**Stats 517 – Final Project Report**

**Abhinav Prabhu Adarapuram**

**Problem Statement, Motivation, Research Goals**

For the final project, I am using the Tennis ATP dataset, which comprises of match details such as tournament date, location, player details, match results, scores, surface used, etc.

This dataset has data which were collected in the past 50 years (1968 – 2018\*). From this I am using the recent 30 years data to make my analysis.

The motive of my project is to analyze whether the left handed players won more matches or the right handed players. This will be the analysis for classification part.

In clustering part, I am going to analyze which country has consistent performance in the past three decades.

## Data Source and Description

I found this interesting dataset from Github – Awesome Public Datasets.

<https://github.com/awesomedata/awesome-public-datasets>

This dataset is updated with recent match’s data (Till September 2018).

## Modeling Process

Initially, this dataset was merged into a single csv file ranging from most recent to last. Few variables which were not required for analysis such asWinner\_seed, Loser\_seed, Tournament\_id, Winner\_entry, Winner\_rank, Loser\_entry, Loser\_rank, Score, surface, etc. The dataset comprised of many missing values. These missing values were filled with the most recurring value of that particular column using R script. Once the imputation is done, the target variable **‘winner\_hand’** was converted into continuous values using the ‘revalue’ function.

After the data cleaning, I imported the csv file to Jupyter Notebook to perform the analysis. To start with, I worked on the classification part, where the machine was trained with the data and various models such as Linear Regression, K nearest neighbors, Decision Tree, Navye Bayes and Random Forest were used.

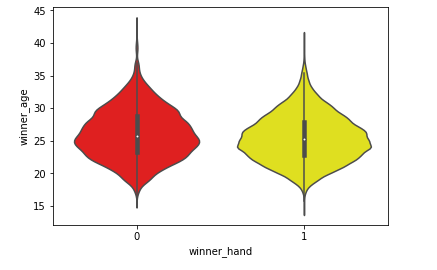
**Obtained Results**

**\*Supervised Learning:**

**Classification:**

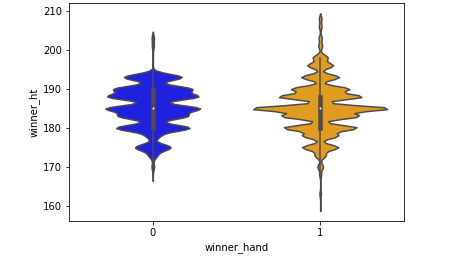
In classification, I have taken the winner hand as the target variable. The values in the winner hand column are considered as binary values where **0** represents “left hand” and **1** represents “right hand”.





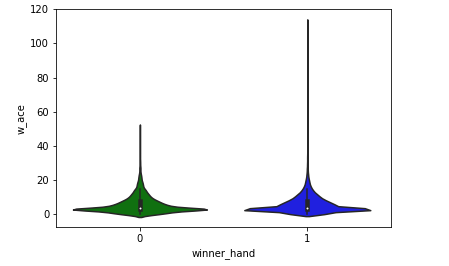
From the above plot, it is visible that most of the right handers and left handers belong to the age group of mid 20s.





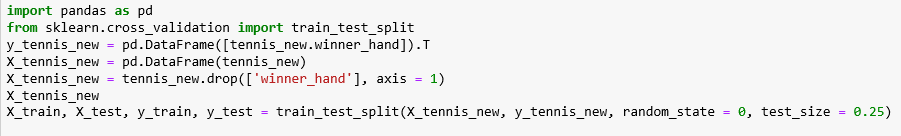
From the above plot, we can understand that most of the winner’s height ranges from 180 to 190.





From the plot, it is clear that the average ace score for the winners is between 5-15 points.

