown in Table 3.9.

All the 6 players picked by our algorithm have bowled well. 10 out of 15 wickets were taken by these 6 bowlers in that match. *JM Anderson* proved to be the top bowler for *England* and was rightly picked by our algorithm. Notice that all of these 6 players are the specialist bowlers for their respective teams and our algorithm was once again able to filter them out among all the 22 players.

3.5.3 Modeling Teams

A team is composed of batsmen and bowlers. The strength of a team is, therefore, a function of the total strength of its bowling and batting units. In Algorithm 3, we calculate $S_{A/B}$, which is the relative strength of team A playing against team B in given match. In Table 3.10, we demonstrate the results of the same algorithm for some sample matches.

While it is difficult to manually judge the results, for some trivial cases, as that of the third one in Table 3.10, played between *Australia* and *Bangladesh*, the results go in sync with the ICC-Ranking of ODI teams. Positive values of $S_{A/B}$ denote that the strength of $Team_A$ is estimated to be greater than that of $Team_B$, and vice-varsa.

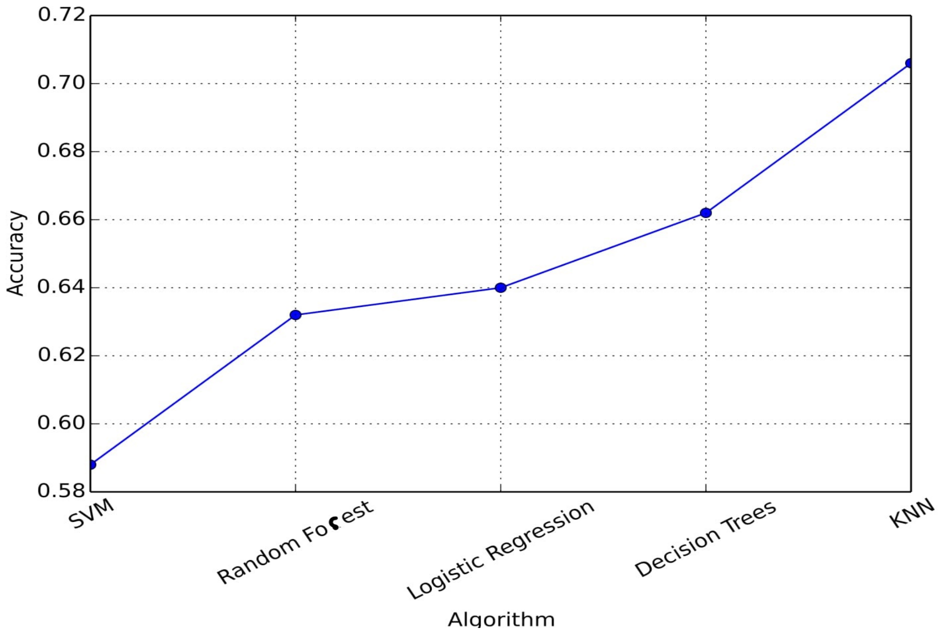

Figure 3.6: Figure depicts the accuracy for different supervised models. As it can be seen, kNN, with k=4, yields best results for our approach to predict the winner of a match.

3.5.4 Prediction Model

This section uses two kinds of features i.e binary and numeric features to create a classification model capable of predicting the binary outcome of a match. The numeric feature, $S_{A/B}$ (*Relative Team Strength*), seems like just one feature, but is truly a blend of many other primitive features which represent the standing of a team as a whole with respect to the other team. It inculcates the strength of a team both from the perspective of the batsmen and the bowlers. Using these features and the outcome of that match as the label, we evaluated a large number of binary classifiers using their scikit-learn implementation ([41]) to generate supervised classification models, including Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision Trees and kNN. The efficacy of kNN algorithm, with k=4, was statistically superior to those obtained by other classifiers, as shown in Figures 3.6 and 3.7. The idea of using the data of future matches to predict the outcome of past matches is absurd. Consequently, we could not carry out any sort of cross-validation procedure as it would interfere with the chronological order of the data.

Although we cannot directly compare these results with the prior state-of-the-art approaches due to differences in the dataset, it is noteworthy that the best accuracy in predicting the outcome of ODI cricket matches reported so far in the literature is between 0.68 and 0.70 ([19]). Team-wise winning accuracy, as predicted by our model, is shown in Figure 3.8. The figure shows that the teams like New Zealand, Sri Lanka and Pakistan depend heavily on their front-line players. If they perform, the team wins, otherwise the team loses most of the time. On the other hand, teams such as South Africa,

Table 3.11: Figure compares our kNN-based model with other baseline models. Our model yields better results as compared to the others.

Model	**Accuracy**
Model_1	0.56
Model_2	0.63
Our Model	0.71

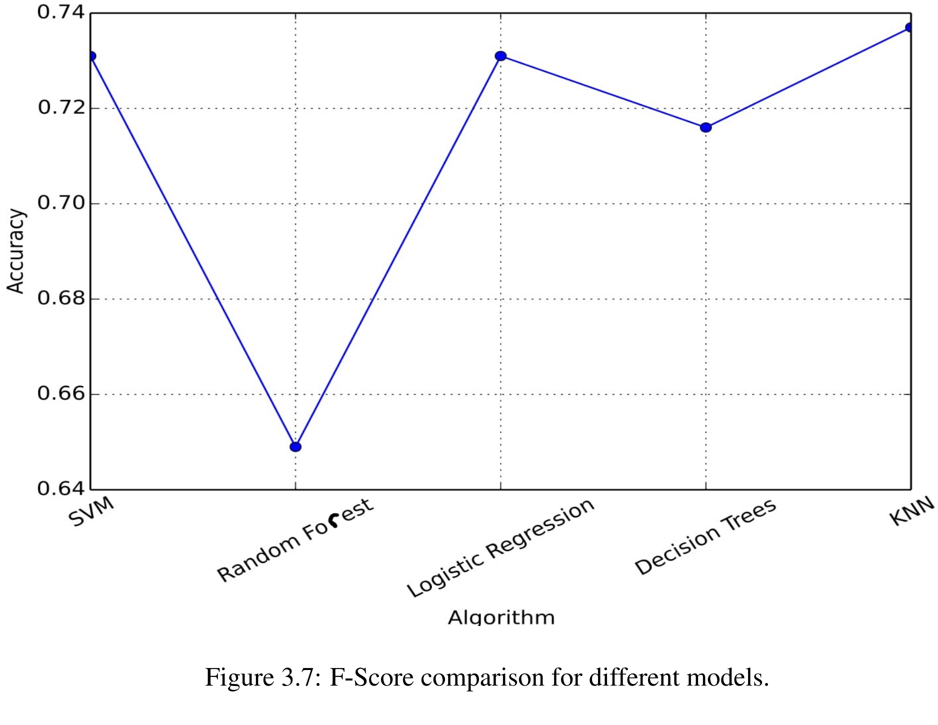

Figure 3.7: F-Score comparison for different models.

