IPL 2020 PREDICTION

Sports Analytics

Sports analytics are a collection of statistics of players, Weather conditions, Team's recent wins/lose,
that when properly applied can provide a competitive advantage to a team or individual.

One such great example is ,
Real Madrid — is using Microsoft technology to transform its operations, performance, fitness, and relationships with 500 million global fans.

DATA SCIENCE

01 Domain Understanding

Understand the characteristics of the sport.

02 Data Pre processing

Data Collection, Data Cleaning & transformation

03 EDA & Feature Selection

Exploratory data analysis and selecting the most important variables

04 Modelling

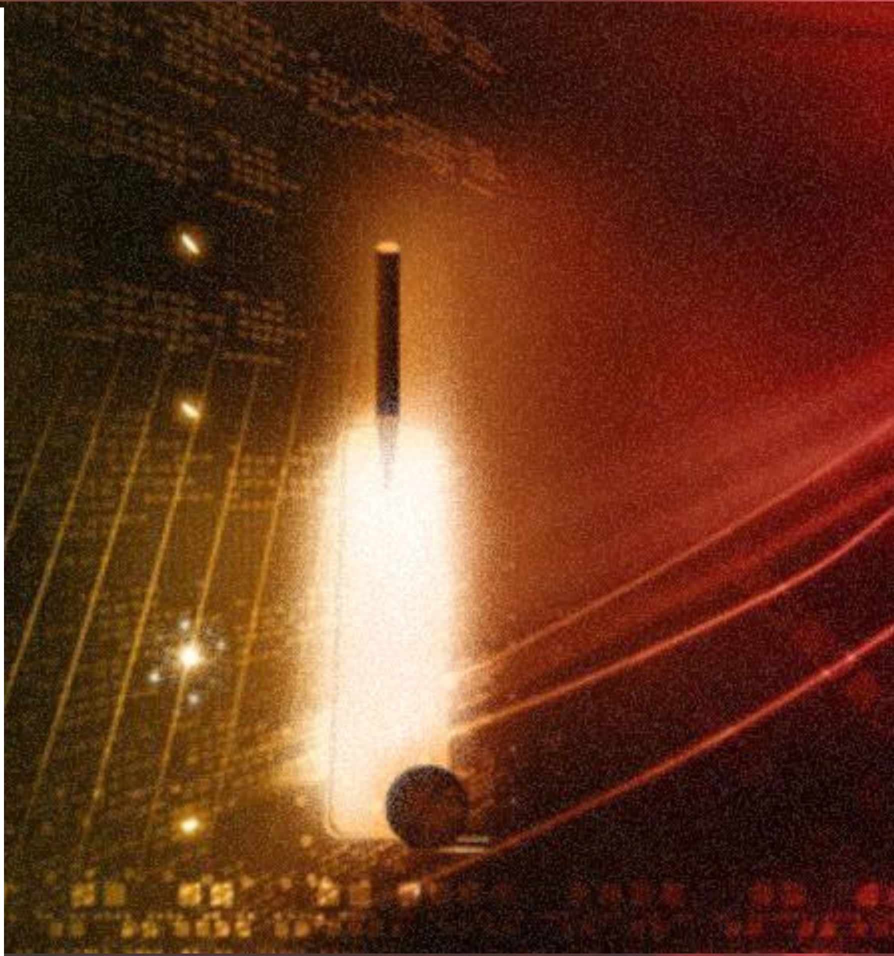
Select candidate models.

05 Model Evaluation

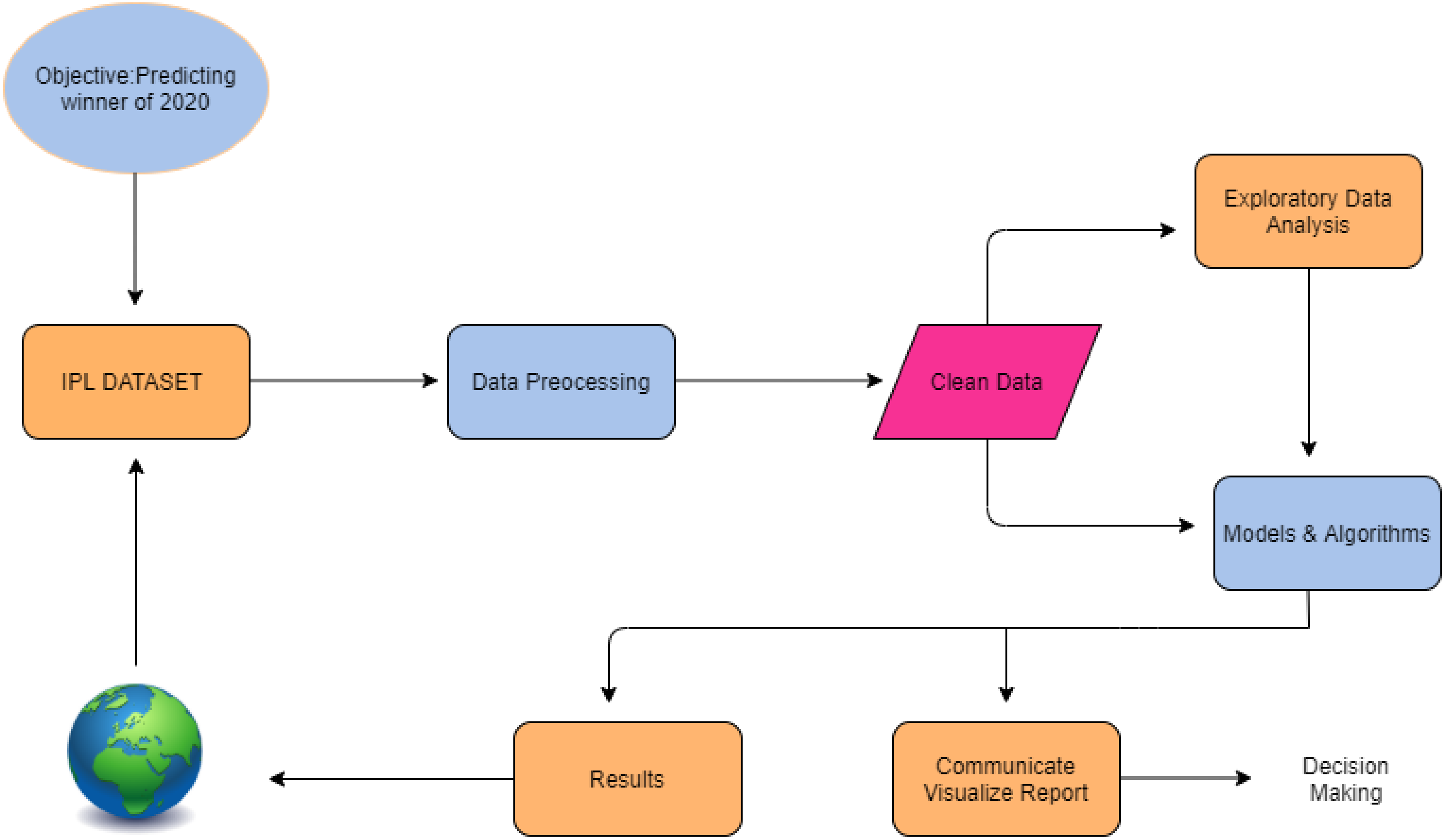
Select measure of model performance

06 Deploy Model

Generate predictions for upcoming matches.



