

T20 cricket match score and winning team prediction using machine learning techniques

**K.A.D.A. Pramoda
2021**



T20 cricket match score and winning team prediction using machine learning techniques

A dissertation submitted for the Degree of Master of Computer Science

**K.A.D.A. Pramoda
University of Colombo School of Computing
2021**



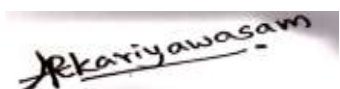
DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: K.A.D.A. Pramoda

Registration Number: 2018/MCS/068

Index Number: 18440687



2021.11.29

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. K.A.D.A. Pramoda under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. M.G.N.A.S. Fernando



2021.11.29

Signature of the Supervisor & Date

I would like to dedicate this thesis to my institution mentors under whose constant guidance I have completed this dissertation. They not only enlightened me with academic knowledge but also gave me valuable advice whenever I needed it the most.

ACKNOWLEDGEMENTS

Thank you to my supervisor, Dr. M.G.N.A.S. Fernando, for providing guidance and feedback throughout this project. Thanks also to my family, for putting up with me being sat in the office for hours on end, and for providing guidance and a sounding board when required.

ABSTRACT

Cricket is one of the famous outdoor sports that contain an outsized set of statistical data in the world. As T20 games rise in popularity, it's necessary to look at the possible predictors that affect the result of the matches. This research aims at analyzing the T20 cricket match results from the dataset collected (2005-2021). It focuses on measuring the result of T20 matches by applying the prevailing machine algorithms learning to the balanced also as an imbalanced dataset. The database used here was an unbalanced database and had to be converted into a balanced database. Therefore, this research was performed on this unbalanced dataset as well as the balanced dataset. Oversampling technique is employed for imbalanced datasets then the algorithm is applied. Accuracy is used as the performance metric and calculated by using machine learning algorithms. It is also considered as evaluation criteria and the percentage will vary consistent with the various algorithms. Three models were created basically, and the third model was a hybrid model created with the output of the first and second models. In the first model, the Random Forest algorithm obtained an accuracy of 84.51% for the imbalance data and 76.61% for the balance data. The Decision Tree algorithm was used to create the third hybrid model, with an accuracy of 97.23% for the imbalanced data and 93.92% for the balanced database. Thus, the hybrid model operates with higher accuracy than the first model. The first model is designed to predict the winning team before the start of the game, the second model is used to predict the score of the team batting in an innings, and the third model is used to predict the winning team in the second inning.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
ABSTRACT.....	iv
TABLE OF CONTENTS	v
LIST OF FIGURES.....	vi
LIST OF TABLES	vii
ABBREVIATIONS	viii
CHAPTER 1: INTRODUCTION	1
1.1 Motivation.....	2
1.2 Statement of the problem.....	2
1.3 Research Aim and Objectives.....	2
1.4 Scope.....	3
1.5 Structure of the thesis	4
CHAPTER 2: LITERATURE REVIEW.....	5
2.1 A Literature Review	5
2.2 Background of used materials.....	12
2.3 Over view of Twenty-Twenty cricket match.....	15
2.4 Tools and Technologies	16
CHAPTER 3: METHODOLOGY	18
2.1 Problem Analysis	18
2.1 Design.....	22
2.1 Methodology	26
CHAPTER 4: EVALUTION AND RESULTS.....	32
CHAPTER 5: CONCLUSION AND FUTURE WORK.....	44
APPENDICES.....	I
BIBILIOGRAPHY	II

LIST OF FIGURES

Figure 1: Two different categories classified using SVM.....	14
Figure 2:Decision tree AND operation.....	15
Figure 3: Example of current score projection	19
Figure 4: Runs vs wickets visualization of West Indies team in a match	20
Figure 5: Runs vs wickets visualization of England team in a match	20
Figure 6: System Architecture	22
Figure 7: Architecture of a model.....	23
Figure 8: High level architecture of the application	24
Figure 9: Python Service running instance.....	24
Figure 10: Home page of the application	24
Figure 11: Prediction of Basic Information (Model 1).....	25
Figure 12: Score predictor (Model 2)	25
Figure 13: Hybrid model predictor (Model 3).....	26
Figure 14: Sample of YAML file	27
Figure 15: Sample CSV file	28
Figure 16: Most of the matches winning teams.....	32
Figure 17: Most of the times get player of the match.....	33
Figure 18: Most of the matches played grounds	34
Figure 19: First Bat and get the win	34
Figure 20: First bat and lost the match	35
Figure 21: Second bat and win the match	35
Figure 22: Second bat and lost the match.....	36
Figure 23: Toss decision of teams	36
Figure 24: Toss decision of top 10 winning team	37
Figure 25: Heat map for Model 01	38
Figure 26: Heat map for model 02.....	39
Figure 27: Density graph for mean absolute error.....	40
Figure 28: Heat map for model 03.....	41

LIST OF TABLES

Table 1: Module wise prediction accuracy((Tekade et al., 2020)	7
Table 2: Truth table for AND operation	14
Table 3:Python Libraries	16
Table 4: Highest innings totals (Source :ESPNcrinfo)	19
Table 5: Match Count against the tournament(Source : https://cricsheet.org/).....	26
Table 6: Example of team names labeling.....	29
Table 7: Example for wicket types labeling	30
Table 8: Example for umpires labeling	30
Table 9: Accuracy for model 1 with different algorithms	38
Table 10: Feature importance for Random Forest.....	38
Table 11: Algorithms mean absolute error for model 02	40
Table 12: Accuracy for model 2 with different algorithms	41
Table 13: Feature importance for model 03	42
Table 14: Data for evaluation (source: CRICSHEET)	42
Table 15: Basic winning team prediction model(Model 1) and Hybrid winning team prediction model (Model 3) evaluation summary	42
Table 16: Batting score prediction model (Model 02) evaluation summary	43

ABBREVIATIONS

SVM	Support Vector Machine
CSV	Comma Separated Values
ICC	International Cricket Council
AI	Artificial Intelligence
ODI	One Day International
IPL	Indian Premier League
LPL	Lanka Premier League
T20	Twenty-Twenty
KNN	K Nearest Neighbor
D/L	Duckworth Luis
MLR	Multiple Linear Regression
ML	Machine Learning
SMS	Short Message Service

CHAPTER 1

INTRODUCTION

Data analysis today, every data analyst needs to examine data sets and draw conclusions from the information to extract useful information from them. Data analytics techniques and algorithms are widely employed by the commercial industry to form accurate business decisions. Verification or refutation of experimental layouts, hypotheses and conclusions are also used by analysts and specialists. In recent years, analytics has been used to predict and draw various insights into the field of sports. Due to money, team spirit, loyalty to the city and the participation of a large number of fans, the results of the competitions are very important to all stakeholders (Jyothisna and Srikanth, 2019).

Sports forecasting can be considered as one of the objectives of sports analysis, which aims to help decision makers to take advantage of competitors. Data analysis is especially common in sports. Cricket is accustomed to using the International Cricket Council (ICC) data analysis results. The barrier to this task depends on the collection of data historical data, the collection of data for future events, the knowledge required to interpret the collected data, and much more. The result of the game has become the center and concentration of the game (Naik et al., 2018).

Most of the past research studies, did their research to develop an AI tool for ODI (One Day International) and test matches. Less availability of the predict the outcome of game using the IPL(Indian Premier League) twenty-twenty matches (Naik et al., 2018). IPL is only one type of club competition of twenty-twenty category. Big Bash League, Caribbean Premier League, T20 Blast, Pakistan Super League and Women's Big Bash League are also in twenty-twenty category club competitions and there are international twenty-twenty matches also. There is a good trend for women's cricket today as well. In 2020 LPL (Lanka premier league) also started as a twenty-twenty club competition. According to the literature review study, any model did not develop consider the all types (Big Bash League, IPL, international, ...) of twenty-twenty cricket matches for their predictions.

However, according to the literature review, there is no proper tool to analyze the twenty-twenty cricket match from start to finish, ball to ball when the match is in progress. The existing tools to predict the match using the batting partnership are not supported to twenty-twenty matches. The ultimate goal of pre-match cricket predictions is to identify key players and prevent them from picking the wrong players at that moment and making statistical predictions based on their

batting performance (SADP, 2018).

1.1 Motivation

After the game of football, the game of cricket is loved and watched by many. Especially the Twenty20 format is watched and loved by many, and nobody can guess who will win the match until the last ball of the ultimate over. For that reason, there is maximum uncertainty, where one over can completely change the momentum of the game. According to that it is very difficult to predict the game. So that we can reduce the complexity of predicting the T20 match score and winning team using Artificial Intelligence.

Following the creation of this new system, its beneficiaries can be shown as follows.

- T20 cricket matches transmitters are able to transmit predicted scores with a very high accuracy value.
- People who bet on these matches can place their bets with very high accuracy.
- During the broadcast of the cricket match, the winning team will have to send a text message before the draw. In this case, the system can be used to select the correct winning team.
- This model can use instead of the Duckworth Lewis method.

1.2 Statement of the problem

Introducing a digital application that uses machine learning technology that provides higher accuracy than the methods currently used to predict the winning team and score of the most popular Twenty20 cricket matches currently being played, T20 International, non-official T20 international, big bash league, Indian premier league, Caribbean premier league, T20 blast, Pakistan super league and women's big bash league.

1.3 Research Aims and Objectives

Nowadays, betting is done during many sports. To do this, they use betting systems. Since there is no tool for predicting Twenty20 matches (according to the literature review), there can be no betting system that uses artificial intelligence. Using this design model can support the development of betting system of the Twenty20 matches.

- Predict the target final batting score in the 1st and 2nd innings.
When the match is in progress we can predict the target score using the ball by ball information in each innings. (Ball by ball information: batting team, bowling team, striker, non-striker, bowler, total runs, extra runs, bowler runs and etc.)
- Predict the target wicket count in the 1st and 2nd innings.
When the match is in progress we can predict the target score using the ball by ball information in each innings same as the previous objective.
- Predict the winning team.
During the 1st innings after the tossing, we can predict the wining team using the main information according to the current match. (Main information are team1, team2, city, venue, umpires, toss winner, toss winner decision and etc.)
During the 2nd innings, we can predict the winning team using main information and ball by ball information. In this stage we need to use hybrid model to predict this.
- Release one automated system to predict the wining team, target final score and target final wicket count.
Released one web based automated system to predict the final score, final wicket counts in each innings and the winning team for the users using outcomes of the above mentioned objectives.

1.4 Scope

According to the literature review, most of the cricket match outcome predictions were done for the ODI and test matches. The research for the Twenty20 matches has also been done only for IPL matches (Naik et al., 2018). There are several tournaments in the world related to the Twenty20 category as well as international competitions. They are T20 international, non-official T20 international, Big bash league, Indian premier league, Caribbean premier league, T20 blast, Pakistan super league, women's big bash league.

In this research study, above mentioned all the tournaments twenty-twenty matches data will be going consider for the predictions. So there are 8 tournaments. According to the past research

studies there is some expanding of the data in this study.

According to my literature review, I identified some parameters that effected to the outcome of a twenty-twenty match. Base on that there are several kinds of parameters are going to use in this research study. Tournament type, city that match held, gender of the players, team name, winning team, win by how many runs/how many wickets, player of the match, toss winner, toss decision, venue and umpires are the basic parameters. Other parameters can be considered in terms of what happens from ball to ball in each over. They are striker, non-striker, bowler, extra runs, extra type, total runs for the delivery, boundaries, wicket, fielder, type of wicket and out player. This research project was carried out using the above mentioned parameters.

1.5 Structure of the Thesis

The thesis covers the research work along with five chapters which outline the research work. The first chapter explains the introduction and the background of T20 cricket match score and winning team prediction with the aims, research goals and the scope of the research. Chapter 2 discusses the literature survey which comprises the previous research works related to the T20 cricket match score and winning team prediction along with the technical aspects. The design and methodology section is detailed in chapter 3. Chapter 4 expresses the evaluation of the results obtained by performing the research methodology. Finally, the chapter 5 presents the conclusion and the impact of the research along with its future works and also the limitations of the research as well.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review

This research, mentioned above in Chapter 1, is correlated with many areas. Some of these fields are the domain of computer science, data mining, machine learning, mathematics, and Twenty-twenty cricket match, and it is essential to conduct a literary survey of recent research conducted in these related fields. It will be a huge advantage during the implementation phase.

