**IPL CRICKET RESULTS REVIEW IN DATA MINING ALGORITHMS**

***Table of Contents***

[**1.** **Introduction** 2](#_Toc33299359)

[**1.1.** **Problem Statement and Domain** 2](#_Toc33299360)

[**1.2.** **List of Requirements** 2](#_Toc33299361)

[**2.** **Related Work** 3](#_Toc33299362)

[**2.1.** **Loading Libraries** 3](#_Toc33299363)

[**2.2.** **Reading input dataset** 3](#_Toc33299364)

[**2.3.** **Get basic information of Data** 3](#_Toc33299365)

[**2.4.** **Data Preparation Steps** 4](#_Toc33299366)

[**2.4.1.** **Checking for null values** 4](#_Toc33299367)

[**2.4.2.** **Cleaning for null values** 5](#_Toc33299368)

[**2.4.3.** **Checking for unique values in each variable** 5](#_Toc33299369)

[**2.5.** **Data Analytics** 7](#_Toc33299370)

[**2.5.1.** **General Analytics** 7](#_Toc33299371)

[**2.5.2.** **Encoding Values** 10](#_Toc33299372)

[**2.5.3.** **Data Frame Describe** 11](#_Toc33299373)

[**2.5.4.** **Significant Correlation among variables** 11](#_Toc33299374)

[**3.** **Methodology** 12](#_Toc33299375)

[**3.1.** **Decision Tree** 12](#_Toc33299376)

[**3.2.** **Random Forest** 13](#_Toc33299377)

[**3.3.** **Naïve Bayes** 13](#_Toc33299378)

[**3.4.** **Stochastic Gradient Descent** 14](#_Toc33299379)

[**3.5.** **K-Nearest Neighbors** 14](#_Toc33299380)

[**3.6.** **Support Vector Machine** 15](#_Toc33299381)

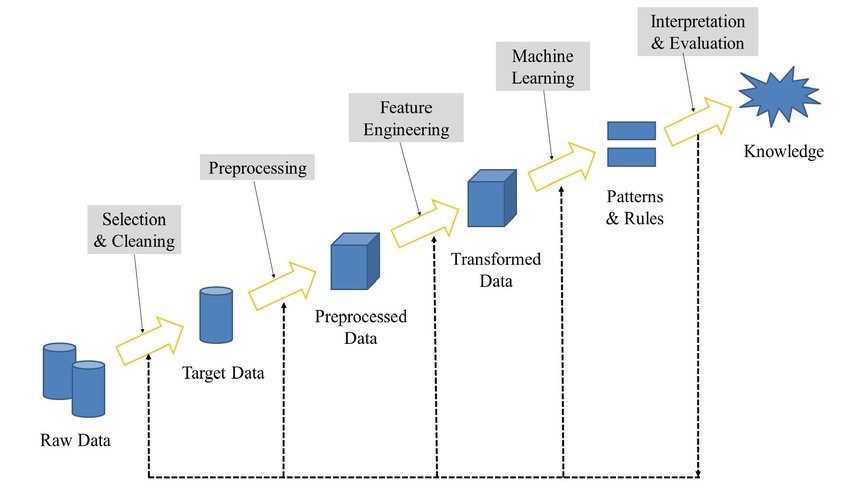
[**4.** **Interpreting and reporting the Results** 15](#_Toc33299382)

[**5.** **Conclusion** 15](#_Toc33299383)

[**6.** **References** 15](#_Toc33299384)

# **Introduction**

Data mining finds valuable information hidden in large volumes of data. Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.



## **Problem Statement and Domain**

The objective of this study is to extract the results of cricket matches from the history, here we have selected T20 format of IPL league from the year 2008 to 2017. We used different attribute selection techniques like: Data mining algorithms used to find out the many results prediction, accuracy to evaluate the how good the prediction performs.

Furthermore, it’s designed to prediction the results of future matches using testing &amp; training data through data modeling. Dataset we used to evaluate was posted on Kaggle dataset [1] sources from cricshet. Well, since in IPL is one of the famous leagues around the world, thus the problem we will solve and the associate questions that we will solving that must be familiar to all cricket loving fans.

## **List of Requirements**

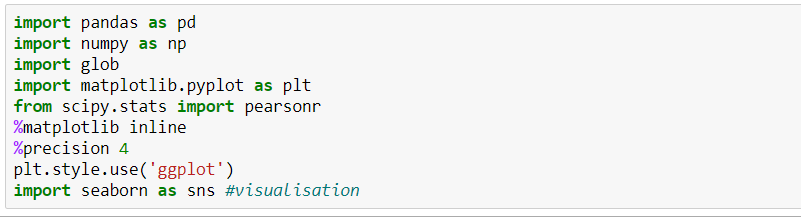
To find out trend or relation among different variables related to diseases. Following is the list of requirements.

1. How many matches we’ve got in the dataset?
2. How many seasons we’ve got in the dataset?
3. Which Team had won by maximum runs?
4. Which Team had won by maximum wicket?
5. Which Team had won by minimum wicket?
6. Which Season had most number of matches?
7. Which IPL Team is more successful?
8. Has Toss-winning helped in winning matches?

