Tech saksham

Case Study Report

Data Analytics with Power BI

“IPL Analysis using Power BI”

Karuppannan Mariappan College

Muthur - 638105

NM ID	NAME
8527A5E0C640CC2A7AA4711AC1CF115D	A. SNIDHU KAMATCHI

Trainer Name: R.UMAMAHESWARI

Master Trainer: R.UMAMAHESWARI

ABSTRACT

Sports Analytics is a blooming sector in the field of Computer Science. Cricket is one of the most popular team games in the world. With this project, we embark on predicting the outcome of Indian Premier League (IPL) cricket match which is the biggest carnival of T20 format in the world of cricket. This project aims at designing an effective result prediction system for a cricket match. The result of a T20 cricket match depends on lots of In-game and pre-game attributes, like venue, Past trackrecords and toss influence the results of the match predominantly. This project also aims to emphasize on exploratory data analysis, modelling and visualization of data regarding the Indian Premier League. Best possible outcome of a given match will be predicted using different supervised machine learning (Random Forest Classifier) and statistical approaches. For easy access and usage of the outcome, this will be hosted on a userfriendly web application that can run on any browser

INDEX

INTRODUCTION

1.1 INTRODUCTION

The game of cricket is played in various formats, i.e., One Day International, T20 and Test Matches. The Indian Premier League (IPL) is a Twenty-20 cricket tournament league established with the objective of promoting cricket in India and thereby nurturing young and talented players. The league is an annual event where teams representing different Indian cities compete against each other. It was started by the Board of Control for Cricket in India (BCCI) and has now become a giant, remunerative cricket venture. The teams for IPL are selected by means of an auction. Players' auctions are not a new phenomenon in the sports world. However, in India, selection of a team from a pool of available players by means of auctioning of players was done in Indian Premier League (IPL) for the first time. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcome of matches is very important for all stake holders. This, in turn, is dependent on the complex rules governing the game, luck of the team (Toss),the ability of players and their performances on a given day. Various other natural parameters, such as the historical data related to

players, play an integral role in predicting the outcome of a cricket match.

1.2 OUTLINE OF THE PROJECT

Statistical Modelling and Data Mining tools are being used in Sports Analytics and prediction vividly now a days. This gives us an opportunity to analyse and predict the outcome of a game (like – Indian Premier League) using different visualization tools and machine learning algorithms.

Cricket has been established as one of the most followed outdoor game in the world; over 1.5 billion people watch cricket worldwide including Asia, Australia, Europe, Africa etc. In India itself cricket has over 766 Million viewers who love to watch the sport.

Cricket has had many evolutions over time; in 2005 Cricket saw the inception of it's shortest and the most entertaining format of the game called T20.

The idea of Indian Premier League was conceived in 2007 after the first successful T20 World Cup with the objective of promoting T20 cricket in India and thereby nurturing young and talented players.

It was started by BCCI (Board of Control for Cricket in India) and now has become a massive, remunerative annual venture and considered as the best of all the T20 Leagues in the world.

In this tournament 8 different teams representing different provinces of India play in a Round Robin fashion for the ulterior motive of winning the prestigious trophy

SERVICES AND TOOLS REQUIRED

Services Used

- **Data Collection and Storage Services**: Banks need to collect and store customer data in real-time. This could be achieved through services like Azure Data Factory, Azure Event Hubs, or AWS Kinesis for real-time data collection, and Azure SQL Database or AWS RDS for data storage.

- **Data Processing Services**: Services like Azure Stream Analytics or AWS Kinesis Data Analytics can be used to process the real-time data.

- **Machine Learning Services**: Azure Machine Learning or AWS SageMaker can be used to build predictive models based on historical data.

Tools and Software used

Tools:

- **PowerBI**: The main tool for this project is PowerBI, which will be used to create interactive dashboards for real-time data visualization.

- **Power Query**: This is a data connection technology that enables you to discover, connect, combine, and refine data across a wide variety of sources.

Software Requirements:

- **PowerBI Desktop**: This is a Windows application that you can use to create reports and publish them to PowerBI.

- **PowerBI Service**: This is an online SaaS (Software as a Service) service that you use to publish reports, create new dashboards, and share insights.

- **PowerBI Mobile**: This is a mobile application that you can use to access your reports and dashboards on the go.

LITERATURE SURVEY

With the evolution of Cricket, it became a very hot topic for sports analysts. A lot of research has been made on cricket but due to inconsistent and complicated data sets, they could not get breakthrough in predicting match winner accurately.

There are many techniques that has been used in predicting match winner like KNN, Logistic Regression, SVM, Naïve Bayes but nobody has achieved the accuracy.

According to Ahmed &Nazir [1] they implemented different statistical approaches for formation of datasets and tried various classification techniques to predict the winner of One Day Cricket (50 over) match.

He has predicted the winner with 80 % accuracy. Shah predicted One Day International match. In Features combination to predict the match outcome, is relative strength of Team B divided by relative strength of Team A is successful in measuring and comparing the strength of the playing teams.

Implemented Logistic Regression on this data and achieved accuracy in predicting the results by using data of ICC match ratings, ICC ranking points for batsmen and bowlers, home factor, ICC rating differences and ground effects on the match.

