A Project report on

IPL MATCH PREDICTIONS

By

Ammar bin Ayaz (aba450)

Aswathy Mohan (am9094)

Rahul (rr3687)

Sesha Sai Sreevani Kappagantula (ssk785)

ABSTRACT

IPL MATCH PREDICTIONS

Cricket, especially the Twenty20 format, has maximum uncertainty, where a single over can completely change the momentum of the game. With millions of people following the Indian Premier League (IPL), analyzing the matches and developing a model for predicting the outcome of its matches is a real-world problem. A cricket match depends upon various factors, and in this work, the factors which significantly influence the outcome of an IPL match are identified. Each player's performance in the field is considered to find out the strengths of each player. Several datasets were modeled based on the identified factors which influence the outcome and performances in an IPL match. Machine learning models were trained and used for predicting the outcome of an IPL match.

Big data tools predict the future trends and behaviours, which gives an opportunity to predict the outcome of an IPL (Indian Premier League) match using data algorithms. Data cleaning and analysis algorithms have been applied to the IPL dataset and the knowledge from each algorithm has been obtained and analyzed thoroughly as the results are obtained with good accuracy performance. Cricket is one of the most popular sports. Indian Premier League (IPL), a sports league was contested during the month of April and May on every year by the teams representing the Indian cities. The result has been predicted using machine learning approaches and have analyzed the results of the IPL match.

This project consists of scrapping the IPL cricket data needed, analyzing the data collected, making match predictions and displaying the results. For Data scraping, we will be using BeautifulSoup library and extract the required player and match
features. For the data analysis part, we used technologies like Pyspark and SparkSQL. For the predictions we will be using suitable Machine learning techniques with SparkML. We have also visualized the tweets from the twitter API.

KEYWORDS: Cricket, IPL, Machine learning, Analytics, players, prediction

Table of Contents

CHAPTER – I

1.1 INTRODUCTION

Indian Premier League (IPL) is a professional cricket league based on Twenty20 format and is governed by the Board of Control for Cricket in India. The league happens every year with participating teams' names 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 espncricinfo, 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. So it is really useful to associate analytics with this field and get conclusions highly vital in the decision-making.

Sports analytics is a promising research field which involves deriving valuable information about the game, based on past games played, or even games in progress. The prediction of the final outcome of the match proves very beneficial to team members, team coaches and also bettors. For example, games tactics can be developed by club managers based on the outcome of previous matches or statistics related to certain players. IPL being a very dynamic league, bettors and bookies are incentivised to bet on the match results or during a game. The sports betting industry is growing at a fast rate.

In this project we are using this method of collecting and analyzing historical game information to derive essential knowledge from it, with the aim that it will promote successful decisions being made. The past data of IPL containing the players' details, match venue details, teams, ball to ball details, etc is taken and analyzed to draw various conclusions which help in the improvement of a player's performance, predict outcomes and bid on a player.

1.2 PROJECT OUTCOMES

- Scrapping the data and cluster players based on the features collected. We will be using separate clustering criteria for bowlers and batsmen while performing the step.

- Simulating the entire IPL match using ball by ball data and clustered players from the previous step. We will be trying to calculate player vs player probability. If the players haven't played before then we will be using clustering probability.

- Each innings is simulated and batsmen/bowlers are interchanged for every 6 balls. Selection of bowlers will also be decided.

- Train a Machine Learning classifier and predict the results. The predictions consist of the number of runs scored by each batsman, the number of wickets taken by each bowler and the final scores for the match.

1.3 PROBLEM STATEMENT

The need to depend on data to draw better conclusion and decisions cannot be stressed enough in today's world. In IPL matches where huge amounts of money and time are invested, it becomes very important to properly understand, analyze and estimate things before concluding on to the things. This has given to the scope of big data analytics in the field of cricket throwing the question "Can big data analytics and machine learning technology derive accurate predictive models for cricket matches related to IPL and help in better decision-making? If so, what kind of analysis plays a role and which machine learning models can perform are the best with respect to accuracy measures?"

1.4 OBJECTIVE

The main objective is to understand the implementations of Big data analytics through the problem statement stated and inculcate the knowledge of the technologies used in the process. Also, the project aims to elaborate the usage of Big data in the field of sports and the advances happening in the field.