It is usually difficult to predict the outcome of a cricket tournament until the end. Therefore, it is necessary to study and make a prediction about the event from its beginning to its end. That is, the data is processed dynamically and must provide a prediction as the game progresses (Naik et al., 2018). Before the start of the match, their prediction depends on the batting, bowling, batting order, the captain of both teams and the batting figures on the field against the opponent, and their prediction after the start of the match depends on the batsman-order and batting order of a particular player. The biggest obstacle to making this prediction is the collection of historical data and the interpretation of that collected data (Naik et al., 2018). Therefore, in such a task, it is necessary to collect these data correctly and interpret them correctly. According to this paper, they study the logistic regression, neural network and K-means clustering algorithms, but they used only the logistic regression algorithm for their implementation. As mentioned here, according to the last data reports, the data of the 11 players belonging to the two playing teams should be obtained. In addition, the stadium where the match will be held must be obtained. The batting performance at the stadium is considered against each team first. Secondly, the two teams are considered to be performing well against the other teams at the same stadium. The following mathematical equation is used to calculate this.

n = Number of available players

x_i = Average of each player

$$\text{Team batting performance} = \sum_{i=0}^n \frac{x_i}{n} \dots\dots\dots (1)$$

Using this equation, the calculated value for the two cases mentioned above is obtained for the two separate teams. This is how the pre-match calculations are done. These data are then entered into the following algorithm and output is obtained (Naik et al., 2018).

The system is designed according to the mathematical representation. In this article they have tried this model only in one match. It was against Australia on 26 March 2015 at the Sydney Cricket Ground against India. Their predictions are sometimes wrong because every ball fluctuates. According to this research paper, logistic regression works accurately. They mentioned some future works as prediction of test matches and the T20 matches (Naik et al., 2018).

Selecting a best team for a cricket match plays an important role in the team's success. It is also stated that the result will be very beneficial to the parties. The reason for that are involvement of money, team spirit, City loyalty and the massive fans (Jyothsna and Srikanth, 2019). "Before prediction we have to explore and visualize the data because data exploration and visualization is an important stage of predictive modeling" (Jyothsna and Srikanth, 2019). In this research, the outcome of the game is predicted using machine learning methods. Machine learning is a branch of the AI. It helps to solving real life engineering problems. The IPL has been introduced to popularize cricket in India and to identify new players. So teams are tempted to get players with some experience at auction (Jyothsna and Srikanth, 2019). It has turned a new page in the game of cricket. This prediction should be made before the start of the competition, and it has become even more important to predict a match that is currently in progress. This is because of the fact that many betting takes place even when the competition is active. Predicting the outcome of a match takes into account the performance of the players as well as the data related to that match (Jyothsna and Srikanth, 2019). For example, match venue details, teams, ball to ball details, umpires, toss winner and toss decision can be shown. There are three main objectives expressed in this research paper. They are to supply the statistical analysis of players supported different characteristics, to predict the performance of a team counting on individual player statistics and to successfully predict the outcome of IPL matches (Jyothsna and Srikanth, 2019). According to their methodology, there are four parts. They are Data pre-processing, Data Cleaning, Data Preparation and Encoding. The main purpose of this article is to analyze IPL cricket data and predict the performance of the players. Here, three classification algorithms are used and compared to find the best correct algorithm. Random forest is observed as the best accurate rating with 89.15% to predict the performance of the best player (Jyothsna and Srikanth, 2019). In this research, it has been numerically shown that teams, venue, toss decision and umpires directly affect the output of a match. The activation tools used are Anaconda Navigator and Jupiter. This knowledge will be used to predict the winning teams for future IPL matches. So the best team can be formed using this prediction.

The instability of a Twenty20 match is such that its outcome is sometimes unpredictable, even on the last ball of the innings. Home ground advantage, past performance on that ground, records at the same venue, the overall experience of the players, record with a particular opposition, over all current form of a team, and individual player form can be considered as influencing factors (Tekade et al., 2020). “Cricket match outcome prediction using machine learning” by (Tekade et al., 2020), they selected 17 key features and 6 machine learning models that give the best possible prediction accuracy. Multiple Linear Regression, support vector regression, Decision Tree Regression, Random Forest Regression, Naïve Bayes and Logistics Regressions are the used machine learning models (Tekade et al., 2020). According to this research paper, they develop three machine learning models and then combine those as a one model. It means the final module is the hybrid model. There three models, they developed. 1st model consider Team Name 1, Team Name 2, Venue, Toss, Decision and Time. Multi linear regression use for this model. 2nd model consider Pitch Conditions, Temperature, Humidity and Precipitation. Support vector regression is used for this model. 3rd model consider team1 batting average, team2 batting average, team1 bowling average, team2 bowling average, team1 economy and team2 economy. Decision Tree regression, Random forest regression, Naïve Bayes and Logistic regression are used machine learning techniques (Tekade et al., 2020). According to these models, finally get hybrid model as outcome. They point out that the accuracy of the output obtained from a hybrid system in this way is high. The table below shows the accuracy of all models.

Table 1: Module wise prediction accuracy((Tekade et al., 2020)

<i>Module</i>	<i>No of features</i>	<i>No of models</i>	<i>Prediction Accuracy</i>
Module 1	6	1	67%
Module 2	4	2	76%
Module 3	7	4	86%
Module 4 (Hybrid)	17	6	90%

IPL is governed by the board of control for cricket in India. Various natural factors affect the sport, the assumptions given by the media, and the fantasy-like 11th big market and the betting pattern on the websites have given great importance to the format in Indian Premier League (Tekade et al., 2020). Technology is evolving and applications such as fantasy 11 and betting sites are gaining popularity, and people are tempted to use the predictions provided by the

machine learning model (Tekade et al., 2020). The use of machine learning makes life easier in many areas.

There are two types of learning in machine learning. They are supervised and unsupervised learning. In supervised learning need labelled data. It means the already knows the patterns according to the past data. But in unsupervised learning, using unlabeled data and should find the patterns of those data. Cricket matches have properly labelled data. Because of that in this application use the supervised learning (Tekade et al., 2020).

Furthermore, Supervised learning can represent in mathematically as follows,

$$y = f(x) \rightarrow \text{Prediction Model} \dots\dots\dots (2)$$

$$D = ((X_1, y_1), (X_2, y_2), (X_3, y_3), \dots, (X_n, y_n)) \rightarrow \text{Data set} \dots\dots\dots (3)$$

Most of the past researches try to develop a model to predict the outcome of the IPL (Indian Premier League) matches. So that, (Lamsal and Choudhary, 2020) followed a research. It is “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”. In this study, IPL. They found seven factors that significantly affect the outcome of a match: the home team, the distance team, the coin winner, the decision of the coin, the stadium and the weight of the respective teams. Here it is necessary to calculate the strength of a group. A multivariate regression based model was formulated multivariate regression based model for each player to calculate points earned based on their past performance (Lamsal and Choudhary, 2020). Accordingly, the points of one player must be calculated first. According to that, every player is awarded points based on these 6 features:

- (i) number of wickets taken
- (ii) number of dot balls given
- (iii) number of fours
- (iv) number of sixes
- (v) number of catches
- (vi) number of stumpings

For this problem with six independent variables, the multivariate regression model takes the following form:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 \dots\dots\dots (4)$$

The weight of the group is then calculated as follows.

$$\text{weight of a team} = \frac{\sum_{i=1}^{11} \text{ith player's points}}{\text{total appearance of the team (ongoing season)}} \dots\dots\dots (5)$$

They tested on that model in 2018 IPL match. Therefore, design a machine learning model (Multilayer Perceptron) to predict the return of the auction based T20 format Premier League tournament with an accuracy of 72.66%. This format makes predictions 15 minutes before the match and after the coin toss.

Machine learning has become a trend in sports analysis, as both live and historical historical data are available as an alternative. Decision-making can be anything from a strategic task such as which player to buy at an auction, which player is on the field for tomorrow's match, or building strategies for upcoming matches based on the players' previous performance (Lamsal and Choudhary, 2020). WASP (Winner and Score Predictor) is one of the few real estate tools implemented in the game of cricket. And so on Sky Sports New Zealand first introduced the device in 2012 at a Twenty20 tournament. Technology like Hawk-Eye monitors the trajectory of a ball and visually shows the most statistically. Now a day similarly, other sports like tennis, badminton, snooker also make use of this computer-assisted intelligent technology.

There is very important to apply existing data mining algorithms to unbalanced datasets to measure the returns of Indian Premier League (IPL) matches. Because most of the time it can be unbalanced (Shimona et al., 2018). Oversampling technique is used for the unbalanced dataset and then the algorithm is applied. Accuracy is used as a performance metric and is calculated using data extraction algorithms. It is also considered as an evaluation criterion and the percentage varies according to different algorithms (Shimona et al., 2018). They used Decision tree, Naive Bayes and linear regression. Before sampling the both balanced and imbalanced data set Naïve Bayes given 96.98% accuracy percentage. After sampling it give the 97.56% accuracy level (Shimona et al., 2018).

According to the research by (Viswanadha et al., n.d.), IPL They suggest a model to predict the winner at the end of each over in the second inning of a cricket match. Their methodology not only incorporates the game context that is dynamically updated as the game progresses, but also includes the relative strength between the two teams playing the game. Assessing the relative

strength between two teams is to demonstrate the potential of the participating players. According to them, there are two main roles of a cricketer. They are batting and bowling. Based on these two factors, they have created the parameters to select the winning team in the second innings. Those parameters are, Runs remaining to be scored to win the match, wickets that batting team still has in hand, balls remaining to be played by batting team in the second inning, and the relative strength of the batting team against to the bowling team (Tekade et al., 2020). They consider 21 states for each match; 1 at the start of the second inning and 20 at the end of each over in the second inning. In these states, predictions can be made by their format. According to the methodology they used Batting average, batting strike rate, batsman score for the batsman rating. Bowling average, bowling strike rate and Bowler score is used to evaluate the Bowlers rating. They used random forest classification algorithms using different dynamics to predict the winner of the competition (Viswanadha et al., n.d.). They got 75.68% accuracy model.

Nowadays, everything is decided by money, so it is no secret that even the second most famous sport in the world is determined by money. Therefore, it is very valuable if the match result can be obtained successfully before the end of the match. There are already used two methods to predict the outcome of the match. The most common method of determining the score by first-batting teams is the current run rate. Second, the standard method of determining the revised target for the pursuit team is the Duckworth-Lewis method (Kalla, et al., 2018). The Duckworth-Lewis (D / L) method is a mathematical formula designed to calculate the goals of a runner-up team during a limited overs match that's interrupted by weather or other conditions (Kalla et al., n.d.).

According to the research project by (Kalla et al., n.d.), they analyzed the work done by several of them as well as the current technology of predicting the performance of the individual as well as the team in a one-day international cricket match. Also, they found the shortcomings of each technique and therefore analyzed them. The inefficiency of the existing system, namely the revised D/L system for determining targets for second-time batting teams, and the current run-rate system for determining the number of points a team scores have led to a comprehensive study of the field (Kalla, et al., 2018). This model aims to use linear regression by taking the balls played by a player in each of his innings and therefore the corresponding runs scored by him because the data set and passing it through the linear regression algorithm (Kalla, et al., 2018).

According to the methodology of this research linear regression algorithm used to develop the model. This gives us the slope and intercept of each player so that we can predict the maximum number of points a player can score in their next match, thereby helping us to predict the total score of the team (Kalla, et al., 2018). Following algorithm developed to implement this model prediction.

$$slope = \frac{(mean(x) \times mean(y) - mean(x \times y))}{(mean(x) \times mean(x) - mean(x \times x))} \dots\dots\dots (6)$$

$$Intercept = mean(y) - mean(x) \times slope \dots\dots\dots (7)$$

Nowadays many people are starting to look for other methods instead of Duckworth Lewis method. They hope to find a more accurate way than the existing Duckworth Lewis method (Ramakrishnan et al., n.d.). To predict the target score and the wicket need extract the key feature from the data set. Current points and wickets are calculated in a game on a ball-by-ball basis. This was then formatted and written to a file and then used to execute the algorithm. They had limited to details - runs, wickets, team, location, batsman, bowler and nature of the over. Other available information, such as batting average and team composition, can be used to make accurate predictions about the progress of a match. (Ramakrishnan et al., n.d.). Using quadratic regression model with KNN and linear regression to develop this prediction.

Data collection for a cricket match shows that over time, the amount of data increases and becomes more complex to analyze. In 2017, (Singh and Kaur, 2017) introduce the Hbase to keep the data related to IPL cricket matches and players. This is helps to the huge data maintain. Past data is needed when the model created over time needs to be updated, even with new data. This profiling system which can be great help for the predictors to predict players to the match. HBase for scalability of application (Singh and Kaur, 2017). And this research paper used KNN algorithm to predict the suitable players. This research exposed a good idea for visualized the past data and the how can it be collected.

2.2 Background of used materials

Multiple Linear Regression (MLR)

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the return of a response variable. The purpose of the multiple regression is to demonstrate the linear relationship between the explanatory (independent) variables and the response (dependent) variables.

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \dots + \alpha_n x_{in} + \varepsilon \dots \dots \dots (8)$$

Where, for $i=m$ observations:

$y_i =$ *dependent variable*

The variable being tested and measured in an experiment, and is 'dependent' on the independent variable

$x_i =$ *explanatory variable*

The variable the experimenter manipulates or changes, and is assumed to have a direct effect on the dependent variable

$\alpha_0 =$ *y - intercept*

The mean for the response when all of the explanatory variables take on the value 0

$y_n =$ *slope coefficient for each explanatory variable*

$\varepsilon =$ *the model's error term*

Ridge Regression

Ridge regression is a format tuning method used to analyze any data that is poly conductive. This method formalizes L2. When there is a problem with multiple conductivity, at least the squares are neutral, and when the variables are large, this causes the expected values to be many of the true values.

$$y = xB + \theta \dots \dots \dots (9)$$

If Y is the dependent variable, then X represents the independent variables, B is the regression coefficient to be estimated, and E represents the remainder.