# **Related Work**

## **Loading Libraries**

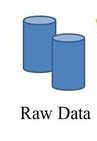
Let us create our analysis by loading Python libraries for our work.

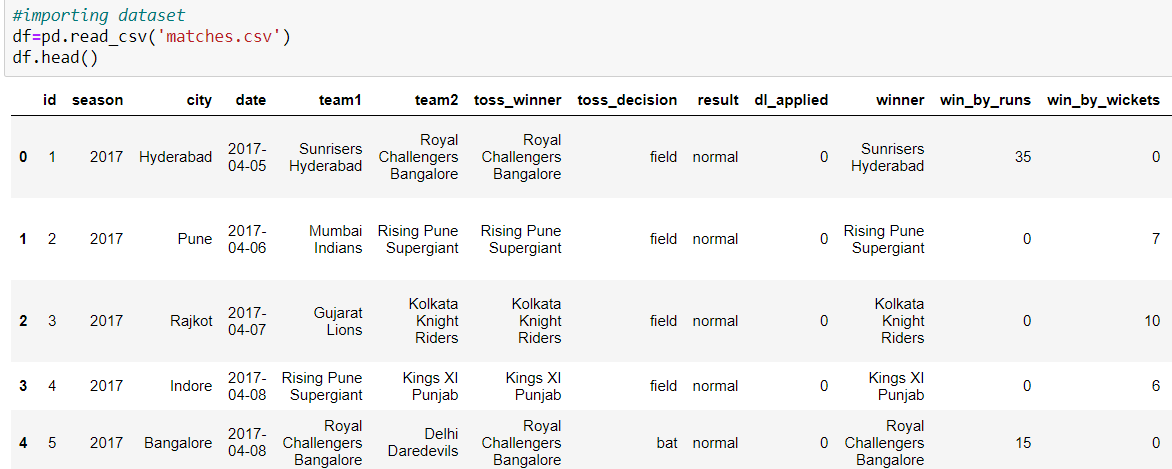


## **Reading input dataset**

In order to read the data set, let us define the data set in the directory in which the input file is present.

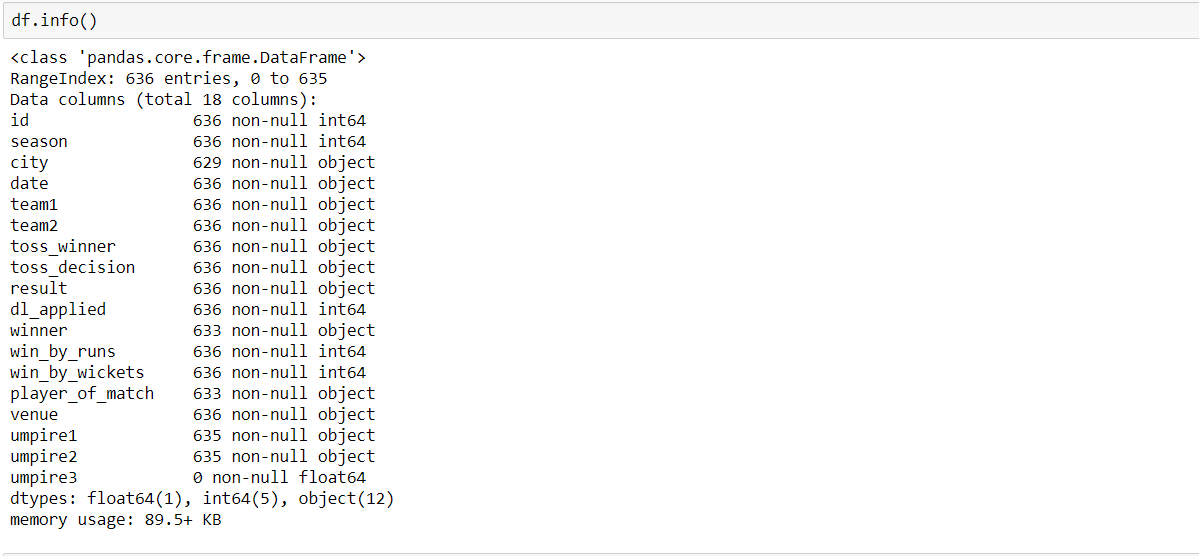
It is important to note that path must be provided in the format of string and the file name to read the input csv using **pd.read\_csv()**function.





## **Get basic information of Data**

After loading the file we get the basic information of our dataset in which the data set defines rows and columns information that’s what we find out with the method info ().



Missing values were detected. Five variables are of integer type, one is float type and remaining one is object.