CHAPTER – II

DATASETS

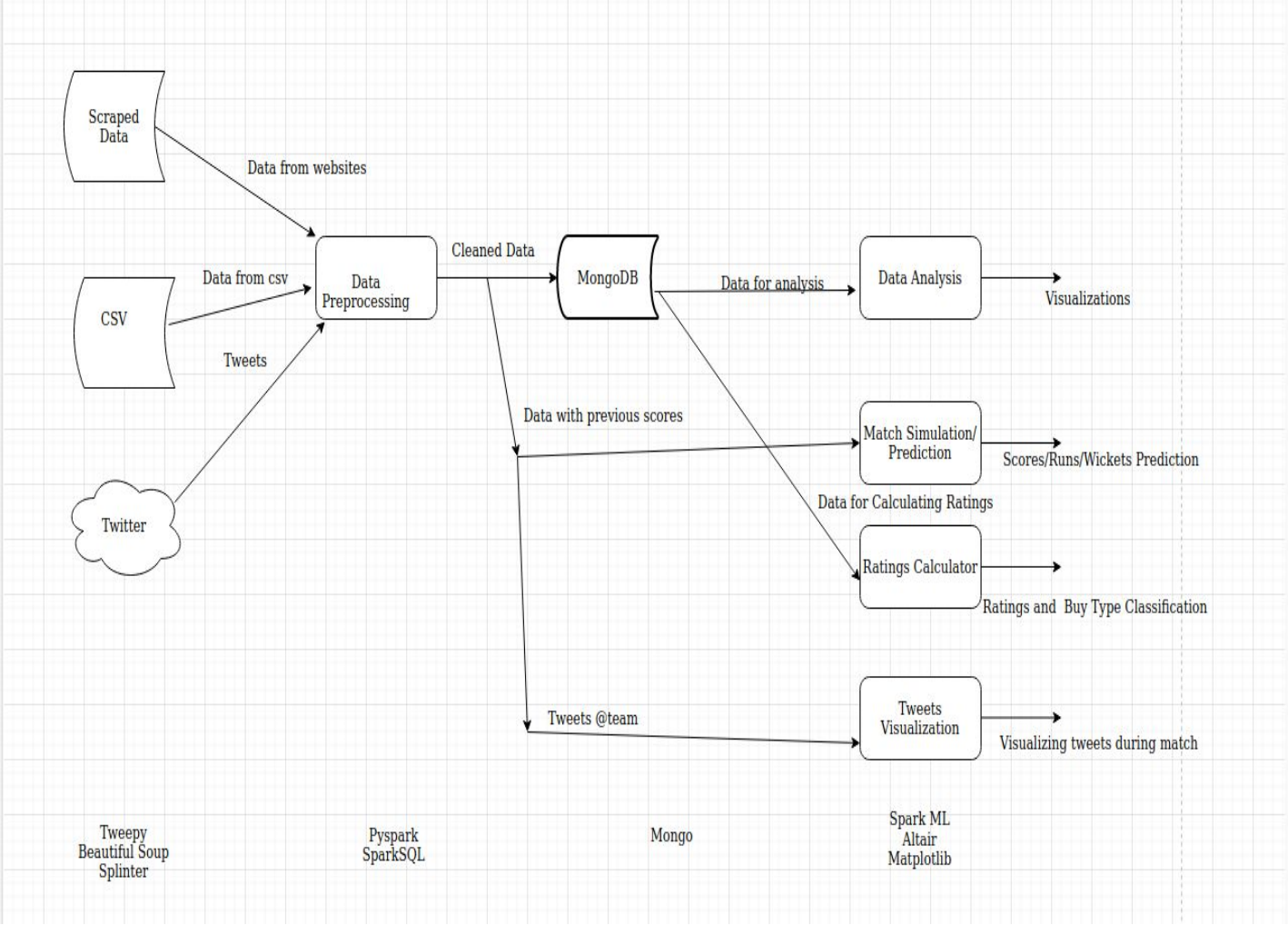
The required data (around 30MB) is used, taken from available csv files and scraped using BeautifulSoup which is a python library used for web scraping. The data typically includes details of players like matches played, batting innings, not outs, runs scored, highest score, average score, balls faced, strike rate, hundreds, fifties, fours, sixes, bowling innings, balls bowled, economy, bowling strike rate and runs conceded.

The data gathered can be categorized as follows:

- Data files used for match simulation
- Deliveries data (ball to ball details of matches played)
- Matches data (details of every match played between teams)
- Players data (player skills and records)
- Auction data (player price and buyer details)
- T20 ODI Data(2010-2019)

CHAPTER – III

3.1 PROJECT ARCHITECTURE:



3.2 DESCRIPTION:

The project architecture basically consists of the following six key components

(a) Data gathering

The data has been gathered in the form of csv files from sites such as kaggle and also scraped

data from the web using splinter and beautifulsoup libraries.

(b) Data preprocessing

The data has been properly cleaned before applying analysis tools on it. The missing values and null values have been handled and data has been prepared to achieve better predictions for the matches.

(c) Data Analysis

Various aspects of IPL matches have been analyzed to get a better picture of the matches. Some of the types of analysis include toss analysis, umpire analysis, venue analysis, players' performance analysis and so on. The conclusions drawn from this analysis can be useful in better decision making on the matches.

(d) Match prediction/simulation

Simulated match between players (various permutations) selected to understand the possible outcomes and manage the batting and bowling orders accordingly.

(e) Ratings calculator

Each player has been assigned a buy rating as (maybe, avoid,excellent, good). The buy ratings help the investors decide on the player's performance scope and decide on the money to be put in the auction of the players.

(f) Tweets visualization

The twitter API has been used using tweepy to get the trending tweets belonging to a player or a team and make tweet cloud or tweet visualization with the top tweets.