The machine learning based approach used in [5] is reached at by an in-depth analysis of T20 cricket features. In order to indicate the players' performance, a novel index, namely Deep Performance Index (DPI) is derived using the characteristics specific to T20 cricket.

The authors extract relevant features using the machine learning algorithm of Recursive Feature elimination for designing the DPI.

It is demonstrated that DPI achieves better results in analysis of performance related data for batsmen as well as bowlers in comparison to some other ranking methods for T20 cricket.

There exist some other approaches [6,7] which have specifically worked upon IPL data

AIM AND SCOPE

3.1 AIM OF THE PROJECT

This project aims at designing an effective result prediction system for a cricket match. The result of a T20 cricket match depends on lots of In-game and pre-game attributes, like venue, Past track-records and toss influence the results of the match predominantly.

This project also aims to emphasize on exploratory data analysis, modelling and visualization of data regarding the Indian Premier League. Best possible outcome

of a given match will be predicted using different supervised machine learning (Random Forest Classifier) and statistical approaches.

For easy access and usage of the outcome, this will be hosted on a user-friendly web application that can run on any browser.

3.2 OBJECTIVEAND SCOPE

To predict the outcome of an IPL match.It also aims to analyse and visualize data using various data visualisation techniques for better understanding.

The data has to be preprocessed and fed to various supervised machine learning algorithms and analysed in accordance to their accuracies.

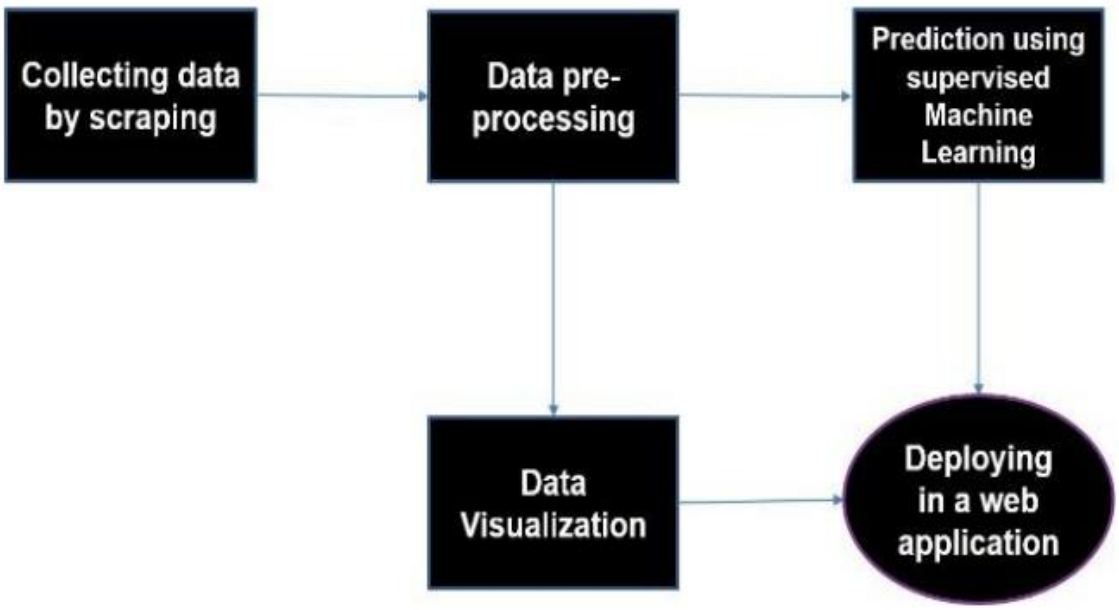
The best possible outcome will be predicted using a perfect model and will be hosted in a user-friendly web application

SYSTEM IMPLEMENTATION

4.1 SYSTEM ARCHITECTURE:

The proposed system aims to analyse the data generated by IPL matches and predict the outcome of the match (one Pre-Toss and then Post-Toss). The steps followed are

- Collecting data by scraping
- Data pre-processing
- Prediction using supervised learning algorithm (Random Forest Classifier)
- Data Visualization
- Deploying in a web application



4.2 METHODS AND MODEL DETAILS:

This project mainly has three parts:

- IPL Data Analytics (Team and Player Stats)
- Pre-Toss Prediction
- Post-Toss Prediction

4.2.1 IPL DATA ANALYTICS:

As the process of analysing raw data to find trends and answer questions, the definition of data analytics captures its broad scope of the field. However, it includes many techniques with many different goals.

The data analytics process has some components that can help a variety of initiatives.

By combining these components, a successful data analytics initiative will provide a clear picture of where you are, where you have been and where you should go. Statistics have always had a significant role in sports.

As I mentioned above, sports analytics is on the rise and will continue to play a significant role in how teams operate, pick their players, how they play the game, etc.

Cricket is no different. The runs scored by a batsman, the wickets taken by a bowler, or the matches won by a cricket team – these are all examples of the most important numbers in the game of cricket.

Maintaining a record of all such statistics has multiple benefits. The teams and the individual players can dig deep into this data and find areas of improvement.

It can also be used to assess an opponent’s strengths and weaknesses.Data analytics is a broad field. There are four primary types of data analytics: descriptive, diagnostic, predictive and prescriptive analytics.

Each type has a different goal and a different place in the data analysis process. These are also the primary data analytics applications in business.

