

Marathwada Shikshan Prasarak Mandal's
Deogiri Institute of Engineering and Management Studies,
Aurangabad

Project Report
on

IPL Score Prediction

Submitted By

Abhishek Jaju (46035)

Vivek Soddy (46039)

Pradeep Maku (46034)

Dr. Babasaheb Ambedkar Technological University
Lonere (M.S.)



Department of Computer Science and Engineering
Deogiri Institute of Engineering and Management Studies,
Aurangabad
(2021- 2022)

Project Report
on
IPL Score Prediction

Submitted By

Abhishek Jaju (46035)

Vivek Soddy (46039)

Pradeep Maku (46034)

In partial fulfillment of
Bachelor of Technology
(Computer Science & Engineering)

Guided By

Mr. Vijay S. Kolte

Department of Computer Science & Engineering
Deogiri Institute of Engineering and Management Studies,
Aurangabad
(2021- 2022)

CERTIFICATE

This is to certify that, the Project entitled “**IPL Score Prediction**” submitted by **Abhishek Jaju, Vivek Soddy, Pradeep Maku** is a bonafide work completed under my supervision and guidance in partial fulfillment for award of Bachelor of Technology (Computer Science and Engineering) Degree of Dr. Babasaheb Ambedkar Technological University, Lonere.

Place: Aurangabad

Date: 23/06/2022

Prof. V.S. Kolte
Guide

Prof. S.B. Kalyankar
Head

Dr. Ulhas D. Shiurkar
Director,
Deogiri Institute of Engineering and Management Studies,
Aurangabad

DECLARATION

This is to certify that, the partial project report entitled, “**IPL Score predictions**” Submitted by **Abhishek Jaju, Vivek Soddy, Pradeep Maku** is a bonafide work completed under my supervision and guidance in partial fulfillment for award of bachelor’s degree in computer science and Engineering of Deogiri Institute of Engineering and Management Studies, Aurangabad under Dr. Babasaheb Ambedkar Technological University, Lonere.

Place: Aurangabad

Date: 23/06/2022

External Examiner

Mr. Vijay Kolte
Guide

Abstract

Cricket is a popular team sport played internationally. It has tremendous spectator support and the masses show great interest in predicting the outcome of games both in their one-day international as well as the modern T-20 format. The game is governed by complex rules and scoring system. Accurate prediction of winning or losing a match faces significant challenges. Multiple parameters, including cricketing skills and performances, match venues can significantly affect the outcome of a game. These diverse parameters, along with their interdependence and variance create a non-trivial challenge to create an accurate prediction of a game. In this paper, we build a prediction system that takes in historical match data, player performance as well as the scores predicted by spectator, and predicts future match events culminating in a victory or loss. Our system predicts match outcome by analyzing pre-stored match data using simple but effective algorithm. We describe our system and algorithms and finally present quantitative results, demonstrating the performance of our algorithms in predicting the number of runs scored, one of the most important determinants of match outcome. Keywords: Sports prediction, analysis. The cricket in the T-20 format is highly unpredictable - many features contribute to the result of a cricket match, and each attribute feature has a weighted impact on the outcome of a game. In this paper, first, a meaningful dataset through data mining was defined; next, essential features using various methods like feature engineering and Analytic Hierarchy Process were derived. Besides, a key issue on data symmetry and the inability of models to handle it was identified, which extends to all types of classification models that compare two or more classes using similar features for both the classes. This concept in the paper is termed as model ambiguity that occurs due to the model's asymmetric nature. Alongside, different machine learning classification algorithms like Naïve Bayes, SVM, kNearest Neighbor, Random Forest, Logistic Regression, ExtraTreesClassifier, XGBoost were adopted to train the models for predicting the winner. As per the investigation, tree-based classifiers provided better results with the derived model. The highest accuracy of 60.043% with Random Forest, with a standard deviation of 6.3% and an ambiguity of 1.4%, was observed.

Contents

List of Abbreviations	i
List of Figures	ii
List of Graphs	iii
List of Tables	iv
List of Screens	v
1. INTRODUCTION	1
1.1 Introduction	
1.2 Necessity	
1.3 Objectives	
1.4 Theme of the Project	
2. LITERATURE SURVEY	7
2.1 Sports and Sports analytics	
2.2 A Multivariate Regression approach	
2.2.1 Dataset	
3. SYSTEM DEVELOPMENT	10
3.1 Requirement Specification	
3.1.1 DFD (level 0,1,2)	
3.1.2 Specification Document/UML Diagrams of all modules	
3.2 User Interface Design	
3.3 Database Design / ER Diagrams	
4. Result Analysis	20
5. Conclusion	27
REFERENCES	
ACKNOWLEDGEMENT	

List of Abbreviation

Sr.No	Acronym	Abbreviation
1	IPL	Indian Premiere League
2	ODI	One Day International
3	DMP	Deep Mayo Predictor
4	BA	Batting Average
5	ICC	Indian Cricket Council

List of Figures

Figure	Illustration	Page
1	Figure 3.1	10
2	Figure 3.2	15
3	Figure 3.3	15
4	Figure 3.4	15
5	Figure 3.5	15

List of Graphs

Figure	Illustration	Page
1	Graph 4.1	10

List of Tables

Figure	Illustration	Page
1	Cross Validation Technique	25
2	IPL Team Names	29

List of Screens

Figure	Illustration	Page
1	Screen 3.1	15
2	Screen 3.2	16
3	Screen 3.3	17
4	Screen 3.4	10
5	Screen 4.1	25
6	Screen 4.2	26

1 Introduction

1.1 Introduction

England first introduced T20 Cricket in 2003. Because of its shorter format, it became very popular. Due to its popularity of high voltage action, T20 came to India also. BCCI initiated a 20-20 cricket tournament Indian Premier League (IPL) in 2008. BCCI has been organizing the IPL T20 cricket tournament every year. The use of analytical methods in various aspects of cricket including results prediction is very important. There is a huge demand for the algorithm that best predicts the result of cricket because of its popularity and huge amount of money involved in the game. Thus, the analysis of IPL results becomes more important. Prediction of outcome of a match using machine learning algorithms is an important aspect in cricket. Records of the past performance of players and other related data can be analyzed to create models that predicts the winning team. This model can be created using the machine learning algorithms and their results can be compared based on the Evaluation Measures as accuracy, precision, recall, sensitivity and error rate.

As of now, data analysis is need for every fields to examine the sets of data to extract the useful information from it and to draw conclusion and as well make decisions according to the information. Data is accessed by the computer programs developed using Machine learning to build models. The algorithm first analyses the data to create a model, specifically for understanding the patterns or trends. For creating the mining model, the model is optimized by selecting parameters and iterating. To extract actionable patterns and detailed statistics, the parameters are then fed into the dataset. This work focuses on finding the meaningful information about the IPL Teams by using the functions of R Package. R reduces the complexity of data analysis as it displays the analysis results in the form of visual representations. The dataset is loaded, and a set of pre-processing is done followed by feature selection. The best of the machine learning techniques is then applied to predict the winner and visualizes the results as graphs.

1.2 Necessity

Indian Premier League (IPL) is a professional cricket league based on Twenty20 format and is governed by Board of Control for Cricket in India. The league happens every year with participating teams name representing various cities of India. There are many countries active in organizing Twenty20 cricket leagues. While most of the leagues are being overhyped and team franchises are routinely losing money, IPL has stood out as an exception. As reported by espnricinfo, with Star Sports spending \$2.5 billion for exclusive broadcasting rights, the latest season of IPL (2018, 11th) saw 29% increment in the number of viewers including both the digital streaming media and television. The 10th season had 130 million people streaming the league through their digital devices and 410 million people watching directly on the TV. The numbers prove that IPL is a successful Twenty20 format-based cricket league.

The amount of data available in today's world because of technology advancements is seemingly unimaginable. Sports teams are able to use this available data to their advantage. When many people think of sports analysis, the movie "Moneyball" often times comes to mind, but the movie only shows a glimpse what it all entails. The sports industry uses sports analysis to increase revenue, improve player performance and a team's quality of play, prevent injury and for many more enhancements.

All this data is a great resource; however, it serves no use without people to interpret and analyze how it may be useful. Sports analysts are currently in high demand as many teams are developing entire departments just to analyze statistics, in order to become the best program in the league. In other words, sports teams are using analytics for a competitive advantage. As technology continues to progress, several new developments are expected to emerge in 2018. Three specific

advancements include integrating data sources to advance competition, communicate why the data is useful and create a different fan experience.

The data is beneficial to many in the industry including coaches, managers, agents, scouts, marketing professionals, medical personnel and the analytics staff. With the current available technology, sports analysts are able to take data and create insightful yet simple visualizations to communicate to other key decision makers of a team. Many teams use a program called SAS to manage and understand their data.