3.3 TECHNOLOGIES USED:

- BeautifulSoup
- Splinter
- Tweepy
- Pyspark
- Spark SQL
- Mongodb (Mongodb atlas)
- Spark ML
- Matplotlib
- Altair

CHAPTER – IV

DATA ANALYSIS

4.1 Toss analysis

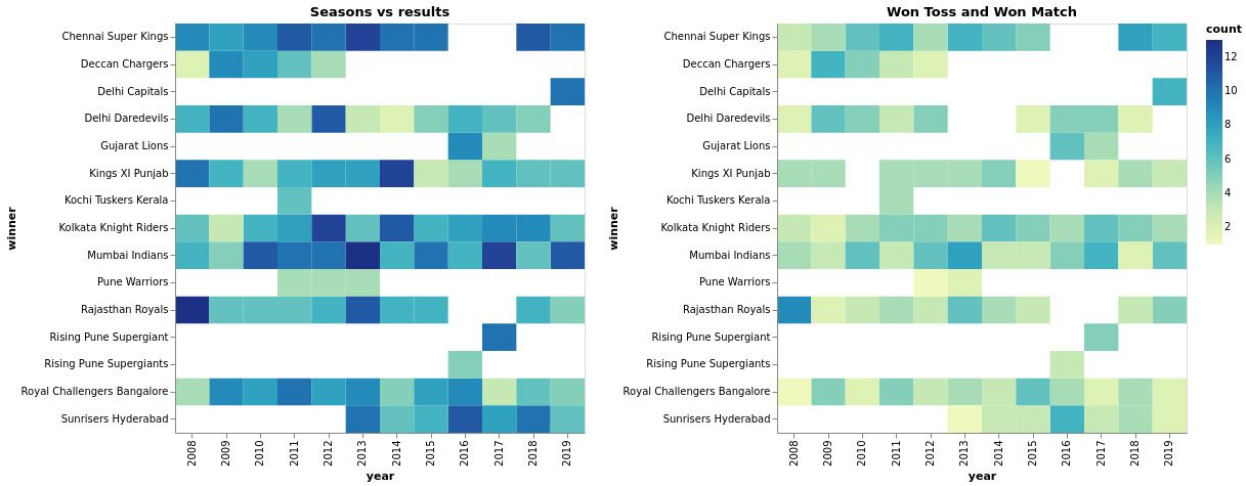


Fig 4.1(a)

Fig 4.1(b)

In Fig 4.1(a), we can see the number of matches each player has played in each season. Chennai Super Kings have won most of the matches during the seasons they have played. During 2016 and 2017 they were not playing the match and hence blank during those years.

Fig 4.1(b) shows the number of matches won , when they got Toss. It's visible that CSK has won around 80% of matches when they got toss first. For Rajasthan Royals, this happened only in 2008

4.2 Match analysis

(a) Umpire analysis

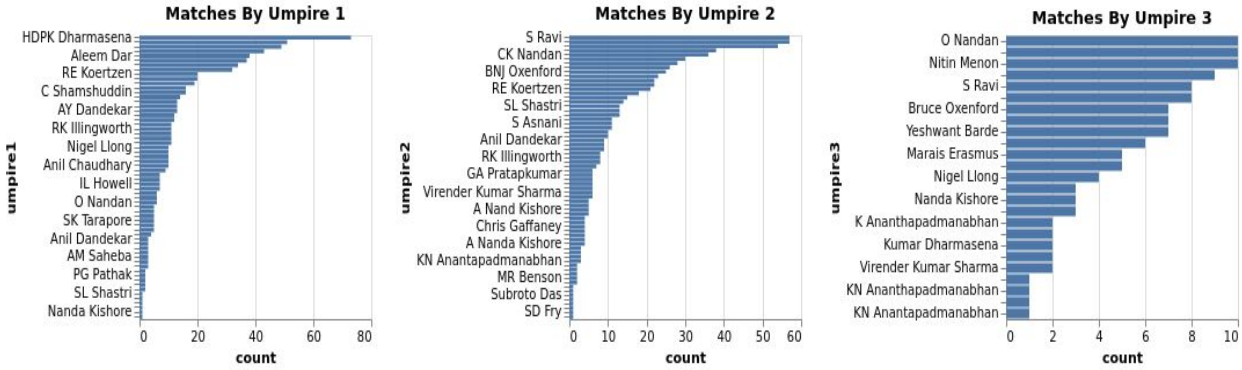


Fig 4.2.(a)

Umpires are often ignored in the matches. So we wanted to find out how is the distribution of umpires in matches. Surprisingly, we can see that Umpire 1 has umpired around 80 matches, but Umpire 2 has 60 and 3 has around 10. It could be because the same person has been assigned as Umpires.