- Descriptive analytics helps answer questions about what happened. These techniques summarize large datasets to describe outcomes to stakeholders. By developing key performance indicators (KPIs,) these strategies can help track successes or failures. Metrics such as return on investment (ROI) are used in many industries. Specialized metrics are developed to track performance in specific industries. This process

- Diagnostic analytics helps answer questions about why things happened. These techniques supplement more basic descriptive analytics. They take the findings from descriptive analytics and dig deeper to find the cause. The performance indicators are further investigated to discover why they got better or worse. This generally occurs in three steps

❖ Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.
❖ Data that is related to these anomalies is collected.
❖ Statistical techniques are used to find relationships and trends that explain these anomalies.

- Predictive analytics helps answer questions about what will happen in the future. These techniques use historical data to identify trends and determine if they are likely to recur. Predictive analytical tools provide valuable insight into what may happen in the future and its techniques include a variety of statistical and machine learning techniques, such as: neural networks, decision trees, and regression.

- Prescriptive analytics helps answer questions about what should be done. By using insights from predictive analytics, data-driven decisions can be made. This allows businesses to make informed decisions in the face of uncertainty. Prescriptive analytics techniques rely on machine learning strategies that can find patterns in large datasets. By analysing past decisions and events, the likelihood of different outcomes can be estimated.

- These types of data analytics provide the insight that businesses need to make effective and efficient decisions. Used in combination they provide a well-rounded understanding of a company's needs and opportunities.

- The primary goal of a data analyst is to increase efficiency and improve performance by discovering patterns in data. The work of a data analyst involves working with data throughout the data analysis pipeline.

- This means working with data in various ways. The primary steps in the data analytics process are data mining, data management, statistical analysis, and data presentation. The 8 importance and balance of these steps depend on the data being used and the goal of the analysis.

- Data mining is an essential process for many data analytics tasks. This involves extracting data from unstructured data sources. These may include written text, large complex databases, or raw sensor data.

- The key steps in this process are to extract, transform, and load data (often called ETL.) These steps convert raw data into a useful and manageable format.

- This prepares data for storage and analysis. Data mining is generally the most time-intensive step in the data analysis pipeline. Data management or data warehousing is another key aspect of a data analyst's job. Data warehousing involves designing and implementing databases that allow easy access to the results of data mining.

- This step generally involves creating and managing SQL databases. Non-relational and NoSQL databases are becoming

- more common as well. Statistical analysis allows analysts to create insights from data.

- Both statistics and machine learning techniques are used to analyse data. Big data is used to create statistical models that reveal trends in data. These models can then be applied to new data to make predictions and inform decision making. Statistical programming languages such as R or Python (with pandas) are essential to this process. In addition, open-source libraries and packages such as TensorFlow enable advanced analysis.

- The final step in most data analytics processes is data presentation. This step allows insights to be shared with stakeholders. Data visualization is often the most important tool in data presentation. Compelling visualizations can help tell the story in the data which may help executives and managers understand the importance of these insights.

4.2.2 MATCH PREDICTION:

- The next part of the project is the prediction part where both the Pre toss and Post toss prediction is done using the Supervised machine learning algorithmssuch as Multiple

Linear Regression and Random Forest Classifier algorithm. Multiple Linear Regression:It's a form of linear regression that is used when there are two or more predictors.Itis the most common form of linear regression analysis.

- As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
- The independent variables can be continuous or categorical Here, Y is the output variable, and X terms are the corresponding input variables. Notice that this equation is just an extension of Simple Linear Regression, and each predictor has a corresponding slope coefficient (β).
- The first β term (βo) is the intercept constant and is the value of Y in absence of all predictors (i.e., when all X terms are 0). It may or may or may not hold any significance in a given regression problem.
- It's generally there to give a relevant nudge to the line/plane of regression. Regression residuals must be normally distributed. A linear relationship is assumed between the dependent variable and the independent variables.
- The residuals are homoscedastic and approximately rectangular-shaped. Absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated. At the centre of the multiple linear regression analysis is the task of fitting a single line through a

scatter plot. More specifically the multiple linear regression fits a line through a multi-dimensional space of data points. The simplest form has one dependent and two independent variables. The dependent variable may also be referred to as the outcome variable or regressand.

- The independent variables may also be referred to as the predictor variables or regressors. There are 3 major uses for multiple linear regression analysis. First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable.
- Second, it can be used to forecast effects or impacts of changes. That is, multiple linear regression analysis helps us to understand how much will the dependent variable change when we change the independent variables.
- Third, multiple linear 10 regression analysis predicts trends and future values. The multiple linear regression analysis can be used to get point estimates. When selecting the model for the multiple linear regression analysis, another important consideration is the model fit.
- Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable (typically expressed as R^2).
- Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model. Using Multiple Linear Regression in this project, the