Once we add the lambda function to this equation, we will consider the variance that is not evaluated by the normal model. Once the data has been processed and identified as part of the L2 formalization, there are steps one can take.

Lasso Regression

In the study of statistics and machinery, "LASSO" is a retrospective analysis method that performs both variable selection and formalization to improve the predictive accuracy and interpretation of the resulting statistical model. The Lasso regression uses L1 formalization technology. Feature selection is done automatically so it is used when we have more features.

Lasso regression is similar to linear regression, but it uses a "shrinkage" technique, where the determination coefficient is reduced to zero. lasso regression allows you to compress or regulate these coefficients.

Neural Network

A neural network can be divided into 3 parts. They are input layer, hidden layers and output layer.

Input layer: This enters past data values into the next layer.

Hidden layer: This is a key component of a neural network. It has complex functions that predict. A group of nodes called neurons in the hidden layer represent mathematical functions that alter input data.

Output layer: The predictions of the hidden layer are added to produce the final layer. It is the prediction of the model.

Because of the hidden layers, neural networks perform better in predictive analysis. Linear regression models use only input and output nodes for prediction. The neural network also uses a hidden layer to make predictions more accurate.

Think of each individual node as a linear regression model, consisting of input data, scales, inclination (or threshold), and output.

Support Vector Machine Algorithm

Support vector machine or SVM is one of the most popular supervised learning algorithms used for classification as well as regression problems. Basically, it is used for classification problems in machine learning.

The objective of the SVM algorithm is to create the best line or decision boundary that can divide n-dimensional space into classes so that the new data point can be easily placed in the correct category in the future. This best decision boundary is called the hyperplane.

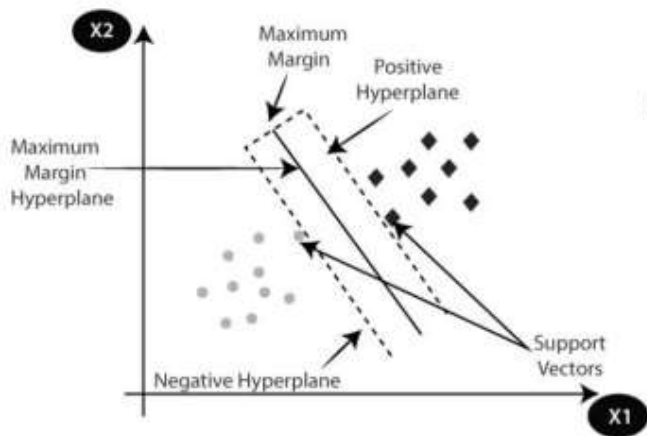


Figure 1: Two different categories classified using SVM

Decision Tree

The decision tree is the most powerful and popular algorithm for classification and prediction. A decision tree is a flow chart, such as a tree structure, where each internal node represents a test on a feature, each branch represents a test result, and each terminal node has a class label.

Table 2: Truth table for AND operation

<i>A</i>	<i>B</i>	<i>A and B</i>
F	F	F
F	T	F
T	F	F
T	T	T

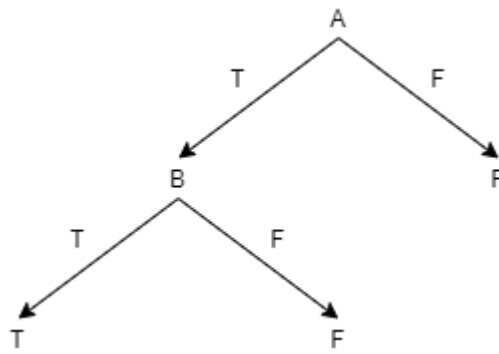


Figure 2: Decision tree AND operation

Random Forest

Random Forest is a machine learning technology used to solve regression and classification problems. It utilizes team learning, a technology that integrates many classification machines to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

MLP Classifier

A multi-layer perceptron is a group of artificial neural networks that provide nutrition. The term MLP is used vaguely, sometimes referring to any loosely feed forward ANN, sometimes referring to networks with several layers of rigid sensors.

2.3 Over View of Twenty-Twenty Cricket Match

- Introduced by the England and Wales Cricket Board in 2003 for the inter country competition.
- Both teams have single innings.
- Maximum 20 overs and 10 wickets for each team.
- A sequence of six balls bowled by a bowler from one end of the pitch is called an over (Viswanadha et al., n.d.).
- An innings is one of the divisions of a cricket match during which one team takes its turn to bat (Viswanadha et al., n.d.).
- Each inning about 90 minutes and 10 minutes break between the innings.
- Nearly 3 hours for a match.
- In a Super Over, each team bats for one extra over after the match. The team with the highest score in the Super Over is declared the winner. Super Over, also known as One Over Eliminator, is a method used to determine the winner of a tie in limited overs cricket

matches.

2.4 Tools and Technologies

Python

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation.

Why python can use in AI?

Implementing the AI and ML algorithms can be confusing and time consuming. Having a well-structured and well-tested environment is essential for developers to be able to come up with the best coding solutions. To reduce development time, programmers turn to a number of Python frameworks and libraries.

Table 3: Python Libraries

Data Analysis and Visualization	NUMPY, SCIPY, PANDAS, SEABORN
Machine Learning	TensorFlow, Keras, Scikit-learn
Computer vision	OpenCV
Natural Language processing	NLTK, spaCy

YAML

The data from the data source available for this research were included in this YAML format. YAML is a human-readable data-serialization language. It is commonly used for configuration files and in applications where data is being stored or transmitted. YAML targets many of the same communications applications as Extensible Markup Language but has a minimal syntax which intentionally differs from SGML.

CSV

For the convenience of this study, the collected data was converted to CSV format. A comma-separated values file is a delimited text file that uses a comma to separate values.

SPYDER

Spyder is an open-source cross-platform integrated development environment for scientific programming in the Python language. This IDE was used to build the python scripts for the research project.

ANGULAR

Angular is a TypeScript-based open-source web application framework led by the Angular Team at Google and by a community of individuals and corporations. Angular is a complete rewrite from the same team that built AngularJS. This technology was helped to develop an user interfaces for this research project.

VISUAL STUDIO CODE

Visual Studio Code is a freeware source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. This IDE was used to angular developments.

CHAPTER 3

METHODOLOGY

This chapter focuses on the problem analysis, design and methodology of research work. The methodology is equipped with solutions for the research purposes outlined in the introductory chapter.

3.1 Problem Analysis

The aim of this research project is to resolve the issue mentioned in the problem statement. According to the literature review, the stakeholders of the twenty20 match are currently not using an AI based application for predicting the winning team and the final score of each inning and all types of matches in the Twenty20 category.

Although the ICC has all the data related to past matches, it requires a great deal of numerical and mathematical knowledge to predict (Naik et al., 2018). A typical Twenty20 game is completed in about three hours, with each innings lasting around 90 minutes and an official 10-minute break between the innings. Accordingly, the match will end in a very short time. During that time, it is impractical to perform calculations and predict the result with a large amount of data. Therefore, this can be achieved in a very short time by using artificial intelligence. In Twenty20 matches, bettors have only limited time period to predict the outcome. Otherwise they cannot win their game. Because of that we need much efficient and high fast system to predict the twenty20 game outcome.

Currently, in Twenty20 matches inning score is predicted on the basis of current run rate which can be calculated as the amount of runs scored per the number of overs bowled. It does not include factors like of wickets fallen, remaining overs and venue of the match. This calculation results in inaccurate interpretations. For example, in the first innings of the innings, when 24 runs for 2 wickets, the projected score at the end of 20 overs is 480 runs. It can't happen according to the Table 4: Highest innings totals (Source :ESPNcricinfo).

Given the following fact, it is clear that this is an inaccurate conjecture and that various factors should appear to be influencing the projected mark. This example gets from an ODI match. The calculation method of the projected score same as the Twenty20. It shows the problem of the current score prediction. For this reason, other factors influencing the decision of the projected score of the match must be found.



Figure 3: Example of current score projection

According to the above figure shows the score prediction of the ODI match with West Indies and England. West Indies scored 66 runs for 4.3 overs with 0 wickets. According to that run rate, projected score is 733 if they maintain the same run rate for the all 50 overs. This value is unbelievable because of the wickets can be fallen in relevant inning. According to the below table in T20 matches the highest score is the 278. So that current projection system is not accuracy with the all the parameters that can be an effect to the final result of the match.

Table 4: Highest innings totals (Source :ESPNCricinfo)

Rank	Score	Team	Opponent	Venue	Date
1	278/3 (20 overs)	Afghanistan	Ireland	Dehradun	2019/02/23
	278/4 (20 overs)	Czech Republic	Turkey	Ilfov Country	2019/08/30
2	267/2 (20 overs)	Trinbago Knight Riders	Jamaica Tallawahs	Kingston Jamaica	2019/09/23
3	263/3 (20 overs)	Australia	Sri Lanka	Pallekele	2016/09/06
	263/3 (20 overs)	Royal Challengers Bangalore	Pune Warriors India	Pallekele	2013/04/23

According to the problem statement, in order to provide a successful solution to it, we need to achieve the objectives mentioned above. In first and second objectives can achieve using the ball by ball information of a match. So that we need to identify the key features to predict the inning final score and the wicket count. Based on the literature review and some statistical analysis, several key parameters were identified to predict the final score and number of wickets according to the status of a match during its existence. They are tournament type, inning, batting

team, bowling team, over, ball of the over, striker, non-striker, bowler, total runs for the ball, runs for the batsman in the delivery, extra runs in the delivery, wide runs in the delivery, no ball runs in the delivery, byes in the delivery, leg byes in the delivery, is that delivery pick up the wicket, wicket type, current runs of the team, current wickets of the team, current runs of the striker, current runs of the non-striker, last 5 overs runs, last 5 overs wicket count.