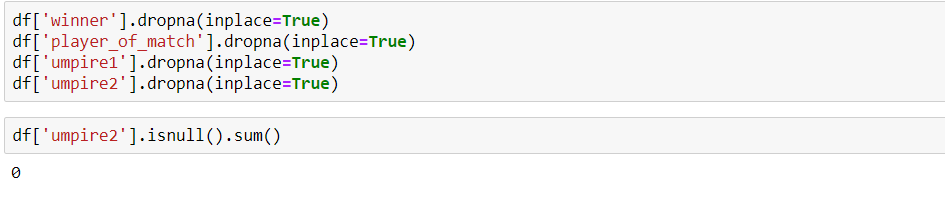
## **Data Preparation Steps**



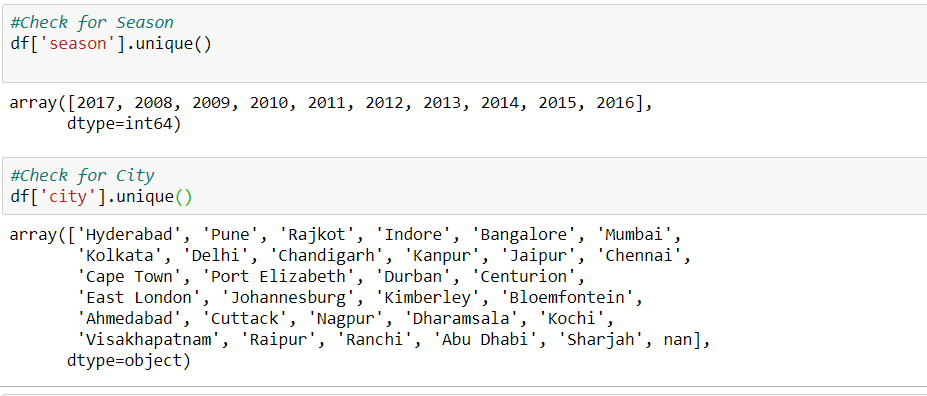
## **Checking for null values**

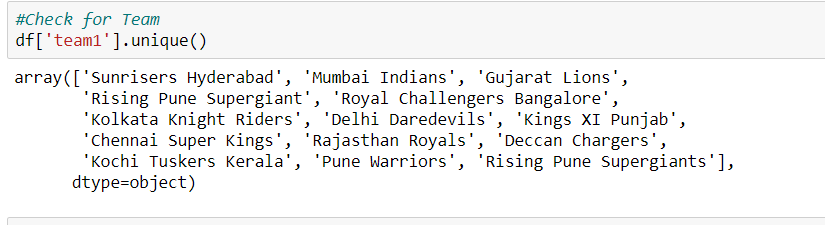
****

## **Cleaning for null values**

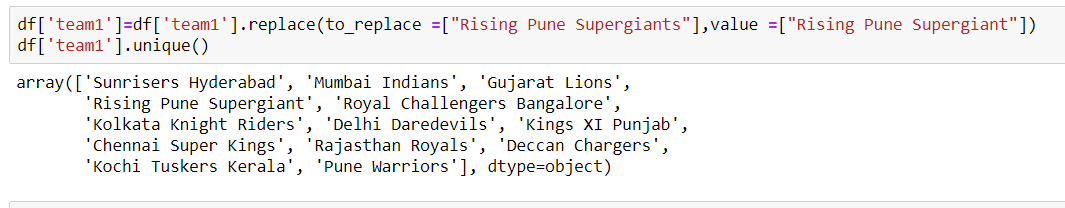
**Note:** Now there are no missing values in any variable.

