International Journal of Innovative Research in Engineering
Volume 3, Issue 4 (July-August 2022), PP: 51-55.
www.theijire.com
ISSN No: 2582-8746

IPL Data Analysis using Python

Jayesh S. Mahajan[1], Dipak R. Purkar[2], Sagar S. Pardeshi[3], Ganesh M. Kharche[4], prof. Rahul Kumar Patel[5].
[1 2 3 4 5]*E&TC, KCE Society's College of Engineering and Management, Jalgaon, India.*

How to cite this paper: Jayesh S. Mahajan[1], Dipak R. Purkar[2], Sagar S. Pardeshi[3], Ganesh M. Kharche[4], prof. Rahul Kumar Patel[5], "IPL Data Analysis using Python", IJIRE-V3I04-51-55.

Abstract: *We all know that we live in an information age, where data plays akey role. If you own the data, you own everything. But what happens after you get the data? Well, it depends on what kind of data you get. You might have some kind of data on your hands that you have to analyse to get valuableinformation. Like if you are working in Zomato, then you have to do data analysis on the data you have. If you are working in some advertisement company then you have to perform data analysis there, too. By analyzing their data, you may provide some valuable information and strategy to the company. Enough thesis here. Now, we all watch cricket generally and we allknow the Indian premier league (IPL) is the biggest cricket league in the world. Let's perform the data analysis of IPL with the data of IPL matches from 2008 to 2020. Firstly, we studied about python programming language on online platform and then studied about Data analysis from E papers , and from all possible thingson that basis we select this project. After selecting this project we decide to analyse the IPL data from 2008 To 2020. IPL dataset we take from Kaggle website and afterwe studied about machine learning algorithms that is linear and logisticsalgorithms under the guidance of our respective Guide.*

I.INTRODUCTION

Data Science is the study of data to extract knowledge and insights from the data and apply knowledge and actionable insights. In this project, we will work on IPL Data Analysis and Visualization Project using Python wherewe will explore interesting insights from the data of IPL matches like most run by a player, most wicket taken by a player, and much more from IPL season 2008-2020.

II.MATERIAL AND METHODS

- We use dataset in the form csv that is take from kaggle website.
- And we use online platform for import our data that name is google colab.

A. Data Analysis:

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

B. Python for Data Analysis

Python is a popular multi-purpose programming language widely used for its flexibility, as well as its extensive collection of libraries, which are valuable for analytics and complex calculations. Python's extensibility means that it has thousands of libraries dedicated to analytics, including the widely used Python Data Analysis Library (also knownas Pandas).For the most part, data analytics libraries in Python are at least somewhat derived from the NumPy library,which includes hundreds of mathematical calculations, operations, and functions.

C. Use Of Python for Data Analytics

There are several ways you can integrate python data analytics into your existing business intelligence and analytics tools.

One of the most common uses for Python is in its ability to create and manage data structures quickly — Pandas, for instance, offers a plethora of tools to manipulate, analyze, and even represent data structures and complex datasets. This includes time series and more complex data structures such as merging, pivoting, and slicing tables to create newviews and perspectives on existing sets. Elsewhere, tools like Scikit-Learn (also known as Sklearn) provides advancedanalytics tools combined with complex machine learning capabilities. This allows you to build more sophisticated models, performing more complex and multivariate regressions, as well as data preprocessing. C ombined with libraries such as iPython and NumPy itself, these tools can form the foundation of a powerful data analytics suite. Additionally, you can use Python to write your own data analysis algorithms that can be directly integrated into yourbusiness intelligence tools via API.

III.LIBRARIES USES

A. Numpy:- Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

B. Pandas:-Pandas is an open-source library that is made mainly for working with relational or labeled data both easilyand intuitively. It provides various data structures and operations for manipulating numerical data and time series

C. CSV:- (Comma Separated Values) is a simple **file format** used to store tabular data, such as a spreadsheet or database.A CSV file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consistsof one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for thisfill format. For working CSV files in python, there is an inbuilt module called csv.

D. Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was createdby John D. Hunter. Matplotlib is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C0 and Javascript for Platform compatibility. Matplotlib is easy to use and an amazing visualizing library in Python.