We watched a video addressing the effects the program had on an NBA team, the Orlando Magic. Because of SAS, the Orlando Magic is among the top revenue earners in the NBA, despite being in the 20th-largest market. Many teams are able to see just how much this effected the Magic and thus are encouraged to pay more attention to their analytics department, which may include investing in a resource like SAS.

Other data collecting resources include wearable technology which can be used for many reasons. One example of this is NFL players wearing devices in their helmets to receive data to help minimize injury. Wearable technology is becoming more and more popular among other teams and leagues to help with injury prevention and player performance. Increasing technology resources are encouraging more leagues to take a closer look at the best resources for analytics to have that competitive advantage teams are looking for.

Another aspect of data analytics in sports is using data to increase revenue and to enhance the fan experience. When ticket sales and attendance are down from the previous seasons, it is the sports analyst job to communicate the numbers and changes from previous seasons. The chart below shows a few MLB teams and the attendance numbers

from the 2001 to 2016 seasons. This is just one example of the data that is collected and analyzed to help increase fan experience and attendance from year to year.

I have read several articles that discussed ticket sales, and based on statistics, many sports analysts say that the most influential factor of ticket sales is the wins and losses record of a sport team. The chart below shows increases and decreases in attendance that change each season, with other available data analysts can determine the probable success each team had the prior season. Sports analysts are able to compare other forms of data with these numbers to determine the biggest cause and further relay the information to sports marketers or other professionals involved in ticket sales and fan experience.

Baseball - Major League Regular Season Attendance (add 000) (Games in parentheses)

Item	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
American League:																
Toronto Blue Jays	1,915	1,637	1,800	1,900 (81)	1,978 (80)	2,302 (81)	2,361 (81)	2,400 (81)	1,876 (81)	1,626 (81)	1,818 (81)	2,100 (81)	2,537 (81)	2,376 (81)	2,795 (81)	3,392 (81)
New York Yankees	3,265	3,462	3,466	3,775 (79)	4,090 (81)	4,201 (81)	4,272 (81)	4,299 (81)	3,719 (81)	3,766 (81)	3,654 (81)	3,542 (81)	3,280 (81)	3,402 (80)	3,194 (80)	3,063 (81)
Los Angeles Angels	2,001	2,306	3,061	3,376 (81)	3,405 (81)	3,407 (81)	3,366 (81)	3,337 (81)	3,240 (81)	3,251 (81)	3,166 (81)	3,062 (81)	3,020 (81)	3,096 (81)	3,013 (81)	3,016 (81)
Boston Red Sox	2,625	2,650	2,724	2,837 (81)	2,813 (80)	2,931 (81)	2,971 (81)	3,048 (81)	3,063 (81)	3,046 (81)	3,054 (81)	3,043 (81)	2,833 (81)	2,956 (81)	2,881 (81)	2,955 (81)

Figure 1.1

Communicating data efficiently is what sports analytics all comes down to. Without people to analyze and interpret these numbers, they have no meaning to other professionals in the industry. Anyone can create a statistic, but if they are not able to explain the meaning behind how it can help improve the team, then the statistic is useless. Sports analytics are crucial to many teams by helping them become their best through interpretation and analysis of statistics gained in practices and games. As technology and resources are progressing for data collection, sports analytics is a growing field as teams are looking to have a competitive advantage against their opponents. In the technology savvy world, we live in, it only makes sense to use data and sports analytics as an advantage in taking sports teams to a new and improved level.

1.3 Objective

Cricket is one of the most popular outdoor sports that has captured everyone's heart. There are many series that are held, and the Indian Premier League (IPL) is one of them which has a long and illustrious tradition in the sports world. IPL is a professional Twenty20 (T20) league started in 2008 which was founded by the Board of Control for Cricket in India (BCCI). The IPL is a 20-over league, which means each team plays 20 overs from both sides. Every year, eight teams from eight Indian cities participate in this league. A cricket match is influenced by a variety of factors, and the factors that have a major impact on the outcome of a T20 cricket match are described in this project. **IPL Score Prediction** project takes several years of IPL data, including player information, match location information, team information, and ball to ball information, and analyzes it to draw different conclusions that help in the enhancement of player's results. It focuses on calculating the results of IPL matches using data mining techniques on both balanced and imbalanced datasets. In T20 Cricket matches, the first innings score is currently estimated based on the existing run rate, which is measured as the number of runs scored per a number of overs bowled.

1.4 Theme of the Project

Various machine learning algorithms have been applied and tested for their efficiency in solving the problems in sports. The relation between machine learning and games dates to the initial days of artificial intelligence when Arthur Samuel, a pioneer in the field of gaming and artificial intelligence studied machine learning approaches using the game of checkers.

A study was performed to predict the outcome (win, lose or draw) of football matches played by a professional English Premier League (EPL) team, Tottenham Hotspurs, based on matches that were played in the year between 1995 to 1997. It was observed that Bayesian networks relatively outperformed other machine learning algorithms which included MC4 - a decision tree learner, Naive Bayesian learner, Data-driven Bayesian, and K-nearest neighbor. The prediction accuracy of the Bayesian nets model was 59.21%. Match outcome prediction and game-play analysis are a prevalent problem that is tackled using machine learning. Another area where machine learning approaches are being used is extracting highlights from an on-going match. A study was performed to extract baseball match highlights on a set-top device. The relative strength of classification algorithms, namely Support Vector Machine (SVM), Gaussian Fitting (GAU) and K-Nearest Neighbors (KNN) was considered for "excited speech" classification, and finally, SVM was applied. Six baseball matches covering 7 hours of game-play time was fed to the algorithm. 75% of the highlights extracted by the algorithm were common with the highlights extracted manually by a human.

Just like in football, supervised machine learning algorithms have also been used in predicting the outcome of baseball matches. A project [27] used two learning methods, i.e., logistic classification and Artificial Neural Network (ANN) to predict the result of the baseball post-season series. Although ANN came up with very poor accuracies, the accuracies out of the logistic model were satisfactory with training and test accuracies of 73.6% and 62.6% respectively. Another project applied four machine learning algorithms to understand career progression in Baseball [28].

The implemented algorithms were Linear Regression (Ridge Model), Multi-Layer Perceptron Regression (Neural Network), Random Forests Regression (Tree Bagging Model), Support Vector Regression (SVR). The dataset which was used to train these algorithms contained match data of the first six seasons of players' career. And the players' value was predicted. The prediction was near 60% for the batters, while for pitchers the accuracy was very poor, i.e., something around 30-40%.

2 Literature Survey

2.1 Sports and Sports analytics

With technology growing more and more advanced in the last few years, an in-depth acquisition of data has become relatively easy. As a result, Machine Learning is becoming quite a trend in sports analytics because of the availability of live as well as historical data. Sports analytics is the process of collecting past matches data and analyzing them to extract the essential knowledge out of it, with a hope that it facilitates in effective decision making. Decision making may be anything including which player to buy during an auction, which player to set on the field for tomorrow's match, or something more strategic task like, building the tactics for forthcoming matches based on players' previous performances.

Machine Learning can be used effectively over various occasions in sports, both on-the-field and off-the-field. When it is about on-the-field, machine learning applies to the analysis of a player's fitness level, design of offensive tactics, or decide shot selection. It is also used in predicting the performance of a player or a team, or the outcome of a match. On the other hand, the off-the-field scenario concerns the business perspective of the sport, which includes understanding sales pattern (tickets, merchandise) and assigning prices accordingly. The main focus is the healthy growth in business and profitability of the team owners and other stakeholders. On-the-field analytics generally make use of supervised machine learning algorithms, example: (i) regression for calculating the fitness of a player, (ii) classification for predicting an outcome of a match; while off-the-field analytics concerns around performing sentiment analysis to understand people's opinion about a player or a team or a sport league. At present, Twitter has become one of the primary sources of data for sentiment analysis.

Sport Lisboa e Benfica, one of Portugal's most successful football clubs advancing in the use of data modeling techniques while making decisions is one real-world example of the use of machine learning in sports science. The club monitors and analyzes almost every aspect of a player, including their sleeping, eating, training habits. Once raw player data is recorded, various models

are designed to analyze the data for optimizing match readiness and defining personalized practice schedules. With the application of machine learning and predictive analysis, the facts coming out of the devised models enable players to improve their performance continually. On the other cards, with those facts at hand, manager/coach gets a better idea about which player to be replaced, which player to be kept in the playing list and which player to be kept in the bench.