## **Checking for unique values in each variable**

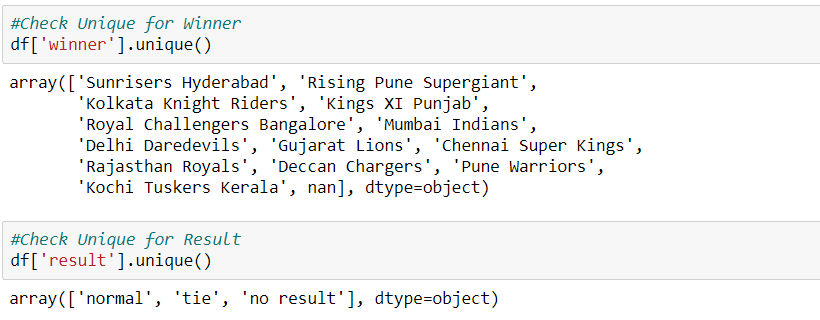
****

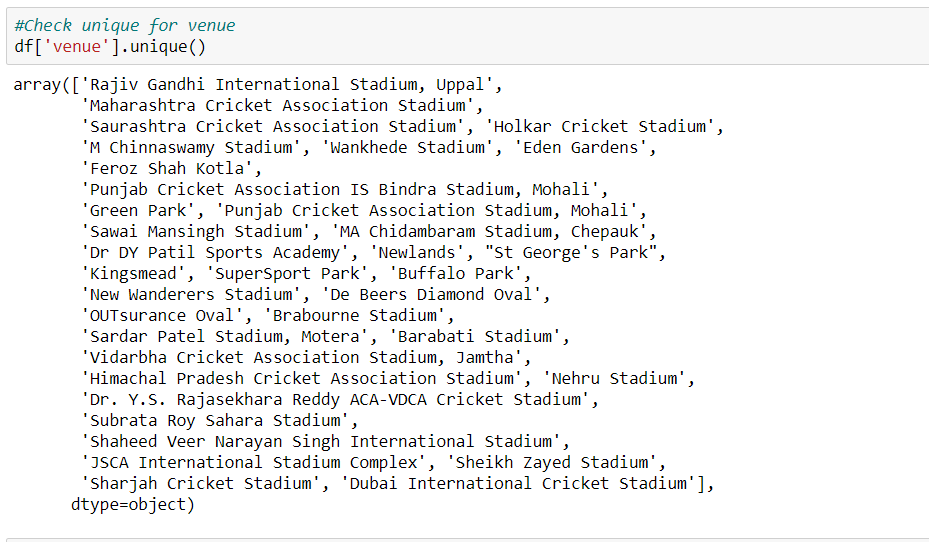
****

**Note: It was found that Rising Pune Supergiant origin was repeated twice. Hence, it was merged to a single value.**

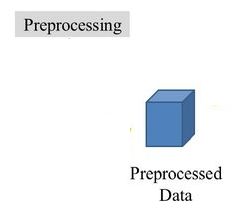








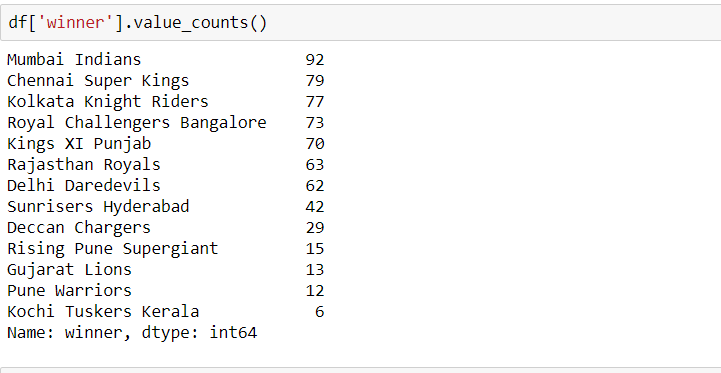
## **Data Analytics**

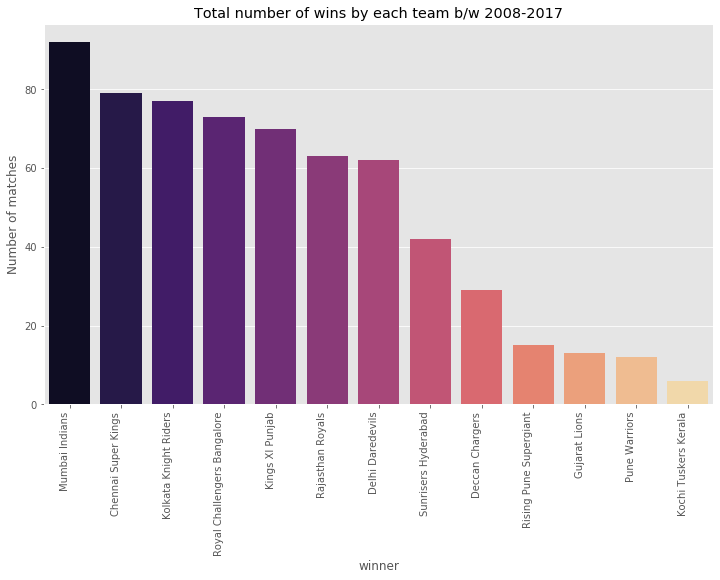


## **General Analytics**

Using our dataset, we can examine the number of winner

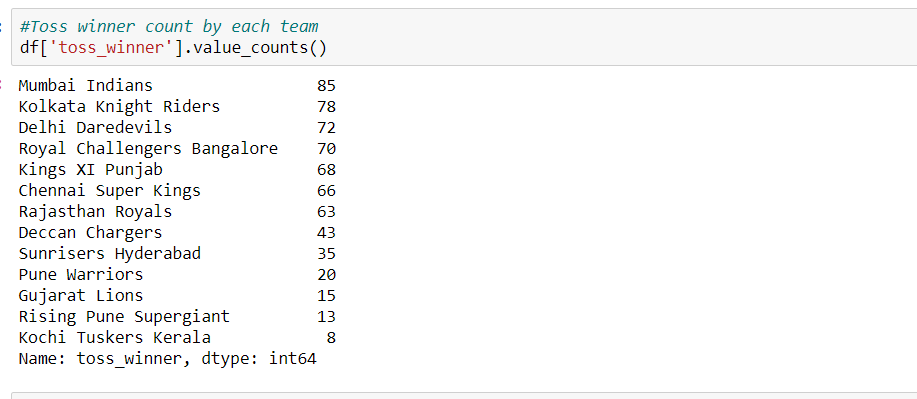
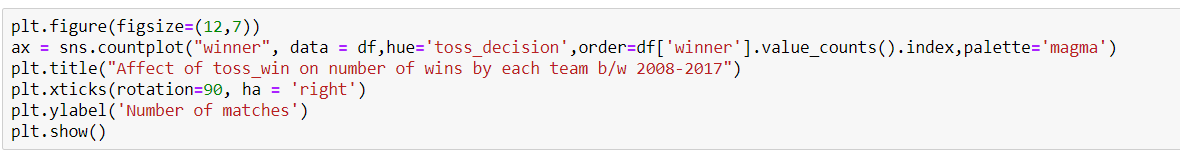
with different season of IPL during 2008 to 2017.