E. Seaborn:- is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

IV.ALGORITHMS

A. Linear Regression:-

In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data. In naïve words, “Regression shows a line or curve that passes through all thedata points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.” It is used principally for prediction, forecasting, time series modeling, and determining the causal-effet relationship between variables.

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independentvariable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, suchlinear regression is called multiple linear regression. The linear regression model gives a sloped straight line describingthe relationship within the variables

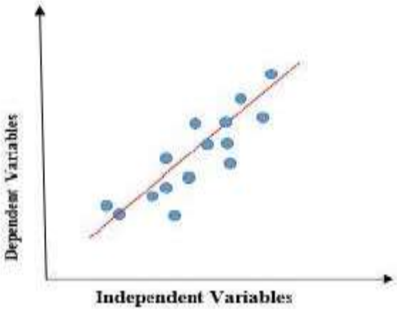
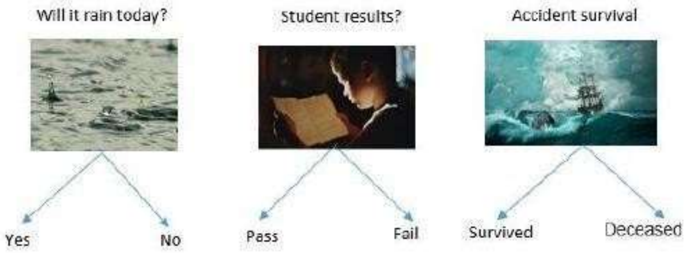


Fig:- Linear Regression Algorithm

The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. Thered line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

B. Logistics Regression:-

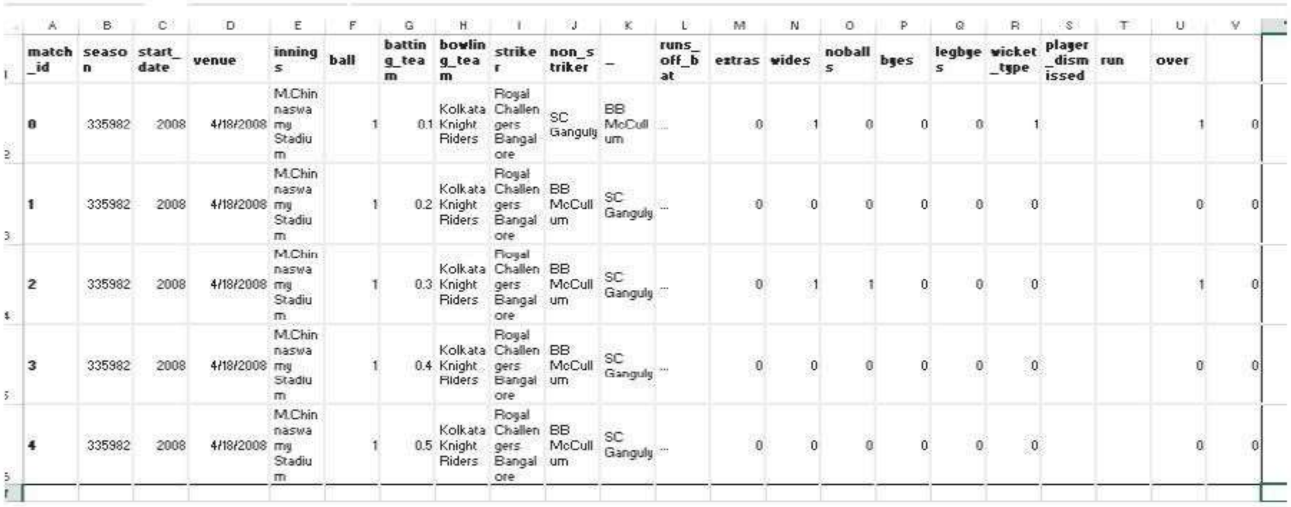
The logistic regression statistic modeling technique is used when we have a binary outcome variable. For example: given the parameters, will the student pass or fail? Will it rain or not? etc. So, though we may have continuous or categorical independent variables, we can use the logistic regression modeling technique to predict the outcome whenthe outcome variable is binary.