Major League Baseball (MLB) has seen enormous growth in the arena of sports analytics in the last few years. Professional MLB teams collect tremendous amount of ball-by-ball data and apply various machine learning approaches to get clear insights into the game, which is usually not visible through human analysis. Predicting the outcome of a match, classifying if a team will intentionally make a player walk at bat or classifying non-fastball pitches according to pitch type, etc. are some of the classification problems dealt using machine learning in baseball world

Similarly, cricket has also been making use of sports analytics to perform prediction of outcome of a match, while the gameplay is in progress or before the match has even begun. Even problem like predicting runs or wickets of a player for a match, based on his/her past performance is an interesting problem to work on. Some real-world tools which have been implemented in cricket include WASP (Winning and Score Predictor) a tool which predicts a score and possible outcome of a limited over cricket match, i.e., One-day or Twenty20. Sky Sports New Zealand first introduced this tool in 2012 during an ongoing Twenty20 match. Technology like Hawk-Eye which tracks the trajectory of a ball and visually displays the most statistically significant path, has also been officially in use as the Umpire Decision Review System since 2009. Similarly, other sports like tennis, badminton, snooker also make use of this computer-assisted intelligent technology.

In cricket, to predict an outcome of a match, the primary task is to extract out the essentials factors (features) which affect result of a match. Interesting works have been done in the field of predicting outcome in cricket. Some interesting machine learning works have been performed on data acquired from Indian Premier League matches. In a study, the Naive Bayesian classifier was used to classify the performance of all-rounder players (bowler plus batsman) into four various non-overlapping categories, viz., a performer, a batting all-rounder, a bowling all-rounder or an underperformer by being based on their strike rate and economy rate. Step-wise multinomial logistic regression (SMLR) was used to extract the essential predictors. When validated, the Naive Bayesian model was able to classify 66.7% of the all-rounders correctly.

The same authors later published a work in which an Artificial Neural Network model was used to predict the performance of bowlers based on their performance in the first three seasons of IPL. When the predicted results were validated with actual performance of the players in season four, the developed ANN model had an accuracy of 71.43%. Although not related to IPL, a study performed at University College London in the area of predicting the outcome of a Twenty20 match would be a healthy addition as the literature work in the Twenty20 domain. The study made use of Naive Bayes, Logistic Regression, Random Forests, Gradient Boosting algorithms to predict the outcome of English County cricket matches.

Two models were developed, each was given input of two different sets of features. The team only related features were input to the first model, while team and players related features were input to the second model. The study was concluded with Naive Bayes outperforming all other algorithms with the first model giving out average prediction accuracy of 62.4% and second model giving average prediction accuracy of 63.9%, i.e., 64% average accuracy with 2009-2014 data and 63.8% average accuracy with 2010-2014 data.

The work done on Data Mining of Cricket dataset describes the various data mining techniques applied on the IPL dataset, the model is built for predicting the results of the matches. The selection of the best team is always required by the management for best outcome. The paper provides the optimal solution to select the best team using Data Mining Techniques rather than following the traditional method which is tedious. When we are declaring a time for the championship it is mandatory to select the best team and so the chance of the team to be the champion becomes easy.

In previous work the authors propose the fuzzy clustering logic. The results of the IPL batting Statistics were grouped into various clusters and it gave efficient and effective accurate results with the Data Mining Technique – Clustering. This work has been done with the help of MATLAB. The concept of clustering is used in order to classify batting statistics of the Indian Premier League which has the fuzzy data into appropriate clusters. Raza Ul Mustafa et al presented a study on the investigation of the feasibility of using the Twitter data to forecast the results of the match. The work has been proposed to check the machine learning techniques' effectiveness when applied on data collected to derive insight obtained from social media networks and other real-world events are predicted. The techniques used in their work are Support Vector Machine, Naive Bayes Classifier and the Linear Regression. The SVM technique holds good.

Live Cricket Score and Winning Prediction work describes about the building of the model which predicts the score for the chasing team and will estimate the score of the second innings of match. The proposed work uses the concepts of Linear Regression, Naive Bayes Classifier and Reinforce Learning Algorithm and gives the idea about building a system of prediction that takes the historical data and predicts the victory or loss of the forthcoming matches.

The work done on Data Mining of Cricket dataset describes the various data mining techniques viz Decision Tree, Naive Bayes, KNN, Random Forest applied on the IPL dataset, the model is built for predicting the results of the matches. The best attributes were selected using the Wrapper and Ranker method and then the classification has been done.

This work was done with the help of WEKA. Gupta et al. says that the selection of the best team is always required by the management for best outcome. The paper provides the optimal solution to select the best team using Data Mining Techniques rather than following the traditional method which is tedious. When we are declaring a time for the particular championship it is mandatory to select the best team and so the chance of the team to be the champion becomes easy. In, the authors proposes the fuzzy clustering logic. The results of the IPL batting Statistics were grouped into various clusters and it gave efficient and effective accurate results with the Data Mining Technique – Clustering. This work has been done with the help of MATLAB. The concept of clustering is used in order to classify batting statistics of the Indian Premier League which has the fuzzy data into appropriate clusters.

Raza Ul Mustafa et al presented a study on the investigation of the feasibility of using the Twitter data to forecast the results of the match. The work has been proposed to check the machine learning techniques' effectiveness when applied on data collected to derive insight obtained from social media networks and other real-world events are predicted. The techniques used in their work are Support Vector Machine, Naive Bayes Classifier and the Linear Regression. The SVM technique holds good. Live Cricket Score and Winning Prediction work describes about the building of the model which predicts the score for the chasing team and will estimate the score of the second innings of match. The proposed work uses the concepts of Linear Regression, Naive Bayes Classifier and Reinforce Learning Algorithm. The factors such as toss result, ranking of the team, home team advantage was considered. Sankaranarayanan gives the idea about building a system of prediction that takes the historical data and predicts the victory or loss of the forthcoming matches.

They used Linear Regression, Nearest Neighboring and clustering methods will present the mathematical results will exhibit the performance of the algorithm for predicting the results of the model. Parag Shah, in his work of predicting outcome of the live match proposed model which predict the match result after each ball. The par score concept has been used Duckworth & Lewis, the probability is calculated, and it provides clarity of who will win the match

Kaluarachchi used artificial intelligent technique specifically Bayes classifiers in machine learning to classify the factors as home game advantage, day/night effect the toss and batting first that affects the result of the match. The outcome of this work is the delivered as software tool called CricAI. The tool gives the probability of winning based on the input factors such as home game advantage is available at the beginning of the match. The CricAI can be used in real-world applications when teams are playing cricket. It is used for modifying certain factors to increase the chances of winning in the real field. The models use Data Analytics methods from machine learning domain. In a rain affected match the prediction of the result may be difficult .In a rain affected match Batting, Bowling, Fielding, Team Selection, Result Prediction, Target Revision is very important. The match prediction can be solved using the mathematical model created based on the insights of the results of prior matches. The Predictor models are created with the help of SVM. This work is proposed using Deep Mayo Predictor

The software developed based on the work by Jayshree Hajgude for Statistical Analysis and Data Mining forms a dream team with the Bayesian Prediction Technique and Parameter based filtering. The database has the details of the current IPL players. This work helps to mine the needed data for using by prediction algorithm in order to obtain the statistical analysis of each player. Predictive model is developed to predict the cricket score and player performance using ODI dataset . Performed Supervised methods like SVM, Naïve Bayes. Clustering methods like KNN and MLP method to classify accurately.

The work done by Kaluarachchi used artificial intelligent technique specifically bayes classifiers in machine learning to classify the factors as home game advantage, day/night effect the toss and batting first that affects the result of the match. The final outcome of this work is the delivered as software tool called CricAI. The tool gives the probability of winning based on the input factors such as home game advantage is available at the beginning of the match. The CricAI can be used in real-world applications when teams are playing cricket. It is used for modifying certain factors to increase the chances of winning in the real field

The models use data analytics methods from machine learning domain. In a rain affected match the prediction of the result may be difficult. In a rain affected match batting, bowling, fielding, team selection, result prediction, target revision is very important. The match prediction can be solved using the mathematical model created based on the insights of the results of prior matches. The Predictor models are created with the help of SVM. This work has been proposed using Deep Mayo Predictor .

The software developed based on the work by Hajgude for Statistical Analysis and Data Mining forms a dream team with the Bayesian Prediction Technique and Parameter based filtering. The database has the details of the current IPL players. This work helps to mine the needed data for using by prediction algorithm in order to obtain the statistical analysis of each player. Predictive model is developed to predict the cricket score and player performance using ODI dataset Performed Supervised methods like SVM, Naïve Bayes. Clustering methods like KNN and MLP method to classify accurately. IPL match winner prediction is carried out using ML techniques . IPL is interesting game each year there is lot of expectations of who will win the prestigious title. IPL is game where the result can be changed in just few seconds so to predict the winner ML algorithms like SVM, Logistic Regression, Naïve Bayes, Decision Tree and KNN. To overcome these existing modules, we developed a predictive technique for predicting the winning team.