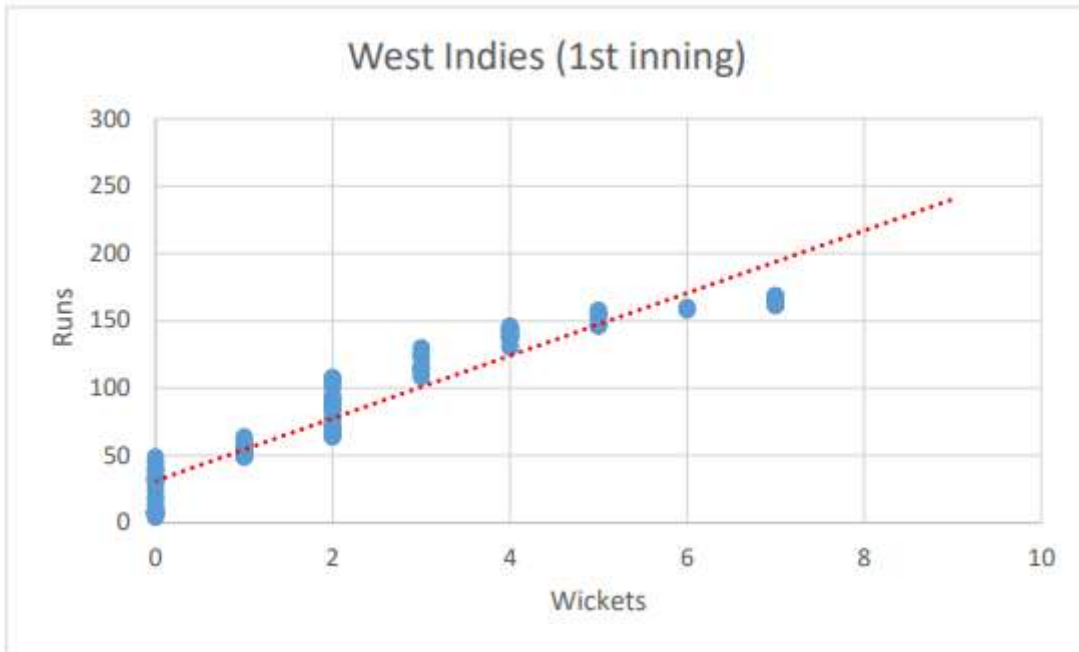


Figure 4: Runs vs wickets visualization of West Indies team in a match

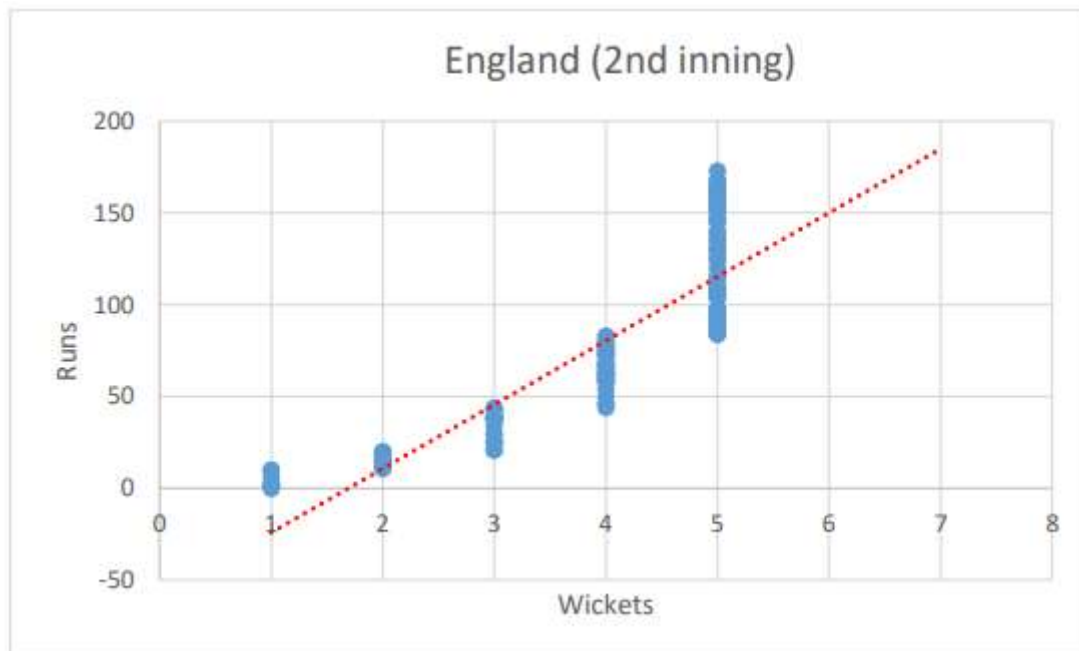


Figure 5: Runs vs wickets visualization of England team in a match

The charts above show how the scores of both teams and the number of wickets taken vary in a Twenty20 cricket match. This varies in this way in this match but in another way in another match. Let us interpret the above two graphs related to this competition. The two teams

participating in the tournament are the West Indies and England. In the first innings, the West Indies batted and lost their first wicket for a total of 49 runs. They also lost 3, 4, 5, 6, 7 wickets for 54 runs. They scored 108 runs for the first two wickets but then had to score 169 runs due to rapid wicket collapse after the second wicket. Playing in the second innings, England lost the first two wickets for 20 runs, but as the pace of wickets slowed they scored 0 50 100 150 200 250 300 0 2 4 6 8 10 Runs Wickets West Indies (1st inning) -50 0 50 100 150 200 0 1 2 3 4 5 6 7 8 Runs Wickets England (2nd inning) 22 173 for 3, 4 and 5 wickets. Thus, it seems that the final score depends on the speed at which the wickets are taken. In this way we can show how the above selected key parameters affect the final score and they can be seen even in the literature review.

Then let us look at the factors that influence the achievement of the third objective. Accordingly, this prediction can be made after the coin toss and when the second inning is played. This prediction can be made based on the basics information of the match after looking at the coin advantage, which also has key parameters. They are tournament type, team1, team2, toss winner, decision of the toss winner, umpire1, umpire2, venue, is Duckworth Lewis applied, is team1 playing at their home ground, is team2 playing at their home ground and final results of the match (can be no result or tie). Thus we see that the parameters such as the venue, the team that wins the toss, and the two teams that play clearly affect the team that wins the match. This is clearly stated in the literary review.

When the second inning is played, the winning team can be selected using not only the basic data but also the ball-to-ball data. This can be done by creating a hybrid model with the help of the models obtained above, i.e. the model obtained based on the core data and the model obtained using ball-to-ball data. Accordingly, in one model the winning team gets directly, while in the other model it looks like the projected score is received. Since the winning team does not receive directly from one model, some additional counting is required. This hybrid format is required in the second inning, so the winning team can be selected based on the value of the projected value and the score obtained in the first inning.

After achieving the first 3 objectives in this way, it can be made available to the desired users using a web based application. Therefore, using this in a very short period of time, users can get the winning team and the projected score.

In this way the solution to the problem mentioned above will enable the bettors to obtain all the

data required for betting and the cricket governing body with a very high degree of success. It is also possible to stop the transmission of projected values that may not occur as mentioned above.

3.2 Design

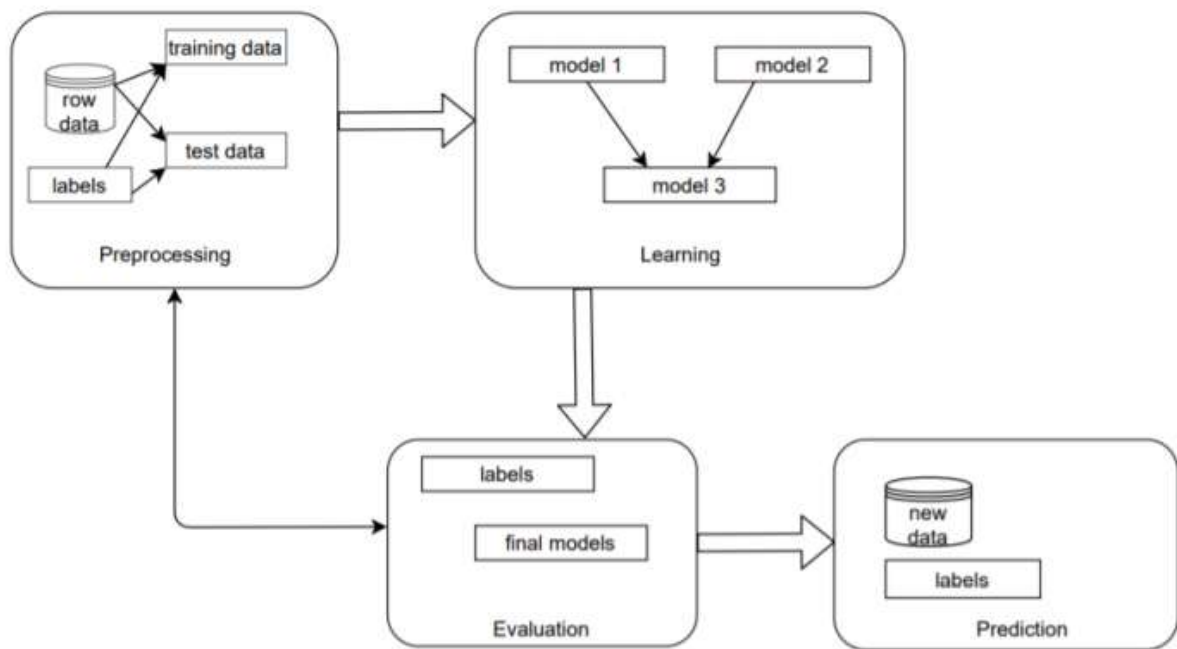


Figure 6: System Architecture

The figure presented above shows an outline of the solution to be given. In the first stage collected data should be preprocessed. At that stage, data extraction, data cleaning, label encoding, and dataset splitting (Test data & Train data) are performed. The second stage involves training machine learning models using training data. Accordingly, one model is designed to analyze key data and determine the winning team, while another model is designed to analyze ball-by-ball and determine the projected score and the projected wickets count. Here the output of the first two models is used to create the third model. In the evaluation stage evaluate all the model with the test data and get the final machine learning models. At the end of the process we can make our predictions using newly fresh data.

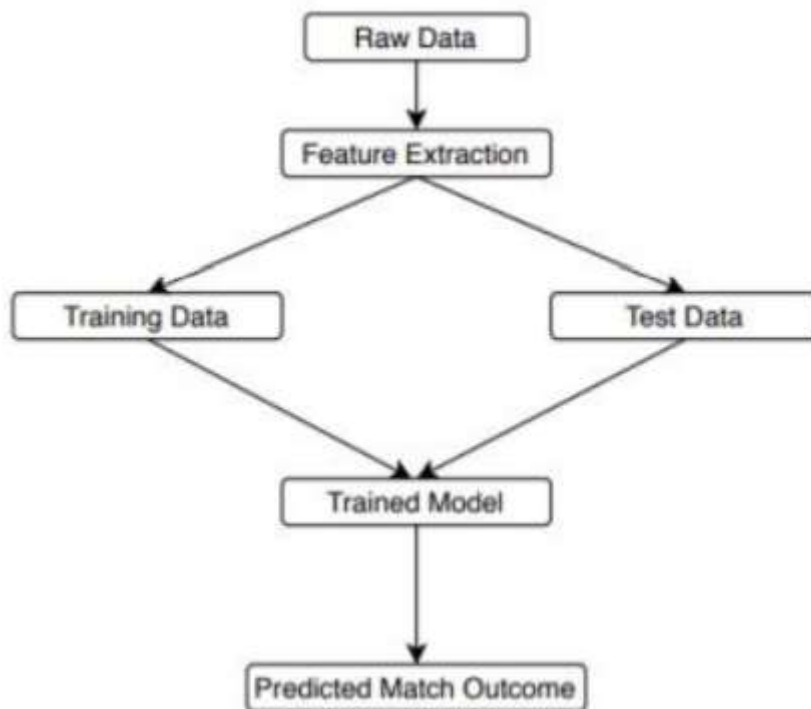


Figure 7: Architecture of a model

This figure shows how one model is developed and the match result is determined. That is, it shows how one model is developed here. Here the features are extracted from the raw data and divided into two parts as training data and test data. The model is then trained and the predictions can be made from it.

The diagram below shows the sequence of actions that take place when a user uses this solution. Accordingly, a user can enter the data while the Twenty20 match is in progress, giving the winning team and the total number of runs and wickets expected in the innings being played using this system. Here the user can enter the data into the interface provided to him or her. The data is transferred to a database through the backend service. If a parameter can be obtained from a calculation, the user does not need to enter it. Take, for example, the number of runs scored in the last 5 overs. Through that backend service, the data is retrieved, the other required parameters are calculated and submitted to the machine learning model. Here only the required data is given to the user to enter. The results obtained by those machine learning models are then brought back to the interface via the backend service. Then user can see the result.