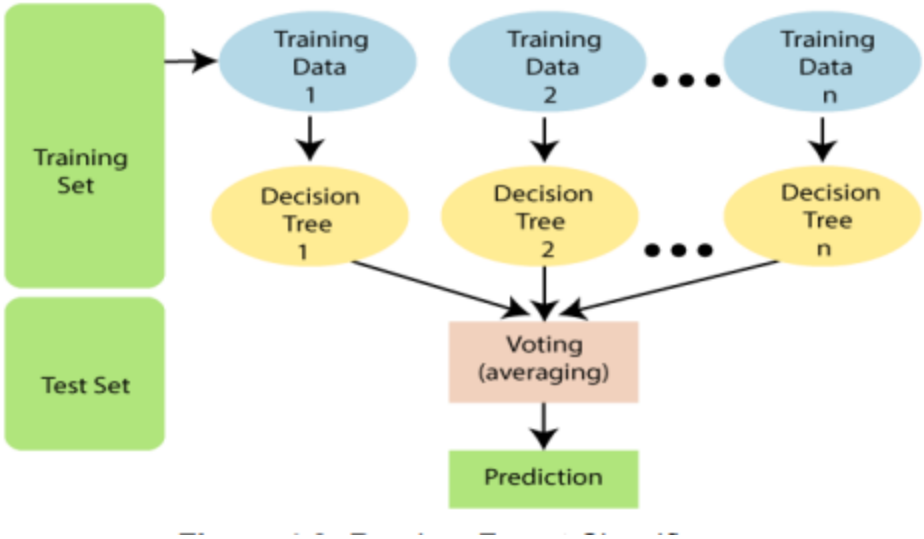
outcome of a match is predicted two times. Once, before the toss, without taking into consideration the toss decision (PreToss).

- The model takes in the Team name as input and create a linear regression model (team names are encoded), to give the output of the prediction. On the other hand, the Post-Toss takes other factors like toss winner and toss decision into consideration for predicting the match outcome. Random Forest Algorithm: Random forest algorithm is a flexible machine learning algorithm that produces great results even without hyper-parameter tuning. Apart from being simple to use, it is extremely accurate also. It is basically a supervised learning algorithm.
- A large number of decision trees operate together; an individual tree in a random forest model gives some prediction and finally the one with most votes becomes the prediction of the model.

thus even if some trees in the group are wrong, all the tree

s

are able to move in correct direction given that many other trees will be right.



Random Forest works in four steps:

- ❖ Select random samples from a given dataset.
- ❖ Construct a decision tree for each sample and get a prediction result from each decision tree.
- ❖ Perform a vote for each predicted result.

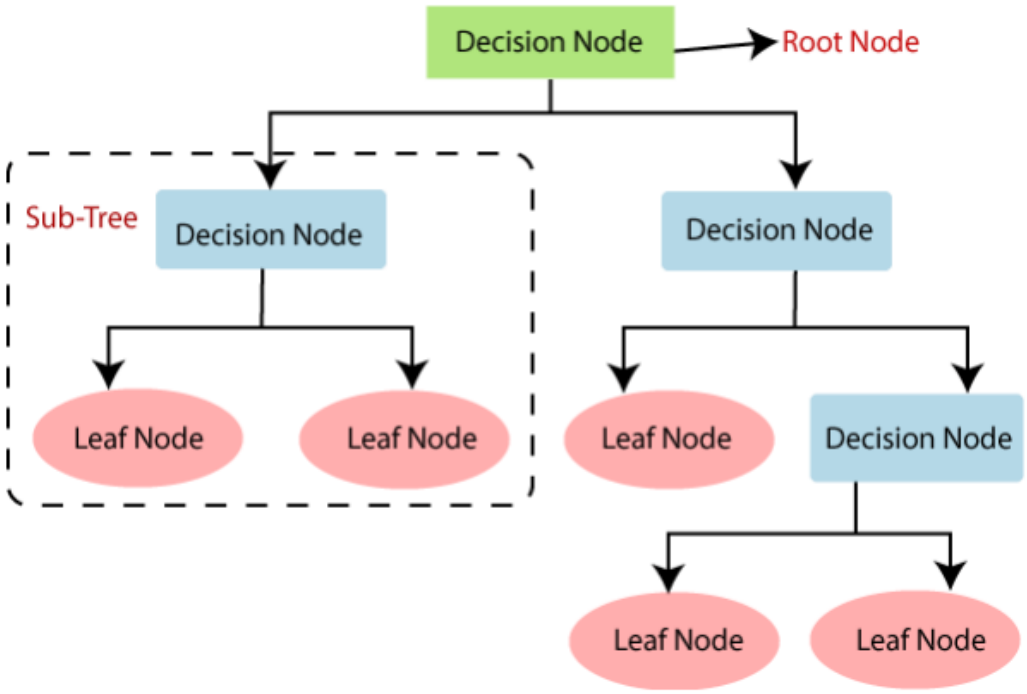
❖ Select the prediction result with the most votes as the final prediction.

Decision Trees:

- The classification technique is a systematic approach to build classification models from an input dataset. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers are different technique to solve a classification problem. Each technique adopts a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. Therefore, a key objective of the learning algorithm is to build predictive model that accurately predict the class labels of previously unknown records.
- Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. 12 Build an optimal decision tree is key problem in decision tree classifier.
- In general, may decision trees can be constructed from a given set of attributes. While some of the trees are more accurate than others, finding the optimal tree is computationally infeasible because of the exponential size of the search space. However,

various efficient algorithms have been developed to construct a reasonably accurate, albeit suboptimal, decision tree in a reasonable amount of time.

- These algorithms usually employ a greedy strategy that grows a decision tree by making a series of locally 17 optimum decisions about which attribute to use for partitioning the data. For example, Hunt's algorithm, ID3, C4.5, CART, SPRINT are greedy decision tree induction algorithms.
- The decision tree inducing algorithm must provide a method for specifying the test condition for different attribute types as well as an objective measure for evaluating the goodness of each test condition.