2.2 A Multivariate Regression approach

The literature survey concluded that there was a need for a machine learning model which could predict the outcome of an IPL match before the game begins. Among all formats of cricket, Twenty20 format sees a lot of turnarounds in the momentum of the game. An over can completely change a game. Hence, predicting an outcome for a Twenty20 game is quite a challenging task. Besides, developing a prediction model for a league which is wholly based on auction is another hurdle. IPL matches cannot be predicted simply by making use of statistics over historical data solely. Because of players going under auctions, the players are bound to change their teams; which is why the ongoing performance of every player must be taken into consideration while developing a prediction model.

In sports, most of the prediction job is done using regression or classification tasks, both of which come under supervised learning. In simple terms, $y=f(x)$ is a prediction model which is learned by the learning algorithm from a set of dataset: $D = ((X_1, y_1), (X_2, y_2), (X_3, y_3), \dots, (X_n, y_n))$. Based on the type of output (y) supervised learning is divided further into two categories, viz., regression, and classification. In Regression, the output is a continuous value; however, classification deals with discrete kind of output. For predicting continuous values, Linear Regression appeared to be quite effective, and for classification problems like predicting the outcome of matches or classifying players, learning algorithms like Naive Bayes, Logistic Regression, Neural Networks, Random Forests were found being used in most of the previous studies

In this work, the various factors that affect the outcome of a cricket match were analyzed, and it was observed that home team, away team, venue, toss winner, toss decision, home team weight, away team weight, influence the win probability of a team. The proposed prediction model makes

use of multivariate Regression to calculate points of each player in the league and compute the overall strength of each team based on the past performance of the players who have appeared most for the team

2.2.1 Dataset

The official website of Indian Premier League was the primary source of data for this study. The data was scraped from the site and maintained in a Comma Separated Values (CSV) format. The initial dataset had many features including date, season, home team, away team, toss winner, man of the match, venue, umpires, referee, home team score, away team score, power play score, overs details when team reached milestone of multiple of 50 (i.e., 50 runs, 100 runs, 150runs), playing 11 players, winner and won by details. In a single season, a team has to play with other teams in two occasions, i.e., once as a home team and next time as an away team. For example, once KKR plays with CSK in its home stadium (Eden Gardens) next time they play against CSK in their home stadium (M Chinnaswamy Stadium). So, while making the dataset, the concept of home team and away team was considered to prevent the redundancy.

Indian Premier League has just been 11 years old, which is why only 634 matches data were available after the pre-processing. This number is considerably less with comparison to the data available relating to the test or ODI formats. Due to certain difficulties with some ongoing team franchises, in some seasons the league has seen the participation of new teams, and some teams have discontinued. Presence of those inactive teams in the dataset was not really necessary, but if the matches data were omitted where the inactive teams appeared, the chances were that the valuable knowledge about the teams which were still active in the league would deteriorate. For better understanding and to make the dataset look somehow cluttered-free, acronyms were used for the teams. Table 1 lists the acronyms used in the dataset.

There are various ways a player can be awarded points for their performance in the field. The official website of IPL has a Player Points section where 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, and (vi) number of stumpings. To find out how IPL management was assigning points to each player based on these 6 features, a multivariate regression was used on the players' points data. Freedman [36] has beautifully explained the mathematics behind the Regression models.

For a team, there can be as many as 25 players. This is a limit put on by IPL governing council to the franchises. To find the average strength of a team, every player of the team is first sorted in the descending order according to their number of appearances in previous matches of the same season. Once players have been sorted, the top 11 players are considered for calculating the weight of the team because these players have played more games for the team and their performance influence the overall team strength. Figuring team weight for all 634 matches was a tedious task. So, for example purpose, the final results of each season were considered, and the team weight for each team was calculated accordingly, and the same score was used for all the matches in that particular season. For better performance of the classifier, the team weight must be calculated immediately after the end of each match. This way, the real-time performance of each team and the newly computed weight can be used in predicting upcoming games.

In this study, Recursive Feature Elimination (RFE) algorithm was used as a feature selection method. As the name suggests, RFE recursively removes

an unessential feature from a set of features, re-builds the model using the remaining features and recalculates the accuracy of the model. The process goes on for all the features in the dataset. Once completed, RFE comes up with top k number of features which influence the target variable (independent variable) at a level of extent. Sometimes, ranking the features and using the top k features for building a model might result in wrong conclusions to prevent this from happening, the dataset was resampled, and RFE was operated in the subsets. The results were the same set of features obtained initially; hence, the initial set of features obtained from RFE did not seem to be biased. Using the RFE model, the number of features was reduced to . Thus, obtained features which highly influenced the target variable were the home team, the away team, the venue, the toss winner, toss decision, and the respective teams' weight.

3 SYSTEM DEVELOPMENT

3.1 Requirement Specification

System Features and Requirements

Functional Requirements

- It should provide schedule without any of clashes among IPL teams, day, time and stadiums that must be visible to all.
- It should generate a report about the registered complaint to the admin and response report to the user who has submitted his queries.
- Secure registration and profile management facilities for different users.
- It should provide detailed guide on software installation and using procedure to users.
- It should generate and notify alerts via SMS and E-mail.

Non- Functional Requirements

1) Safety Requirements

If there is extensive damage to a wide portion of the database due to catastrophic failure, such as a disk crash, the recovery method restores a past copy of the database that was backed up to archival storage (typically tape) and reconstructs a more current state by reapplying or redoing the operations of committed transactions from the backed up log, up to the time of failure.

2) Security Requirements

Security systems need database storage just like many other applications. However, the special requirements of the security market mean that vendors must choose their database partner carefully.

3) Software Quality Attributes

- **AVAILABILITY:** Since we are hosting our project on the server it will be available all the time.
- **CORRECTNESS:** The system should generate an appropriate report about different activities of the IPL season and should keep track of all records.
- **MAINTAINABILITY:** The system should maintain correct schedules of date, match and timings of all matches.
- **USABILITY:** The system should satisfy the maximum number of user's needs.

Software Requirements:

- Flask==1.1.1
- gunicorn==19.9.0
- itsdangerous==1.1.0
- Jinja2==2.10.1
- MarkupSafe==1.1.1
- Werkzeug==0.15.5
- numpy>=1.9.2
- scipy>=0.15.1
- scikit-learn>=0.18
- matplotlib>=1.4.3
- pandas>=0.19

Hardware Requirement

- 1GHz+ CPU.
- 512MB RAM.
- 20MB database space.
- 300MB disk space

3.1.1 DFD Diagram

Data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination. Data flowcharts can range from simple, even hand-drawn process overviews, to in-depth, multi-level DFDs that dig progressively deeper into how the data is handled. They can be used to analyze an existing system or model a new one. Like all the best diagrams and charts, a DFD can often visually “say” things that would be hard to explain in words, and they work for both technical and nontechnical audiences, from developer to CEO. That’s why DFDs remain so popular after all these years. This DFD diagram is an overview of the process, data store and data flow in the project.

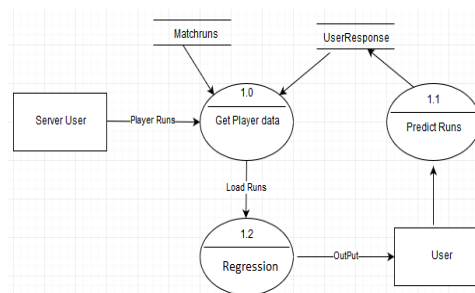


Figure 3.1 DFD Diagram

3.1.2 Specification Document / Uml diagram of all modules

The **Unified Modeling Language (UML)** is a language used in the field of software engineering that represent the components of the Object-Oriented Programming concepts. It is the general way to define the whole software architecture or structure. In Object-Oriented Programming, we solve and interact with complex algorithms by considering themselves as objects or entities. These objects can be anything. It can be the bank or a bank manager too. The object can be a vehicle, animal, machine, etc. The thing is how we interact and manipulate them that they can perform tasks and they should. The tasks can be interacting with other objects, transferring data from one object to another, manipulating other objects, etc.