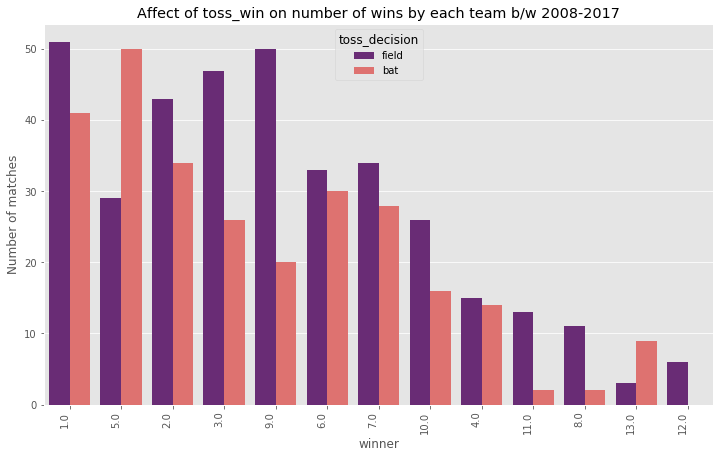




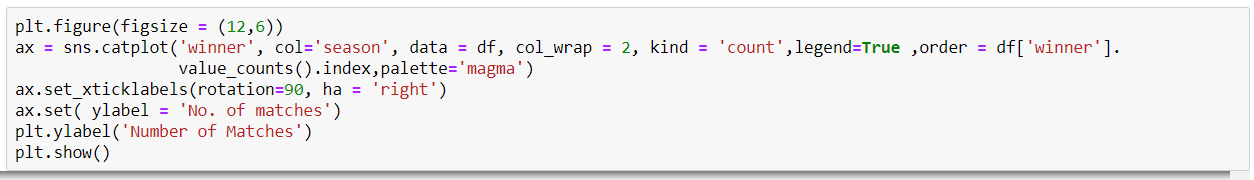
As shown in the above figure, team Mumbai Indians were won most of the matches in the duration of 2008-2017