Block Diagram



Objective:Predicting winner of 2020

IPL DATASET

Data Preocessing

Clean Data

Exploratory Data Analysis

Models & Algorithms

Results

Communicate Visualize Report

Decision Making

01

DOMAIN UNDERSTANDING

Indian Premier League

The IPL is a **professional Twenty20 cricket league** in India contested during March or April and May of every year by the Board of Control for Cricket in India(**BCCI**) in 2008.

TEAM

Total of 8 teams particate each year,representing different cities of India

SQUAD COMPOSITION

The squad strength must be between 18 and 25 players, with a maximum of 8 overseas players.

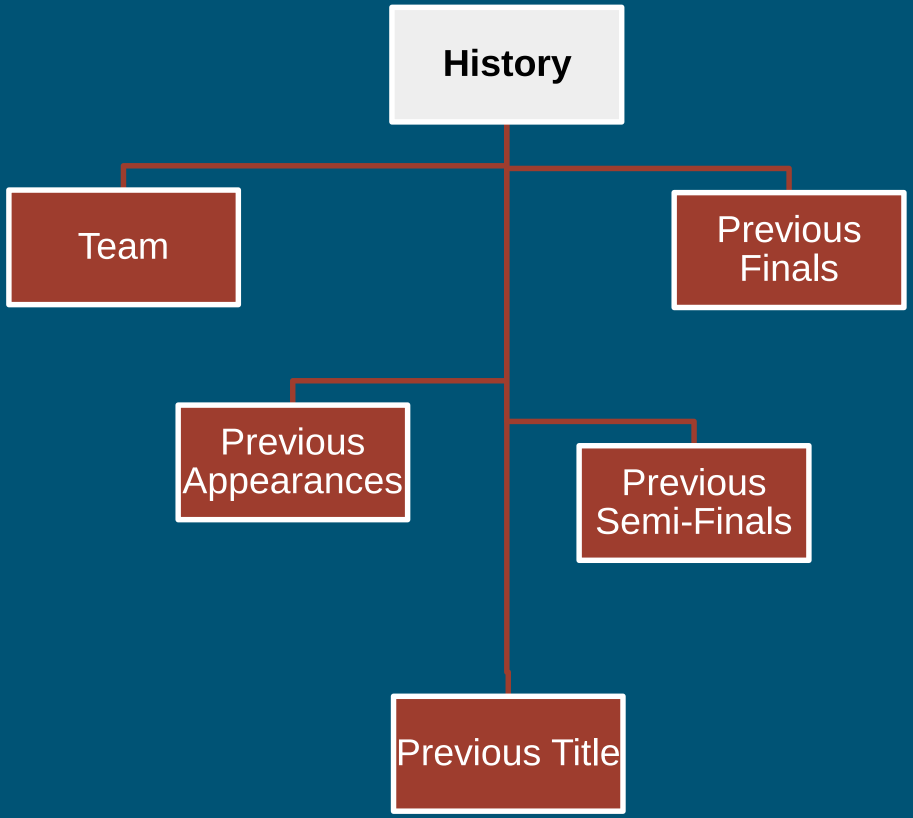
FACT

According to BCCI, the 2015 IPL season contributed ₹11.5 billion (US$160 million) to the GDP of the Indian