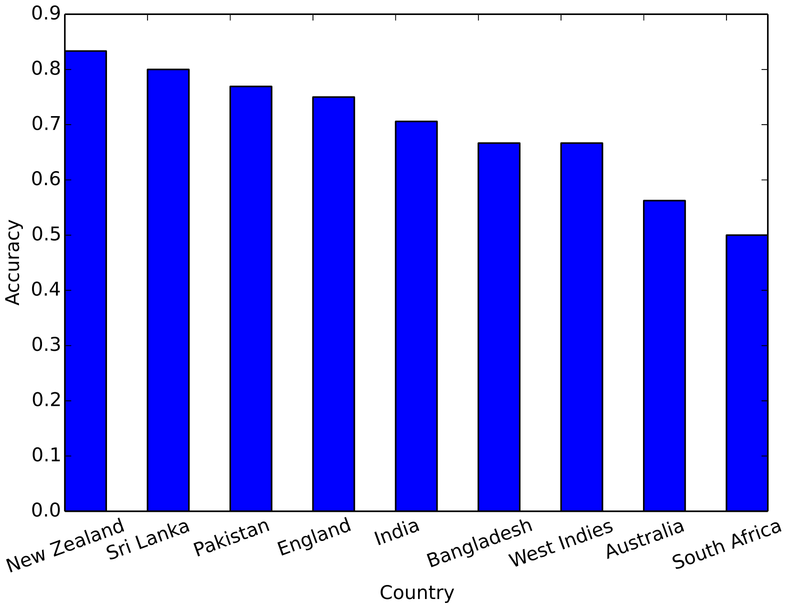

Figure 3.8: Figure shows the accuracy of our model for different countries. It depicts that some countries like New Zealand depend heavily on their front-line players, while other countries like South Africa do not.

Australia and West Indies have players where any one of them could win matches for them. These teams do not depend solely on the performances by their lead players.

3.6 Discussions

This chapter addresses the problem of predicting the outcome of an ODI cricket match using the statistics of 366 matches. The novelty of our approach lies in addressing the problem as a dynamic one, and using the participating players as the key feature in predicting the winner of the match. We observe that simple features can yield very promising results.

Chapter 4

Honest Mirror: Quantitative Assessment of Player Performances

Players have multiple roles in a game of cricket, predominantly as batsmen and bowlers. Over the generations, statistics such as batting and bowling averages, and strike and economy rates have been used to judge the performance of individual players. These measures, however, do not take into consideration the context of the game in which a player performed. Furthermore, these types of statistics are incapable of comparing the performance of players across different roles. In this paper, we present an approach to quantitatively assess the performances of individual players in single match of ODI cricket. We have developed a new measure, called the *Work Index*, which represents the amount of work that is yet to be done by a team to achieve its target. Our approach incorporates game situations and the team strengths to measure the player contributions. This not only helps us in evaluating the individual performances, but also enables us to compare players within and across various roles on a common scale. Multiple applications of our approach, including determining the *player of the match* and ranking the best players over a period of time, have been experimentally evaluated.

4.1 Introduction

A game of cricket is played between two teams of 11 players each, where a team comprises of batsmen and bowlers. The batsmen of one team work against the bowlers of the other team, and vice-versa, in order to win the match. Therefore, evaluating the performances of individual players in a game of cricket becomes very critical. It helps in segregating the players who are contributing to the team from the ones who are failing to deliver on the ground. However, evaluating the performances of players is not a straight-forward task. Traditionally, statistics such as batting and bowling averages, and strike and economy rates have been used to assess the performance of individual players. However, these statistics fail to incorporate several important aspects of the game. Runs scored or wickets taken under pressure at crucial stages are of more value as compared to scoring more number of runs or taking more number of wickets. Furthermore, assessing the overall performance of an individual cricketer requires a comprehensive evaluation of his contributions to the team, both in terms of his batting and bowling contributions. However, combining and comparing the batting and bowling performances of a player, on a common scale, is a challenging task and often becomes a subjective decision.

Therefore, in this chapter, we propose a methodology to quantitatively assess the performances of individual players in a single game of ODI cricket match. We introduce a new measure, called the *Work Index*, which represents the amount of work yet to be done by a team to reach their expected target score. Work Index incorporates several important aspects of a game, including the current stage of the match, the progress so far as compared to the initial estimations, the two competing teams' strengths, etc. We measure the Work Index for both the batting as well as bowling teams, namely, the *Batting Work Index* and the *Bowling Work Index*. The former denotes the amount of work to be done by the batting team to reach the target, while the latter represents the amount of work to be done by the bowling team to restrict the batting team from reaching the target. Using these two work indices, we calculate a *utility score* for each player which represents his batting and bowling contributions towards achieving the team's overall goal.

Apart from estimating the player contributions in a match, assessing player performances has several other applications. Akin to many other sports, the *player of the match* title is awarded to the player who played the most significant role in the match. Today, in cricket, it is chosen by the match committee and the commentators which makes it a subjective decision. Therefore, we propose a methodology to determine the player of the match using the player utility scores calculated by our approach. Moreover, comparing and ranking the players over time, for their varying roles, has been of great interest in the cricketing realm. We, further, demonstrate the adaptability of the player utility scores to help find the best batsmen, bowlers and all-rounders of all time.

4.2 Dataset

The basic details of each match including the two competing teams, the venue, the winner and the margin of victory, along with the participating players and their career statistics for each match have been scrapped from the cricinfo website [40]. This data has been scrapped for all the ODI cricket matches played during the period of 1st January, 2000 to 30th June, 2016. Ball-by-ball data for each match has been taken from the cricsheet database [42]. However, the ball-by-ball data is available only for the matches which have been played on or after 1st January, 2006, although with some missing matches. Therefore, we have restricted our study only to the period of January, 2006 to June, 2016. But, several statistics from the matches played during January, 2000 to December, 2005 have been used to determine certain parameters in our work.

We have focused our study to only the top 9 ODI-playing teams, namely, India, Australia, South Africa, England, Sri Lanka, Pakistan, New Zealand, Bangladesh and West Indies. Since the impact of nature on the game cannot be foreseen, a total of 216 matches which were either interrupted by rain or ended up in a draw/tie, have been removed from the dataset. In all, we studied a total of 786 ODI cricket matches.

4.3 Methodology

Our methodology to assess the player performances for a given ODI cricket match involves estimating a new measure, called the *Work Index*. As mentioned previously, Work Index incorporates several crucial qualitative and quantitative aspects of the game. The three parameters considered in calculating the work index are as follows:

- The progress, in terms of runs scored, towards chasing the set target.
- The current stand, in terms of scoring rate, of the batting team relative to the initial estimations.
- The remaining batting and bowling potentials of the batting and bowling teams, respectively.

The first and second parameters capture the quantitative aspects of the game situation in terms of the runs scored and the required run rate as compared to the initial required run rate for the batting team. On the other hand, the third measure captures the quality of the batsmen and bowlers remaining for the batting and bowling teams, respectively. Therefore, work index, a blend of these features, successfully captures the context at a given stage of the match.

Calculating the first and second parameters require us to know the target the team is trying to achieve. In an ODI cricket match, the team batting second has a predefined target, set by the opponent team, to chase in order to win the match. On the other hand, the team batting first does not have a fixed target to score. Estimating the target score for the first innings in itself is a research problem. Ideally, they aim at scoring as many runs as possible. But, as explained in [34], trying to score runs at a high rate increases the risk of losing wickets. Therefore, the team batting first keeps an achievable target in mind which they consider to be a defendable score, and try to score at least that many runs. Discussed in detail in Subsection 4.3.1, we use a possibly sub-optimal, yet a reasonable solution to estimate the target score and use it to measure the work indices for the first innings.