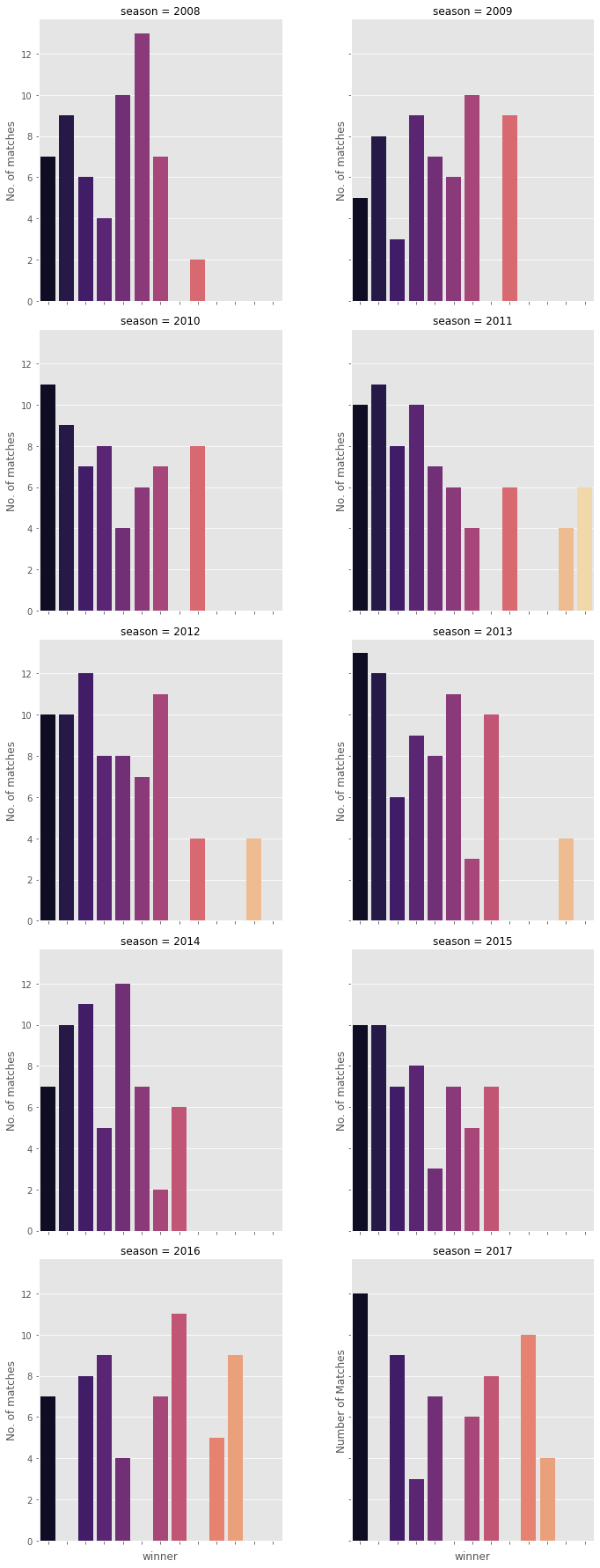
**Toss Winner**





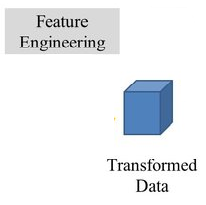
**Winner vs Season**



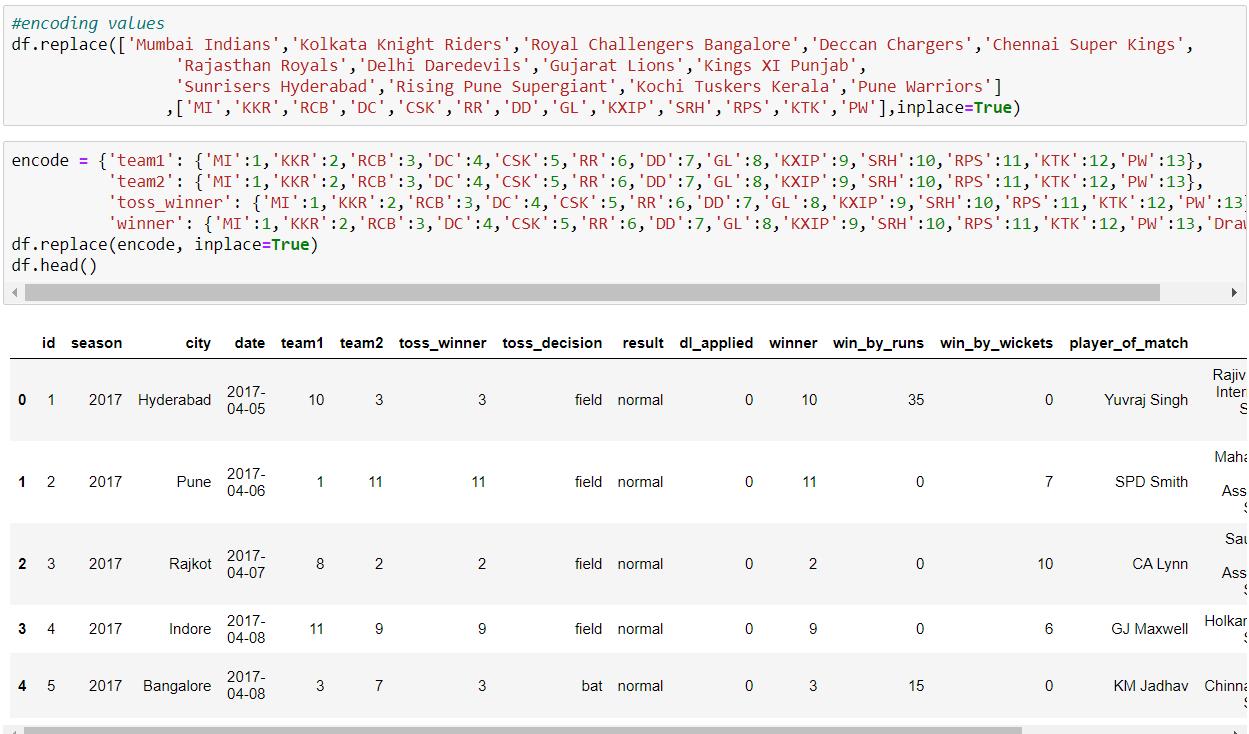


**Year wise top team**

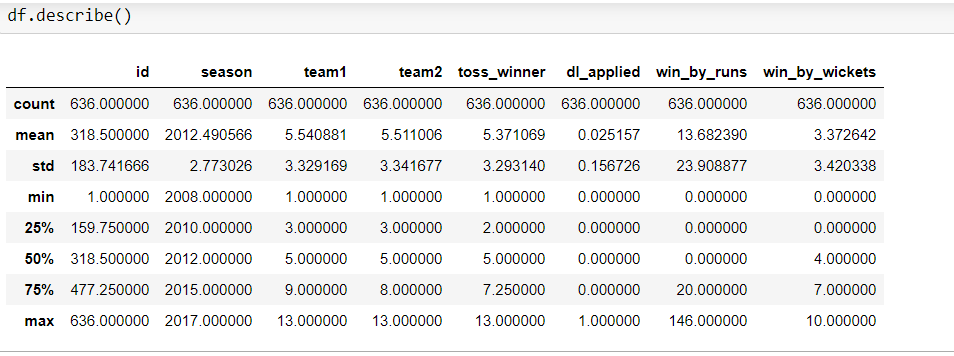
****



## **Encoding Values**

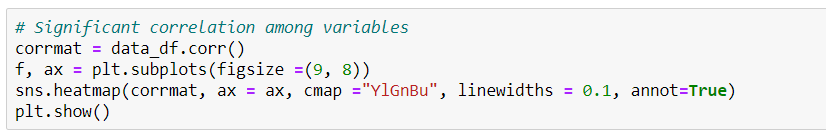