(b) Venue analysis

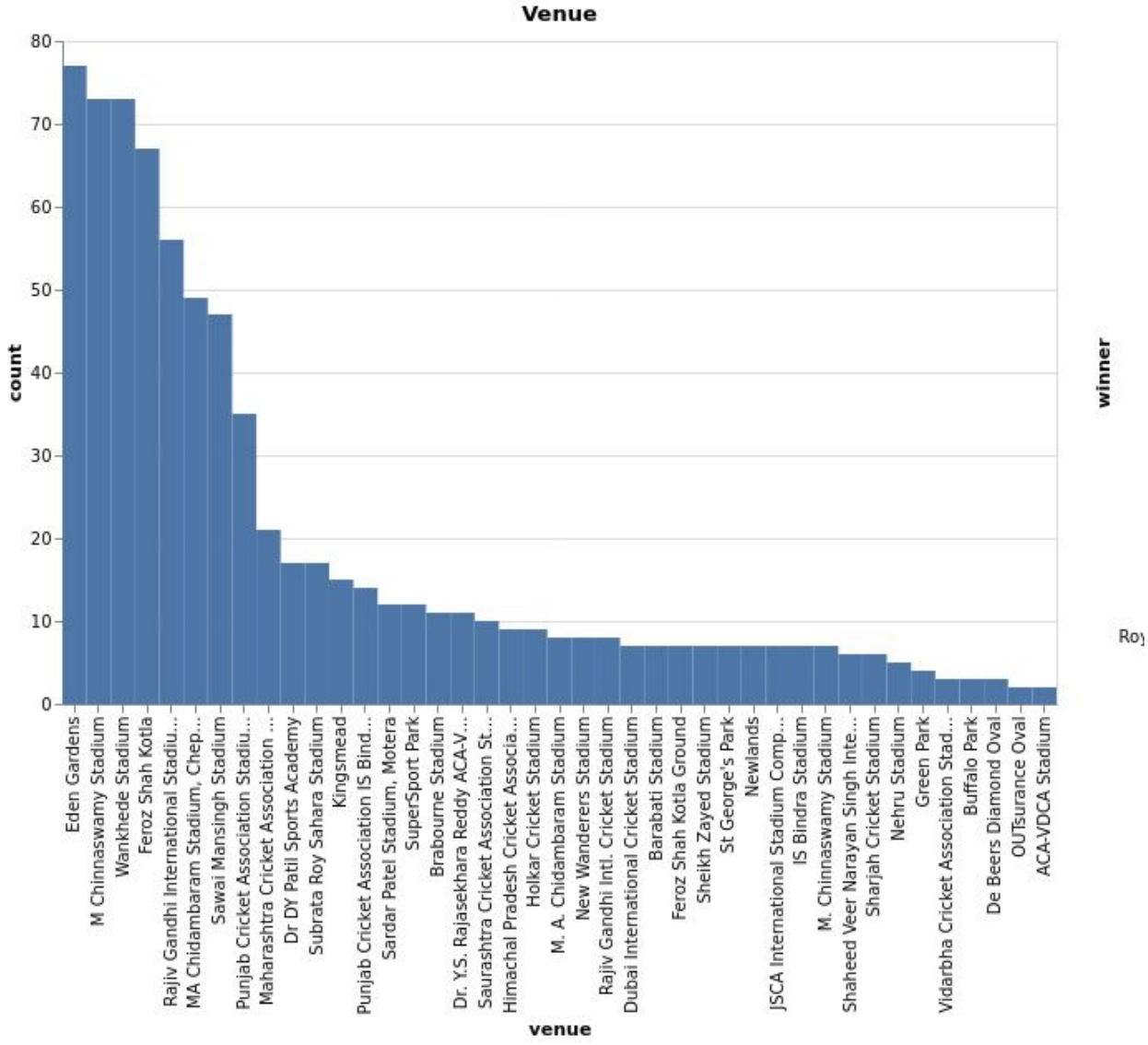


Fig 4.2 a.(b)

Most wins By Matches in Venues:

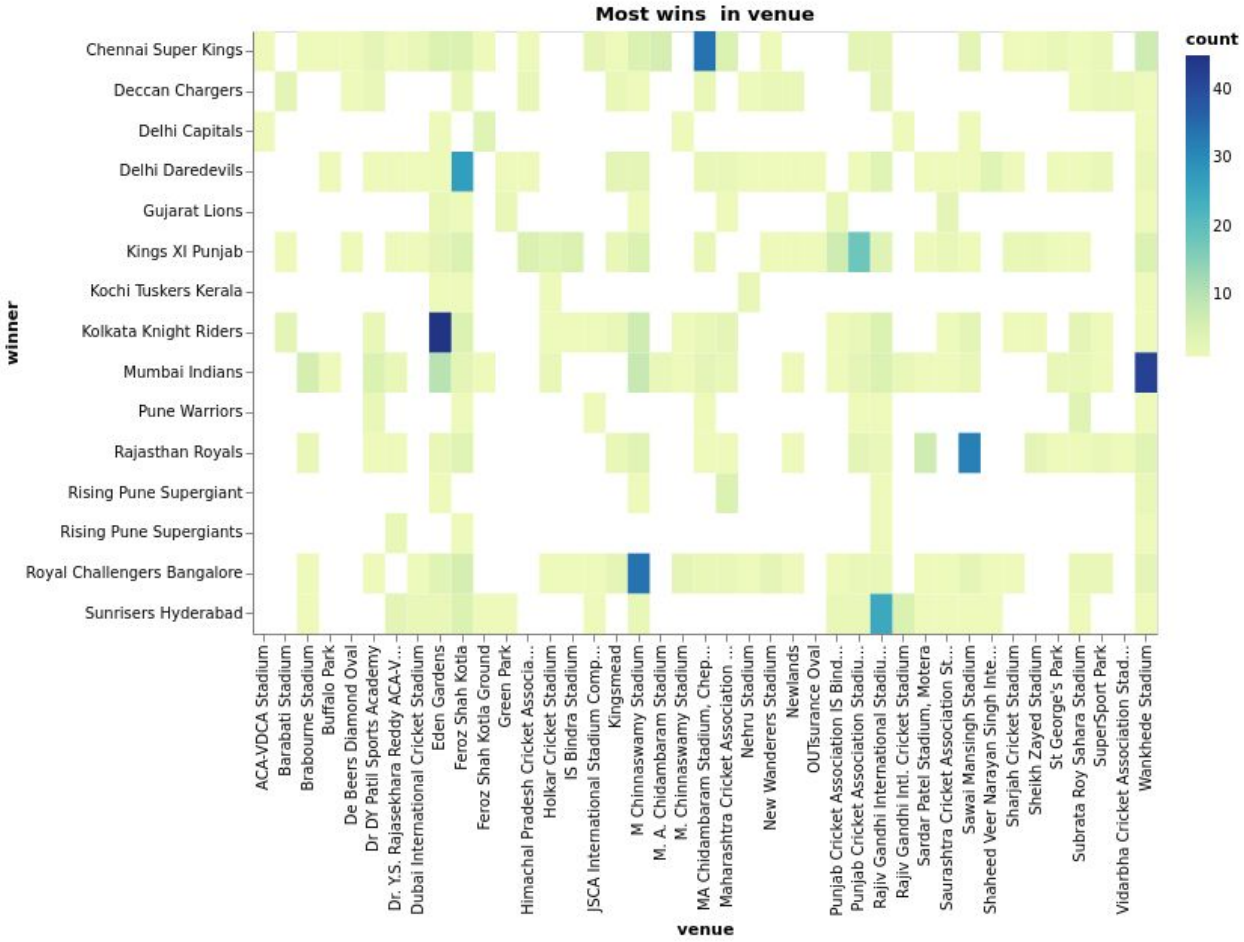


Fig 4.2.b(b)

Eden Gardens, Kolkata has the most matches played in IPL seasons. As we can see in the 4.2.b(b), all the teams has highest wins in their home grounds. Like CSK-Chidambaram stadium, RCB has most wins in Chinnaswamy stadium and so on.

(c) Batsmen performance analysis

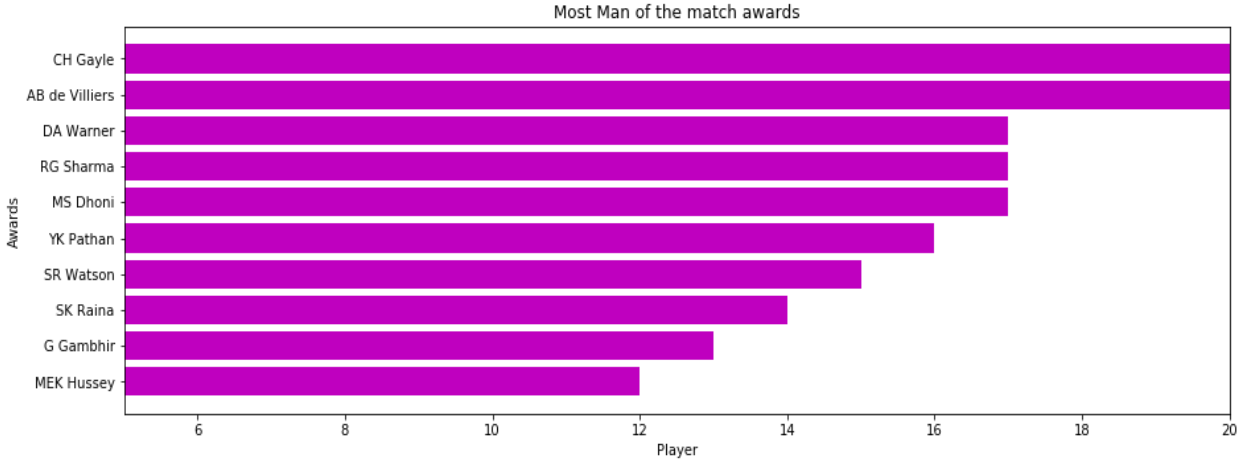


Fig 4.2.c(a)

This graph explains the plates who has won the most man of match awards. The top player is not from West Indies, Chris Gayle.

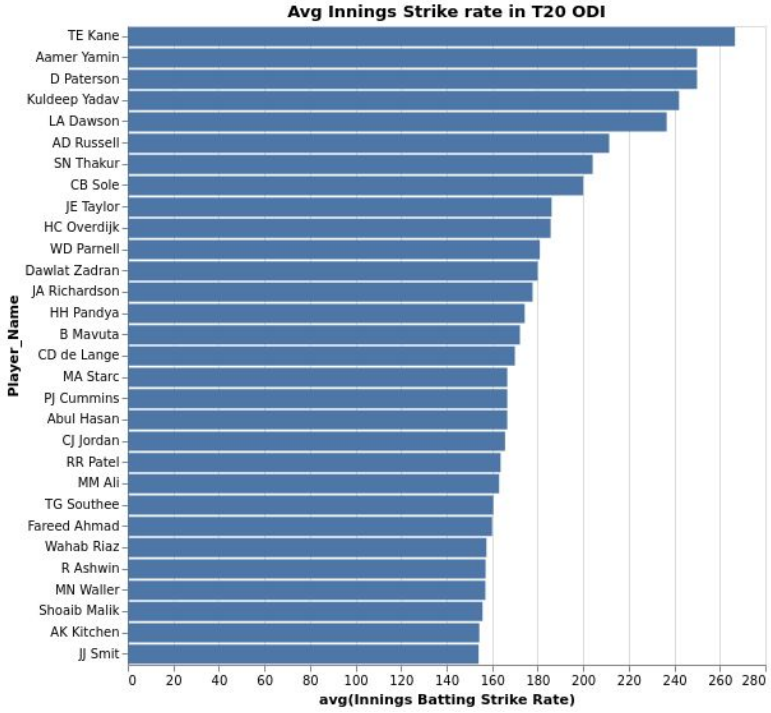


Fig 4.2.c(b)

