

## Project Description:

Overview: Since 2008, most of top stars in the cricketing world have been taking a collective break from international and domestic cricket to take part in the Indian Premier League (IPL) that happens mostly in a two-month window between March and May. The purpose of this project will be to predict the outcome of the match based on the team's previous success and failure.

Dataset: The IPL dataset will be taken from <https://www.kaggle.com/nowke9/ipldata>. Understanding the dataset, we get to know that there are two files which are:

1. Matches.csv: Having every match detail from 2008-2019
2. Deliveries.csv: Having ball by ball detail from 2008-2019

I will be cleaning the dataset as some of the values in the data set are duplicate values in some columns that is making the dataset Redundant.

Problem Statement: As many people play fantasy leagues on many websites. They always think that if they at least have an idea about who is going to win they can be more confident. This algorithm will make the prediction based on the previous 11 years data that which team is going to win.

List of Factors that affect the cricket match that is going to be considered.

1. Pitch.
2. Weather.
3. Toss
4. Bat or Field first
5. Home Record and Away Record
6. Star Batsmen and Star Bowler
7. Average Innings Record on particular ground

## Scope of the project:

- Predicting the winner of next season based on the past performances.
- The project uses the data of players and grounds that have been using from 2008-2019. If any new player or ground is added then the data might give less accuracy because of the lack of information.
- In this project, we can see that features like toss or grounds play huge role in the decision factor of whether which team is going to win.
- In the previous projects, particular player statistics was also very important in the outcome of the match.

## **Algorithms we will be using in this project**

**Random Forest Classifiers:** Random Forest Classifier is a supervised learning algorithm. In this project, it will mainly deal with the feature selections.

**Logistic Regression:** It is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

**Support Vector Machine:** It is a predictive analysis data – classification algorithm that assigns new data elements to one of the labelled categories.

**Naïve Bayes:** The problem of classification predictive modelling can be framed as calculating the conditional probability of a class label given a data sample.

## **Tasks:**

1. **Data Cleaning:** The data we have contains NULL values in several columns. We will be using Imputation on Column. We will be using Numpy library to fill in the NULL value.
2. **Converting textual data into Numerical:** We all know that machine only understand binary values so we need to convert the numerical data into binary values. For example, Win becomes 1 and Loss becomes 0.
3. **Use of Label Encoding:** Once the data is converted into the numerical format, we can start analysing the data.
4. **Data Split into Train & Test:** We will be using K fold cross validation to split the data into train and test after building the data. But initially we will be 70–30 split ratio
5. **Train the Model:** After data split for training the model we use Logistic regression, SVM, Naïve Bayes and also Random Forest Testing Model. After training model using various methods, we test them on test data and achieve best precision and recall values for SVM regression.
6. **Feature Importance:** The Features listed above in the proposal play a major role in outcome of the match as we need to select several different features and apply machine leaning algorithms on those features.
7. **Tuning:** To tune the model better we will be using K fold cross validation method to increase the accuracy of the model

Name: Vinit Girkar  
Student Id: 801194118

**References:**

- [http://www.ijirset.com/upload/2019/april/69\\_Analyzing.pdf](http://www.ijirset.com/upload/2019/april/69_Analyzing.pdf)
- [http://ceur-ws.org/Vol-1842/paper\\_06.pdf](http://ceur-ws.org/Vol-1842/paper_06.pdf)