02

DATA PRE_PROCESSING

DATA COLLECTION



History

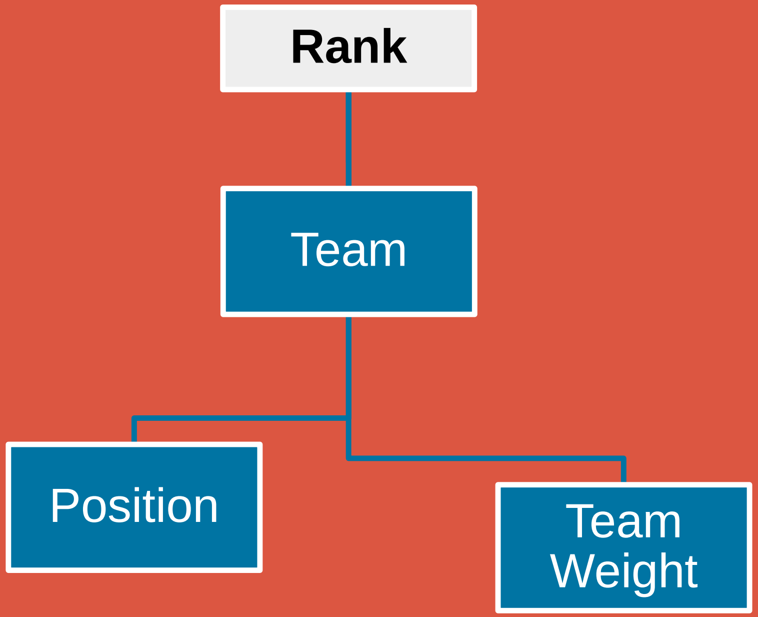
Team

Previous Finals

Previous Appearances

Previous Semi-Finals

Previous Title



Rank

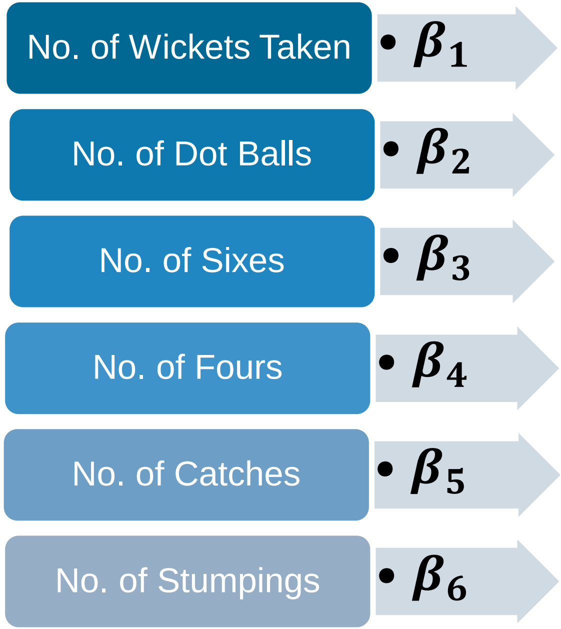
Team

Position

Team Weight

PLAYER POINTS CALCULATION

$$PlayerPoints(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$



No. of Wickets Taken
- β_1

No. of Dot Balls
- β_2

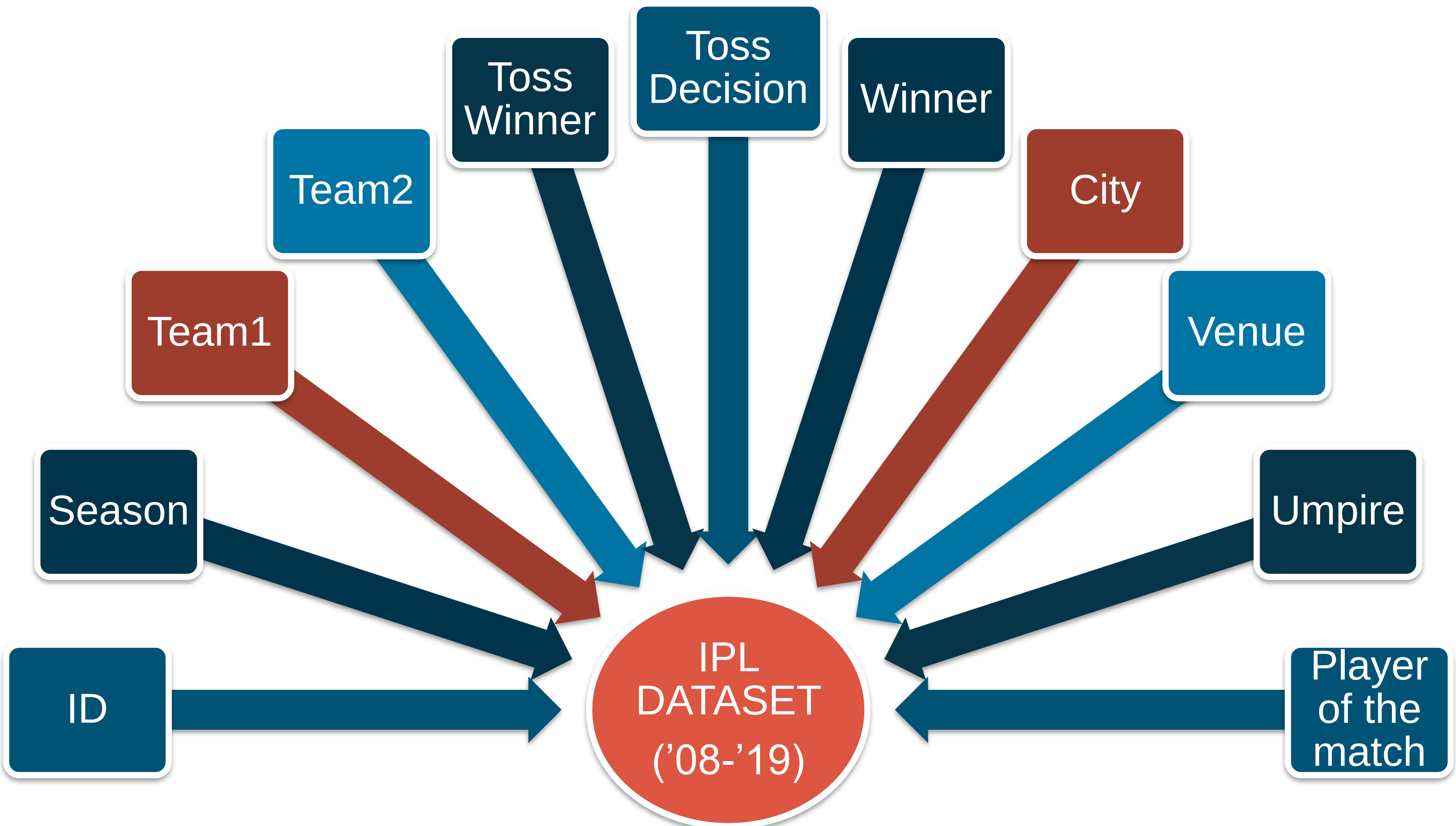
No. of Sixes
- β_3

No. of Fours
- β_4

No. of Catches
- β_5

No. of Stumpings
- β_6

Weight of the team$= \dfrac{\sum_{i=1}^{11} i^{th} Player\ Points}{Total\ apperance\ of\ the\ team}$



ID

Season

Team1

Team2

Toss Winner

Toss Decision

Winner

City

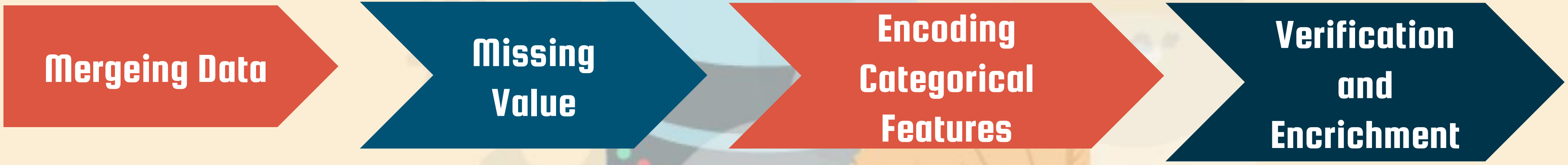
Venue

Umpire

Player of the match

IPL DATASET ('08-'19)

DATA CLEANING



Mergeing Data

History dataset was merged with IPL dataset with the function “concat”

Missing Value

The rows with Missing values were eliminated from the data

Encoding Categorical Features

“Ordinal Encoder and Label Encoder” were used to covert categorical data to intergers