****

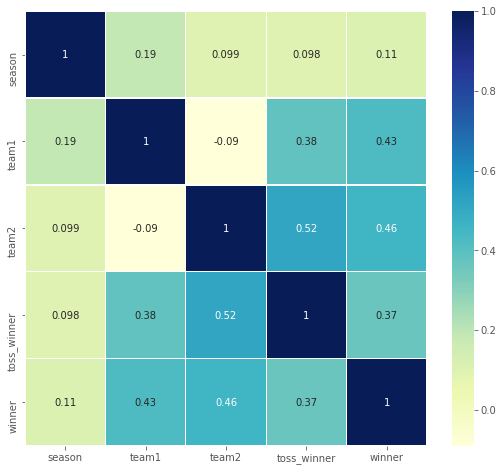
## **Data Frame Describe**

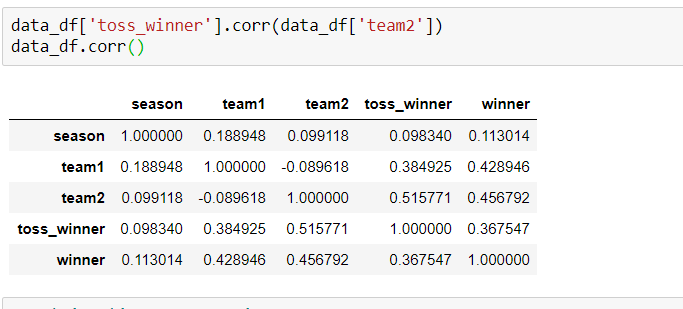
****

Analyzes both numeric and object series, as well as Data Frame column sets of df.describe () numeric count, mean, STD, min, max etc.

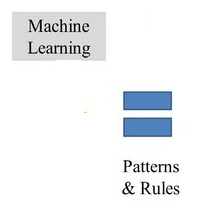
## **Significant Correlation among variables**