This graph is the distribution of T20 ODI match scores for each player. As the performance of a player is decided based on all matches including IPL,T20 we thought of

analysing strike rate of each player.

Batting strike rate (s/r) : The average number of runs scored per 100 balls faced. The higher the strike rate, the more effective a batsman is at scoring quickly. Strike rates of over 150 are becoming common in T20 cricket. Strike rate is probably considered by most as the key factor in a batsman in one day cricket. Accordingly, the batsmen with the higher strike rate, especially in Twenty20 matches, are more valued than those with a lesser strike rate [5].

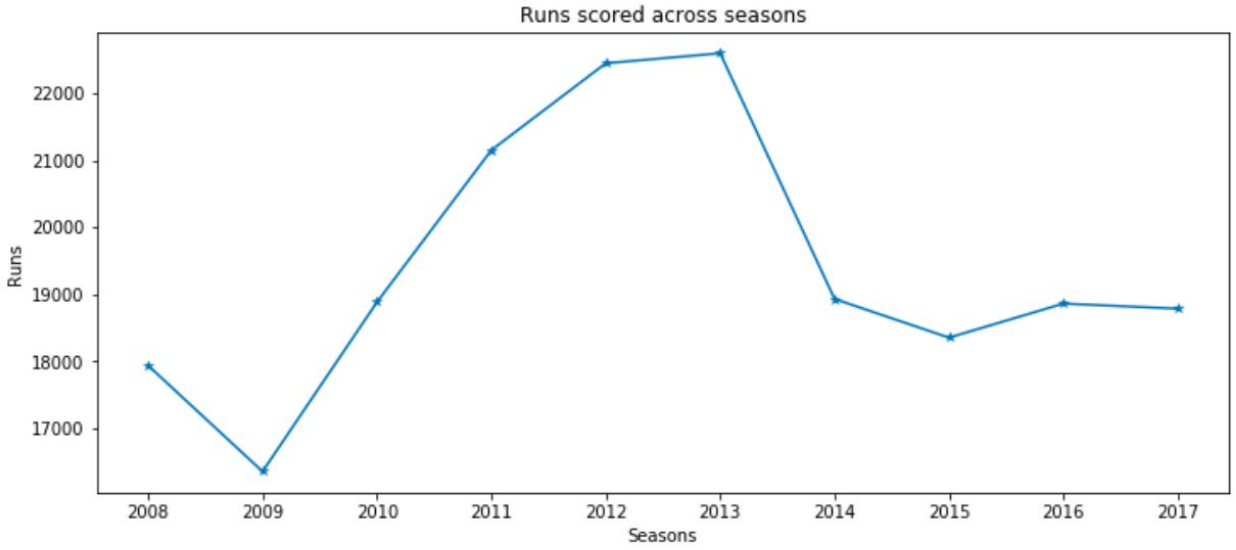


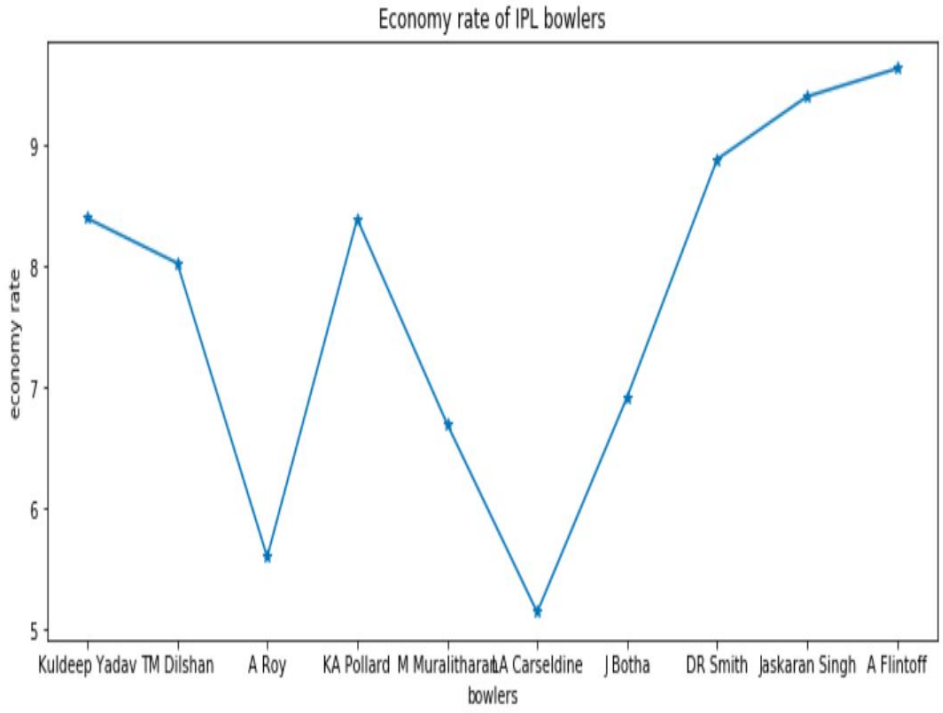
Fig 4.2.c(c)

This graph shows the trend of the number of runs scored in each match. There is a spike in 2012 and 2013, this is because there were more matches played during that season.

