Muhammad Ibrahim

**Group Members : 18SW24 & 18SW26**

PSL Win Predictor

Exploratory Data Analysis

## **Introduction**

The Pakistan Super League (PSL) is a professional Twenty20 cricket league contested during February and March of every year by six teams representing six cities in Pakistan. The league was **founded on 9 September 2015** with five teams by the Pakistan Cricket Board. PSL is most important Event of Pakistan Cricket and the competition in PSL is growing everyday as technology advance and expectation increase. Teams and Players are expected to perform at their best with their skills.

The goal of this analysis is to identify these factors that can potentially effect Teams performance on the Winning the tournament and if it does what can we predict base on these results.

## **Objectives**

* Predict the winning Team of The Tournment.

## **Key**

* Identify Relation between variables and target value.
* Test different models to determine the best predictions.
* Validate the model results with our assumptions.

## **Approach**

* Conduct Exploratory Data Analysis
* State our Assumptions based on the Analysis
* Pre-Processing the data for modeling(cleaning the data set).
* Compare results from different models
* Validate and conclude our results

## **List of useful metadata**

**Nominal Variables**

* Team : Name of the teams.
* Ground : location where match is played.
* Opposition : opponent team.

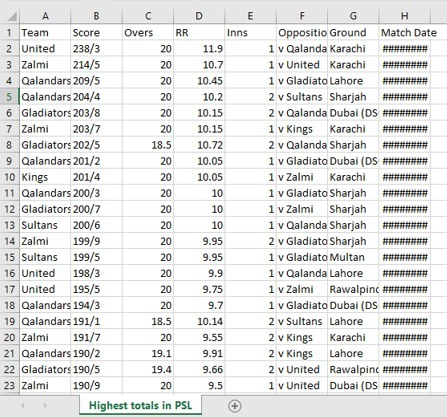
**Numerical Variables**

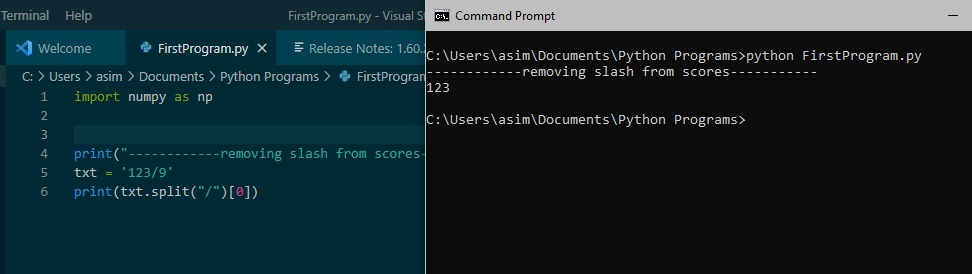
* Score : The total of the score of teams.
* RR : Run rate in the match.
* Overs: total numbers of overs played.
* Wickets: wickets.
* Inns : first or second innings of the match.

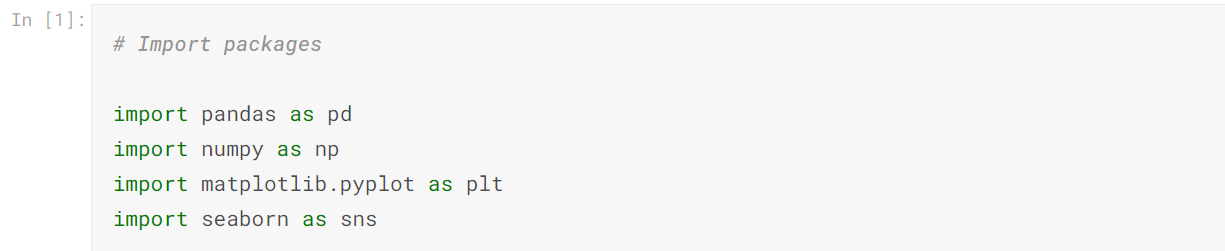
**Target**

* Win : Winning of random top 50 and low 50 matches of the teams out of 100 Matches