****

****

****

**N**ote**:** In our data no strong correlation found



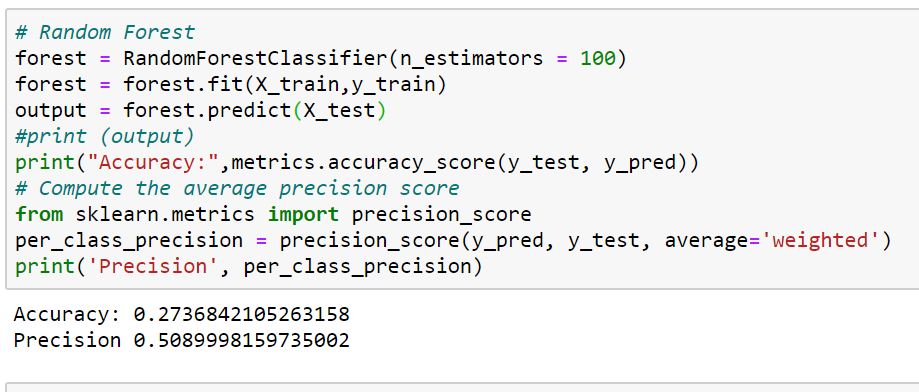
# **Methodology**

We apply the algorithm using 70% of training data and 30%

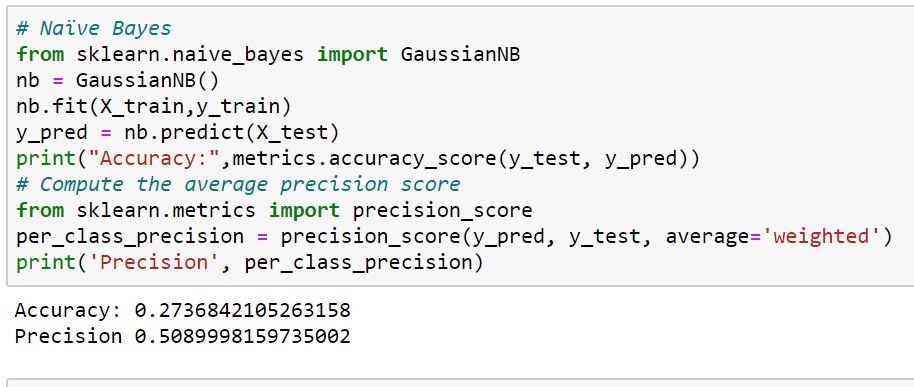
of testing data in our IPL dataset

## **Decision Tree**

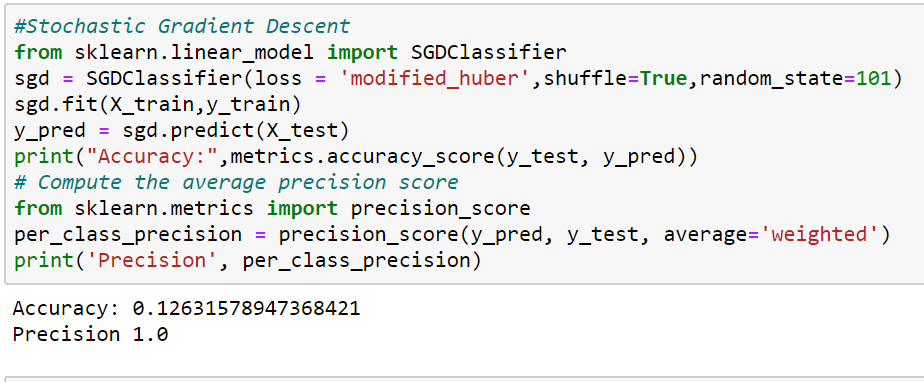
## **Random Forest**

****

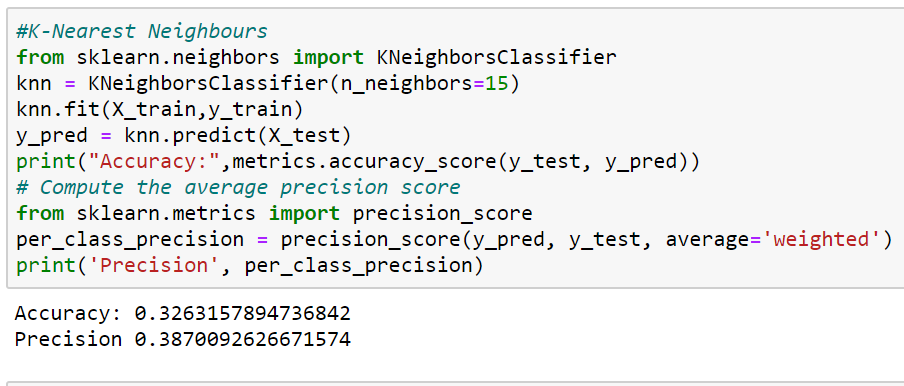
## **Naïve Bayes**

****

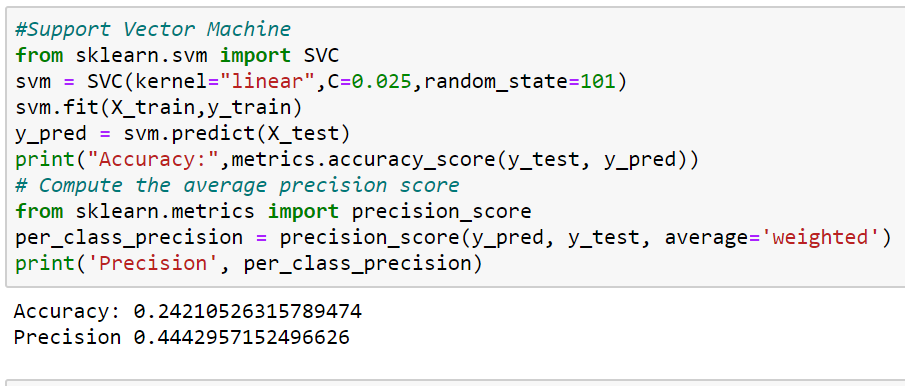
## **Stochastic Gradient Descent**

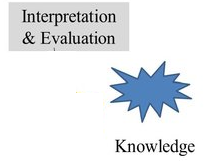
****

## **K-Nearest Neighbors**

****

## **Support Vector Machine**

****



# **Interpreting and reporting the Results**

|  |  |  |
| --- | --- | --- |
| **Classification Algorithms** | **Accuracy Percentage** | **Precision** |
| Naïve Bayes | **27.36%** | **0.50899** |
| Stochastic Gradient Descent | **12.63%** | **1.0** |
| K-Nearest Neighbors | **32.63%** | **0.38700** |
| Decision Tree | **44.21%** | **0.44873** |
| Random Forest | **33.68%** | **0.46029** |
| Support Vector Machine | **24.21%** | **0.44429** |

# **Conclusion**

The model which we have built to predict the results of the matches was successful with an accuracy of 44.41%. We used up to 6 data mining algorithms [3] that were used on models were Decision tree, Random forest, Naïve Bayes, Support Vector Machine, Stochastic Gradient Decent and K Nearest Neighbor. Decision Tree validation method was so effective than percentage split based on the prediction results we found. We got much better result with Decision Tree Algorithm, up to 44.21% accuracy and 0.44873 precision.

# **References**

[1] <https://www.kaggle.com/aadilmalik94/ipl-cricket-dataset>

[2] <https://www.kaggle.com/vaishaligarg/analysing-ipl-data>

[3] <https://analyticsindiamag.com/7-types-classification-algorithms/>