- First, the specification of an attribute test condition and its corresponding outcomes depends on the attribute types. We can do two-way split or multi-way split, discretize or group attribute values as needed. The binary attributes lead to two-way split test condition. For nominal attributes which have many values, the test condition can be expressed into multi way split on each distinct value, or two-way split by grouping the attribute values into two subsets.
- Similarly, the ordinal attributes can also produce binary or multi way splits as long as the grouping does not violate the order property of the attribute values. For continuous attributes, the test condition can be expressed as a comparison test with two outcomes, or a range query. Or we can discretize the continuous value into nominal attribute and then perform two-way or multi-way split. Since there are many choices to specify the test conditions from the given training set, we need use a measurement to determine the best way to split the records.
- The goal of best test conditions is whether it leads a homogenous class distribution in the nodes, which is the purity of the child nodes before and after splitting. The larger the degree of purity, the better is the class distribution. To determine how well a test condition performs, we need to compare the degree of impurity of the parent before splitting with degree of the impurity of the child

nodes after splitting. The larger their difference, the better is the test condition. The measurements of node impurity/purity are:

- Gini Index
- Entropy
- Misclassification Error
- In this project, the outcome of a match is predicted two times. Once, before the toss, without taking into consideration the toss decision (Pre-Toss).
- The model (Random 14 Forest Classifier) takes in the Team name as input and creates an ensemble of decision trees usually trained with “bagging” method, to give the output of the prediction.
- On the other hand, the Post-Toss takes other factors like toss winner and toss decision into consideration for predicting the match outcome in a more accurate fashion

RESULTS AND DISCUSSION

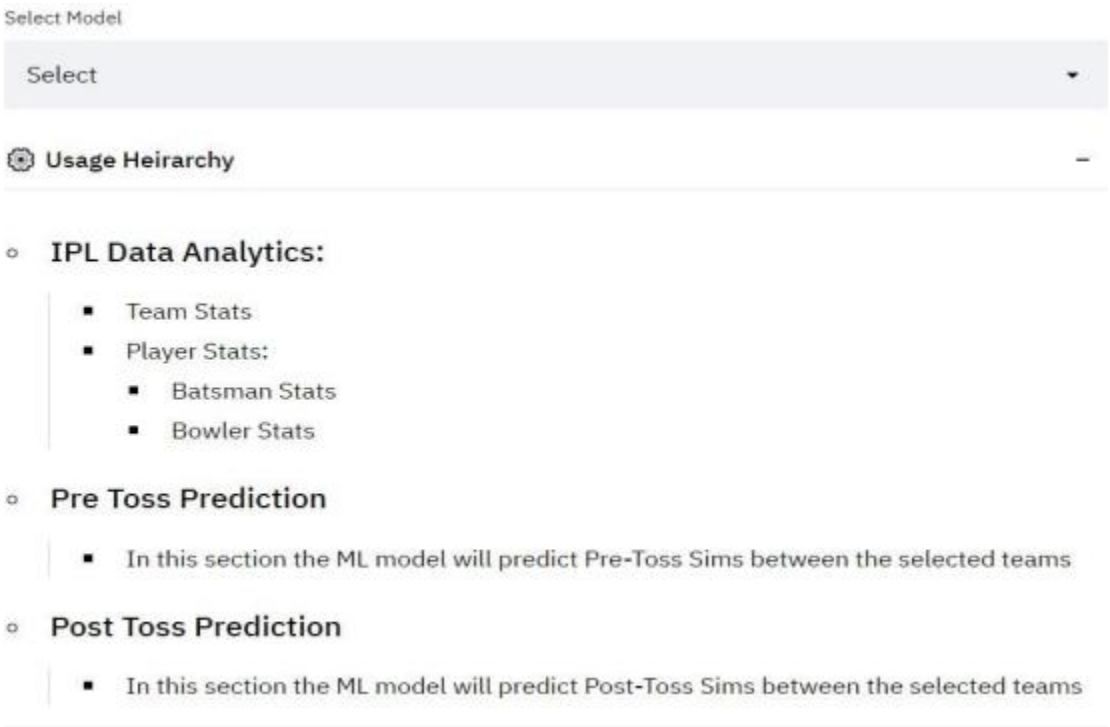
5.1 DATA ANALYTICS:

This paper focuses on predicting the outcome of an IPL match by taking factors like Toss, Toss Decision into consideration along with Data analytics and Visualization of teams and players.