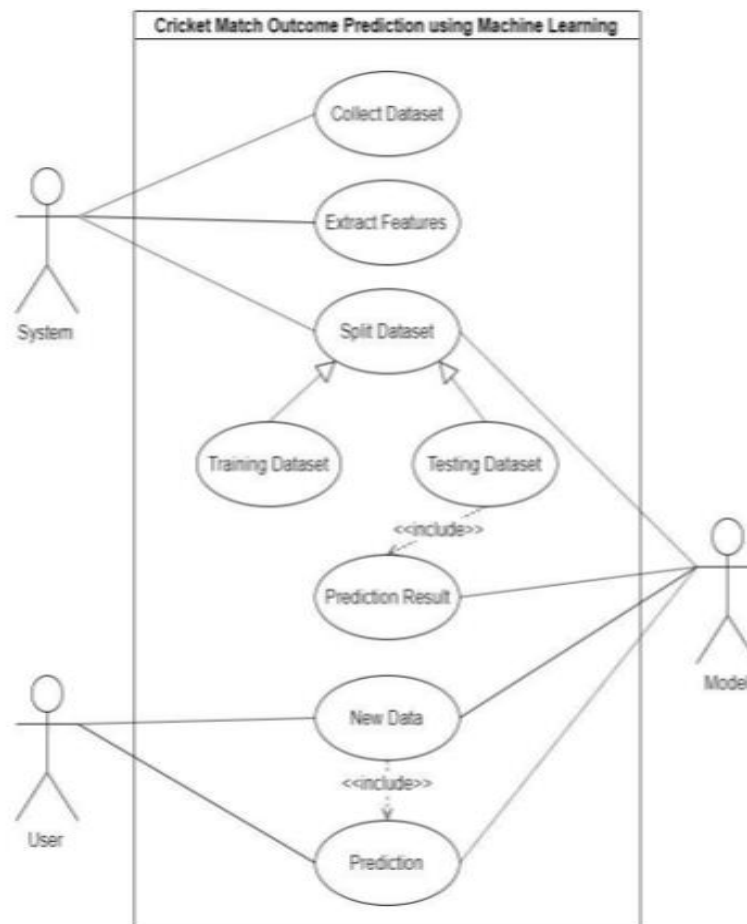
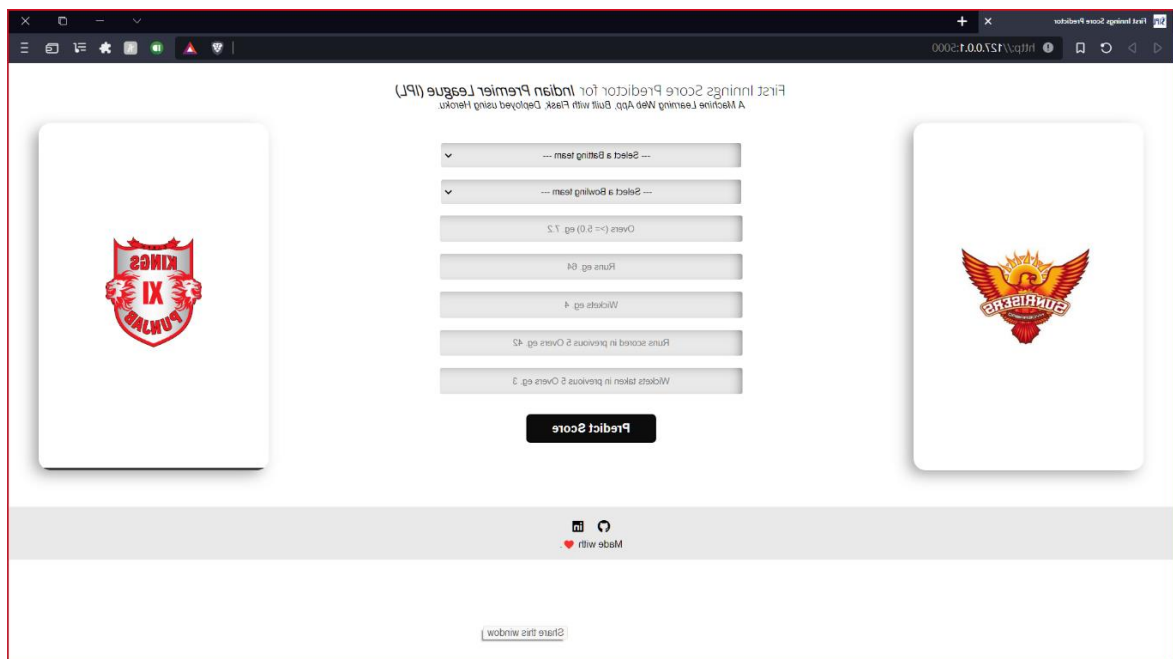


Figure 3.2 Use case diagram

The single software could have hundreds or even thousands of objects. So, **UML** provides us a way to represent and track those objects in a diagram to become a blueprint of our software architecture. A Use Case Diagram can be regarded as a good starting point for discussing project key actors and processes without going into too many implementation details. This UML diagrams is also the most popular type of the Behavioral UML diagram category, and is used to analyze the functionality (the use cases) and the interactions with different types of agents (actors) of a system. This UML diagram is the blueprint of the software architecture .

3.2 User Interface Design

The Graphical User Interface is developed for the machine learning models using the Flask Framework. For the backend of the site Python is used. The site can be used to predict the IPL match score with the help of last 5 overs of the data. We can also predict the Winner of the match with the data of just Toss Winner and Toss Decision. All the input information necessary for the model for the prediction is provided to the model. The calculation is not stored in the system because all calculations computed at real time. We implemented it that way as we can add change more attribute to the system with minor changes to the program. A. Score Prediction. The GUI required at least 5 overs of the data to predict the score as shown in the Fig. Model require the input data of Batting team, Bowling team, Over, Runs, Wickets, Run Scored in last 5 overs, Wickets fall in last 5 overs to predict the score of the match as shown in the Fig



The screenshot shows a web browser window displaying the 'IPL Score Predictor' application. The interface is clean and modern, with a white background and a dark header. On the left and right sides, there are large, rounded rectangular boxes containing the logos of the Mumbai Indians (left) and Sunrisers Hyderabad (right). In the center, there is a form with several input fields and a 'Predict Score' button. The form fields are labeled as follows:

- Select a Batting team (dropdown menu)
- Select a Bowling team (dropdown menu)
- Over (1-5) of 1.5
- Runs of 10
- Wickets of 4
- Runs scored in previous 5 Overs of 45
- Wickets taken in previous 5 Overs of 3

Below the form fields is a prominent black button with the text 'Predict Score' in white. At the bottom of the page, there is a small footer area with a 'Made with' logo and a 'Share this window' link.

Screen 3.1

First Innings Score Predictor for *Indian Premier League (IPL)*
A Machine Learning Web App. Built with Flask. Deployed using Heroku.

--- Select a Batting team ---
--- Select a Bowling team ---
Overs (>= 5.0) eg. 7.2
Runs eg. 64
Wickets eg. 4
Runs scored in previous 5 Overs eg. 42
Wickets taken in previous 5 Overs eg. 3

Predict Score

CRICKET MERI JAAN

WHISTLE BOO

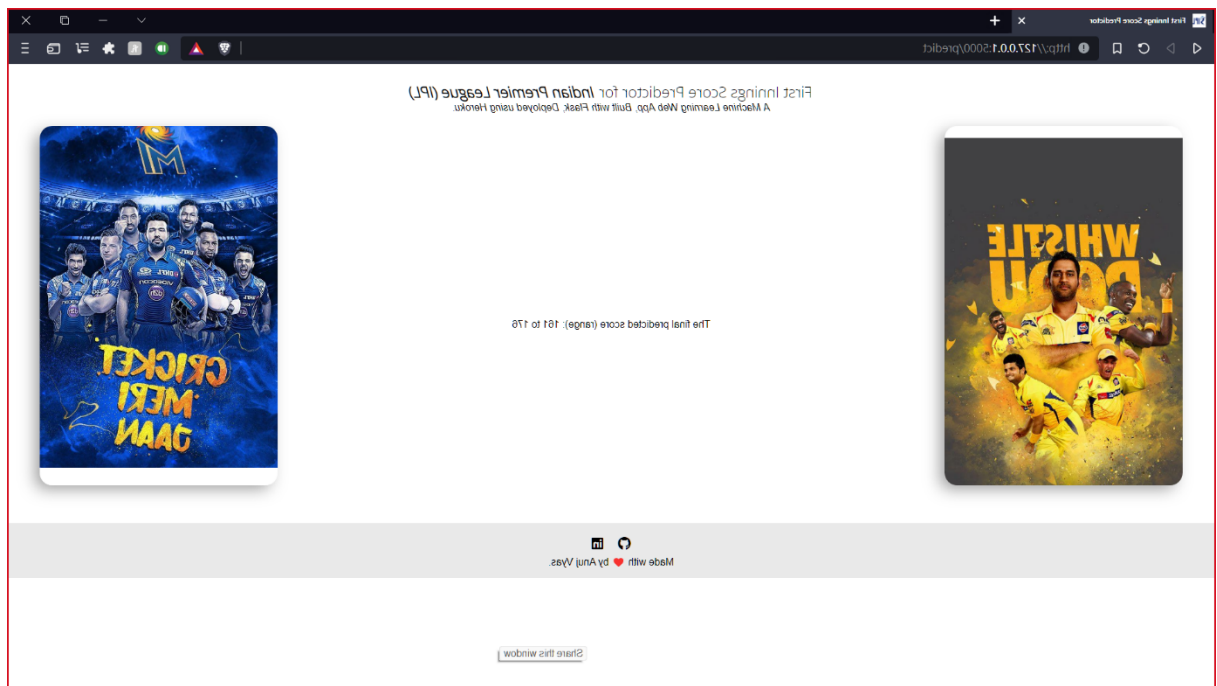
Made with ❤️

[Share this window](#)