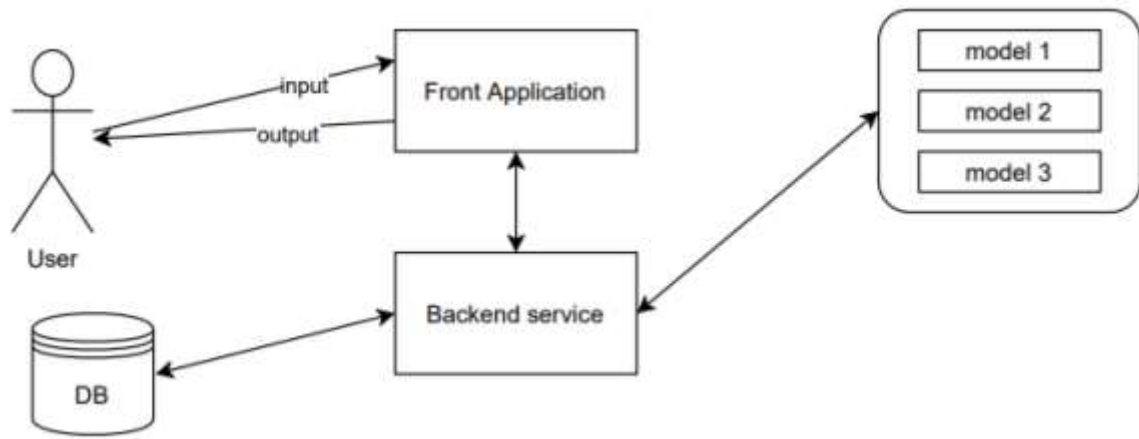


Figure 8: High level architecture of the application



Figure 9: Python Service running instance

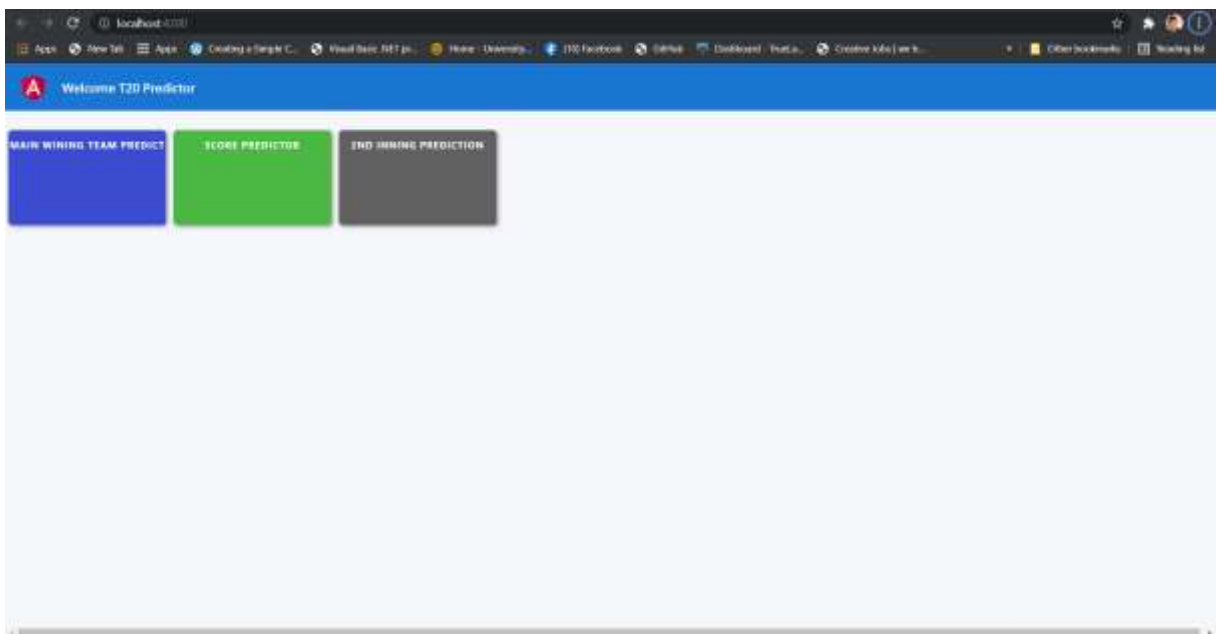


Figure 10: Home page of the application

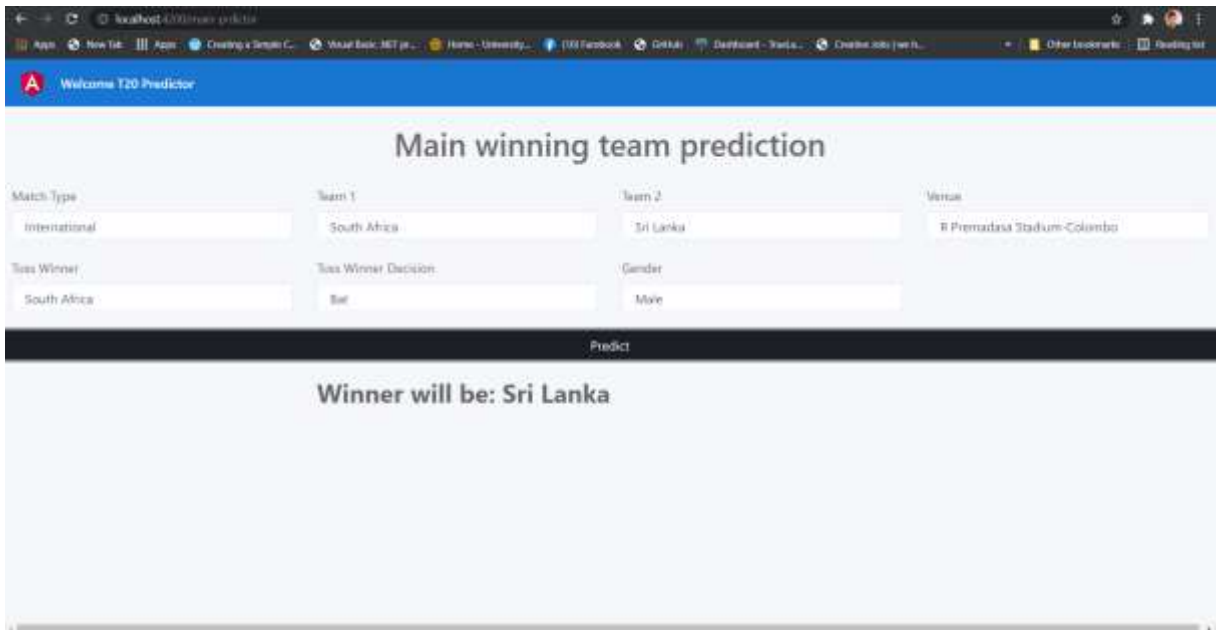


Figure 11: Prediction of Basic Information (Model 1)

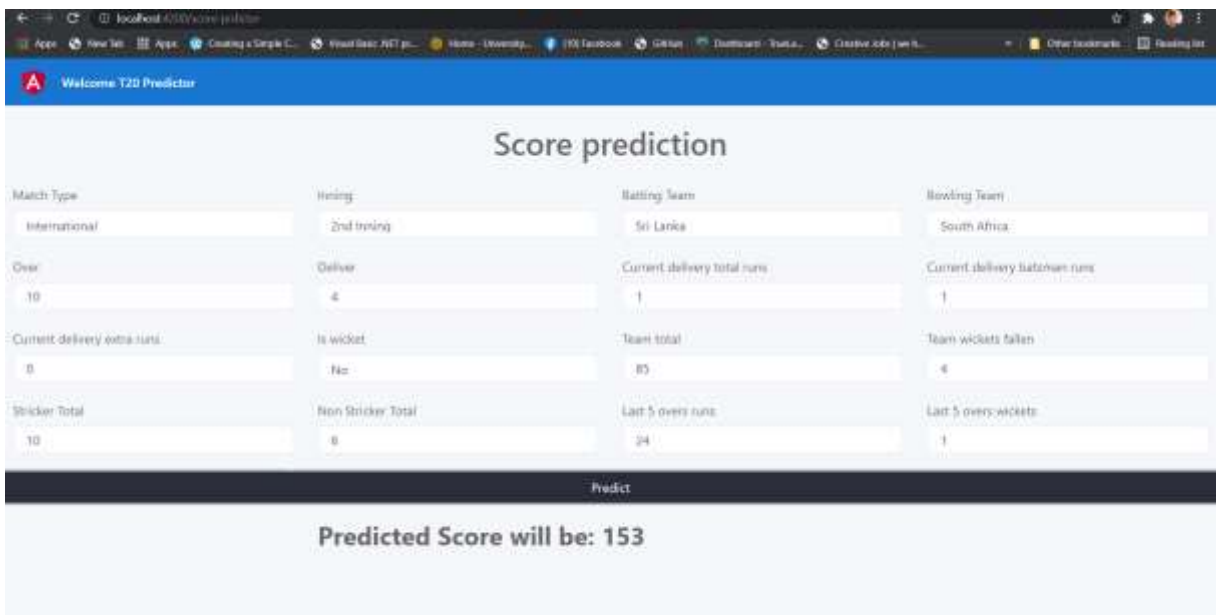


Figure 12: Score predictor (Model 2)

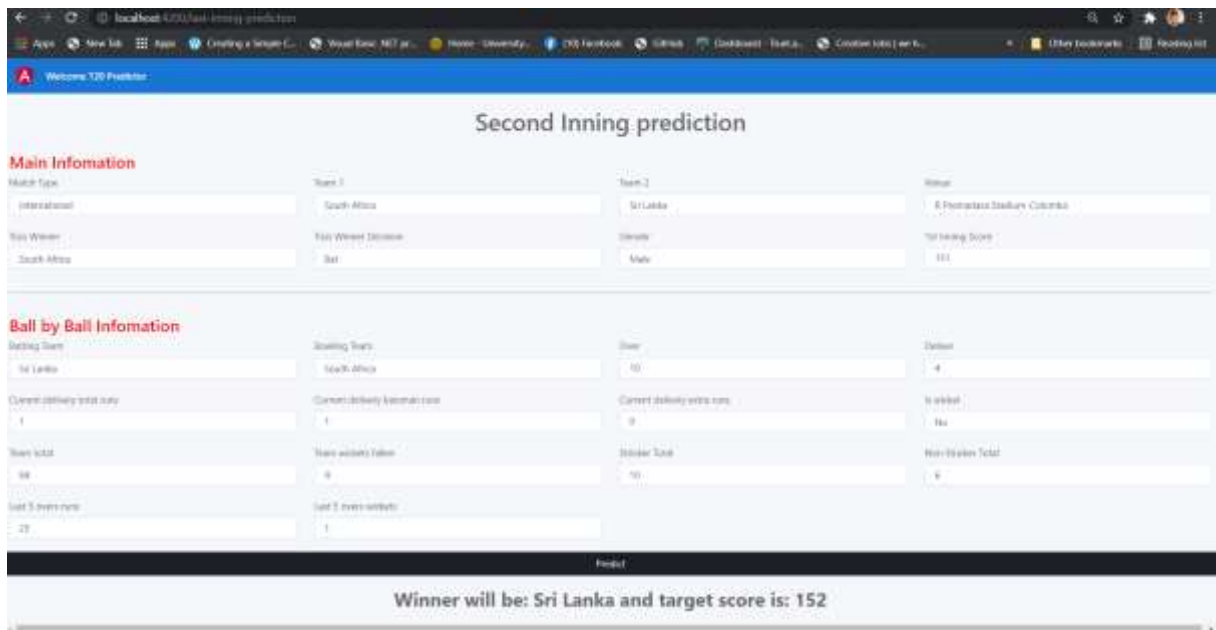


Figure 13: Hybrid model predictor (Model 3)

3.3 Methodology

According to the project the first goal is to collect the data. The “CRICSHEET” website provides all the data that required to the research. It is Powered by tea, obsessiveness, and a fascination for cricket stats. The web page provides access to all data from 2004 to 2021. Accordingly, the following Twenty20 matches were used for research.

Table 5: Match Count against the tournament(Source : <https://cricsheet.org/>)

Tournament	Match Count
T20 International	1482
Non official T20 International	328
Big Bash League	414
Indian Premier League	845
Caribbean premier League	244
T20 Blast	815
Pakistan Super League	160
Women’s Big Bash League	264

The data on this web page is in YAML format. YAML is a human readable data serialization language. There is one YAML file for each match. It was converted to CSV format because it was difficult to deal with this YAML format and the machine learning algorithm together. All

the YAML files related to one tournament type were taken and all the basic match data required as mentioned above were obtained to one CSV file. Further, all the ball-by-ball data of one match of the tournament was obtained to another CSV file. In that way, ball-by-ball data for all matches related to that tournament were retrieved into separate CSV files. The data for all of the tournaments mentioned above were then processed into CSV files. In that way, 8 CSV files containing the basic data were obtained. In addition, ball-by-ball data for each match in each tournament was retrieved into separate CSV files. A Python script was used for this. The YAML and CSV libraries were used in that Python script.

```
1 | ---
2 | meta:
3 |   data_version: 0.9
4 |   created: 2013-02-22
5 |   revision: 1
6 | info:
7 |   city: Southampton
8 |   dates:
9 |     - 2005-06-13
10 |   gender: male
11 |   match_type: T20
12 |   outcome:
13 |     by:
14 |       runs: 100
15 |       winner: England
16 |     overs: 20
17 |   player_of_match:
18 |     - KP Pietersen
19 |   teams:
20 |     - England
21 |     - Australia
22 |   toss:
23 |     decision: bat
24 |     winner: England
25 |   umpires:
26 |     - NJ Llong
27 |     - JW Lloyds
28 |   venue: The Rose Bowl
29 | innings:
30 |   - 1st innings:
31 |     team: England
32 |     deliveries:
33 |       - 0.1:
34 |         batsman: ME Trescothick
35 |         bowler: B Lee
36 |         non_striker: GO Jones
37 |         runs:
38 |           batsman: 0
39 |           extras: 0
40 |           total: 0
41 |       - 0.2:
42 |         batsman: ME Trescothick
43 |         bowler: B Lee
```

Figure 14: Sample of YAML file

All the CSV files obtained above should be merged. In doing so, the basic data was transferred to one CSV file, and the ball-to-ball data was retrieved into another CSV file according to the date. At this stage, the relevant data has been collected into two CSV files.

	A	B	C	D	E	F	G
1	match_no	match_type	team1	team2	toss_winner	toss_winner_decision	umpire_1
2	1001349	International	Australia	Sri Lanka	Sri Lanka	field	MD Martell
3	1001351	International	Australia	Sri Lanka	Sri Lanka	field	SD Fry
4	1001353	International	Australia	Sri Lanka	Sri Lanka	field	MD Martell
5	1004729	International	Ireland	Hong Kong	Hong Kong	bat	R Black
6	1007655	International	Zimbabwe	India	India	field	TJ Matibiri
7	1007657	International	Zimbabwe	India	Zimbabwe	bat	L Rusere
8	1007659	International	Zimbabwe	India	Zimbabwe	field	TJ Matibiri
9	1019979	International	New Zealand	Bangladesh	Bangladesh	bat	SB Haig
10	1019981	International	New Zealand	Bangladesh	Bangladesh	field	CM Brown
11	1019983	International	New Zealand	Bangladesh	Bangladesh	field	CM Brown
12	1020029	International	New Zealand	South Africa	New Zealand	field	CM Brown
13	1031431	International	England	South Africa	South Africa	bat	RJ Bailey
14	1031433	International	England	South Africa	England	field	RJ Bailey
15	1031435	International	England	South Africa	South Africa	field	MA Gough
16	1031665	International	West Indies	England	England	field	RT Robinson
17	1034825	International	India	England	England	field	AK Chaudhary
18	1034827	International	India	England	England	field	AK Chaudhary
19	1034829	International	India	England	England	field	AK Chaudhary
20	1040485	International	Afghanistan	Ireland	Ireland	bat	Ahmed Shah Durrani
21	1040487	International	Afghanistan	Ireland	Afghanistan	bat	Ahmed Shah Durrani
22	1040489	International	Afghanistan	Ireland	Ireland	field	Ahmed Shah Durrani
23	1041615	International	India	West Indies	India	field	N Duguid
24	1041617	International	India	West Indies	India	field	LS Reifer
25	1042080	International	Australia	New Zealand	New Zealand	field	CA Abroad

Figure 15: Sample CSV file

For convenience, consider the CSV file containing the basic data as the first CSV file, and the CSV file containing the ball-by-ball data as the second CSV file. Also, Let the machine learning model designed to achieve the first and second objectives is the second machine learning model, the machine learning model designed to find the winning team in the 1st inning is the 1st machine learning model, and the hybrid machine learning model is designed to find the winning team in the 2nd inning as the 3rd model of machine learning.