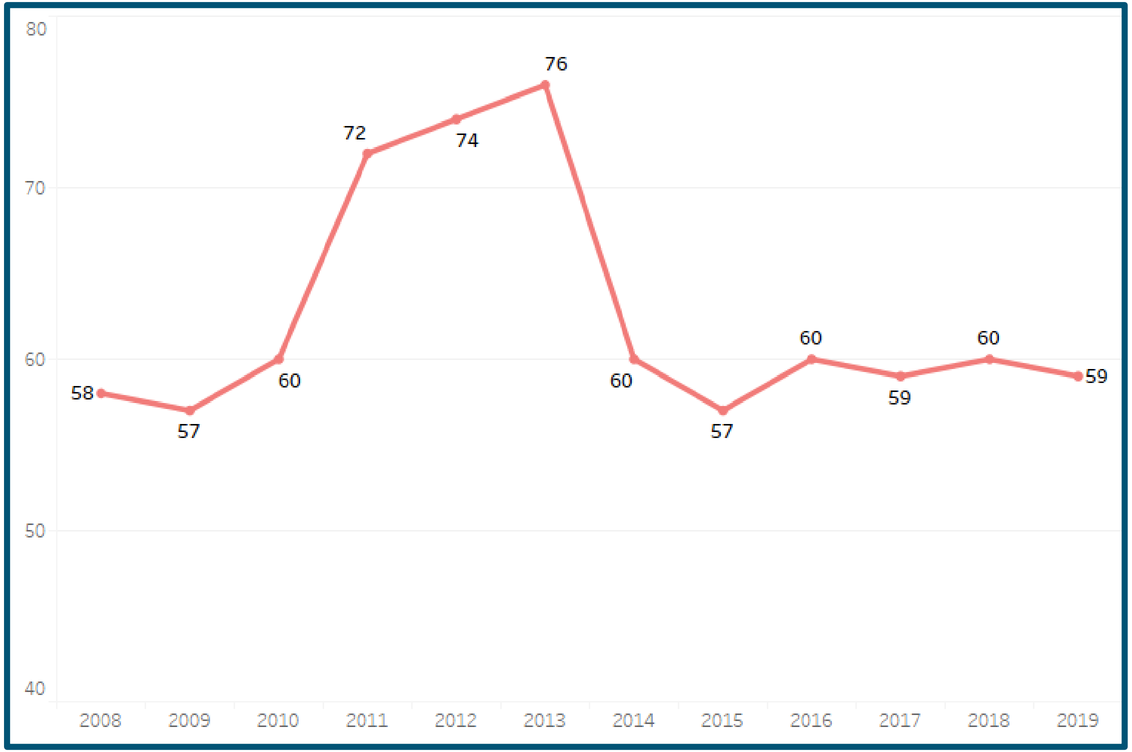
Verification and Encrichment

The final ready data was inspected and Uniformity was maintained

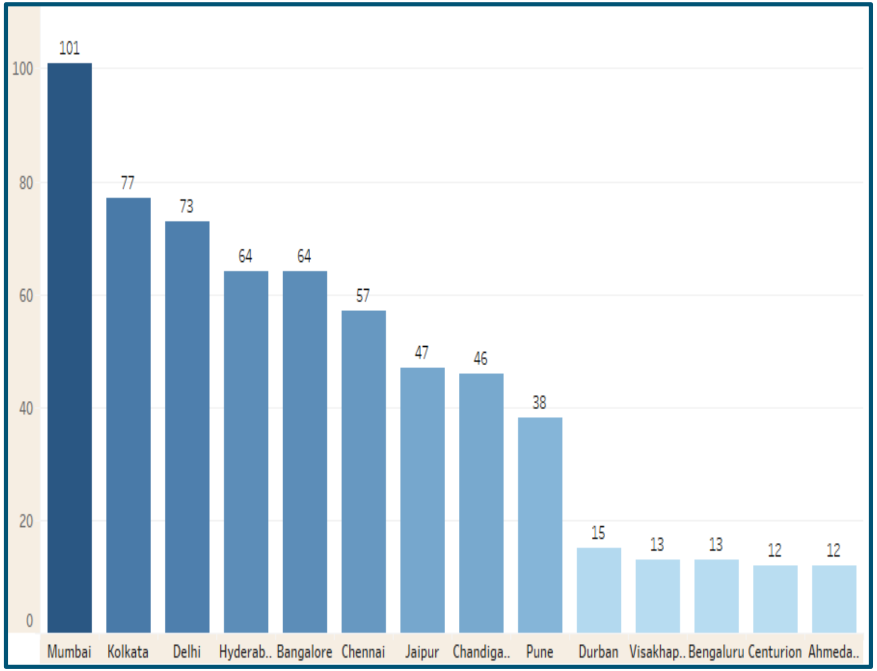
03

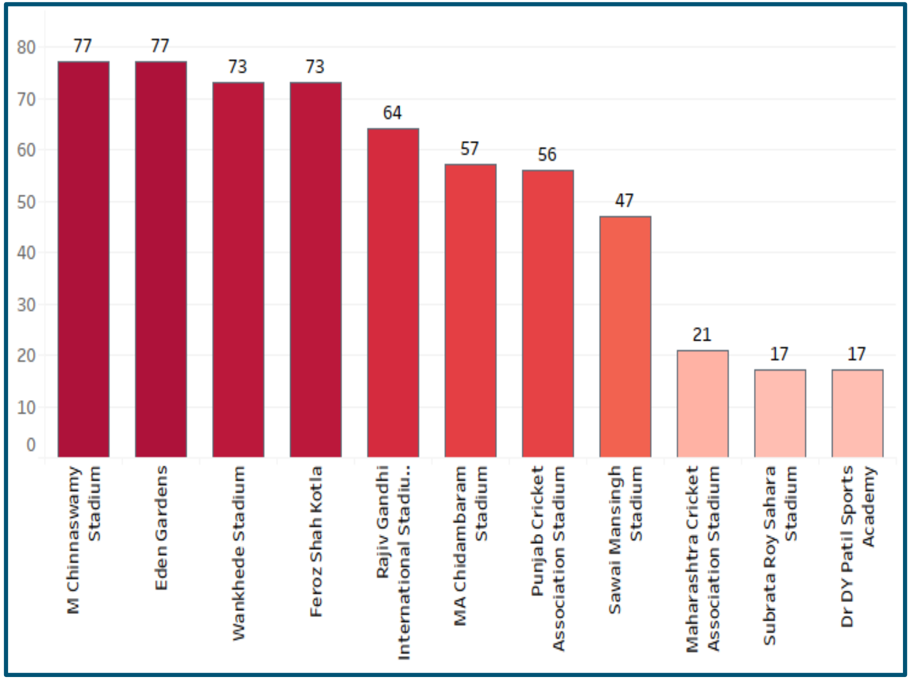
EDA & Feature Selection

MATCHES EVERY SEASON

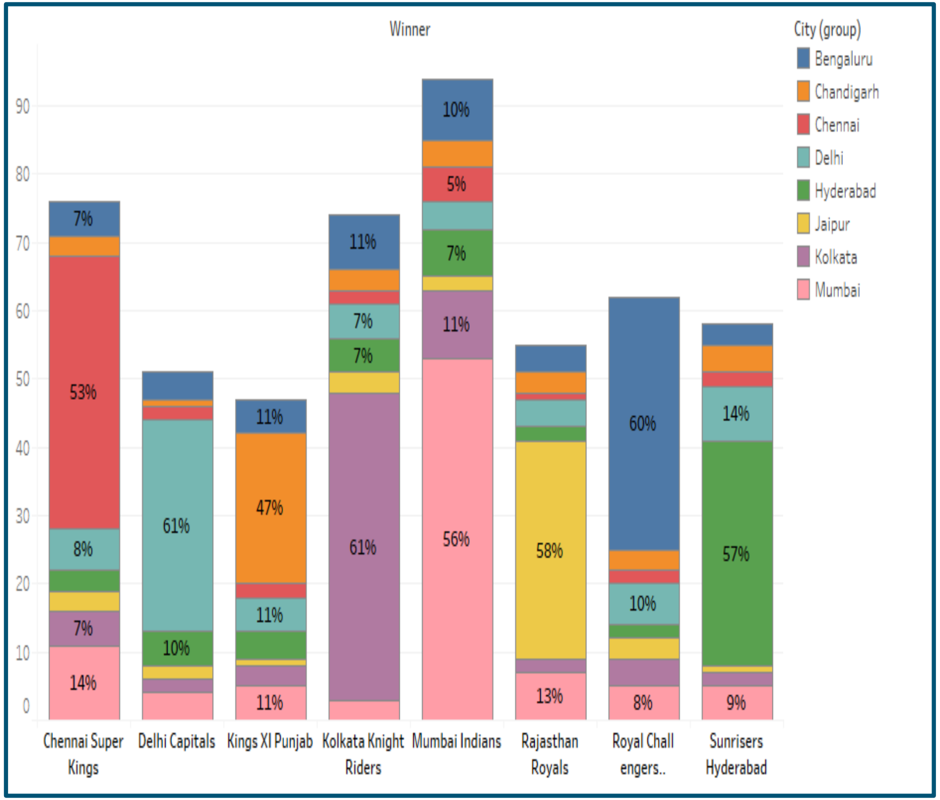