V.IPL DATASET

Importing IPL Dataset

We have imported the CSV dataset below with the help of pandas read csv functions We can see the content of the dataset by using head() function.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	match_id	season	start_date	venue	innings	ball	batting_team	bowling_team	striker	non_striker	_	runs_off_bat	extras	wides	noballs	byes	legbyes	wicket_type	player_dismissed	run	over	
2	0	335982	2008	4/18/2008	M.Chinnaswamy Stadium	1	0.1	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	...	0	1	0	0	0	1			1	0
3	1	335982	2008	4/18/2008	M.Chinnaswamy Stadium	1	0.2	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	...	0	0	0	0	0	0			0	0
4	2	335982	2008	4/18/2008	M.Chinnaswamy Stadium	1	0.3	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	...	0	1	1	0	0	0			1	0
5	3	335982	2008	4/18/2008	M.Chinnaswamy Stadium	1	0.4	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	...	0	0	0	0	0	0			0	0
6	4	335982	2008	4/18/2008	M.Chinnaswamy Stadium	1	0.5	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	...	0	0	0	0	0	0			0	0

VI.IPL DATA ANALYSIS AND VISUALIZATION WITH PYTHON

i) General Analysis of IPL Matches

1. List of Seasons

We can get the list of seasons from the dataset by applying unique() function on the season column which confirms thatour dataset contains data of matches played from season 2008-2020.The data set we have includes the data of each and every match played from season 2008 to 2021.

2. First ball of IPL history

Each data point describes the match_id, season, start_date, venue, innings, ball, batting_team, bowling_team, striker, non_striker, bowler, runs_off_bat, extras, wides, no balls, byes, leg byes, wicket_type, player_dismissed, run which areself-explanatory

3. First ball of IPL history

Each data point describes the match_id, season, start_date, venue, innings, ball, batting_team, bowling_team, striker, non_striker, bowler, runs_off_bat, extras, wides, no balls, byes, leg byes, wicket_type, player_dismissed, run which areself-explanatory out of it by dropping the first index layer that is the match_id. We can see the visualization of the IPL matches using the Matlotlib library.

4. Most IPL Matches played in a Venue

The analysis shows most of the IPL matches were played in Chennai, Mumbai, Kolkata, Banglore, and Delhi.

5. IPL Matches Played by Each Team

We can find out the matches played by each team by the same process which is grouping the batting_team and the match_id column and counting the data and then dropping the first index layer which is match_id.

ii) IPL Batting Analysis

6. Most Run Scored by IPL Teams

To calculate the most run scored by a team across all seasons we have grouped by Team and have summed up the runscored by them. And finally, sort them in descending order. Without any surprise, MI is at the top of the list.

7. Most IPL Runs by a Batsman

From the below visualization we can see that the Run-Machine, Virat Kohli is at the top of this list with more than 6,000 runs followed by Suresh Raina and Shikhar Dhawan.

8. Avg Run by Teams in Powerplay

Team Delhi Capital has the best average in the powerplay with an average of 48 runs followed by SRH and RPS.

9. Most IPL Century by a Player

The Universe Boss Chris Gayle is at the top of the list in scoring the most number of centuries in IPL history. He has hitsix tons and has scored 4804 runs in IPL.His former teammate Virat Kohli has scored five hundred's and he is at the second spot in the list followed by Watson, AB de Villiers, Brendon McCullum, and David Warner. This can be calculated by grouping the columns striker and match_id and then calculating the sum.

10. Most IPL Fifty by Player

When a number of the fifties comes Warner is top in the list followed by Virat Kohli and Shikhar Dhawan. This willalso be calculated by the same method as above, plus we have shown a bar graph visualization for better representation.

11. Orange Cap Holder Each Season

The batsman with the most runs in the tournament during the course of the season would wear the Orange Cap while fielding, with the overall leading run-scorer at the conclusion of the tournament winning the actual Orange Cap awardon the day of the season's final. Shaun Marsh became the first winner of the award in 2008, the complete list is presented below from the dataset.

12. Most Sixes in an IPL Inning

Chris Gayle has hit the highest number of sixes in an inning with the number being 17 in the entire IPL history.

13. Most Boundary (4s) hit by a Batsman

The Indian Gabbar, Shikhar Dhawan is at the top of the list with more than 600 boundaries followed by Virat Kohli andDavid warner.