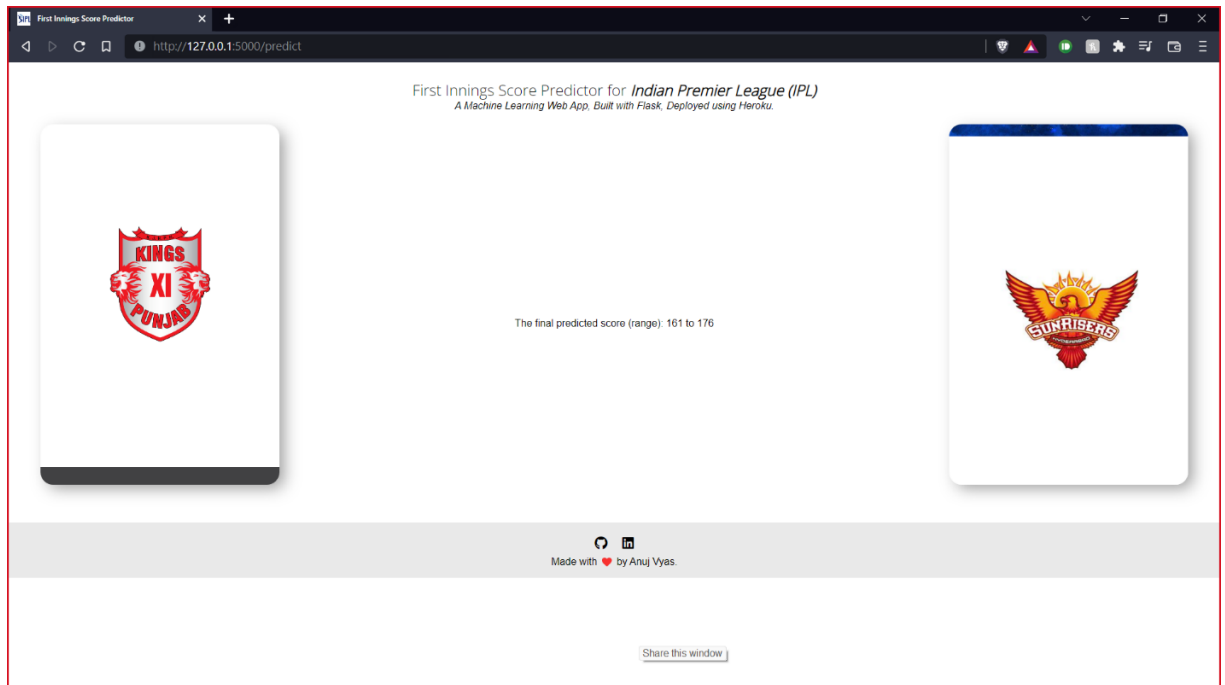
Screen 3.2

In both above screenshots(screen 3.1 and screen 3.2) we are entering the required information required for prediction of the scores. You have to give input that is Batting team and Bowling team

You have to enter the overs that have been bowled, current runs and then runs scored in the last five overs of the batting team and wickets taken in the last five overs.



Screen 3.3



Screen 3.4

In this screenshots (screen 3.3 and screen 3.4) It is showing the final score it has predicted from the given user input.

3.3 Database Design / ER Diagrams

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research. The ER diagram is the pictorial representation of the objects and their relationships. The ER diagram is the subclass of the UML diagram. It is used to design and implement the databases. They use a defined set of symbols such as rectangles, diamonds, ovals and connecting lines to depict the interconnectedness of entities, relationships and their attributes. They mirror grammatical structure, with entities as nouns and relationships as verbs. This E-R diagram is the representation of various entities and relationships and their attributes.

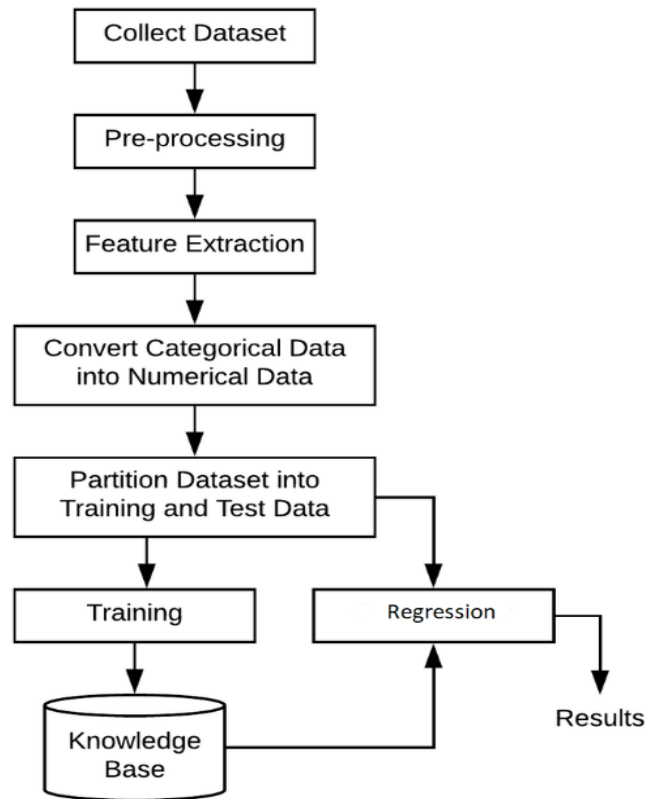


Figure 3.3

3.4 Process Diagram

Process diagrams, called “flow diagrams” by TOGAF, are used to model the sequence of activities within a process. Flow diagrams represent process participants, activity sequences, information exchanged during a process, and trigger events. A process diagram consists of activities, events, and gateways, which a sequence flow puts in a *flow sequence*. The following process diagram describes the overall processes includes in this architecture.

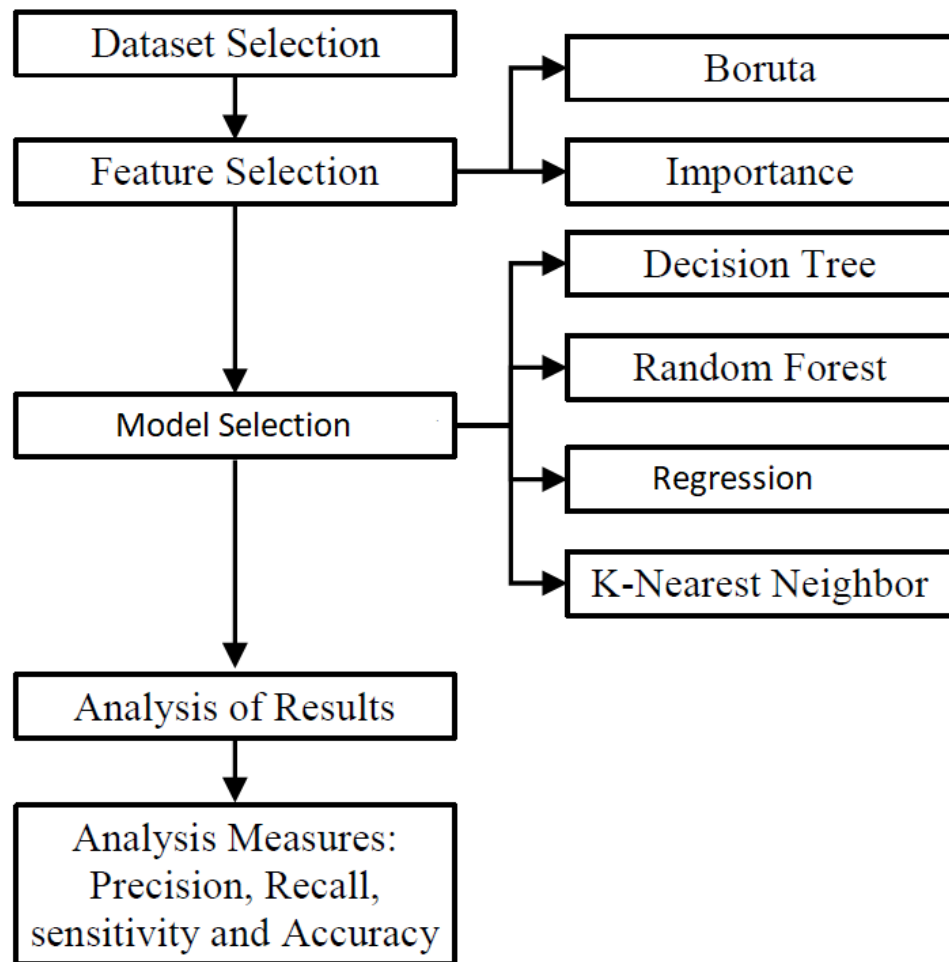


Figure 3.4

4. Result Analysis

IPL Analysis:

- I. From the analysis most IPL matches took place in Eden Garden and Wankhede stadium and it is also an important factor for winning the match.
- II. Mumbai Indians are the most successful group who has won 109 games out of 11 times (2008-2020). Chennai Super Kings is the second most successful group in the 100 wins, followed by the Kolkata Night Riders with 92 wins. Mumbai Indians won the most cups.
- III. Most of the time the man of the match of IPL game goes to Chris Gayle who owns the top number followed by AB de Villiers and David Warner and Indians MS Dhoni is top on the list.
- IV. In this toss win or loss has least effect to win a match. After winning the toss, team wins the match is higher than when team win the toss but did not manage to win the match but however the decision to balling or batting first play a major role.
- V. In a different models of machine learning, we find that Linear Regression, Random Forest has a very accurate model at a rate of 88% accuracy.