The “**Score”** column has different types of format and the “**Match Date”** is corrupt column

* 



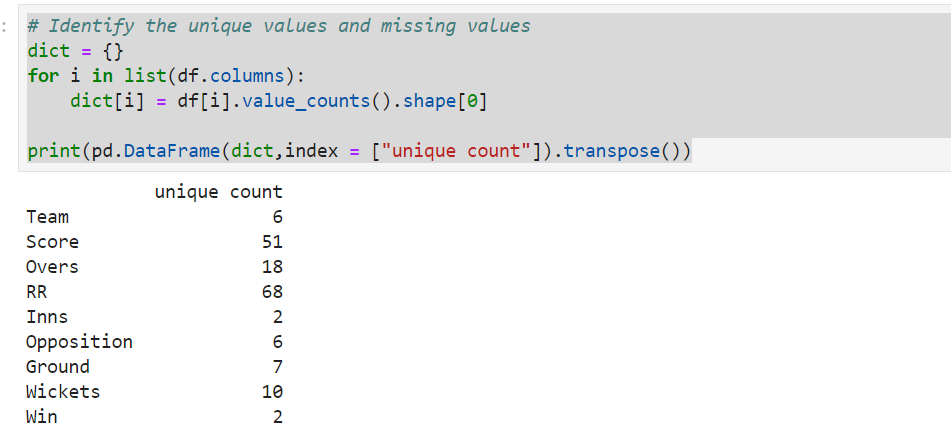


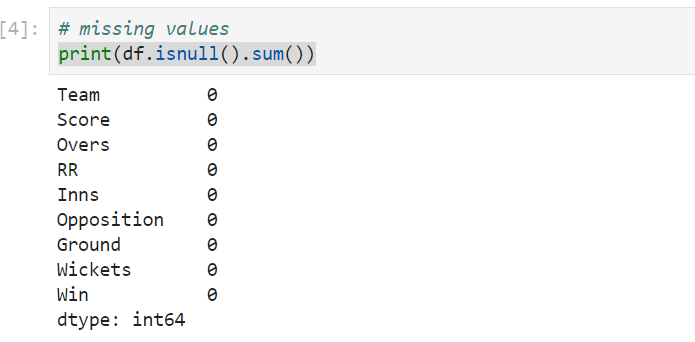


### **Missing values**

There are no missing values found in this datasets

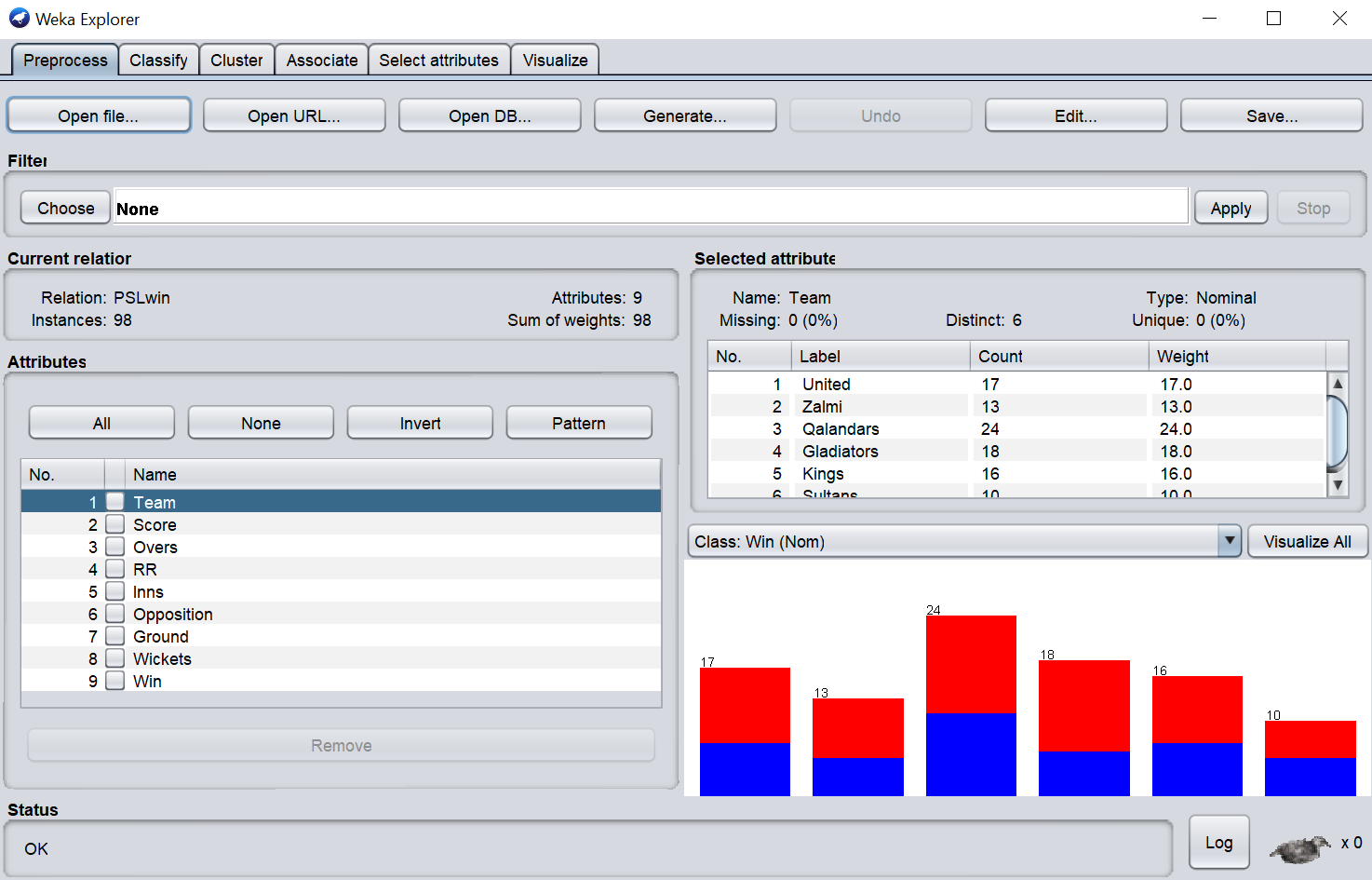
Identifying the unique and missing values





**Using WEKA**

Weka (**Waikato Environment for Knowledge Analysis**) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.



Our Data set is contains both Nominal and Numeric values and it is always very difficult question that which algorithm we should we select for the analysis of our Data.