Accordingly, the 1st CSV file is required for the second machine learning model. The first CSV file contains all the input and output parameters needed to create this second machine learning model. So it was very easy to extract that data. The parameters required to create the 1st machine learning model are not contained in the 2nd CSV file, and all other data were extracted.

Accordingly, the 1st CSV file is required for the second machine learning model. The first CSV file contains all the input and output parameters needed to create this second machine learning model. So, it was very easy to extract that data. The parameters required to create the 1st machine learning model are not contained in the 2nd CSV file, and all other data were extracted.

Current runs: The total number of runs scored from ball by ball in that innings was added Up to now

Current wicket counts: The wickets fallen from ball by ball in that innings was added Up to now.

Last five overs run: The total number of runs scored from ball by ball in that last five overs were added. In that case, not consider the current over runs. Only considering the last five overs before the current over. In example the current over is 7, then consider the 6th ,5th ,4th ,3rd and 2nd overs.

last five overs wickets: The wickets fallen from ball by ball in that last five overs were added. Same as here not consider the current over wickets. Only considering the last five overs before the current over. In example the current over is 10, then consider the 9th ,8th ,7th ,6th and 5th overs.

Total score: The total number of runs scored from ball by ball in that innings was added.

Next, the data was cleaned up. It did not require much data cleaning. But had to be cleaned up a bit. For example, one player may be named in two ways.

Ex: KP Pietersen / Kevin Pietersen
B Lee / Brett Lee

It was also necessary to remove unnecessary special characters in certain string values. This data was also cleaned using a Python script.

After the data cleaning part, data labeling should be done. Because the categorical data should be labeled for the machine leaning algorithms. Some examples shown in the below figures.

Table 6: Example of team names labeling

<i>Team</i>	Sri Lanka	India	Pakistan	...	Australia
<i>Labeling</i>	1	2	3	...	345

Table 7: Example for wicket types labeling

<i>Wicket Type</i>	Caught	Bowled	Hit Wicket	Run Out	Stumping
<i>Labeling</i>	1	2	3	4	5

Table 8: Example for umpires labeling

<i>Umpires</i>	AJ Neill	RT Robinson	K. Dharmasena	...	Nitin Menon
<i>Labeling</i>	1	2	3	...	124

All of the following classification data are labeled as above.

- Tournament Type
- Team Names
- Players
- Wicket types
- Toss decision
- Umpires
- Gender
- Venues
- Results of the match

After that need to split the data as the training data and the test data. According to a tournament select 70% for training and the 30% for test. That 70% should select from first 70% matches according to the Date. Latest matches data as the test data.

According to the literature review study, the multiple linear regression algorithm is used to create the first machine learning model. And the neural network is used to create the second machine learning model. After that 3rd machine learning model created using the outputs of the 1st and 2nd algorithm. That output uses as input of the 3rd machine learning model. This 3rd model develop using support vector machine algorithm.

Finally, all the three models should be integrating into one single application. According to the design that application user interfaces implement using the angular, service layer design using implement using the node.js and the database should be a MySQL. Because certain parameters are not captured according to the user interface, the data is calculating and rendered to the database. This is performed by the backend service. Doing so makes it easier to input data into machine learning models.

All the Python scripts run using the Spyder tool and the Angular developments done using the visual studio code tool.

CHAPTER 4

EVALUATION AND RESULTS

During this chapter, the collected data will be analyzed and the models created using that data will be evaluated. Here, to evaluate those models, he used data from sports that had not been used to design the model.

Model 01 (Basic winning team predictor)

This model selects the winning team based on the data available before the start of the game. So that, the data obtained from the previous matches of the teams participating in the tournament, the stadium where the match will be played as well as the information on the toss advantage are presented here. Accordingly, an analysis of what the inputs are for the first model is done here.

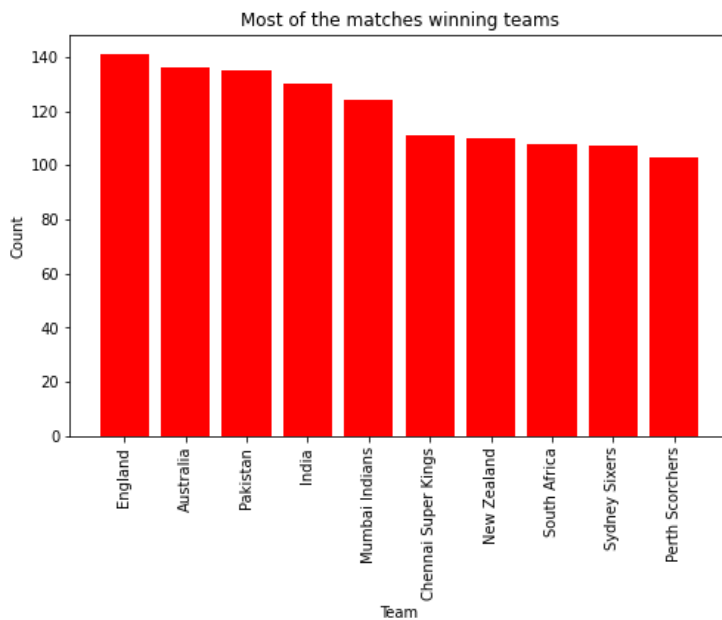


Figure 16: Most of the matches winning teams

Above figure shows the teams that got the most wins in T20 Category. These results obtained from the all the T20 matches records to do this research. This figure shows the top 10 winning teams.

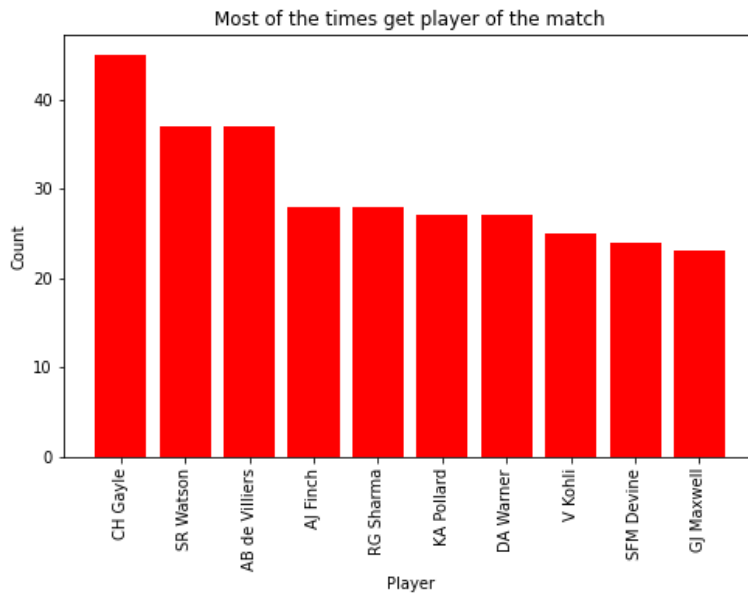


Figure 17: Most of the times get player of the match

The Figure 17: Most of the times get player of the match shows the who has get the player of match most of the times. This is will be a good parameter for the outcome of the final match. If a batsman gets a high score for an inning, that inning score will get high. According to that, in this research we have to check the effectiveness of the striker and non-striker scores.

The below figure shows most of the matches played grounds. In here shows only the top 10. According to that Dubai International Cricket stadium is in the highest level. According to the ground, conditions of the pitch will change. So that some team can get that advantage to win this game. As an example, Sri Lanka and India has good spinners. Because of those two countries have slow pitches. But Australia and England have good fast bowlers in their teams. Because they have fast pitches. When Sri Lanka and India Playing a game in slow pitch, they will get that advantage with spinners. If Australia and England playing at fast pitch, they will get that advantage with fast bowlers.