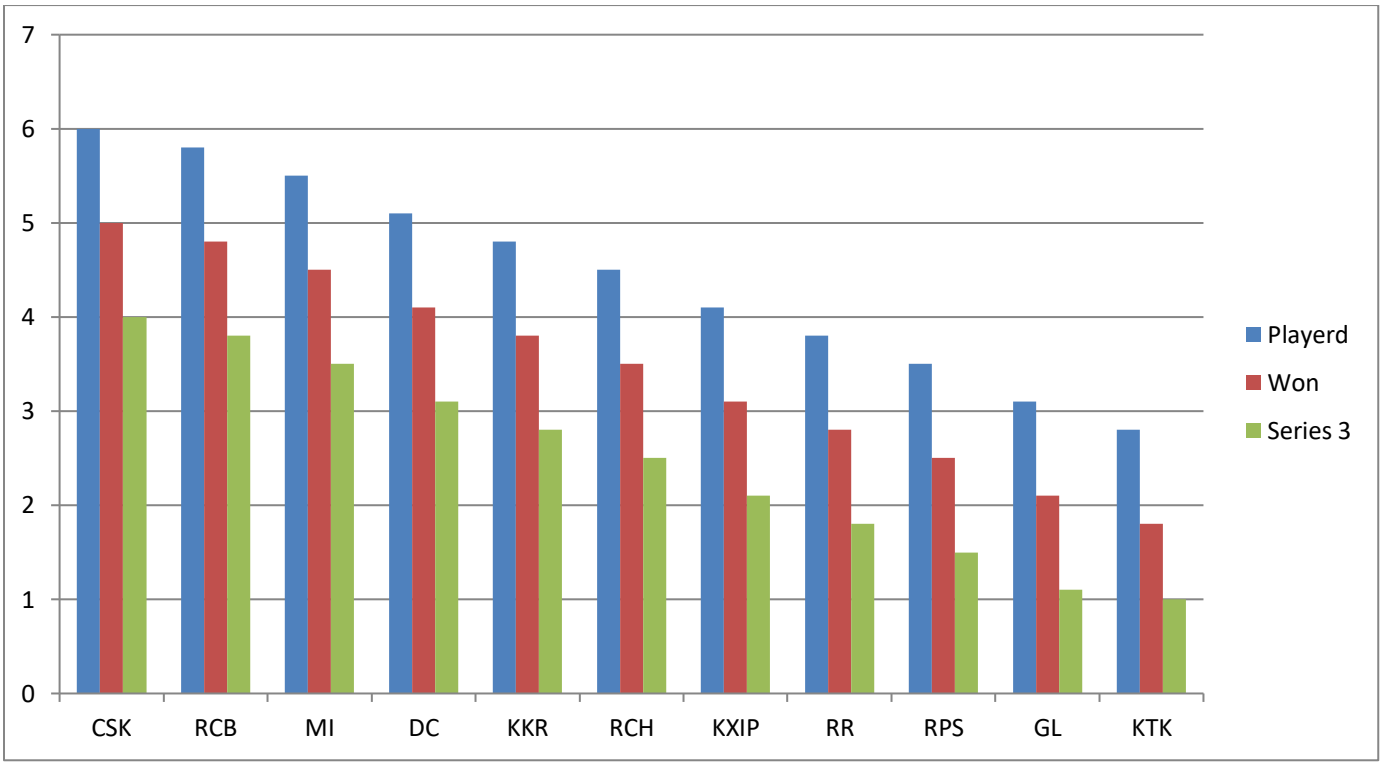
Efficient prediction accuracy of about 84% is achieved in this model with the help of Random Forest algorithm.

All the results and outcomes of the project are hosted in a web application that is user friendly and can run on any web browser.

Welcome to ipl predictions



Represents the Home Page of the web application that can be used by the user for checking the outcome of a particular match as well as visualizing the team stats and player stats.