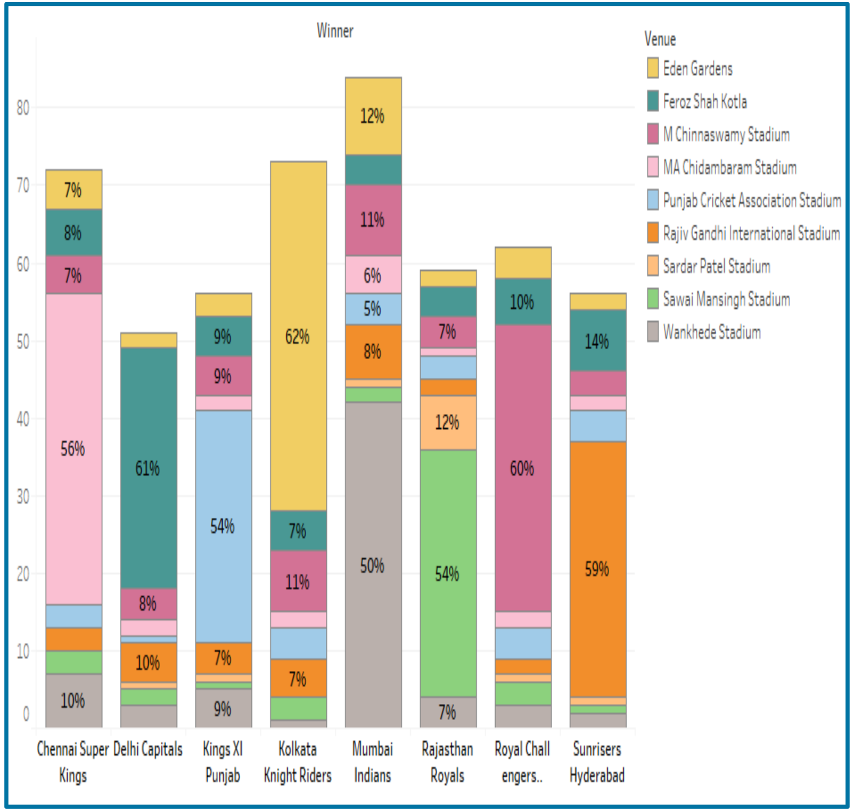
VENUES AND CITIES



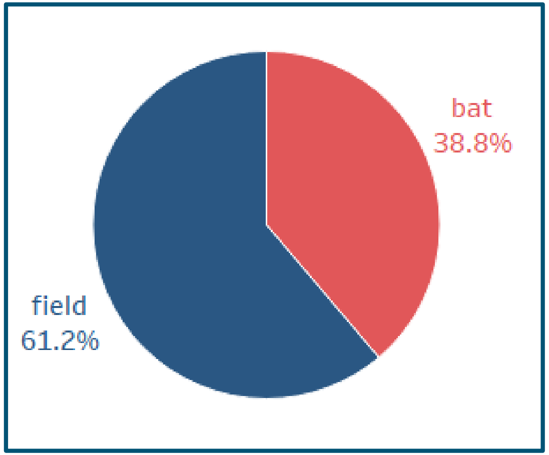


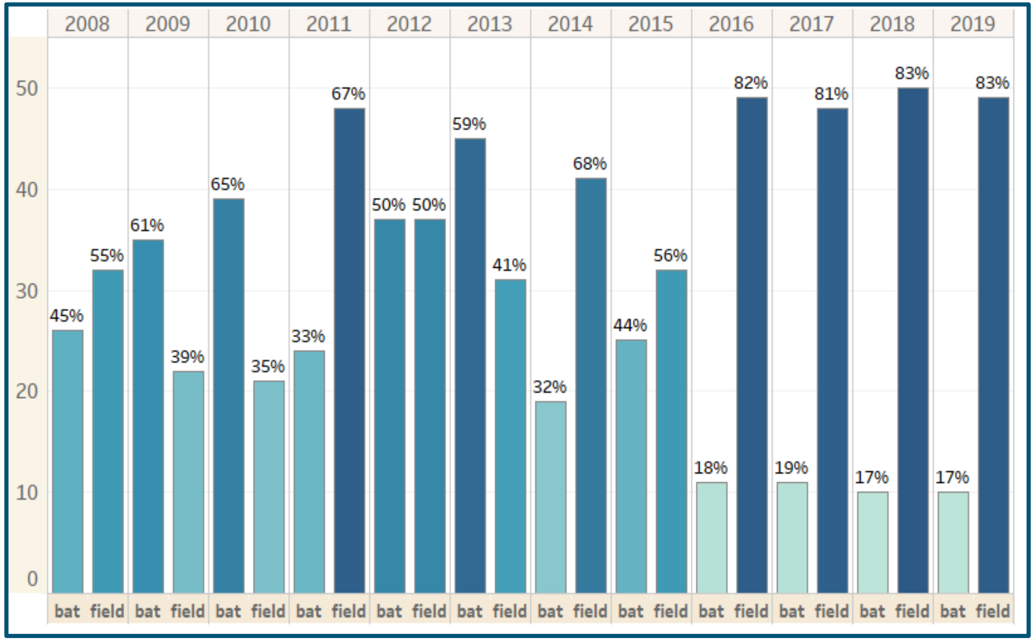
VENUES AND CITIES



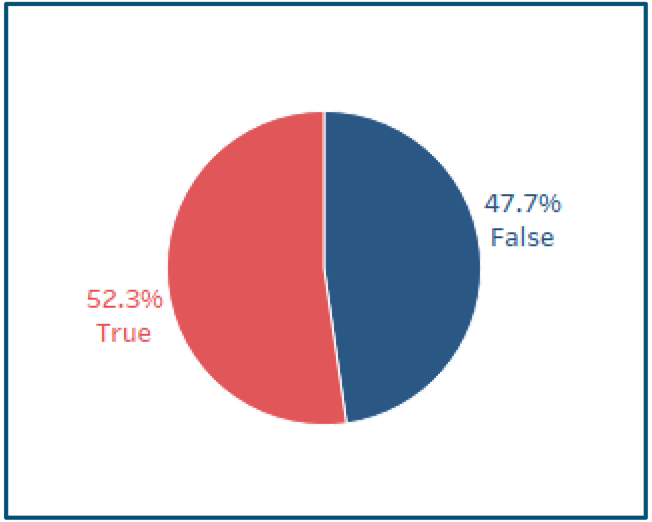


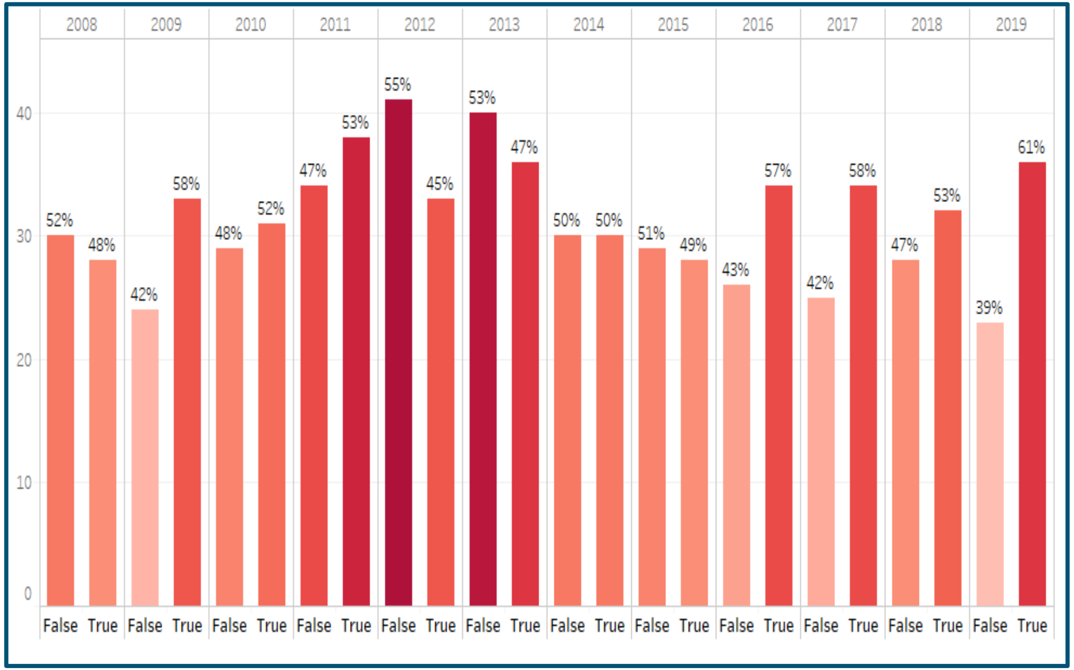
TOSS


bat
38.8%
field
61.2%

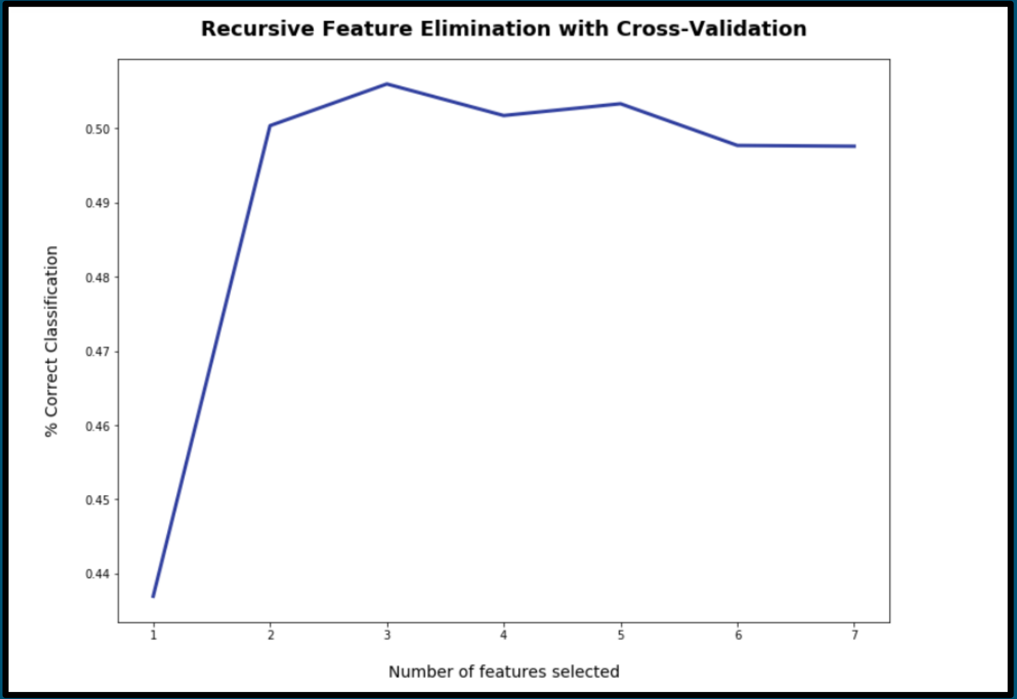

2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
50
40
30
20
10
0
45%
55%
61%
39%
65%
35%
33%
67%
50%
50%
59%
41%
32%
68%
44%
56%
18%
82%
19%
81%
17%
83%
17%
83%
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field
bat
field

Does winning the toss means winning the game?