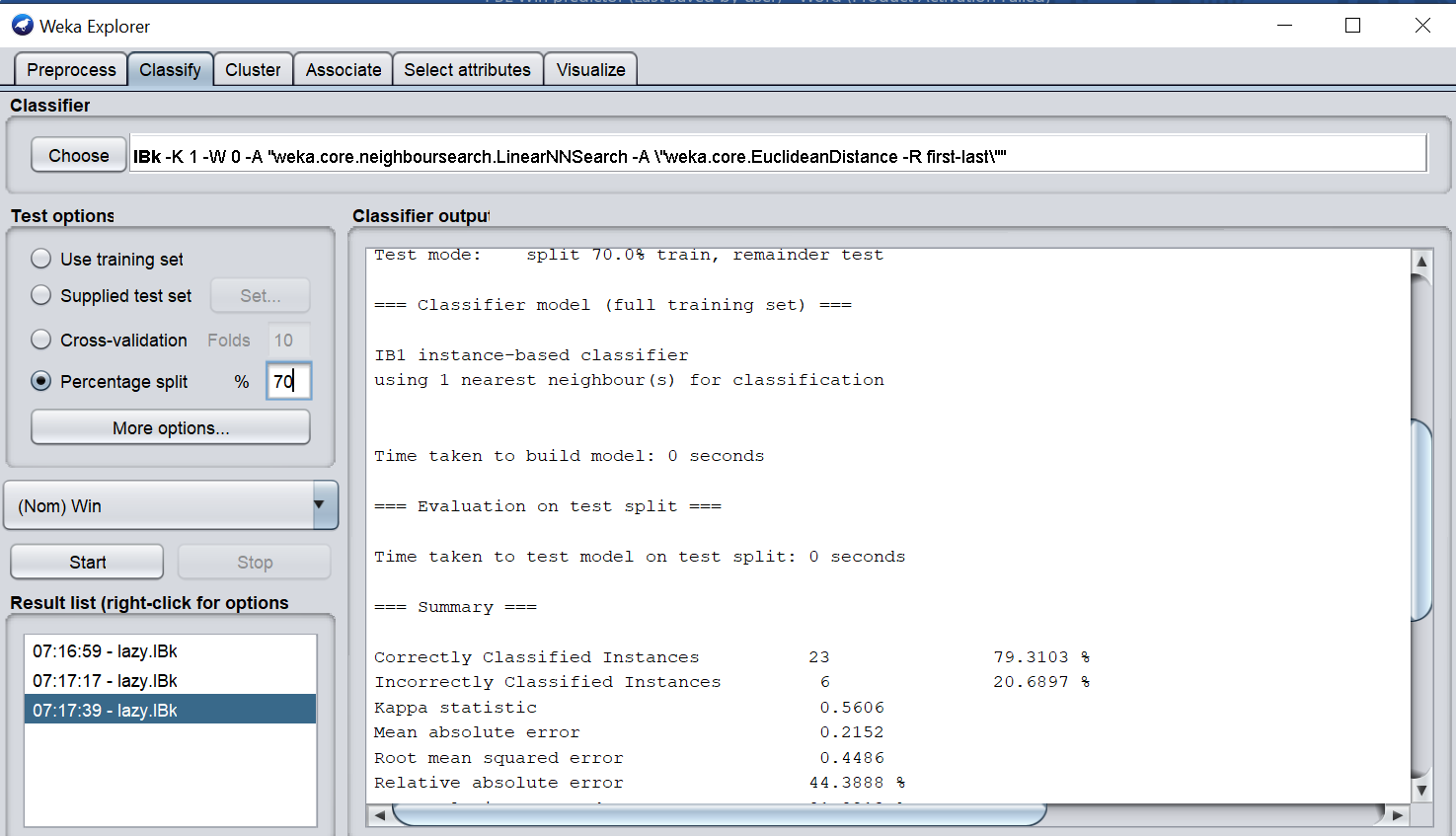
That being said, here are some buzz words to start looking up, to get your head around the possibilities:

* random forest / CART / decision tree (different algorithms, but similar in concept)
* Naive Bayes
* SVM (likely not helpful with the nominal parameters you have)
* Neural Net
* Clustering
* KNN, as Inti suggests
* many more...

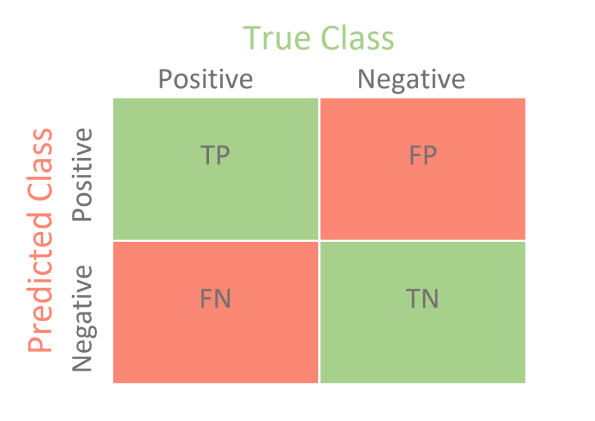
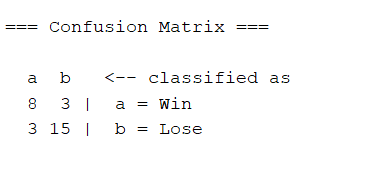
The world of potential options in machine learning algorithms is pretty huge, nothing works perfectly, and nothing works equally well in all situations.