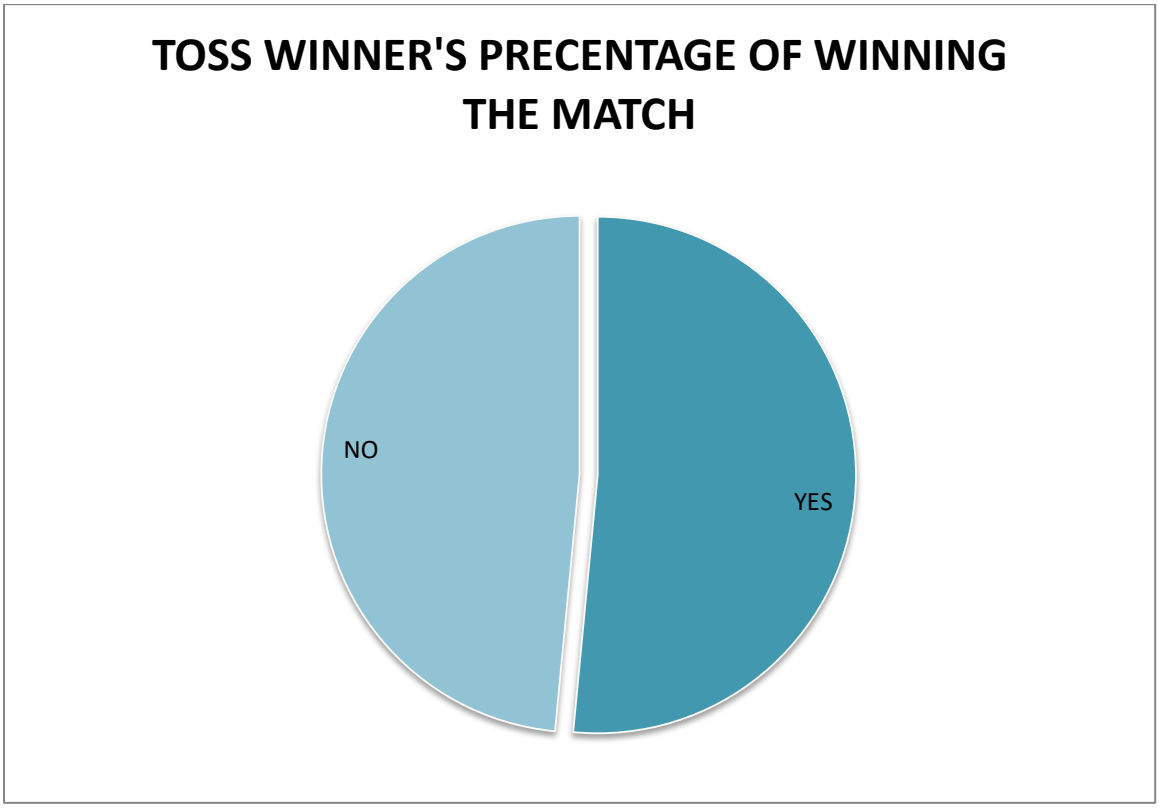


TEAMS

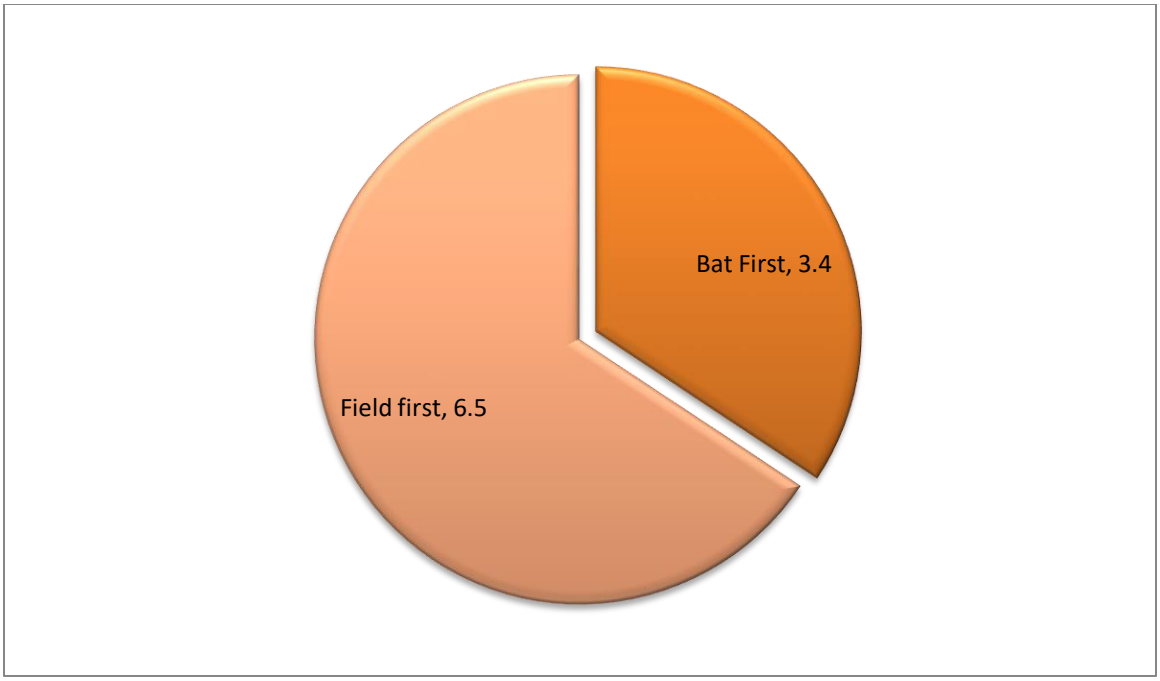
Represents the teamwise analysis with number of matches played, matches won and win percentage of each team in the Y-Axis against the Team names in the X-Axis

Teamwise analysis is very important when it comes to any team sports. The same is true for IPL. Here, through this analysis we can see that MI is the most successful team in IPL.

It has played the most no of matches throughout the IPL. The yellow bar represents that MI has the highest win percentage as well. Similarly, Kochi Tuskers Kerala have played the least matches in IPL, this data is also gives us this insight.



Represents teams that win the toss has 51.2% record of winning the match whereas teams that lose the toss has 48.8% record of winning the match since IPL 2008.