Similarly, calculating the third parameter requires us to model the player and team potentials for a given match. We will discuss, in detail, our approach towards modeling teams and players in Subsection 4.3.2.

4.3.1 Target estimation for first innings

Estimating the target score for the first innings is a difficult but crucial aspect of modeling the game. We use statistics from previous matches to estimate the target score. Figure 4.1 shows the average defendable target scores in different countries from the entire dataset. The difference in the average total could be attributed to varying stadium sizes, different pitch conditions, etc. The overall average defendable score, irrespective of the venue, is 264 runs, with a standard deviation of 30 runs. Therefore, at the start of the match, the initial target for the team batting first, denoted by $initTarget$, is set to the average number of runs scored in the first innings of all the matches played in the same country in the past, where the team batting first was able to successfully defend their score. In some cases where no match has been played in the same country previously, we set the estimated target to the average number of runs scored in the first innings in the past matches where the team batting first was successfully able to defend their score, irrespective of the venue of the match.

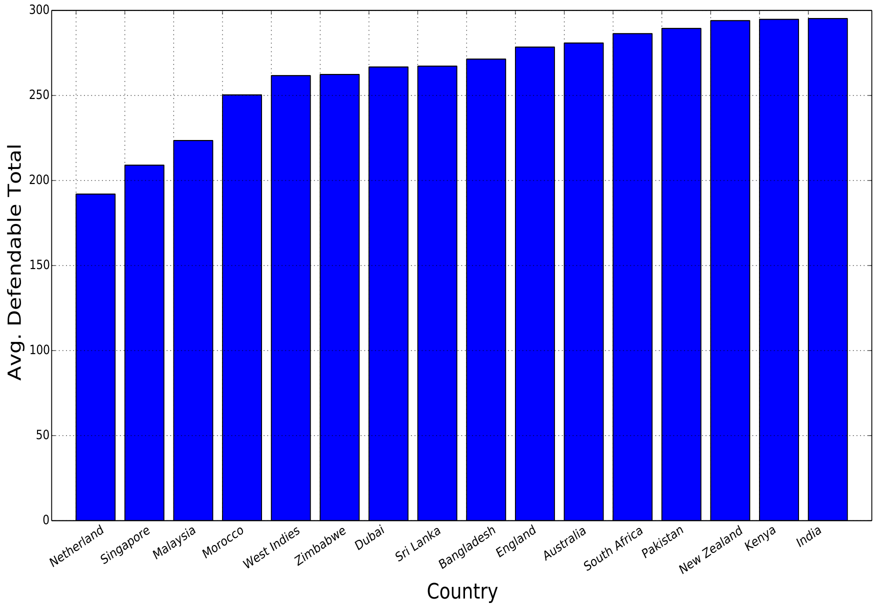

Figure 4.1: Average defendable scores in different countries.

Although, the approximated target serves as a good estimate at the start of the match, it can be improved, depending upon the actual situation of the match, while the game is in progress. That is, at any stage of the game, if the batting team is doing better than the initial estimate, we need to update this estimated target as they now would be aiming for a bigger target. Therefore, as proposed in [18], we update the target score for the team batting first as given in Algorithm 4. We use the DL resources [39], introduced in [2], to estimate the number of runs the batting team could score from the given state of the game. That combined with the number of runs already scored by the team makes up the new target. However, we never let the updated target to be less than the initial estimated target. This is so because a poor batting performances does not result in reduction in the defendable target but corresponds to increased pressure on the team.

With a defined target score to achieve for both the innings, Equations 4.1 and 4.2 represent the mathematical formulation of the first and second parameters (as mentioned at the start of Section 4.3), respectively.

$$k \leftarrow runsRemaining/Target \tag{4.1}$$

$$r \leftarrow reqRunRate/initRunRate \tag{4.2}$$

Algorithm 4 Updating estimated target for the first innings using D/L resources

```
**Input:** initTarget, DLTable, runsScored, ballsBowled, wktsOut
**Output:** newTarget
$ballsRem \leftarrow 300 - ballsBowled$
$DLrem \leftarrow DLTable[ballsRem][wktsOut]$
$newTarget \leftarrow runsScored + initTarget * DLrem$
**if** $newTarget < initTarget$ **then**
    $newTarget \leftarrow initTarget$
**end if**
```

where $runsRemaining$, $reqRunRate$ and $initRunRate$ represent the number of runs yet to be scored, the average number of runs required per over from the current stage and the average number of runs require per over at the start of the match, respectively, by the batting team to achieve its target score. The higher values of k and r intuitively tell us that the batting team has a lot of work to do to reach the target score, and vice-versa.

4.3.2 Modeling Players

Modeling a player refers to estimating the batting and bowling potentials of the player for a given match. Generally, career statistics are used to get necessary insights about a player. As proposed in [34], we use equations 4.3 and 4.4 to estimate the batting and bowling scores of a player, respectively.

$$batScore = batStrikeRate * batAverage \tag{4.3}$$

$$bowlScore = \frac{1}{bowlStrikeRate * bowlEconomy} \tag{4.4}$$

where $batStrikeRate$, $batAverage$, $bowlStrikeRate$ and $bowlEconomy$ are characteristic statistics of a player, and are described in Table 1.1.

4.3.2.1 Modeling Teams

Players form the fundamental unit of a team. Therefore, we define the total batting and bowling scores of a team as the weighted sum of the individual player's batting and bowling scores, respectively. Equation 4.5 represents the total batting score of the batting team. Similarly, Equation 4.6 represents the total bowling score of the bowling team.

$$totalBatScore = \sum_{p}^{p\in Team} \sqrt{\frac{batInngs}{matchesPlayed}} * batScore \tag{4.5}$$

$$totalBowlScore = \sum_{p}^{p \in Team} \sqrt{\frac{bowlInngs}{matchesPlayed}} * bowlScore \tag{4.6}$$

However, as a match proceeds, some of the players from the batting team get out and some of the players from the bowling team have bowled a part of their quota of maximum 10 overs (60 balls). Therefore, at a given stage in the match, akin to Equation 4.5, we define $remBatScore$ of the batting team as the weighted sum of only those players who haven't gotten out yet. Similary, the $remBowlScore$ of the bowling team is calculated by further weighing an individual bowler's score by the number of balls he has remaining to bowl in this match to the maximum number of balls he can bowl in an ODI cricket match, i.e., 60.

With a defined method to estimate the total and remaining batting and bowling potentials of the batting and bowling teams, respectively, Equations 4.7 and 4.8 represent the mathematical formulation of the third parameter.

$$b \leftarrow remBatScore/totalBatScore \tag{4.7}$$

$$w \leftarrow remBowlScore/totalBowlScore \tag{4.8}$$

Where $totalBatScore$ and $remBatScore$ represent the total and remaining batting potentials of the batting team. And similarly for the $totalBowlScore$ and $remBowlScore$. The higher values of b and w tell us that the batting and bowling teams have got good players to perform who can win the match for them, respectively.