But for our case we choose to use KNN (K nearest neighbor).In fact the algorithm is quite simple and the results are good. The only problem is that KNN is a lazy classifier; hence, training stage is fast (almost empty) and classification stage is more slow. Now, that is only important if your training set is really big because the algorithm is O(N\*M) where N is number of training instance and M is number of attributes. In the worst case you can perform some filtering in your data. Weka has some methods to do that.

**Note : In Weka the algorithm has a different name, IBK**

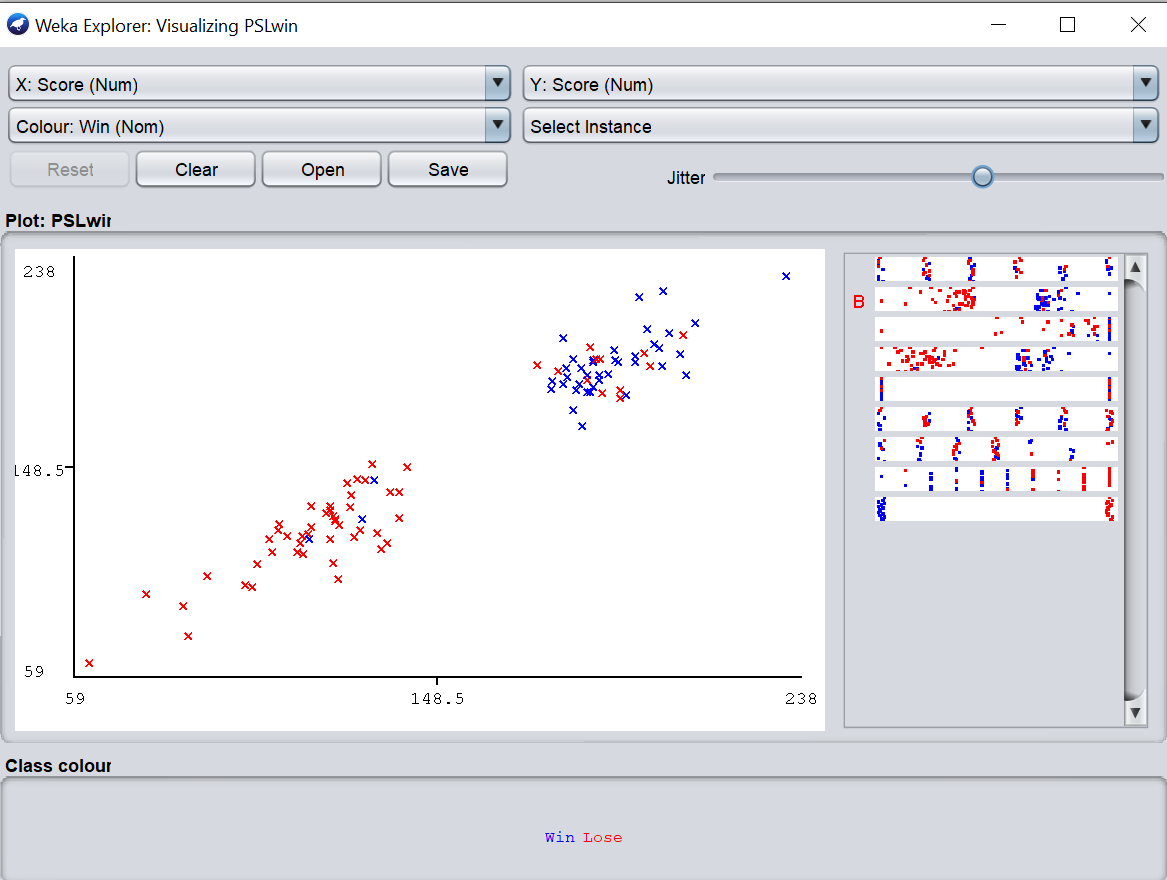