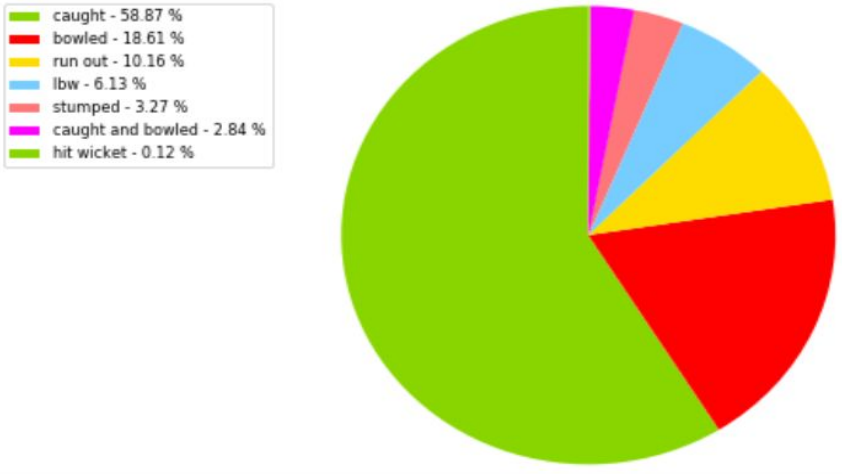
(d) Bowler Performance:

Bowlers Performance was assessed mainly using economy rate. Economy rate is the average number of runs conceded for each over bowled. A lower economy rate is seen as preferable – it means that the bowler is able to get more batsmen out with fewer

balls. The statistic is considered to be more important in shorter games than longer test matches. The graph below shows economy rate for top bowlers



The following graph shows the reason for each wicket. We can see that the majority of wickets were catches .



4.3 Ratings

One of the integral parts of the IPL is the auction process where each team is given a chance to bid for a player. Through this we tried to aid the process for a team where they can use our outcomes when picking up new players for the team. We used multiple factors like batting average, strike rate, age etc and developed a calculator to map the ratings generated into categories such as Excellent buy, Very good buy, Avoid etc. To automate the process we fed our data into a Random forest classifier and got an accuracy of 73% which was low as we were expecting a higher accuracy for generating labels.

	Player	Runs	Wickets	Age	Batting Skill	Bowling Skill	Country	Type	Rating	Buy type
1	SK Raina	5415	2.5	33	Left_Hand	Right-arm offbreak	India	Batsman	10.00	Excellent Purchase
0	V Kohli	5434	0.4	31	Right_Hand	Right-arm medium	India	Batsman	9.18	Excellent Purchase
2	RG Sharma	4914	1.5	33	Right_Hand	Right-arm offbreak	India	Batsman	9.18	Excellent Purchase
29	RA Jadeja	1951	10.8	31	Left_Hand	Slow left-arm orthodox	India	All Rounder	9.18	Excellent Purchase
11	SR Watson	3614	9.2	38	Right_Hand	Right-arm fast-medium	Australia	All Rounder	8.32	Excellent Purchase
138	B Kumar	190	13.3	30	Right_Hand	Right-arm medium	India	Bowler	8.32	Excellent Purchase
79	PP Chawla	587	14.9	31	Left_Hand	Legbreak	India	Bowler	8.32	Excellent Purchase
68	AR Patel	806	7.1	26	Left_Hand	Slow left-arm orthodox	India	Bowler	8.32	Excellent Purchase
267	Sandeep Sharma	26	8.3	26	Right_Hand	Right-arm medium	India	Bowler	8.32	Excellent Purchase
69	SP Narine	803	12.2	31	Left_Hand	Right-arm offbreak	West Indies	Bowler	8.32	Excellent Purchase
247	JJ Bumrah	37	8.2	26	Right_Hand	Right-arm medium	India	Bowler	8.32	Excellent Purchase
233	Kuldeep Yadav	46	3.9	25	Left_Hand	Slow left-arm chinaman	India	Bowler	7.50	Excellent Purchase
67	Harbhajan Singh	834	15.0	39	Right_Hand	Right-arm offbreak	India	Bowler	7.50	Excellent Purchase
7	RV Uthappa	4446	0.0	34	Right_Hand	Right-arm medium	India	Batsman	7.50	Excellent Purchase
16	KA Pollard	2784	5.6	32	Right_Hand	Right-arm medium-fast	West Indies	All Rounder	7.50	Excellent Purchase
3	DA Warner	4741	0.0	33	Left_Hand	Legbreak	Australia	Batsman	7.50	Excellent Purchase
280	YS Chahal	22	10.0	29	Right_Hand	Legbreak googly	India	Bowler	7.50	Excellent Purchase
4	S Dhawan	4632	0.4	34	Left_Hand	Right-arm offbreak	India	Batsman	7.50	Excellent Purchase
50	HH Pandya	1118	4.2	26	Right_Hand	Right-arm medium-fast	India	All Rounder	7.50	Excellent Purchase

CHAPTER – V

5.1 MATCH PREDICTIONS

We used various ML algorithms to predict the outcome of the matches.