4.3.3 Work Index

Having formalized all the three parameters of work index, as mentioned at the start of Section 4.3, we will now explain the Work Index in detail. Work Index is a dynamic measure which is updated as the innings progresses and takes into account the current state of the match to estimate the work yet to be done. With the help of variables defined in Equations 4.1, 4.2, 4.7 and 4.8, the batting and bowling work indices are calculated as per the Equations 4.9 and 4.10, respectively.

$$\textbf{battingWorkIndex} \leftarrow 100 * k * (r + \alpha * (1 - b) + \beta * w) \tag{4.9}$$

$$\textbf{bowlingWorkIndex} \leftarrow 100 * k * (1/r + \alpha * b + \beta * (1 - w)) \tag{4.10}$$

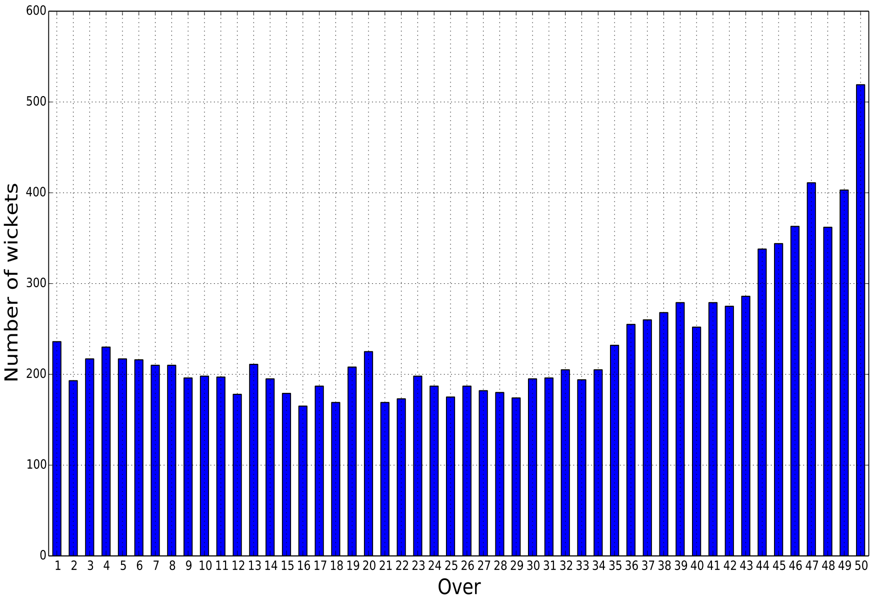

Figure 4.2: Total number of wickets fell in each over in our dataset. As a game reaches its final stages, the batsmen adopt a riskier strategy to score runs at a higher rate, with less concern about losing wickets.

Variable k (defined in Eqation 4.1), defined as the ratio of the runs remaining to score with respect to the target, is used as a bias in calculating the work index. The lower values of k, generally found at the end of the innings, directly reduces the impact of the other factors in determining the work index. As it can be seen from Figure 4.2, as a game reaches its final stages, the batsmen adopt a riskier strategy to score runs at a higher rate, with less concern about losing wickets. The value of a wicket reduces as scoring runs becomes the sole purpose. Similarly, higher values of k, found at the start of the innings, boosts the impact of other factors. This is because, at the initial stages of the innings, the wickets of the batsmen carry a lot more importance. Losing early wickets at the start of an innings puts the batting team into tremendous pressure, as they lose their key batsmen and face the threat of getting all-out, before even completing the quota of 50 overs. Variable r (Equation 4.2) captures how well is the batting team scoring as compared to the initial estimations. It is directly proportional to the batting work index, as increased required run rate, increases the amount of work to be done, and similarly it is inversely propotional to the bowling work index.

Variables b and w (Equations 4.7 and 4.8) represent for the remaining batting and bowling potentials of the corresponding teams. They account for the amount of batting and bowling resources remaining with the batting and bowling teams, respectively. Incorporating individual player's skills into these variables enables us to assess the current game scenario in a detailed way. They enable us to capture those scenarios where a team has lost several wickets yet still has good players remaining in the batting line-up, who can potentially change the game's direction. The parameters, α and β, represent the relative weightage of the remaining batting potential and the remaining bowling potential, respectively. Also, bowlers have 300 balls to possibly get the batsmen out, whereas the batting team possesses only 10 wickets to score runs. Therefore, losing wickets has significant impacts on the batting team. The values of the parameters, α and β, have been discussed in the experiments Section 4.5.

In all, work index is a combination of many important aspects of the game that enables it to capture the overall game scenario.

4.3.4 Assessing player performances

Cricket is a game of bat and ball. Bowlers from the bowling team take turns in overs to bowl their quota of 50 overs, where, in an over, one of the players from the bowling team bowls 6 deliveries to the batsman at strike. We calculate the *Batting Work Index* and the *Bowling Work Index* after every ball is bowled. The difference between the two consecutive batting work indices is contributed to the batsman who played the corresponding bowl. Similarly, the difference between the two consecutive bowling work indices is contributed to the bowler who bowled the ball. We repeat this process for each ball bowled in the entire match. At the end of the match, the scores attributed to an individual player represents his all-round performance in the entire match, and is called the player's *utility score*. The utility score of a player does not just capture the quantitative amount of runs scored or wickets taken, but also the context in which he made the contributions.

Phrases like *"Catches win Matches"* are very popular in the game of cricket and therefore, the fielding efforts of players could also integrated into our model. However, we leave it for the future work as of now.

4.4 Applications

Quantitative assessment of the player performances has multiple applications in the cricketing realm. While deciding the *player of the match* and ranking the players for different roles have been discussed and experimentally evaluated in the upcoming Sections 4.4.1 and 4.4.2, respectively, there are many other possible applications too. [13, 43] used player performances as a measure of rating players and used them for optimal team selections. Further, utility scores can be used to determine the performances of a player over his career.

4.4.1 Player of the Match

Player of the Match title is awarded to the player who played the most significant role in a particular match. Today, player of the match award for an ODI cricket match is decided by the match committee, including the match referees and the commentators. However, comparing and combining the performances of batsmen and bowlers on a common scale is very difficult and often becomes a subjective decision. Further, the number of runs scored or the number of wickets taken cannot directly be used for deciding the player of the match, as a relatively poor but a game changing performance by a player would make him a better candidate for the same. Therefore, we propose a method where the utility scores calculated by our approach for each player can be used to determine the player of the match award.

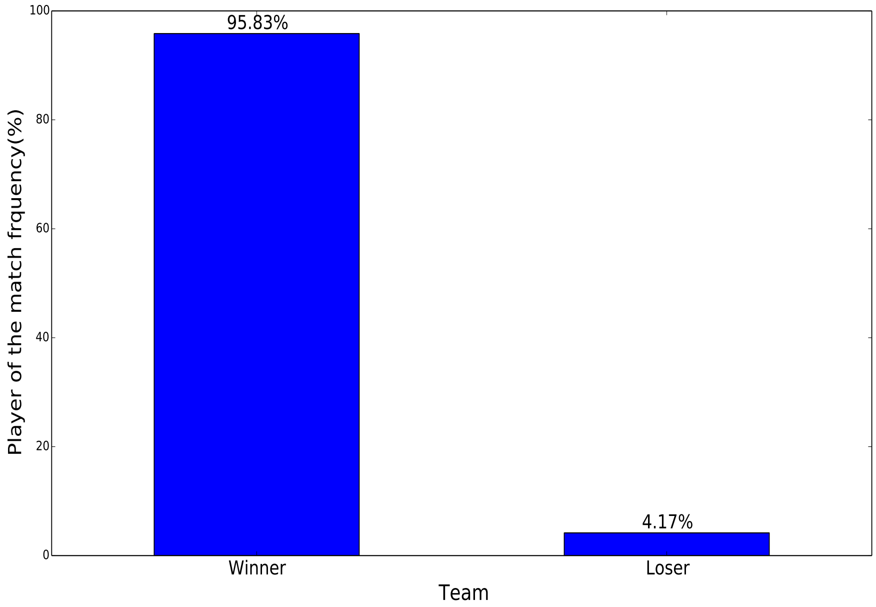

Figure 4.3: Frequency (in %) of selecting the player of the match from the winning and losing teams. This shows that the player of the match is almost always chosen from the winning team.