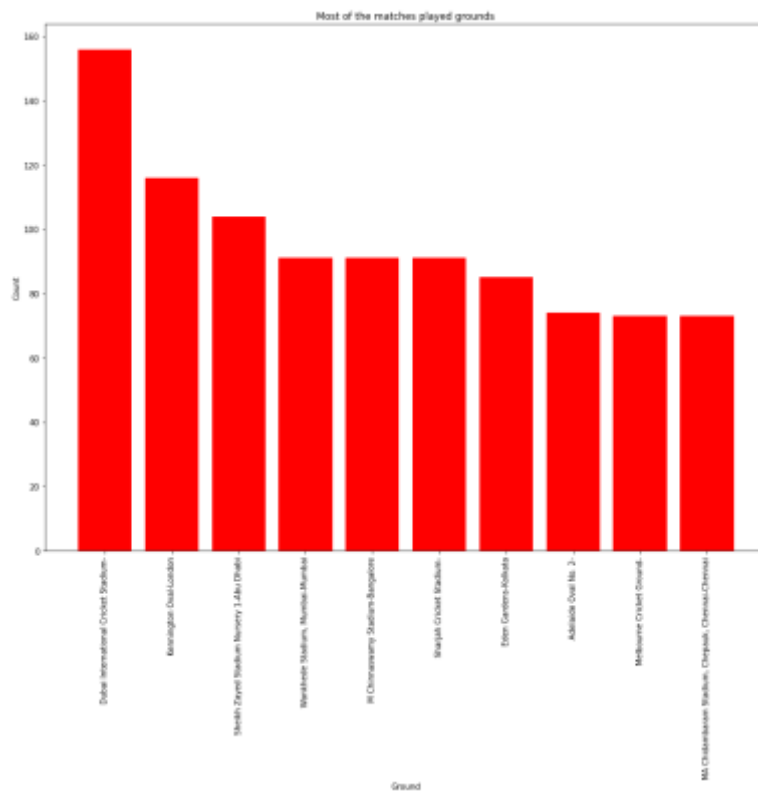


Figure 18: Most of the matches played grounds

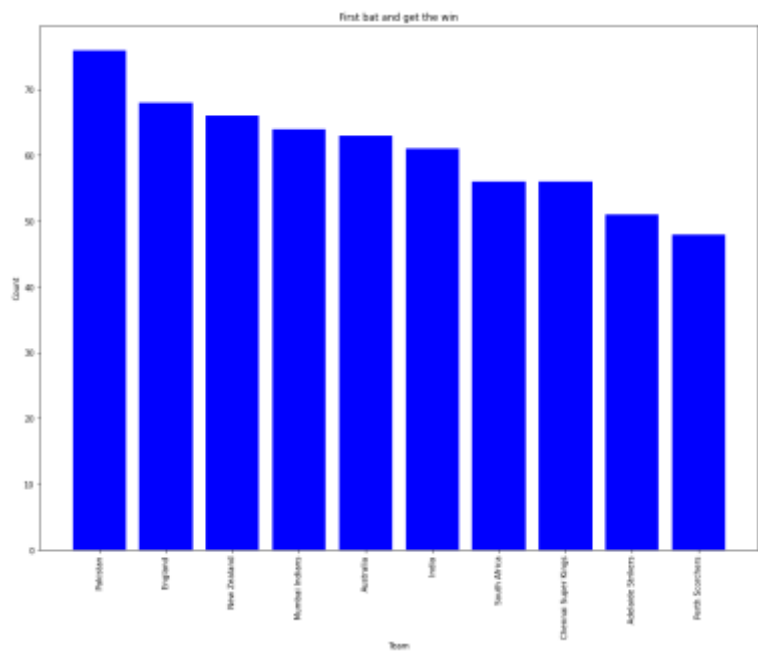


Figure 19: First Bat and get the win

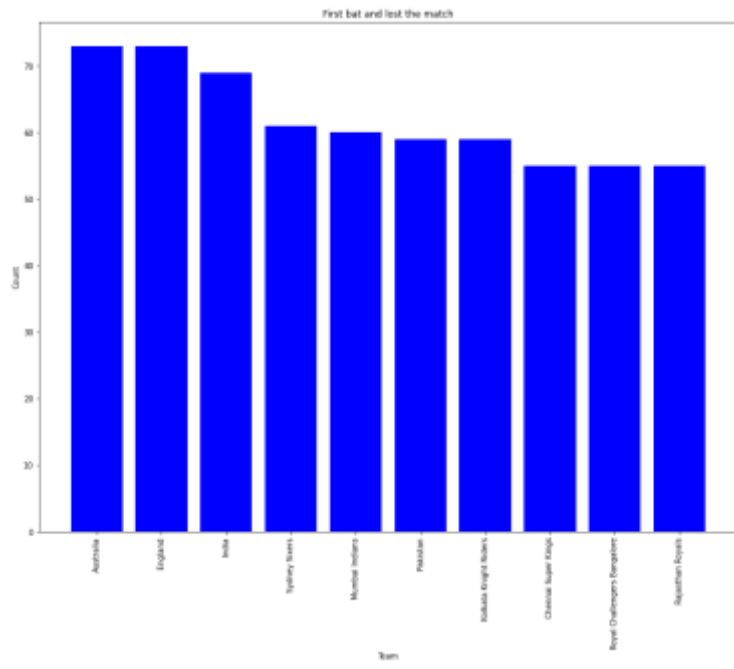


Figure 20: First bat and lost the match

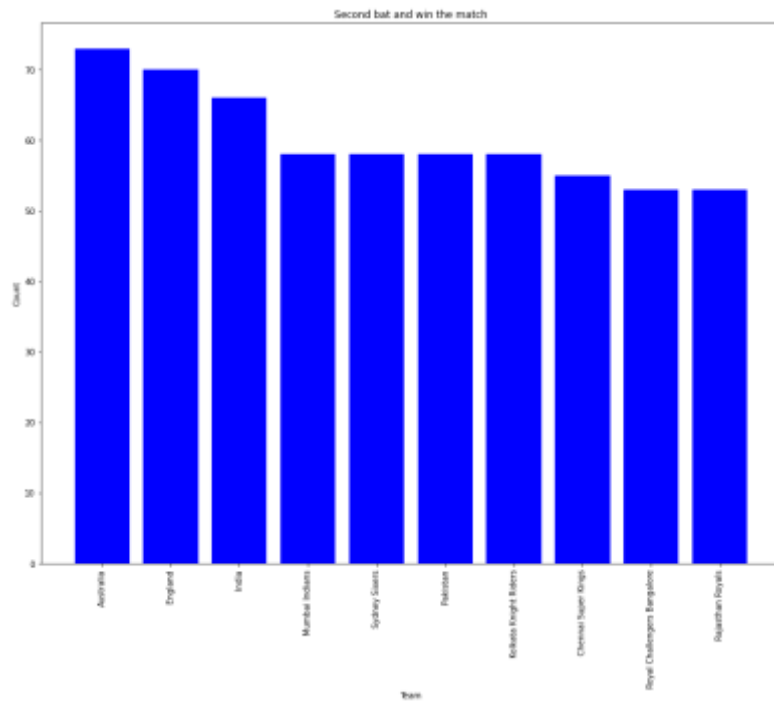


Figure 21: Second bat and win the match

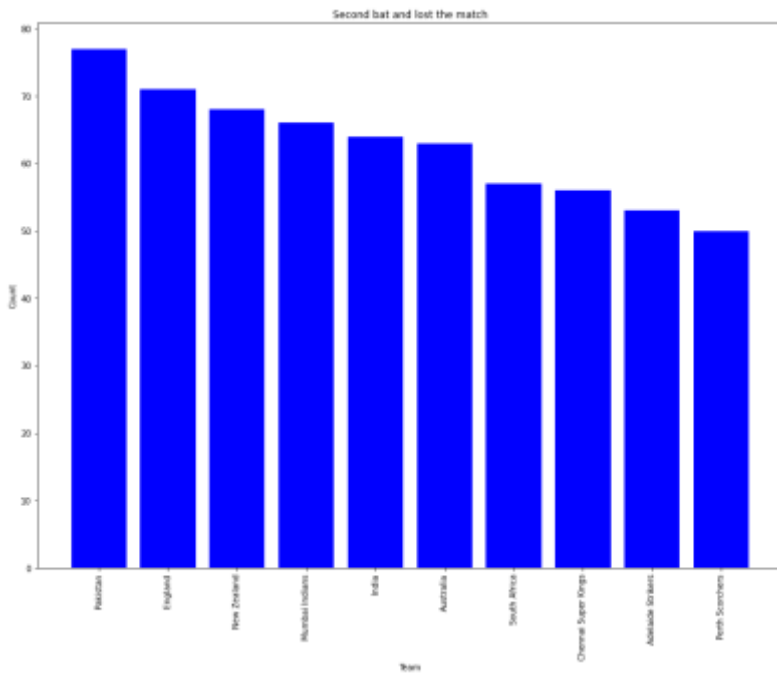


Figure 22: Second bat and lost the match

Another special feature is that the way the first bat and the second bat affect different situations as well as the team changes. In this way, it seems that it is very important for the winning team to change according to the opportunity to bat when playing under different conditions. The team that wins the coin toss is lucky to decide whether or not to bat first.

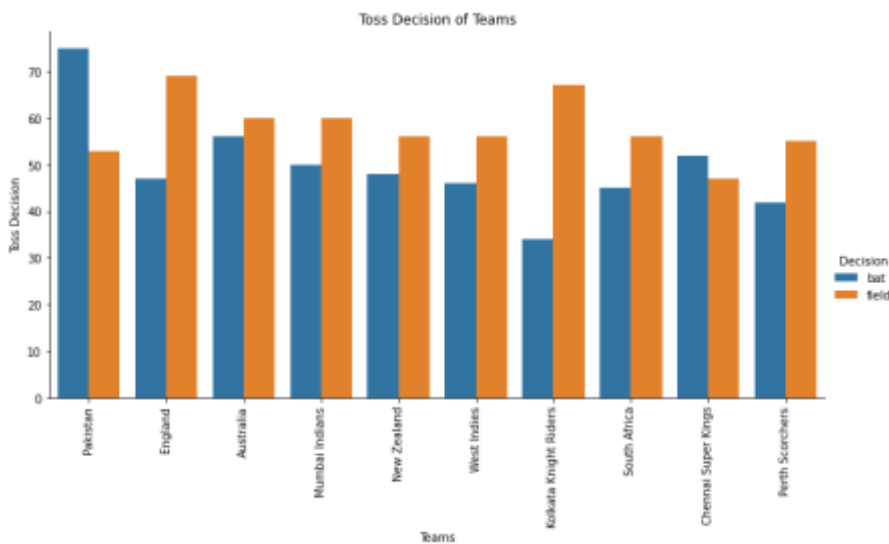


Figure 23: Toss decision of teams

The above figure shows how the toss decision of the teams that have won the most coin toss changes. The image below shows the decision made by the top 10 teams that won the most matches after the coin toss victory.

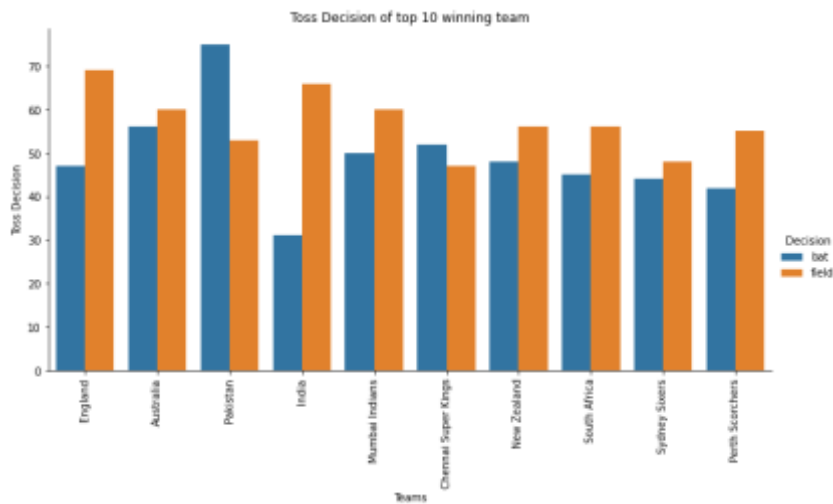


Figure 24: Toss decision of top 10 winning team

Therefore, according to these two images, we can see that the decision of the coin advantage is also changed by different groups at different times. Therefore, the advantage of the coin as well as the decision to make it after the victory is very important for the victory of the competition.

Shown below is a heat map created using all the parameters obtained from the literature review. It can easily present the relationship of each parameter to the final decision.

According to the below heat map, following parameters used for the model 01 inputs.

- Match Type
- Team 1
- Team 2
- Ground
- Toss Winner
- Toss Winner Decision
- Gender



Figure 25: Heat map for Model 01

Table 9: Accuracy for model 1 with different algorithms

Model	Accuracy Imbalanced	Accuracy Balanced
Random Forest	84.51%	76.61%
Decision Tree	86.23%	75.53%
Neural Network	60.01%	54.79%
MLP Classifier	71.14%	69.12%
SVM Classifier	79.58%	74.11%

According to the above table, the Random Forest algorithm gives the highest accuracy for model 1. The importance of the features of the model with that maximum accuracy can be obtained as follows. Accordingly, the highest accuracy of 76.61% is obtained by random forest, which is the value obtained by converting the imbalance database into a balance database. Over sampling methods were used for this conversion.

Table 10: Feature importance for Random Forest

Feature	Importance
Match Type	7.23%
Team1	22.30%
Team2	23.64%
Ground	20.84%
Toss winner	20.62%
Toss winner Decision	4.01%
Gender	1.35%

Model 02 (Score Predictor)

This model predicts the maximum number of runs a team can score in an innings in a match. In order to obtain the relevant input parameters, in addition to the data in the database, some data had to be retrieved computationally. Those categories were presented in the chapter above. Accordingly, the following results were obtained for all the data.

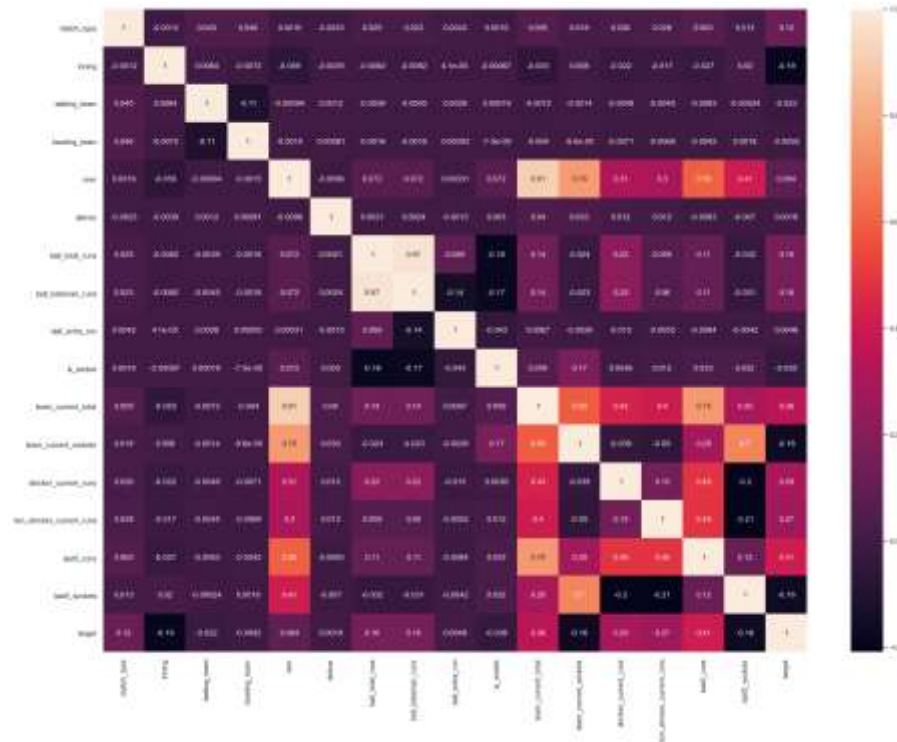


Figure 26: Heat map for model 02

The heat map above was used to check the parameters required to predict the target score. According to that following parameter were selected as the inputs.

- Match Type
- Inning
- Batting team
- Bowling Team
- Over
- Deliver
- Total Runs for ball
- Runs for batsman
- Extra Runs
- Is Wicket
- Team Current Total
- Team Current Wickets

- Striker Current Runs
- Non Striker Current Runs
- Last 5 overs Runs
- Last 5 Overs Wickets

The following results were obtained using these input parameters in the following machine learning models. Lasso regression model give the minimum Mean absolute error 18.63. So the Lasso regression is a good reward for this.

Table 11: Algorithms mean absolute error for model 02

<i>Model</i>	<i>Mean Absolute Error</i>
Linear Regression	25.62
Ridge Regression	20.23
Lasso Regression	18.63

Following density graph for distance between actual and predicted score. It is normally distributed. Because it is symmetric around 0, it is best to predict target scores.