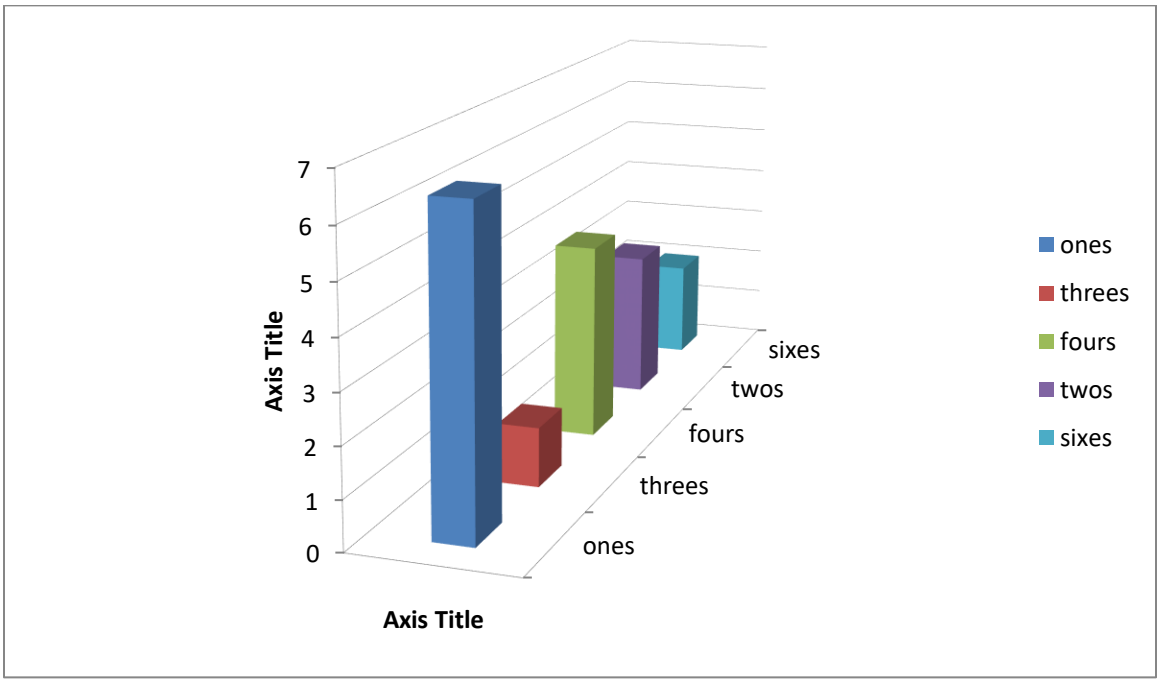
Impact of toss decision

Represents teams that win the toss and elect to bat first has 34.5% record of winning the match whereas teams that win the toss elect to field has 65.5% record of winning the match since IPL 2008. Toss or flip of the coin is one of the most important factors in a cricket match. Unlike other sports Toss plays a huge role in determining the final outcome of the match.

Toss is so important that sometimes the result of whole game is depending upon the Toss and the team that wins the toss wins the match

as well(provided that the captain made the correct decision after winning the toss).

The teams mostly choose the option that is best suited to them (unless the pitch conditions are entirely different) after winning the toss. For example, a team whose strength lies in batting will opt to bowl first after winning the toss most of the times. If the team has a destructive bowling line-up then the toss can be decisive factor in the match.



Represents the runs split of a particular batsman throughout his IPL career (till 2020). The example mentioned here represents the runs split of V Kohli from 2008 to 2020.

CONCLUSION

6.1CONCLUSION:

Statistical Modelling and Data Mining tools are being used in Sports Analytics and prediction vividly now a days. This gives us an opportunity to analyse and predict the outcome of a game (like – Indian Premier League) using different visualization tools and machine learning algorithms. This paper focuses on predicting the outcome of an IPL match by taking factors like Toss, Toss Decision into consideration along with Data analytics and Visualization of teams and players.

To conduct the analysis and predicting the winner of IPL various branches of Data Science has been converged including PreProcessing of data, Visualizations of data, preparation of data, feature selection and implementing different machine learning models for the predictions. SEMMA methodology has been selected for conducting the analysis of IPL T20 match winner dataset.

Future scope

Preprocessing has been done on the dataset to make it consistent by removing missing value, encoding variables into numerical format. Best features were selected by visualizing attributes of data with target variable. On selected features several machine learning models has been applied on the to predict the winner and the results were outstanding. First of all, after the data is cleaned and pre-processed, that data is used to do different data visualization like Team Statistics, Batsman Statistics, Bowler Statistics. The user gets to use the webpage to access any kind of data they need for IPL.

The Data Analysis part is important as it gives insights about the data generated by Indian Premier League. The second part of the project deals with the prediction of the outcome of a match based on factors like previous win record, toss result, toss decision. Firstly, Multiple Linear Regression was used to predict the outcome of a particular match. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. After using Multiple Linear Regression,

predicted the winner with 65% accuracy which was not good enough, so Random Forest Model was also tuned by parameter's tuning and results got better with 73 % accuracy

Models	Accuracy
Multiple Linear Regression	30%
Random Forest Classifier	65%
Random Forest Classifier (Tuned)	73%

Thus finally, both the modules of this project Data Analysis and Outcome Prediction perform well and serve the objective it was supposed to.

References

[1]. Daniel MagoVistro, Faizan Rasheed, Leo Gertrude David, “The Cricket Winner Prediction With Application of Machine Learning And Data Analytics” International Journal of Scientific & Technology Research (2019)

[2]. Madan Gopal Jhanwar and VikramPudi, “Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach” International Institution of Information Technology (2017)

[3]. I. P. Wickramasingheet. al, "Predicting the performance of batsmen in test cricket," Journal of Human Sport & Exercise”, vol. 9, no. 4, pp. (2017)

[4]. R. P. Schumaker, O. K. Solieman and H. Chen, "Predictive Modeling for Sports and Gaming” in Sports Data Mining, vol. 26, Boston, Massachusetts: Springer, (2016)

[5]. J. McCullagh, "Data Mining in Sport: A Neural Network Approach," International Journal of Sports Science and Engineering, vol. 4, no. 3 (2016)


Microsoft
edunet
foundation
SAP
TAMIL NADU SKILL DEVELOPMENT CORPORATION
SKILLS FOR PROGRESS AND PROSPERITY

LINK