Feature Selection



TEAM 1

TEAM 2

CITY

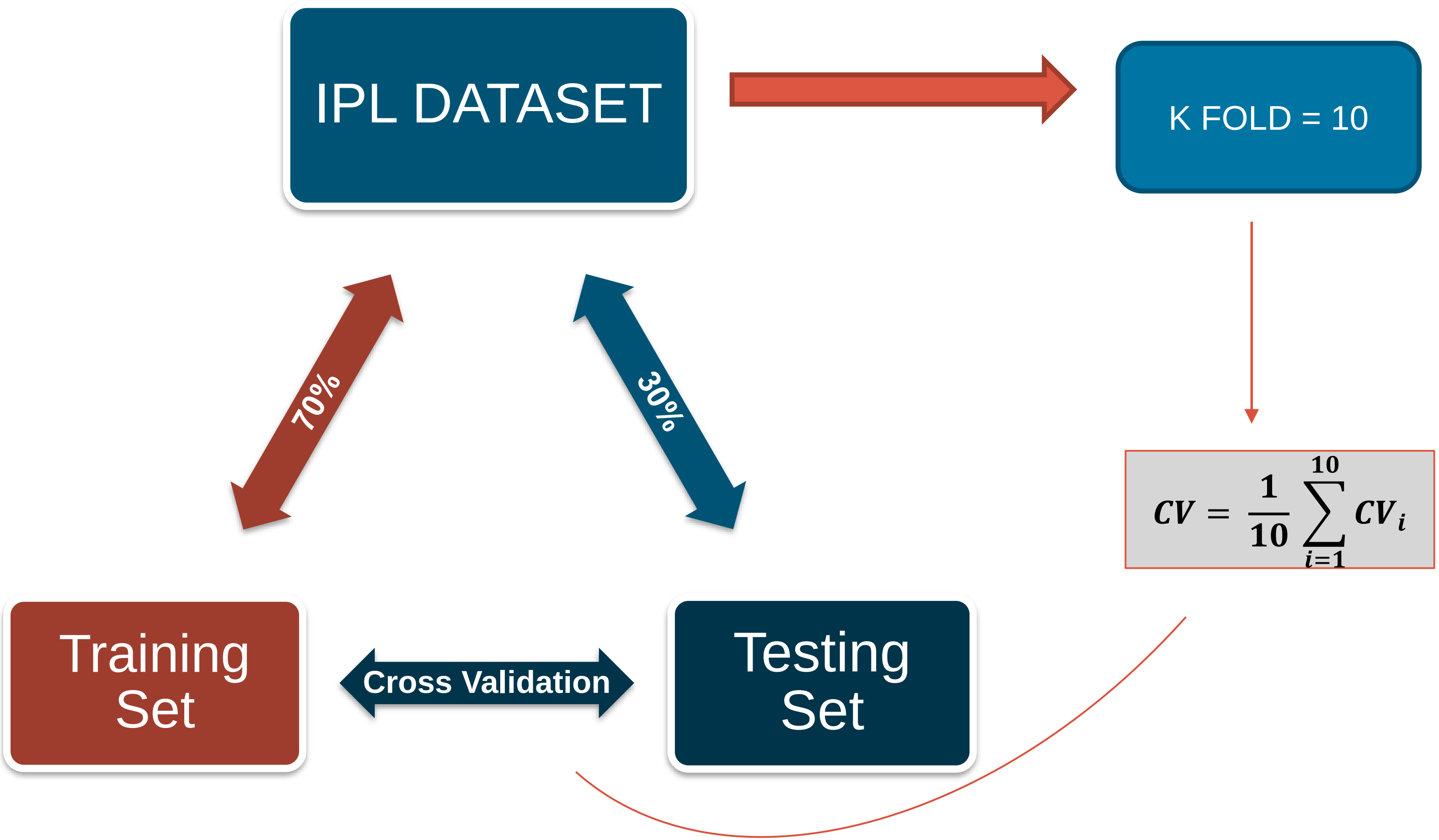
TOSS WINNER

VENUE

TOSS DECISION

04

Modelling



IPL DATASET

K FOLD = 10

70%

30%

Training Set

Cross Validation

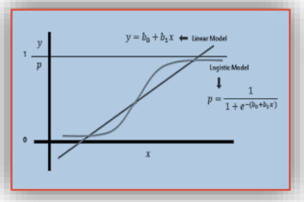
Testing Set

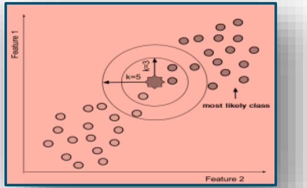
$$CV = \frac{1}{10}\sum_{i=1}^{10} CV_i$$

Model Selection

Logistic Regression

It uses Logistic function to the regression to get the probabilities of it belonging in either class(winner/loser).



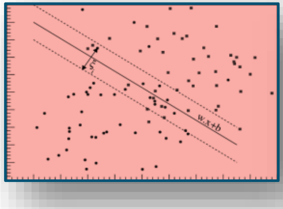


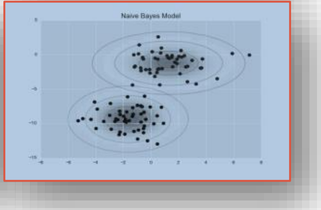
K Nearest Neighbors

It is used to identify the data points that are separated into several classes to predict the classification of a new sample point.

Support Vector Machine

It performs classification by finding the hyperplane that maximizes the margin between classes



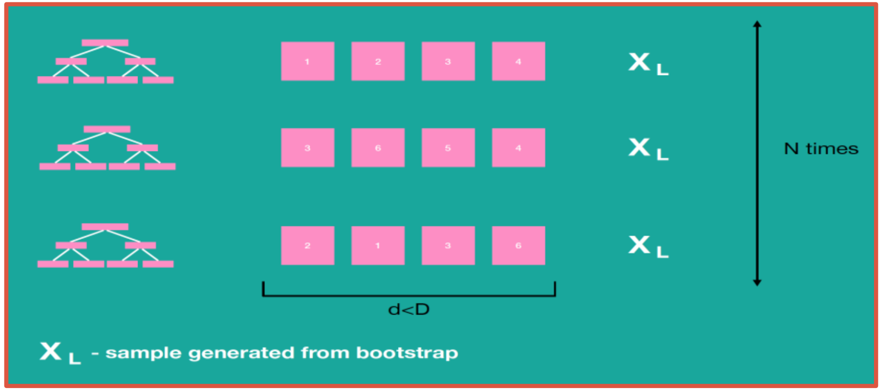


Gaussian Naive Bayes

Based on naive Bayes, Gaussian naive Bayes is used for classification based on the binomial (normal) distribution of data. The probability of a data point having either class, given the data point.