At first, it seems that the player who has the maximum score at the end of the match should be awarded the player of the match. However, as shown in Figure 4.3, player of the match award is almost always (95.83% of the times) given to a player from the winning side. A player from the losing side bags the player of the match award only if his performance is significantly better than the others and has contributed to an almost win for the losing side. Therefore, as of now, we choose the players from the winning side only as the potential candidates for the player of match award. We rank them directly based on their utility scores and the player with the maximum score is awarded the player of the match.

Note that some heuristics can be used to capture the very rare player of the match awards from the losing side also. One of the possible heuristics could be to bias the utility scores of the players based on the team they belong to, i.e., winning or losing. In this case, the player from the losing side could be awarded the player of the match award. However, for now, we have tabled it for future work.

4.4.2 Player Rankings

Akin to other sports, in cricket, players are ranked over time. All time best players and the best performers for a calender year are the most popular rankings amongst them. Usually, the batsmen, bowlers and the all-rounders are ranked separately, where a player is an all-rounder if he can bat as well as bowl.

To rank the batsmen, we use only the *Batting Work Index* as the contributing factor in assessing player performances. For a given time window, we add up the utility scores of a player for all the matches he has played. Finally, we rank the players based on their total utility scores. Similarly, to rank the bowlers, we use only the *Bowling Work Index*.

However, to rank the all-rounders, we first need to filter the players who are potential all-rounders. Emperically, we consider only those players as the all-rounders who have bowled as well as batted in at least 70% of the matches they have played. We consider both the batting and the bowling work indices as the contributing factors in assessing player performances. Thus, we rank the players based on their total utility scores.

4.5 Experiments and Results

In this section, we demonstrate the working and the results of previously discussed parameters and applications of our methodology, namely, real-time target updation for the first innings, the working of the work index, accuracy of the player of the match predictions and player rankings.

4.5.1 Target estimation for first innings

In Section 4.3.1, we discussed the use of D/L resources to re-adjust the estimated target scores for the first innings of a cricket match using Algorithm 4. We demonstrate the real-time updation of the target score using a sample match played between India and Australia on January 23, 2016. Australia batted first and we estimated a target score of 281 runs at the start of the innings using statistics from the past matches. As the innings proceeded, Australian batsmen performed better than the initial expectations and ended up scoring a huge total of 330 runs. The estimated target score and the current score for the entire Australian batting innings has been shown in Figure 4.4.

Since, by the completion of 47 overs (282 balls), the Australian team had performed better than the initial estimations, the dynamic target score got updated from 281 runs to a higher value of 341 runs to reflect the improved expectations. However, they could not perform well in the last 3 overs, adding only 18 runs. Hence, the estimated target score kept decreasing thereafter, eventually converging at 331 runs. This demonstrates that the target score successfully adjusts itself with respect to the current state of the batting team.

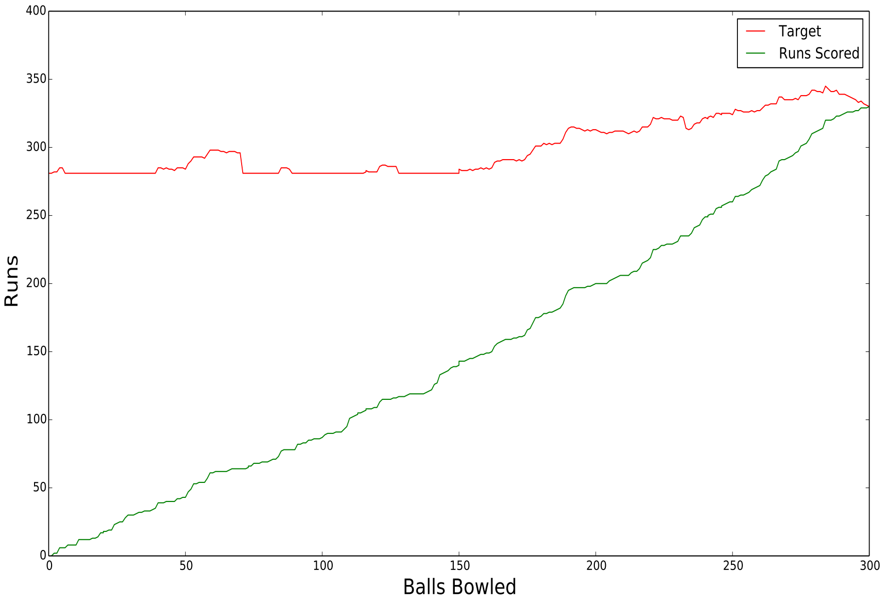

Figure 4.4: Target and the current score for the Australian innings, Ind. vs Aus., Jan. 2016. The plot depicts the self-adjustment of the estimated target score as per the current status of the team after every ball.

4.5.2 Player of the match

For the potential player of the match candidates, as discussed in Section 4.4.1, we consider only the players from the winning team. For a given match of cricket, our model outputs a list of player ranked in the descending order of their contribution in the match. Hence, to measure of the accuracy of our model, we use the exponentially-decaying scoring metric. For a given match, the match score is calculated to be $1/2^{r-1}$, where $1 \leq r \leq 11$ is the rank at which the player of the match has been predicted by our model. Therefore, the accuracy of our model is calculated as the summation of the match scores for all matches. The choice of an exponentially decaying function has been made here to capture the fact that the predictions at the top of the list are of much more significance as compared to the ones lower in the order, as we are only looking for the best performance.

With the defined accuracy metric, we use a validation set containing all the matches played between January, 2006 and December, 2012 to find the most suitable values of the parameters α and β, defined in Equations 4.7 and 4.8. Table 4.1 tabulates the accuracy of our model on the validation set for multiple values of the parameters. As it can be seen, $\alpha = 2.0$ and $\beta = 1.5$ yields the best results. Therefore, these values will be considered for the further discussions.

Total number of predictions (in %) for player of the match for top 5 ranks are shown in Figure 4.5. A non-increasing curve proves that the players who are performing better are placed higher in the rankings than the others. We have achieved 59.39% accuracy for the first rank and an accuracy of 78.29% and 86.80% for the top two and top three ranks respectively. On comparison with past statistics, we can see

Table 4.1: Accuracy score for multiple values of the parameters, α and β, on the validation set. $\alpha = 2.0$ and $\beta = 1.5$ yields the best results.

β α	*0.5*	*1.0*	*1.5*	*2.0*	*2.5*
0.5	362.1	357.9	346.2	334.3	330.1
1.0	369.9	371.2	367.8	354.7	344.2
1.5	349.8	379.6	375.9	373.1	360.8
2.0	328.7	363.7	**384.8**	378.7	371.9
2.5	306.1	344.7	372.6	379.8	377.5

that the player of the match has consistently been one of the players ranked in the top 3 by our approach, with an accuracy of greater than 86%. This validates our claim that *Honest Mirror* is indeed able to assess the individual performances in an ODI cricket match across multiple roles.