Analysis is a type of statistical evaluation that enables three things:

Description: Relationships among the dependent variables and the independent variables can be statistically described by means of regression analysis.

Estimation: The values of the dependent variables can be estimated from the observed values of the independent variables.

Prognostication: Risk factors that influence the outcome can be identified, and individual prognoses can be determined.

Regression analysis employs a model that describes the relationships between the dependent variables and the independent variables in a simplified mathematical form. There may be biological reasons to expect a priori that a certain type of mathematical function will best describe such a relationship, or simple assumptions have to be made that this is the case (e.g., that blood pressure rises linearly with age). The best-known types of regression analysis are the following :-

Univariable linear regression studies the linear relationship between the dependent variable Y and a single independent variable X . The linear regression model describes the dependent variable with a straight line that is defined by the equation $Y = a + b \times X$, where a is the y-intersect of the line, and b is its slope. First, the parameters a and b of the regression line are estimated from the values of the dependent variable Y and the independent variable X with the aid of statistical methods. The regression line enables one to predict the value of the dependent variable Y from that of the independent variable X . Thus, for example, after a linear regression has been performed, one would be able to estimate a person's weight (dependent variable) from his or her height

In many cases, the contribution of a single independent variable does not alone suffice to explain the dependent variable Y . If this is so, one can perform a multivariable linear regression to study the effect of multiple variables on the dependent variable. It is better practice, however, to give the corrected coefficient of determination, as discussed in Box 2. Each of the coefficients b_i reflects the effect of the corresponding individual independent variable X_i on Y , where the potential influences of the remaining independent variables on X_i have been taken into account, i.e., eliminated by an additional computation. Thus, in a multiple regression analysis with age and sex as independent variables and weight as the dependent variable, the adjusted regression coefficient for sex represents the amount of variation in weight that is due to sex alone, after age has been taken into account. This is done by a computation that adjusts for age, so that the effect of sex is not confounded by a simultaneously operative age effect. The study of relationships between variables and the generation of risk scores are very important elements of medical research. The proper performance of regression analysis requires that a number of important factors should be considered and tested:

1. Causality Before a regression analysis is performed, the causal relationships among the variables to be considered must be examined from the point of view of their content and/or temporal relationship. The fact that an independent variable turns out to be significant says nothing about causality. This is an especially relevant point with respect to observational studies (5).

2. Planning of sample size The number of cases needed for a regression analysis depends on the number of independent variables and of their expected effects (strength of relationships). If the sample is too small, only very strong relationships will be demonstrable. The sample size can be planned in the light of the researchers' expectations regarding the coefficient of determination (r^2) and the regression coefficient (b). Furthermore, at least 20 times as many observations should be made as there are independent variables to be studied; thus, if one wants to study 2 independent variables, one should make at least 40 observations.

3. Missing values Missing values are a common problem in medical data. Whenever the value of either a dependent or an independent variable is missing, this particular observation has to be excluded from the regression analysis. If many values are missing from the dataset, the effective sample size will be appreciably diminished, and the sample may then turn out to be too small to yield significant findings, despite seemingly adequate advance planning. If this happens, real relationships can be overlooked, and the study findings may not be generally applicable. Moreover, selection effects can be expected in such cases. There are a number of ways to deal with the problem of missing values

4. The data sample A further important point to be considered is the composition of the study population. If there are subpopulations within it that behave differently with respect to the independent variables in question, then a real effect (or the lack of an effect) may be masked from the analysis and remain undetected. Suppose, for instance, that one wishes to study the effect of sex on weight, in a study population consisting half of children under age 8 and half of adults. Linear regression analysis over the entire population reveals an effect of sex on weight. If, however, a subgroup analysis is performed in which children and adults are considered separately,

an effect of sex on weight is seen only in adults, and not in children. Subgroup analysis should only be performed if the subgroups have been predefined, and the questions already formulated, before the data analysis begins; furthermore, multiple testing should be taken into account (7, 8).

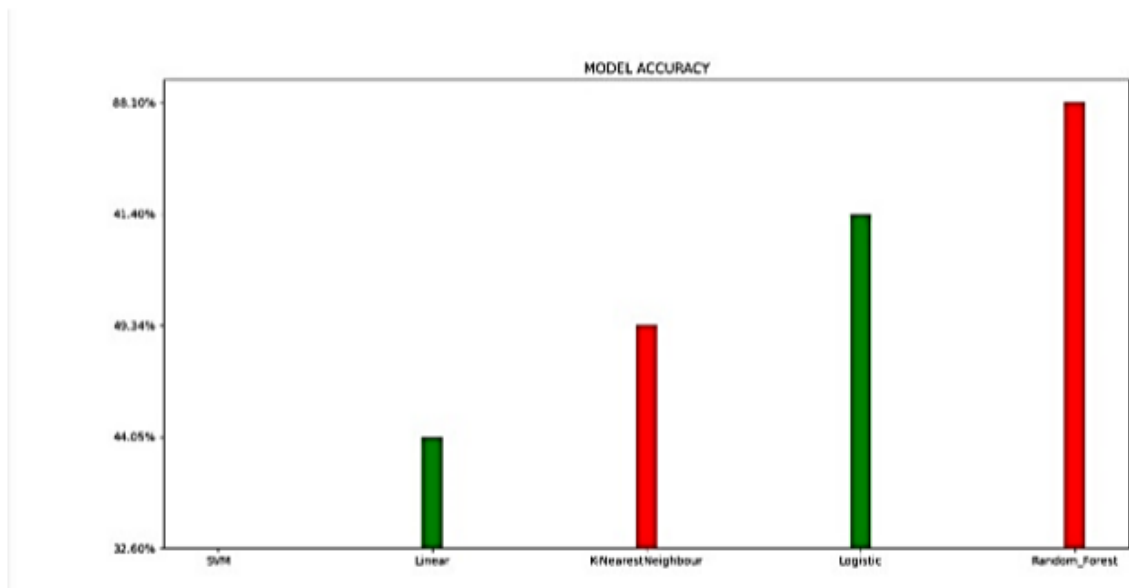
5. The selection of variables If multiple independent variables are considered in a multivariable regression, some of these may turn out to be interdependent. An independent variable that would be found to have a strong effect in a univariable regression model might not turn out to have any appreciable effect in a multivariable regression with variable selection. This will happen if this particular variable itself depends so strongly on the other independent variables that it makes no additional contribution toward explaining the dependent variable. For related reasons, when the independent variables are mutually dependent, different independent variables might end up being included in the model depending on the particular technique that is used for variable selection.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Linear regression has many practical uses. Most applications fall into one of the following two broad categories: If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response. Ordinary linear regression usually isn't enough to take into account all of the real-life factors that have an effect on an outcome.

Performance Evaluation:

1) Accuracy: The fraction of correct forecast in all predictions is known as accuracy. In this experiment, random forest classifier outruns all the algorithms by predicting the result with highest accuracy of 88.10%.

Figure 10 shows the accuracies of the various algorithms implemented.



Graph 4.1 Figure10 Accuracy of various algorithms.