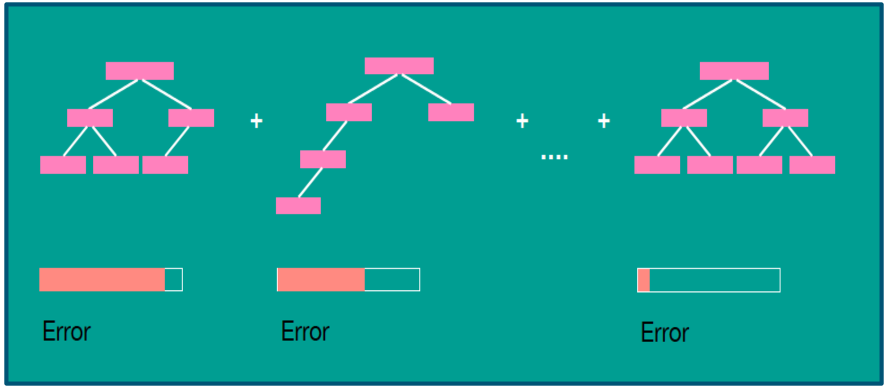
Ensemble Methods

Bagging



RANDOM FOREST CLASSIFIER

Boosting

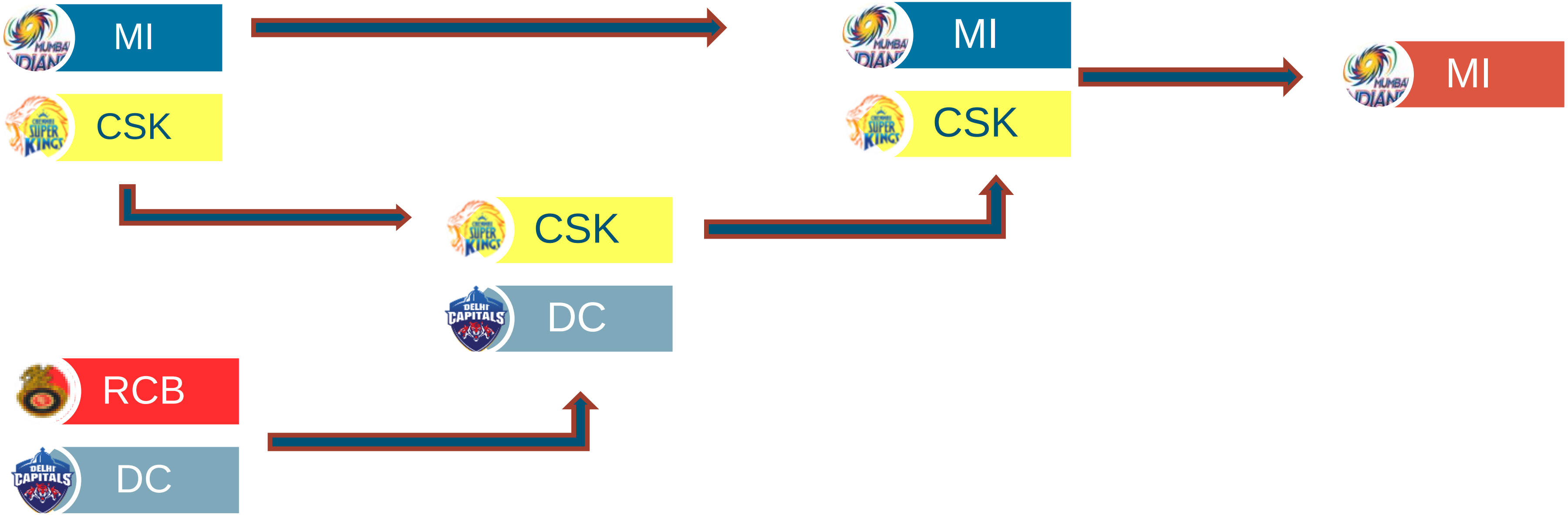


EXTREME GRADIENT BOOST CLASSIFIER

Classifier	Correct Prediction(out of 59)	Model Accuracy (2019)
Logistic Regression	19	35.8%
Gaussian Naive Bayes	21	40.1%
K Nearest Neighbour	26	64.9%
Support Vector Machine	32	85.3%
Random Forest	40	90.0%
XG Boost	42	90.9%

TEAM	PREDICTED WINS	ACTUAL WINS
Mumbai Indians	12	9
Chennai Super Kings	11	9
Royal Challengers Bangalore	**8**	**5**
Delhi Capitals	7	9
Sunrisers Hyderabad	6	6
Kolkata Knight Riders	6	6
Kings XI Punjab	5	5
Rajasthan Royals	4	5

PREDICTIVE PLAYOFFS AND WINNER 2019



MI

CSK

RCB

DC

CSK

DC

MI

CSK

MI

05

Model Evaluation

Hyper Tuning of XG Boost Classifier

An ensemble learning strategy that trains a series of weak models, each one attempting to correctly predict the observations the previous model got wrong.

Parameter	From	To
Learning Rate	0.1	0.3
Max_depth	3	6

Before Tuning

80.3%

After Tuning

83.2%

Testing Accuracy – 71.8%

Assumptions

- It is assumed that the entire squad of the team would be available for selection in every match.
- Player injuries have not been taken into consideration.
- It is assumed every match will result in an outcome i.e. external forces such as rain will not have any impact on the outcome of the match.
- Any kind of fixing involving players has not been considered.

06

Deploy Model

POINTS TABLE

TEAM	MATCHES	WON	LOST	POINTS
CSK	14	11	3	22
MI	14	9	5	18
KKR	14	8	6	16
RCB	14	8	6	16
SRH	14	6	8	12
RR	14	6	8	12
DD	14	5	9	10
KXIP	14	3	11	6

Winner of Vivo IPL 2020

LIMITATIONS

- The model is not real-time, as a result the toss factor could not be used for predicting IPL 2020 results
- The model does not takes into consideration player injuries and washouts which occur due to external forces
- IPL is just a 12 year old league, therefore the sample size of matches is comparatively less

FUTURE SCOPE

- Converting this model in a real-time model will improve the accuracy as essential factor like Toss Winner , Toss Decision and Changing Player Points can be considered.

- Going even further and making a model based on player statistics alone with give an idea on each player performance.

- Using this model for predicting other leagues like Test Match, World Cup and even Dream 11.

CONCLUSION

In cricketing field, to achieve the full convergence into data science world, it would require a lot of additional data to meet full picture of analysis. The prediction of winner produced through this project required a lot of domain information and observation .

The Twenty20 format of cricket carries a lot of **randomness**, because a single over can completely change the ongoing pace of the game.

Hence, designing a machine learning model for predicting the match outcome of an auction-based Twenty20 format premier league with a testing accuracy of **71.8%** is highly satisfactory.

THANK YOU

Presented by-

Ananya P

Dhruv Gupta