In literature, to the best of our knowledge, we could not find any previous work on predicting the player of the match for ODI cricket matches. However, [38] proposed a model, the *PI Model*, to assess player performances in a game of limited overs cricket match using pressure index, but only for the second innings of a match. We extended their method for the first innings by estimating the target score using the same approach as discussed in Section 4.3.1, and add up the player's batting and bowling contributions for both the innings to calculate a player's overall performance in a match. The players from the winning team are considered to be the potential player of the match in the order of their overall contribution in the match. We implemented their work, to the best of our abilities, to compare their approach against our model.

Apart from that, in a game of cricket, the number of runs scored by a player and the number of wickets taken by him are the two major criterians to judge a player's performance. Therefore, we further compared our approach with the two following baseline models which take into account a player's overall contribution in a match–

- **Model_1**: The overall contribution of a player is the summation of the ratio of the runs he has contributed to the teams total batting score and the ratio of the wickets he has taken to the total of wickets taken by his team.
- **Model_2**: To be able to combine the runs scored and wickets taken by a player, we map one of these into another, i.e., we calculate the weight of a wicket taken by a bowler in terms of the runs scored by a batsman. The weight of a wicket, denoted by ω, is calculated as the total number of

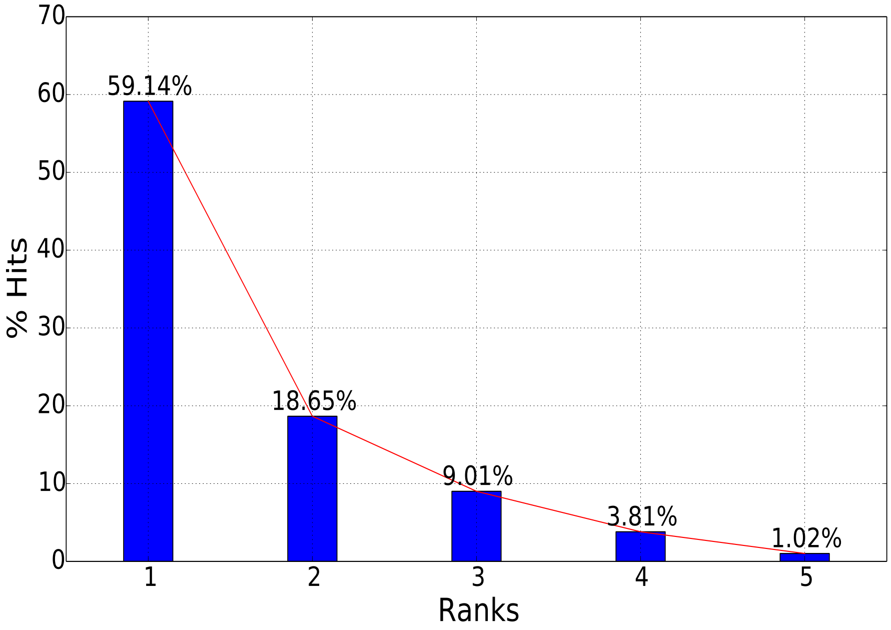

Figure 4.5: Rank frequencies (in %) for player of the match for the top 5 ranks. A non-increasing curve proves that the players who are performing better are placed higher in the rankings than the others.

runs scored in the match divided by the total number of wickets fell down. Therefore, the total contribution of a player in a match is the summation of the number of runs scored by him and ω times the number of wickets taken by him.

The comparison of the accuracy of our approach with other models is shown in Figure 4.6. As it can be seen, our model outperforms the other models with a good margin. Figure 4.7 demonstrates the accuracy comparison for the top-5 ranks for the four models. The number of right predictions, i.e., at the first rank, is higher by our model as compared to the others. Hence, the superiority of our model against the others validates our approach.

4.5.3 Work Index

We demonstrate the working of the *Work Index* using the same sample match, played between India and Australia on the 23rd January, 2016. The batting and the bowling work indices for the first 100 balls of the first innings, batted by Australia, is shown in Figure 4.8.

At the start of the innings, Indian bowlers had a lot of *work* to do against a strong Australian batting line up. However, as the game proceeded, Australia lost early wickets at quick successions (6th, 71st and 89th ball) and by the end of the first 15 overs, India was dominating. As displayed in Figure 4.2, wickets carry a lot of weightage at the start of an innings, and we can see major jumps in both the curves as the wickets fell.

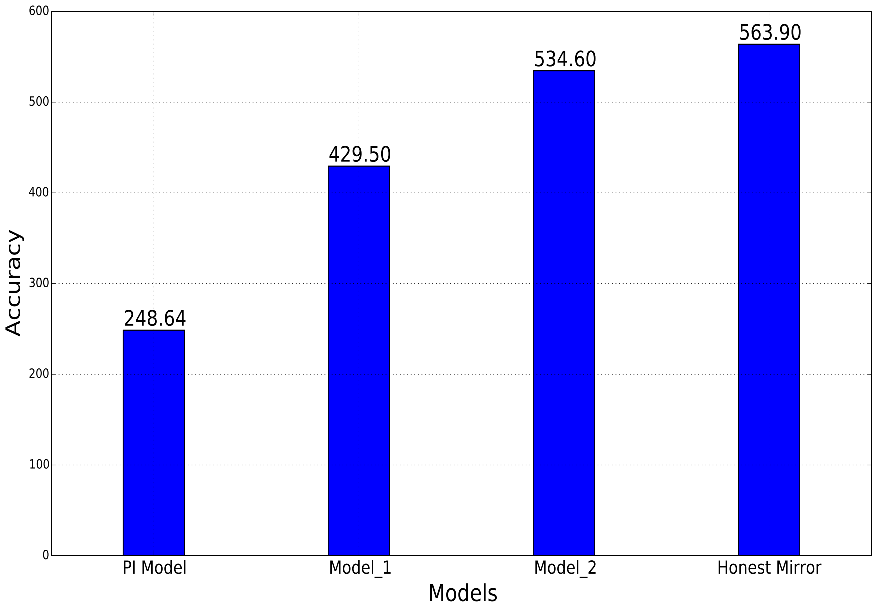

Figure 4.6: Comparison of the accuracy of *Honest Mirror* with the three other models. As it can be seen, our model outperforms the other models with a good margin.

Notice that as the batting team scores runs, the bowling work index also reduces. This is because we do not expect the bowlers to ball maiden overs all the time or take up wickets on every ball. As long as they do not concede too many runs and take wickets at regular intervals, they are doing a fine job. On the other hand, if they do concede too many runs without taking wickets, the variable r in the bowling index (Equation 4.10) would come into play and the graph would go up.

4.5.4 Ranking Players

In section 4.4.2, we explained how our method could be used for ranking players for multiple roles such as batsmen, bowlers and all rounders. The top 10 players for all the three domains for the period of Jan, 2006 to June, 2015, as calculated by our method, have been tabulated in Table 4.2. Tillakaratne Dilshan (Sri Lanka), Mitchell Johnson (Australia) and Shahid Afridi (Pakistan) have been rated as the best batsman, bowler and all rounder respectively.