From the table below we can clearly see that Random forest and Gradient Boost had the same training accuracy, but to avoid overfitting of the model we use K-Fold cross validation and we can clearly say that Gradient Boost performs the best having a cross validation score of around 55%. We feel that to predict the outcome of a match with around 90% accuracy we need to take into account various other factors such as weather which can heavily impact playing conditions. We tried to scrape or find the weather data for the days when the matches were played but we were not able to do it due to various limitations.

Classifier	Accuracy	Cross-validation score
Logistic Regression	23.270%	21.852%
Random Forest	89.151%	47.806%
AdaBoost	19.811%	16.663%
Support Vector Machine	25.786%	18.707%
Multi layer perceptron	28.302%	25.154%
Gradient Boost	89.151%	54.251%

5.2 MATCH SIMULATION

This was one of the parts of the project which was really exciting. Through this we wanted to simulate the actual 20 overs of an ipl match. This would be really helpful in deciding the order of batsman and bowlers for a particular match because the teams can try different combinations and decide which order to pick for a favorable outcome.

Although we were not able to get a very good accuracy in simulating a ball-by-ball match but we still were able to do a pretty good job in predicting the final score of a 20-20 match.

```
['AS Yadav', 'SR Tendulkar', 'DL Chahar', 3, 2]
['SR Tendulkar', 'RG Sharma', 'SN Thakur', 10, 0]
['RG Sharma', 'SR Tendulkar', 'DL Chahar', 10, 0]
['SR Tendulkar', 'RG Sharma', 'SR Watson', 8, 0]
['RG Sharma', 'SR Tendulkar', 'DL Chahar', 8, 0]
['SR Tendulkar', 'RG Sharma', 'SN Thakur', 8, 0]
['RG Sharma', 'SR Tendulkar', 'SR Watson', 8, 0]
['SR Tendulkar', 'RG Sharma', 'Harbhajan Singh', 7, 0]
['RG Sharma', 'SR Tendulkar', 'Imran Tahir', 7, 0]
['SR Tendulkar', 'RG Sharma', 'Harbhajan Singh', 7, 0]
['RG Sharma', 'SR Tendulkar', 'Imran Tahir', 8, 0]
['SR Tendulkar', 'RG Sharma', 'Harbhajan Singh', 8, 0]
['RG Sharma', 'SR Tendulkar', 'DJ Bravo', 8, 0]
['SR Tendulkar', 'RG Sharma', 'SR Watson', 7, 1]
['RG Sharma', 'JP Duminy', 'DJ Bravo', 25, 1]
['JP Duminy', 'KH Pandya', 'SN Thakur', 6, 1]
['KH Pandya', 'HH Pandya', 'DJ Bravo', 15, 0]
['HH Pandya', 'KH Pandya', 'SR Watson', 14, 1]
['KH Pandya', 'BCJ Cutting', 'SN Thakur', 16, 0]
['BCJ Cutting', 'KH Pandya', 'Imran Tahir', 16, 2]

1st Innings:  191 / 6

2nd Innings:  185 / 8
```

CHAPTER – VI

TWEETS VISUALIZATION

Visualized the IPL tweets posted on Twitter, an online social network that allows users to upload short text messages—tweets—of up to 140 characters. This restriction encourages users to construct focused, timely updates.

We have specifically used the tweepy to gather the twitter API that helped in accumulating the tweets related to any team or player.

This tweet visualization is helpful especially for merchandisers and generally done during the matches.

Fig 6.

The figure above depicts the tweet visualization for the IPL team Chennai Super Kings.

CHAPTER – VII

7.1 SUMMARY

- Analysed different aspects of an IPL Match
- Calculated ratings of every player based on certain parameters
- Predicted Outcome of the Match using Machine Learning
- Simulated a ball by ball match between two team and analyzed the tweets.

7.2 IMPLICATIONS

- This analysis can be used by potential bidders and make better investment decisions (team and players bidding)
- Simulation of the Match with different permutations can help in getting a realistic idea about the batting/bowling order
- Tweets cloud can be used in merchandising. It helps in accessing the public opinion on the matches,teams and players.

7.3 REFERENCES

[1].https://www.dexlabanalytics.com/blog/how-stat-this-ipl-season-embrace-big-data-analysis-and-predict-it-right

[2]. Analyzing and predicting outcome of IPL cricket data, International Journal of Innovative Research in Science, Engineering and Technology - Vol. 8, Issue 4, April 2019

[3]. Predictive Analysis of IPL Match Winner using ML, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S, December 2019

[4].https://towardsdatascience.com/analysing-ipl-data-to-begin-data-analytics-with-python-5d2f610126a

[5]. https://en.wikipedia.org/wiki/Strike_rate