Table 7 shows about the various algorithms and their accuracies obtained. It is clear from the table that the random forest classifier performed better than other algorithm. Table 7. Accuracy achieved by the algorithms

Algorithm	Accuracy
Random Forest	88.10%
K-Nearest Neighbor	49.34%
Logistic Regression	51.40%
SVM	32.6%
Linear Regression	44.05%

Furthermore, 2-fold, 5-fold, 10-fold cross validation for rfc is also implemented for having better insights in table 8

Atlantis Highlights in Computer Sciences, volume 4404

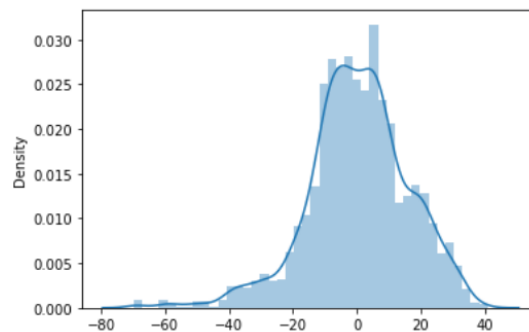
Table 1. Cross validation Technique

NO. OF FOLDS	TRAIN SET	TEST SET	ACCURACY
FOLD 1	680	76	50.0 %
FOLD 2	680	76	47 %
FOLD 3	680	76	51 %
FOLD 4	680	76	58 %
FOLD 5	680	76	46 %
FOLD 6	680	76	49 %
FOLD 7	681	75	55 %
FOLD 8	681	75	43 %
FOLD 9	681	75	47 %
FOLD 10	681	75	53 %

MEAN VALUE FOR ABOVE K-FOLD IS 50 %

```
In [27]: import seaborn as sns
sns.distplot(y_test-prediction)
```

```
Out[27]: <AxesSubplot:ylabel='Density'>
```



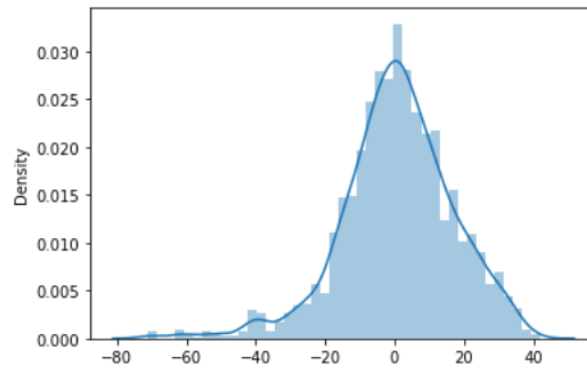
```
In [28]: from sklearn import metrics
import numpy as np
print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

```
MAE: 12.117294527004976
MSE: 251.03172964112574
RMSE: 15.843980864704607
```



```
In [34]: import seaborn as sns  
sns.distplot(y_test-prediction)
```

```
Out[34]: <AxesSubplot:ylabel='Density'>
```



```
In [33]: from sklearn import metrics  
import numpy as np  
print('MAE:', metrics.mean_absolute_error(y_test, prediction))  
print('MSE:', metrics.mean_squared_error(y_test, prediction))  
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

```
MAE: 12.214053814850246  
MSE: 262.37973664007154  
RMSE: 16.198139912967523
```

Screen 4.2

The above image depicts the various measures used to check the performance of the Ridge Regression Lasso Regression model

Conclusion

The IPL score prediction system works properly. All of the attribute values had been preprocessed correctly. The model was applied and trained using training data after all the preprocessing was done. The Linear Regression model accuracy was found to be 82%. The GUI of IPL score prediction was made with Hypertext Markup Language (HTML). The coding was done in Jupiter Notebook and VsCode. After completing all the processes, we have linked the front-end (HTML) with the back-end (Python).

This paper will give the important information regarding IPL score prediction and winning prediction system, hat which parameters are required also the classifiers and algorithms. it helps in mathematical operation. Using all the information we have developed a website. for that the important work we have to do for the model is comparative analysis of machine learning techniques that is for score prediction the regressions and for winning prediction the analysis of classifiers. In Score Prediction analysis accuracy of Linear Regression is more than Ridge and Lasso Regression and in winning prediction analysis among SVC, Decision tree classifier and Random forest classifier, we got Random forest classifier accuracy more than other 2, with all 90%, 80%, 75%, 70% training data.

In this study, the various factors that influence the outcome of an Indian Premier League matches were identified. The seven factors which significantly influence the result of an IPL match include the home team, the away team, the toss winner, toss decision, the stadium, and the respective teams' weight.

Team names and their acronym

Team Name	Acronym
Delhi Daredevils	DD
Kings XI Punjab	KXIP
Kolkata Knight Riders	KKR
Mumbai Indians	MI
Rajasthan Royals	RR
Royal Challenger Bangalore	RCB
Sunrisers Hyderabad	SRH
Rising Pune Supergiant	RPS*
Deccan Chargers	DC*
Pune Warriors India	PWI*
Gujrat Lions	GL*
Kochi Tuskers Kerala	KTK*

Table 2

*Inactive teams as of 2018

References

- [1] IPL _Dataset | Kaggle (12/29/2021)
- [2] Duff & Phelps Launches IPL Brand Valuation Report 2019 | Duff & Phelps (duffandphelps.com) (12/29/2021)
- [3] IPL Advertising: All You Need to Know About the Game of Revenue (kreedon.com) (5/12/2021)
- [4] Sport Analytics | A data-driven approach to sport business and manage (taylorfrancis.com) (5/12/2021)
- [5] Sport Analytics | A data-driven approach to sport business and manage (taylorfrancis.com) (5/12/2021)
- [6] Survey on deep learning with class imbalance | Journal of Big Data | Full Text (springeropen.com) (14/1/2022)
- [7] Parag Shah, "Predicting Outcome of Live Cricket Match Using Duckworth-Lewis Par Score", Publisher: International Journal of Latest Technology in Engineering, Management & Applied Science, Volume VI, Issue VIIS, July 2017.
- [8] Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Hussain Dawood, and Yixian Yang , "Prediction of Rising Stars in the Game of Cricket", Publisher: IEEE Access, Issue March 4 2017.
- [9] Mehvish Khan, Riddhi Shah, "Role of External Factors on Outcome of One Day International Cricket (ODI) Match and Predictive Analysis", Publisher: International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015.
- [10] https://en.wikipedia.org/wiki/2016_Indian_Premier_League.
- [11] <https://www.sportskeeda.com/cricket/ipl-2016-teams-full-players-list>
- [12] <http://statisticstimes.com/sports/all-ipl-points-table.php>
- [13] <https://www.icc-cricket.com/rankings/mens/team-rankings/test>
- [14] <https://www.kaggle.com/manasgarg/ipl>
- [15] <https://www.researchtrend.net/ijet/pdf/Comprehensive%20Data%20Analysis%20and%20Prediction/IPLusingMachine/LearningAlgorithmsValarmathiB%202113j1.pdf>
- [16] <https://www.ijser.org/researchpaper/ANALYZING-IPL-MATCH-RESULTS-USING-DATA-MINING-ALGORITHMS.pdf>
- [17] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015
- [18] International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, India, 2015, pp. 60-66, doi: 10.1109/ICSCTI.2015.7489605.

- [19] J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.
- [20] A. Kaluarachchi and S. V. Aparna, "CricAI: A classification based tool to predict the outcome in ODI cricket," 2010 Fifth International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 2010, pp. 250-255, doi: 10.1109/ICIAFS.2010.5715668.
- [21] A. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT), Dhaka, Bangladesh, 2018, pp. 500-505, doi: 10.1109/CEEICT.2018.8628118.
- [22] N. Rodrigues, N. Sequeira, S. Rodrigues and V. Shrivastava, "Cricket Squad Analysis Using Multiple Random Forest Regression," 2019 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India, 2019, pp. 104-108, doi: 10.1109/ICAIT47043.2019.8987367.
- [23] M. Jhawar and V. Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Riva del Garda, 2016
- [24] A. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," 2018 4th International Conference on Electrical Engineering and

Information & Communication Technology (iCEEICT), Dhaka, Bangladesh, 2018, pp. 500-505, doi:

10.1109/CEEICT.2018.8628118.

[25] Rameshwari Lokhande, P. M. Chawan "Live Cricket Score and Winning Prediction" Published in

International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-5 | Issue1 , February 2018, URL: <http://www.ijtrd.com/papers/IJTRD12180.pdf>

[26] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier

League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7,

doi: 10.1109/INCET49848.2020.9153972.

[27] A. Basit, M. B. Alvi, F. H. Jaskani, M. Alvi, K. H. Memon and R. A. Shah, "ICC T20 Cricket World Cup 2020

Winner Prediction Using Machine Learning Techniques," 2020 IEEE 23rd International Multitopic

Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-6, doi: 10.1109/INMIC50486.2020.9318077.

[28] A. Bandulasiri, "Predicting the Winner in One Day International Cricket", Journal

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to **Prof. S. B. Kalyankar**, HOD-Dept. of Computer Science and Engineering, Deogiri Institute of Engineering, and Management Studies Aurangabad, for his generous guidance, help and useful suggestions.

We express our sincere gratitude to **Prof. Vijay Kolte**, Dept. of Computer Science and Engineering, Deogiri Institute of Engineering, and Management Studies Aurangabad, for his stimulating guidance, continuous encouragement, and supervision throughout the course of present work.

We are extremely thankful to **Dr. Ulhas Shiurkar**, Director, Deogiri Institute of Engineering, and Management Studies Aurangabad, for providing us infrastructural facilities to work in, without which this work would not have been possible.

Signature(s) of Students

Abhishek Vishal Jaju (46035)

Vivek Sivanand Soddy (46039)

Pradeep Shankar Maku (46034)