Confusion Matrix

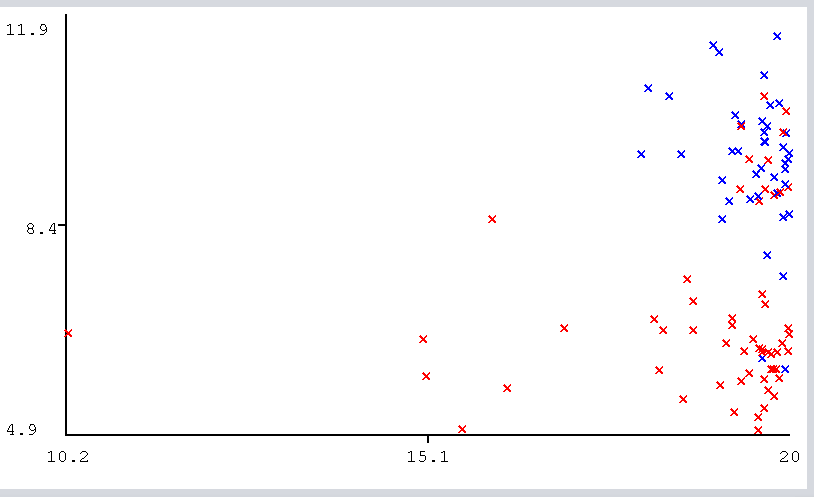


**Data Visualization**

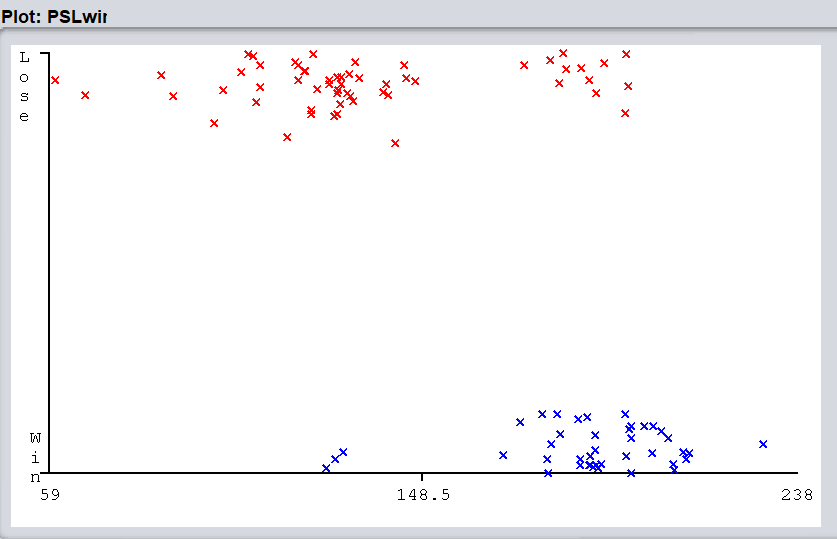
Score



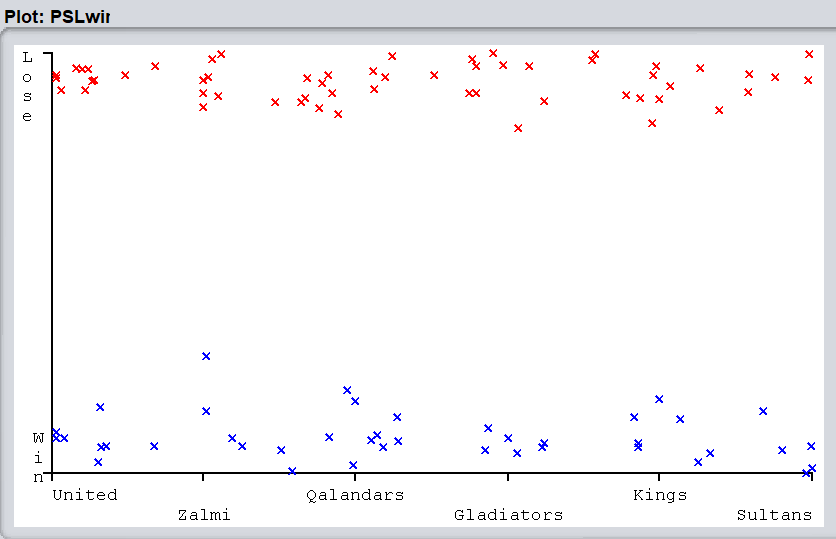
**Over vs RR**



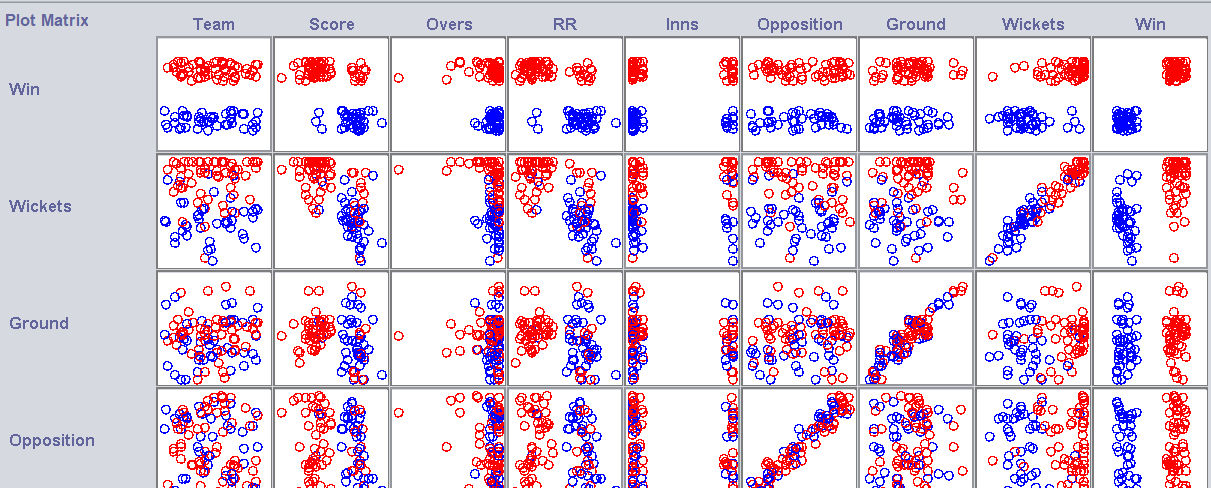
**Score vs Win**

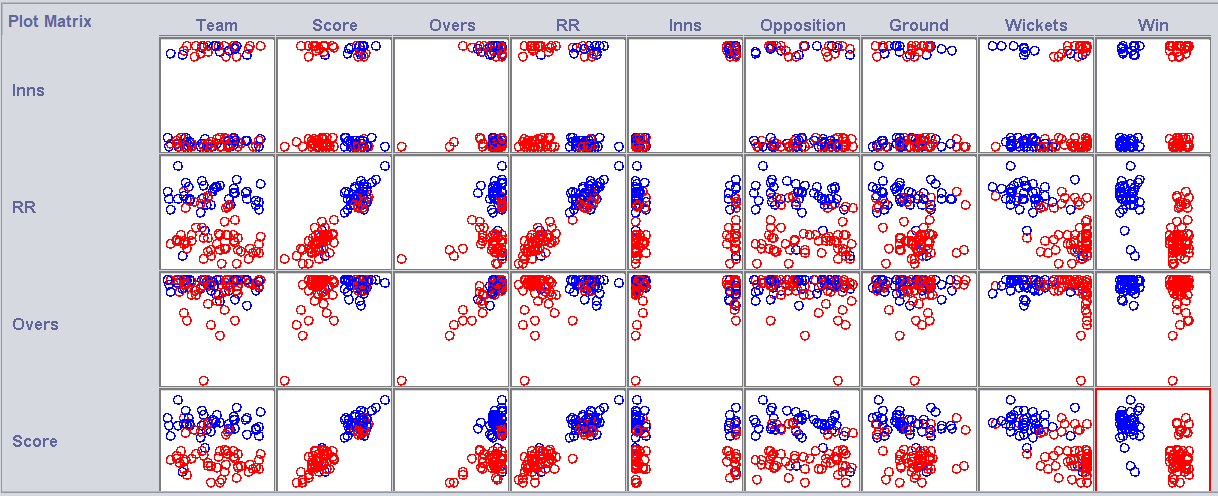