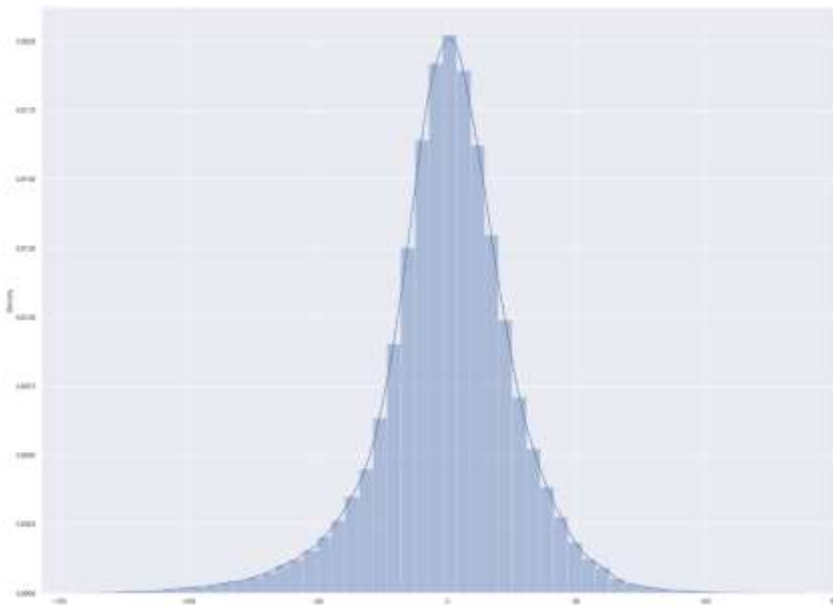


Figure 27: Density graph for mean absolute error

Model 03 (Hybrid winning team predictor)

This model is used to determine which team wins when the second inning is active. Furthermore, the third model is a hybrid model, for which the outputs of the first and second

models were used as input for this hybrid model. Accordingly, the heat map was obtained as follows.

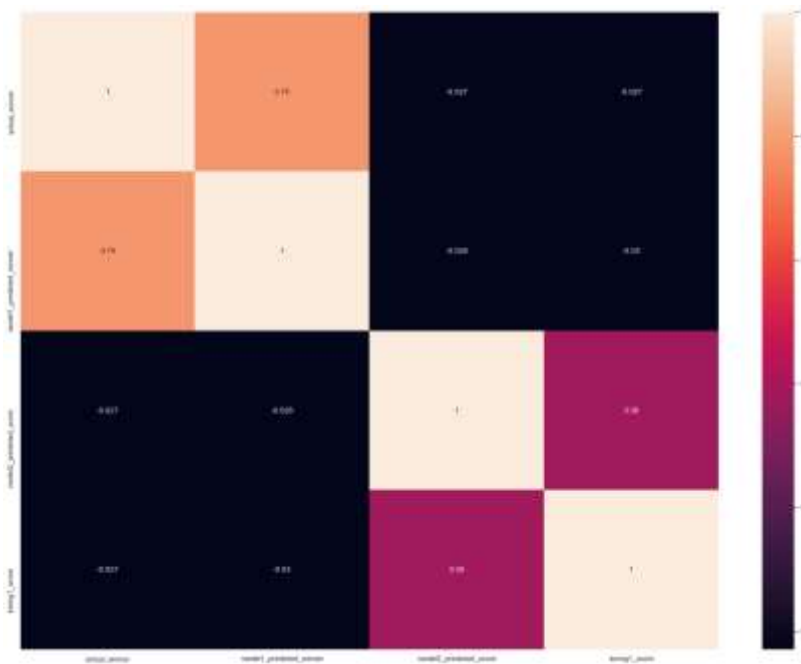


Figure 28: Heat map for model 03

According to the heat map these are the inputs,

- Model 1 output –winning team
- Model 2 output – Target score
- 1st Inning score

Accordingly, the following input parameters were tested using different algorithms and the following results were obtained. Accordingly, the decision tree showed very good high accuracy, that is 93.92 for balanced data set. In this case, too, the oversampling method was used to convert the unbalanced database into a balanced database.

Table 12: Accuracy for model 2 with different algorithms

<i>Model</i>	<i>Accuracy Imbalanced</i>	<i>Accuracy Balanced</i>
Random Forest	96.28%	91.86%
Decision Tree	97.23%	93.92%
Neural Network	80.11%	79.79%
MLP Classifier	84.83%	82.33%
SVM Classifier	78.58%	77.16%

Table 13: Feature importance for model 03

<i>Feature</i>	<i>Importance</i>
Model 1 Predicted Winner	72.90
Model 2 Predicted Score	2.78
Inning 1 score	24.32

This is the feature importance to the model 3. It shows the 72.90% highest importance of the model 1 prediction. However, this hybrid model gets a high accuracy compared to model 1 we developed.

According to the evaluation plan, I collected following mentioned data from the CRICSHEET website. This data is from 2021 May to 2021 July. There is no included all the match types. In future try to collect all this and rerun the evaluation for this. There is less matches due to the Covid 19 pandemic. These data also in YAML format. Using previous python scripts, I had to get the final appropriate data for our models.

Table 14: Data for evaluation (source: CRICSHEET)

<i>Match Type</i>	<i>Count</i>
T20 International	25
Non-official T20 International	10
Pakistan Super League	26

Model 1 and Model 3 gives the winning team of the match. In here I have the actual result of the game. So, I can try to evaluate my model with predicted value and the actual value. According to that, Model 1 gives 77.05% accuracy and the model 3 gives 92.56%. Model 2 evaluate according to the Duckworth Lewis method and current run rate method. In here also I have the final actual result. According to the model 2, it has 18.63 mean absolute error. I allow to that mean absolute error for the Duckworth Lewis and current run rate method and get the accuracy according to that.

Table 15: Basic winning team prediction model(Model 1) and Hybrid winning team prediction model (Model 3) evaluation summary

<i>Model</i>	<i>Number of tested occurrences</i>	<i>Correct prediction</i>	<i>Incorrect predictions</i>	<i>percentage</i>
Model 1	61	47	14	77.05%
Model 3	5000	4628	472	92.56%

The resulting hybrid model exhibits a higher degree of accuracy than the first model. Accordingly, the predictions made in certain innings in the second innings are more accurate than the predictions made at the start of a match.

So It's give 90.12% accuracy for Duckworth Luis method and 45.62% accuracy for current run rate method, then our model gives 92.58% accuracy. All the accuracies checked according to the mean absolute error 18.63. Based on the below calculations, the second model created in this research shows a higher accuracy value than the Duckworth Luis method and current run rate method methods currently in use.

Table 16: Batting score prediction model (Model 02) evaluation summary

<i>Type</i>	<i>Number of tested occurrences</i>	<i>Correct prediction</i>	<i>Incorrect predictions</i>	<i>percentage</i>
D/L	5000	4556	444	90.12%
Current run rate	5000	2281	2719	45.62%
Model 2	5000	4629	471	92.58%

Finally, I give this model to 25 people and get the feedback from them. The feedback was obtained after first explaining how the system works. This was followed by awareness and feedback through Zoom technology. In there most of them says need to develop this with more user friendly.

The third model has the potential to predict the winning team at a higher accuracy value than the first model created. Also, the second model predicts scores with higher accuracy than the Duckworth Luis method and current run rate method methods currently in use. There was also the need to create a user-friendly front end.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The final chapter focuses on providing an overview of the research and summarizing the final results that coincide with the research objectives introduced at the beginning of the research.

The goals of this research are predict the target final batting score in the 1st and 2nd innings, predict the target wicket count in the 1st and 2nd innings, predict the winning team and release one automated system to predict the winning team, target final score and target final wicket count. These objectives are achieved by three machine learning models that we develop in this research.

Twenty20 cricket matches are very popular nowadays. This is since nowadays people are spending more time and not being able to watch a match and they are focused on quenching their cricket thirst in less time. Therefore, there is a tendency to predict those matches as well as to bet by those predictions. Therefore, according to the literary review, a tool was needed to make such a prediction in less time. This report has intended on analyzing the results of the T20 match during the year 2005-2021 by applying the machine learning algorithms on both the balanced as well as imbalanced dataset.

The “CRICSHEET” web page was used to retrieve this data from 2005 up to now. Its data was in YAML format and was converted to row data using Python. Three models were created using that data. The first model uses basic data from the match to predict which team will win the match. Algorithms Random Forest, Decision Tree, Neural Network, MLP classifier and SVM classifier were used for this, and the Random Forest algorithm showed high accuracy. The B algorithm showed high accuracy results for the unbalanced dataset as well as the balanced database. 84.51% accuracy for the imbalanced data set and the 76.61% accuracy for the balanced dataset. Match Type, Team 1, Team 2, Ground, Toss Winner, Toss Winner Decision and Gender are parameter for this model. The oversampling method was used to convert the unbalanced data into equilibrium data.

The second model predicts the number of runs a team can score in an innings based on ball-by-ball data. The Linear Regression, Ridge Regression and Lasso Regression algorithms were used for this, and the Lasso Regression algorithm showed a minimum mean squared error. It's 18.63.

Match type, inning, batting team, bowling team, over, deliver, total runs for ball, runs for batsman, extra runs, is wicket, team current total, team current wickets, striker current runs, non-striker current runs, last 5 overs run and last 5 overs wickets are the parameters for this model.

The third model is a hybrid model that allows you to predict the winning team in the second inning. The same algorithm used for the first model was used for this as well. This gave the Decision Tree algorithm a highly accurate result. Here too the unbalanced dataset received an accuracy of 97.23% and the balanced dataset an accuracy of 93.92%. Model 1 output, Model 2 output, 1st inning total score are the parameters of this hybrid model. The outputs of the first and second models were used for the inputs of this third hybrid model.

The user application was created using the Angular technology and the Python technology so that a user could use the three models created.

The second chapter describes the technical aspects that can be applied to the literature survey and research objectives relevant to the domain. Uses design and methodology. The research is described in Chapter Three with an analysis of the exploratory data. Chapter Four outlines the evaluation of the created model.

FUTURE WORKS

- This is expected to extend to ODI and TEST matches.
- It hopes to create a model to calculate the number of tears available in an inning.
- Using the “CRICINFO” webpage, match data is retrieved during the competition, entered into the system, and fully automated predictions are made. Here the user does not need to enter data, only the relevant match selection.
- Create more user friendly GUI s’

APPENDICES

- URL for data set
https://drive.google.com/drive/u/0/folders/1snuoP6qJJgjqjaVJZ0DAWwF8j_4pcaD_
- URL for results
<https://drive.google.com/drive/u/0/folders/13pOdkSbli5xbrNBi5ElPqAaWuEjRuXNn>
- URL for Source codes
https://drive.google.com/drive/u/0/folders/1C4QyXHnabrrGeuxQy_jV7Db-z823YO6C

BIBLIOGRAPHY

- Awan, M.J., Gilani, S.A.H., Ramzan, H., Nobanee, H., Yasin, A., Zain, A.M., Javed, R., 2021. Cricket Match Analytics Using the Big Data Approach. *Electronics* 10, 2350. <https://doi.org/10.3390/electronics10192350>
- Basit, A., Alvi, M.B., Jaskani, F.H., Alvi, M., Memon, K.H., Shah, R.A., 2020. ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques, in: 2020 IEEE 23rd International Multitopic Conference (INMIC). Presented at the 2020 IEEE 23rd International Multitopic Conference (INMIC), IEEE, Bahawalpur, Pakistan, pp. 1–6. <https://doi.org/10.1109/INMIC50486.2020.9318077>
- Jyothsna, D., Srikanth, K., 2019. Analyzing and Predicting outcome of IPL Cricket Data 8, 6.
- Kalla, A., Karle, N., Wagle, S., Utala, S., n.d. AutoPlay - Cricket Score Predictor 2.
- Kampakis, S., Thomas, W., n.d. Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches 17.
- Lamsal, R., Choudhary, A., 2020. Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning. arXiv:1809.09813 [cs, stat].
- Naik, A., Pawar, S., Naik, M., Mulani, S., 2018. Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques 3, 8.
- Patil, N., Sequeira, B.H., Gonsalves, N.N., Singh, A.A., 2020. Cricket Team Prediction Using Machine Learning Techniques. *SSRN Journal*. <https://doi.org/10.2139/ssrn.3572740>
- Ramakrishnan, V., K, S., R, P., n.d. Target Score Prediction in the game of Cricket 7.
- SADP, S., 2018. ODI Cricket Match Winning Prediction Using Data Mining Techniques. University of Moratuwa.
- Shimona, S., Nivetha, S., Yuvarani, P., 2018. ANALYZING IPL MATCH RESULTS USING DATA MINING ALGORITHMS 9, 5.
- Singh, S., Kaur, P., 2017. IPL Visualization and Prediction Using HBase. *Procedia Computer Science* 122, 910–915. <https://doi.org/10.1016/j.procs.2017.11.454>
- Singhvi, A., Shenoy, A.V., Racha, S., Tunuguntla, S., n.d. Prediction of the outcome of a Twenty-20 Cricket Match 8.
- Sinha, A., 2020. Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020 (preprint). other. <https://doi.org/10.20944/preprints202010.0436.v1>
- Tekade, P., Markad, K., Amage, A., Natekar, B., 2020. CRICKET MATCH OUTCOME PREDICTION USING MACHINE LEARNING 5, 7.
- Viswanadha, S., Sivalenka, K., Jhawar, M.G., Pudi, V., n.d. Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths 10.