14. Most runs in an IPL season by Player

The run machine, Virat Kohli is at the top of the list with 973 runs in 2016 season followed by David Warner and Kane Williamson with 848 and 735 runs in the 2016 and 2018 season respectively.

15. No. of Sixes in IPL Seasons

2018 is the season with the most number of sixes hit. Followed by season 2019 and 2020 in the list of most sixes in aseason.

16. Highest Total by IPL Teams

Royal Challengers Bangalore is at the top of the list of highest run by a team. The match was played against Pune Warrior in the 2019 season.

17. Most IPL Sixes Hit by a batsman

The universe Boss, Chris gale is at the top of the list in the most hitting sixes followed by AB De Villiers and MSDhoni

18. Highest Individual IPL Score

Chris Gayle playing against Pune Warrior has hit the highest individual score in the 2013 season. Brendon Mc Cullum and Ab de Villiers are in the second and third positions on the list.

iii) Bowling Statistics

19. Most run conceded by a bowler in an inning

Basil Thampi playing for SRH against RCB in the 2008 season has conceded 70 runs and is at the top of the listfollowed by Bangladesh player Mujeeb Ur Rahman and Ishant Sharma.

20. Purple Cap Holders

The bowler with the most wickets in the tournament during the course of the season would wear the Purple Cap while fielding, with the overall leading wicket-taker at the conclusion of the tournament winning the actual Purple Cap awardon the day of the season's final. Below is the list of bowlers with purple caps.

21. Most IPL Wickets by a Bowler

Srilankan bowler Malinga is at the top of the list with 170 wickets followed by Amit Mishra and Push Chawla with 160and

156 wickets respectively.

22. Most Dot Ball by a Bowler

The Indian bowler Harbhajan Singh has bowled the most number of Dot balls followed by R. Ashwin and BhuvneshwarKumar

23. Most Maiden over by a Bowler

Indian right-hand medium-pacer bowler Praveen Kumar is at the top of the list with the most maiden overs followed by Irfan Pathan and Dale Stain.

24. Most Wickets by an IPL Team

The Mumbai Indian has taken the most number of wickets in IPL followed by Royal Challengers Banglore andChennai Super Kings

25. Most No Balls by an IPL team

Royal Challengers Bangalore has given most no balls followed by Mumbai Indians and Chennai Super Kings

26. Most No Balls by an IPL Bowler

Indian bowler S Sreesanth has bowled the most number of no balls followed by Jasprit Bumrah and Amit Mishra

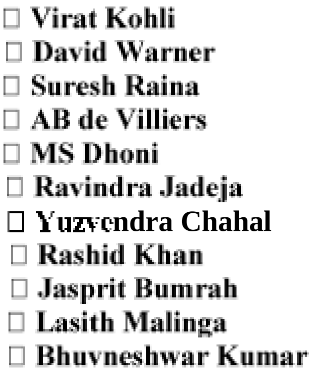
27. Most run given by a team in Extras

Mumbai Indians have given the most number of extras (byes, no balls, wides) followed by Kolkata Knight Riders andKings XI Punjab.

VI.RESULT

In This project we have analyse the ipl data from 2008 to 2020 and on that basis we make our team that's name is KCE 11 by using linear and logistics regression Algorithm. We will predict that which player is best for our team. Suppose in future we take any ipl franchise for that franchise we want best ipl team, by using this we can create newstrongest team by using linear and logistic regression algorithm graph.

KCE 11 BY USING LINEAR REGRESSION



□ Virat Kohli
□ David Warner
□ Suresh Raina
□ AB de Villiers
□ MS Dhoni
□ Ravindra Jadeja
□ Yuzvendra Chahal
□ Rashid Khan
□ Jasprit Bumrah
□ Lasith Malinga
□ Bhuvneshwar Kumar

VII.CONCLUSION

In this project we learn about Python language for interpreted to the data and that Excel sheet we converted into csv by using Python language Syntax and short this csv file by using library in python like Pandas numpy seaborn matlblib etc, and we saw thedata like venue of matches number of teams of IPL most sixes by player in IPL total Seasons played and also bowling data.our project done 80% because lack of time.