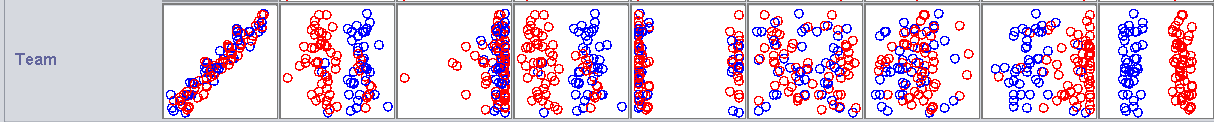
**Team Vs Win/Lose**



**Summary**

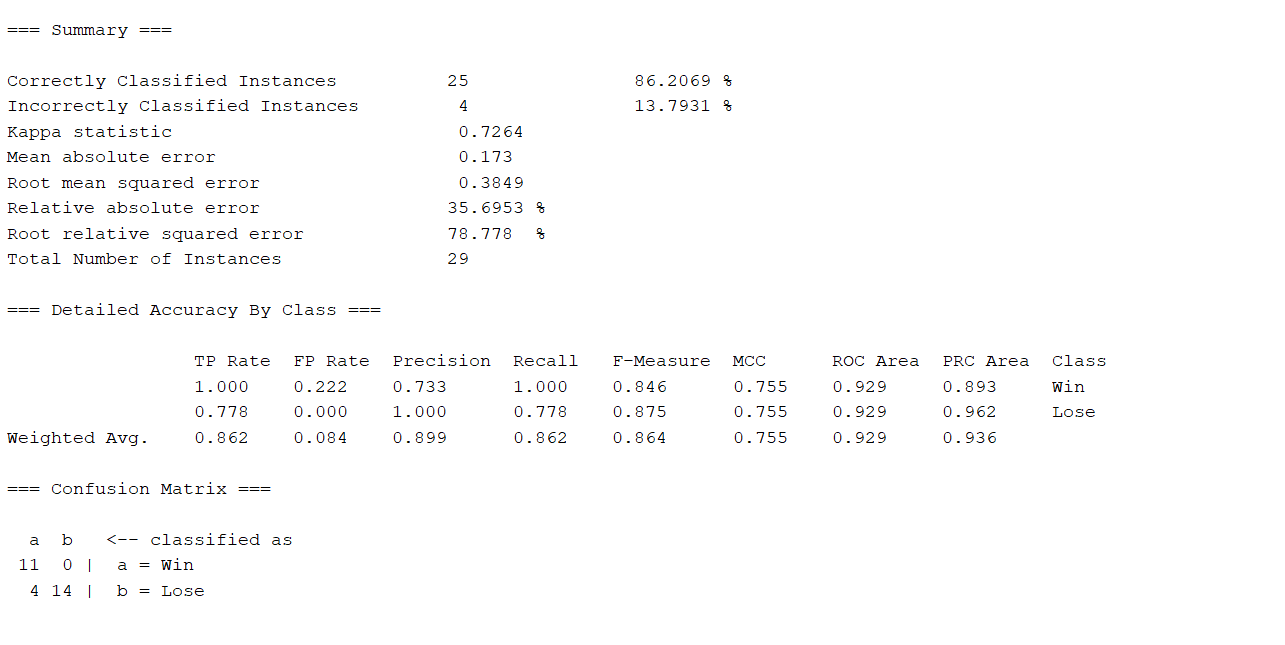






Analyzing the Data set with other Algorithms to find the best option among the chosen KNN and other Algorithm

**Naïve Bayes**



**In this more accurate as compare to KNN.**

**ZeroN**