4.6 Discussions

In this chapter,we proposed a methodology to quantitatively assess the individual player performances in an ODI cricket match, where we take the context of the game into account. Our method also helps us in combining and comparing the performances of various players across different roles. Further,

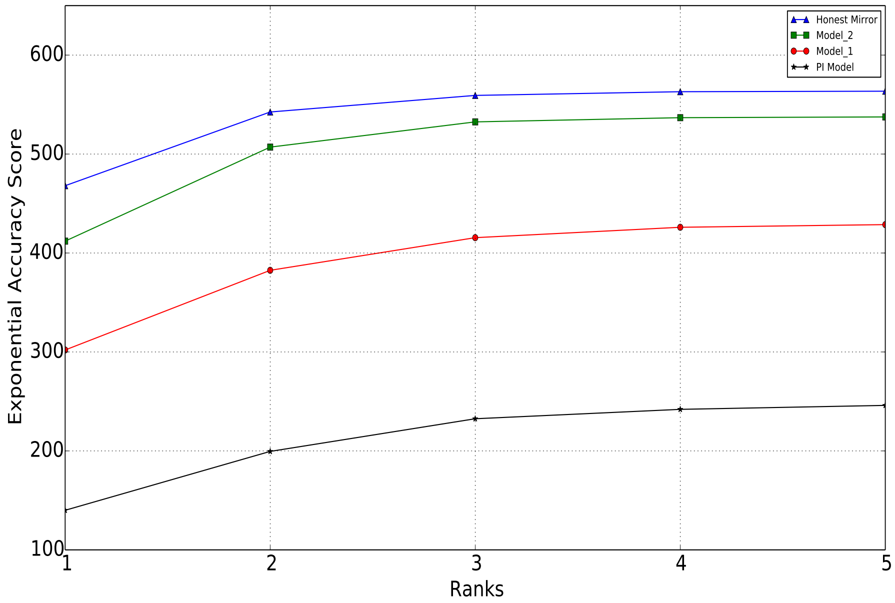

Figure 4.7: Exponential accuracy score for the the top 5 ranks for all the four models. Our model beats the remaining models at all the ranks. Also, significant difference at the first rank proves that our model is able to pick the player of the match by taking the game situations into account.

Table 4.2: Top Players (2006-2016)

Rank	**Batsmen**	**Bowlers**	**All Rounders**
1	TM Dilshan	MG Johnson	Shahid Afridi
2	Yuvraj Singh	SL Malinga	AD Mathews
3	MS Dhoni	KMDN Kulasekara	SR Watson
4	SK Raina	JM Anderson	Mohammad Hafeez
5	AB de Villiers	RA Jadeja	DJ Bravo
6	MJ Clarke	SCJ Broad	TT Bresnan
7	KC Sangakkara	B Lee	Shakib Al
8	JP Duminy	KD Mills	NLTC Perera
9	V Kohli	DL Vettori	DJG Sammy
10	MEK Hussey	Umar Gul	PD Collingwood

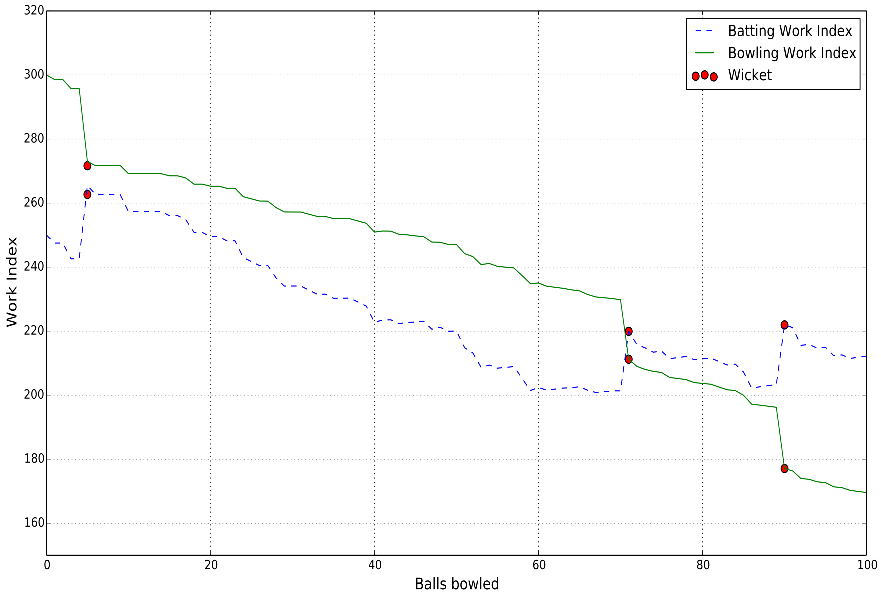

Figure 4.8: Work Index for the first 100 balls of the Australian innings, Ind. vs Aus., 2016. The change in the amount of work to be done, as wickets fell, could be seen with major jumps in the two curves.

the player performances are used to predict the player of the match award. Our method outperforms the previous works and other baseline models.

Chapter 5

Conclusions and Future Work

In this thesis, we addressed two problems related to the ODI format of the game of cricket. First, we proposed a novel method to predict the winner of ODI cricket matches using a team-composition based approach. Second, we presented a method to quantitatively assess the performances of individual players in a match of ODI cricket which incorporates the game situations under which the players performed. The player performances are further used to predict the player of the match award.

5.1 Winner Prediction

To predict the winner of an ODI cricket match, we used the relative team strength between the two competing teams as the key feature. Calculating the relative strength between two teams boils down to modeling the strength of individual players based on their batting and bowling career statistics. Using two other player-independent features, namely, toss outcome and the venue of the match, along with the relative team strength, we used a supervised learning approach to predict the winner of the match. We observe that our approach performs better than previous works even with the use of very simplistic features.

To model a batsman, we used both the career statistics as well as his recent performances. However, we could not use the recent performances of a bowler to model him because the match-wise performance statistics of bowlers were not available. In future, this could be overcome by including the recent performance of bowlers as well in modeling him. Furthermore, our approach could be extended to predict the winner of cricket matches while the game is in progress.

5.2 Honest Mirror

In the second-half of the thesis, we proposed a methodology to quantitatively assess the individual player performances in an ODI cricket match, where we take the context of the game into account. Our method also helps us in combining and comparing the performances of various players across different roles. Using the player performances in a match, we predicted the player of the match award for the ODI matches played between 2006 and 2016. We achieved an accuracy of 86.80% for the top-3 positions in predicting the player of the match award, which is superior to previous works and other baseline models. This further proved the validity of our approach.

However, as of now, we only used the players from the batting team as the potential candidates for player of the match award. In future, adding an artificial bias to the utility scores for the winning and losing teams could be done to capture the player of the match awards from the losing side as well. Also, phrases like *"Catches win Matches"* are very popular in the cricketing world. Therefore, the fielding contributions of the players could also be taken into account in future for estimating his total contribution in the match.

Overall, the thesis proposes solutions to two of the major issues related to ODI cricket.

Related Publications

Workshop

1. Madan Gopal Jhanwar, Vikram Pudi. **Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach.** Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA) @ The European Conference on Machine Learning & Principles and Practice of Knowledge Discovery, ECML-PKDD, 2016.

2. Madan Gopal Jhanwar, Vikram Pudi. **Honest Mirror: Quantitative Assessment of Player Performances in an ODI Cricket Match.** Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA) @ The European Conference on Machine Learning & Principles and Practice of Knowledge Discovery, ECML-PKDD, 2017. (*Submitted*)

Bibliography

[1] "Sporteology." http://sporteology.com/top-10-popular-sports-world/. 1, 8

[2] F. C. Duckworth and A. J. Lewis, "A fair method for resetting the target in interrupted one-day cricket matches," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 220–227, 1998. 1, 8, 12, 14, 38

[3] F. Duckworth and A. Lewis, "A successful operational research intervention in one-day cricket," *Journal of the Operational Research Society*, vol. 55, no. 7, pp. 749–759, 2004. 9

[4] V. Jayadevan, "A new method for the computation of target scores in interrupted, limited-over cricket matches," *Current Science*, vol. 83, no. 5, pp. 577–586, 2002. 9

[5] J. Thomas, "Rain rules for limited overs cricket and probabilities of victory," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 51, no. 2, pp. 189–202, 2002. 9

[6] M. Carter and G. Guthrie, "Cricket interruptus: fairness and incentive in limited overs cricket matches," *Journal of the Operational Research Society*, vol. 55, no. 8, pp. 822–829, 2004. 9

[7] I. G. McHale and M. Asif, "A modified duckworth–lewis method for adjusting targets in interrupted limited overs cricket," *European Journal of Operational Research*, vol. 225, no. 2, pp. 353–362, 2013. 9

[8] T. B. Swartz, P. S. Gill, D. Beaudoin, *et al.*, "Optimal batting orders in one-day cricket," *Computers & operations research*, vol. 33, no. 7, pp. 1939–1950, 2006. 9

[9] J. M. Norman and S. R. Clarke, "Optimal batting orders in cricket," *Journal of the Operational Research Society*, vol. 61, no. 6, pp. 980–986, 2010. 9

[10] H. Gerber and G. D. Sharp, "Selecting a limited overs cricket squad using an integer programming model," *South African Journal for Research in Sport, Physical Education and Recreation*, vol. 28, no. 2, pp. 81–90, 2006. 9

[11] G. Sharp, W. Brettenny, J. Gonsalves, M. Lourens, and R. Stretch, "Integer optimisation for the selection of a twenty20 cricket team," *Journal of the Operational Research Society*, vol. 62, no. 9, pp. 1688–1694, 2011. 9

[12] H. H. Lemmer, "Team selection after a short cricket series," *European Journal of Sport Science*, vol. 13, no. 2, pp. 200–206, 2013. 9

[13] B. S, S. RP, Abhijeet, and R. S, "A self-adapting intelligent optimized analytical model for team selection using player performance utility in cricket," in *MIT Sloan Sports Analytics Conference*, 2015. 9, 42

[14] B. M. De Silva and T. B. Swartz, *Winning the coin toss and the home team advantage in one-day international cricket matches*. Department of Statistics and Operations Research, Royal Melbourne Institute of Technology, 1998. 9

[15] H. H. Lemmer, "A method to measure strangling, a dramatic form of choking in cricket," *International Journal of Sports Science & Coaching*, vol. 10, no. 4, pp. 717–728, 2015. 9

[16] S. Akhtar and P. Scarf, "Forecasting test cricket match outcomes in play," *International Journal of Forecasting*, vol. 28, no. 3, pp. 632–643, 2012. 10

[17] F. Munir, M. K. Hasan, S. Ahmed, S. Md Quraish, *et al.*, *Predicting a T20 cricket match result while the match is in progress*. PhD thesis, BRAC University, 2015. 10

[18] M. Bailey and S. R. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress," *Journal of sports science & medicine*, vol. 5, no. 4, p. 480, 2006. 10, 38

[19] V. V. Sankaranarayanan, J. Sattar, and L. V. Lakshmanan, "Auto-play: A data mining approach to odi cricket simulation and prediction.," in *SDM*, pp. 1064–1072, SIAM, 2014. 10, 32

[20] M. Khan and R. Shah, "Role of external factors on outcome of a one day international cricket (odi) match and predictive analysis," 2015. 11

[21] A. Kaluarachchi and A. S. Varde, "Cricai: A classification based tool to predict the outcome in odi cricket," in *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, pp. 250–255, IEEE, 2010. 11

[22] F. T. Bozbura, A. Beşkese, and T. S. Kaya, "Topsis method on player selection in mba," 11

[23] C.-C. C. Y.-T. Lee and C.-M. Tsai, "A hybrid assessment method for evaluating the performance of starting pitchers in a professional baseball team," 2013. 11

[24] P. K. Dey, D. N. Ghosh, and A. C. Mondal, "A mcdm approach for evaluating bowlers performance in ipl," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 2, no. 11, pp. 563–73, 2011. 11

[25] P. K. Dey, A. C. Mondal, and D. N. Ghosh, "Statistical based multi-criteria decision making analysis for performance measurement of batsmen in indian premier league," *International Journal of Advanced Research in Computer Science*, vol. 3, no. 4, 2012. 11

[26] M. Tavana, F. Azizi, F. Azizi, and M. Behzadian, "A fuzzy inference system with application to player selection and team formation in multi-player sports," *Sport Management Review*, vol. 16, no. 1, pp. 97–110, 2013. 11

[27] S. Akhtar, P. Scarf, and Z. Rasool, "Rating players in test match cricket," *Journal of the Operational Research Society*, vol. 66, no. 4, pp. 684–695, 2015. 11

[28] D. Beaudoin and T. B. Swartz, "The best batsmen and bowlers in one-day cricket," *South African Statistical Journal*, vol. 37, no. 2, p. 203, 2003. 11

[29] 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. 12

[30] 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. 12

[31] A. Kimber, "A graphical display for comparing bowlers in cricket," *Teaching Statistics*, vol. 15, no. 3, pp. 84–86, 1993. 12

[32] P. J. Bracewell and K. Ruggiero, "A parametric control chart for monitoring individual batting performances in cricket," *Journal of Quantitative Analysis in Sports*, vol. 5, no. 3, 2009. 12

[33] P. J. Van Staden *et al.*, "Comparison of cricketers bowling and batting performances using graphical displays," 2009. 12

[34] G. Barr and B. Kantor, "A criterion for comparing and selecting batsmen in limited overs cricket," *Journal of the Operational Research Society*, vol. 55, no. 12, pp. 1266–1274, 2004. 13, 37, 39

[35] M. I. Johnston, S. R. Clarke, D. H. Noble, *et al.*, "Assessing player performance in one-day cricket using dynamic programming," 1993. 13

[36] A. Lewis, "Towards fairer measures of player performance in one-day cricket," *Journal of the Operational Research Society*, vol. 56, no. 7, pp. 804–815, 2005. 13

[37] P. Shah and M. Shah, "Pressure index in cricket," *IOSR Journal of Sports and Physical Education*, vol. 1, pp. 09–11, 2014. 13

[38] D. Bhattacharjee and H. H. Lemmer, "Quantifying the pressure on the teams batting or bowling in the second innings of limited overs cricket matches," *International journal of Sports Science & Coaching*, p. 1747954116667106, 2016. 14, 46

[39] "D/L Table." http://www.tcuandsa.org/Doc/dldocs/DLResourceChartNew.pdf. 14, 38

[40] "ESPN Cricinfo." http://www.espncricinfo.com/. 16, 36

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 32

[42] "Cricsheet." http://cricsheet.org/. 36

[43] S. Mukherjee, "Quantifying individual performance in cricket - a network analysis of batsmen and bowlers," *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 624–637, 2014. 42