**Data Partitioning:**

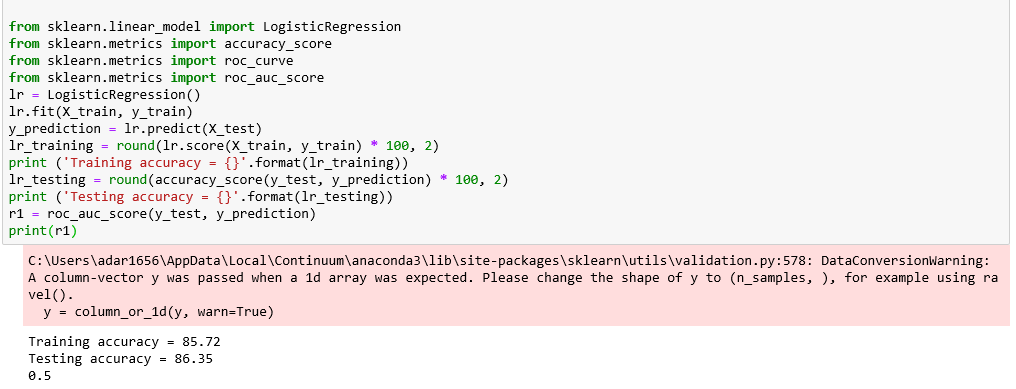


For the analysis, I have taken the test size as 25% of the original dataset. The winner\_hand is considered as ‘y’ variable and dropped from the ‘x’ variable. Al other columns are considered as ‘x’ variables.

**Results:**

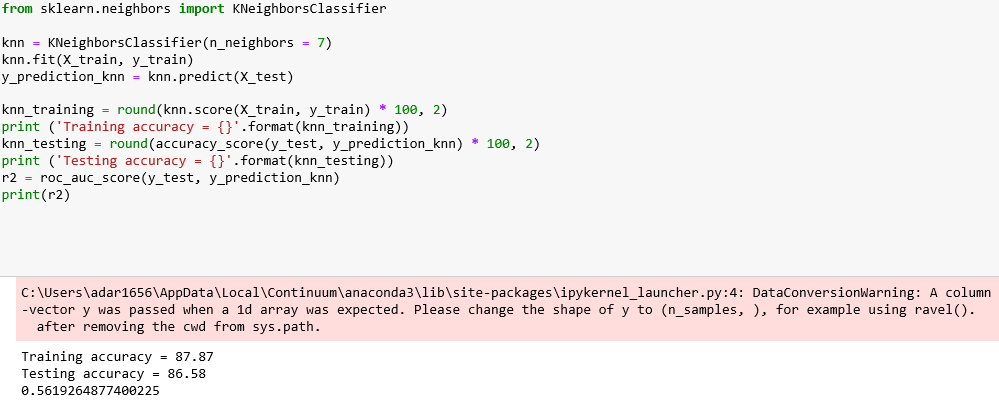
**Logistic Regression:**

Logistic regression is used to explain the data and to view the relationship between a dependent binary variable and many independent variables. Here, my target variable “winner\_hand” has the binary values – ‘0’ – left hand and ‘1’ – right hand.



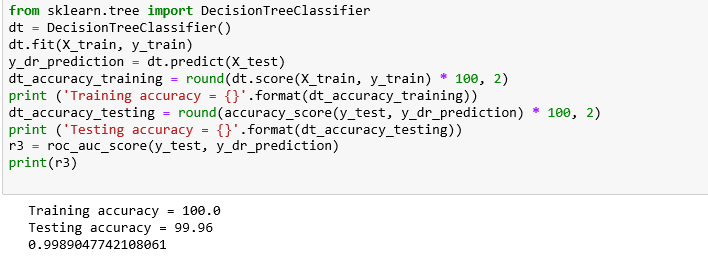
**K Nearest Neighbor:**

It is one of the simplest machine learning method. In this method, we use the K nearest training data points and get the most frequently occurring class.



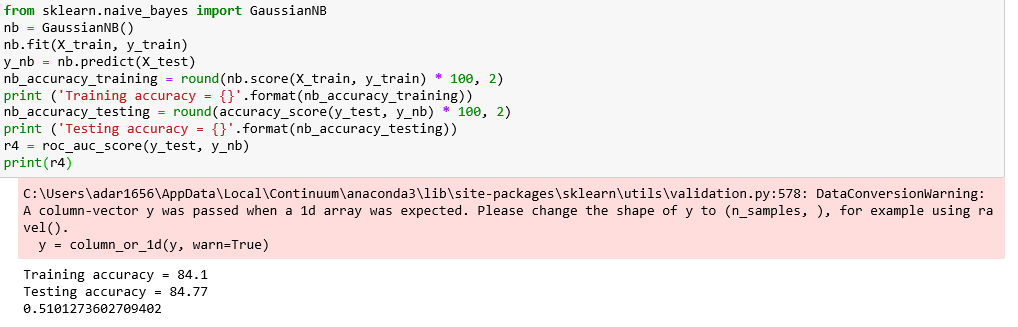
**Decision Tree:**

A decision tree often starts with a single node, which is further sub divided into various nodes or outcomes. It goes on and forms like a tree like structure. It is useful in describing an algorithm with the actual flow of events.



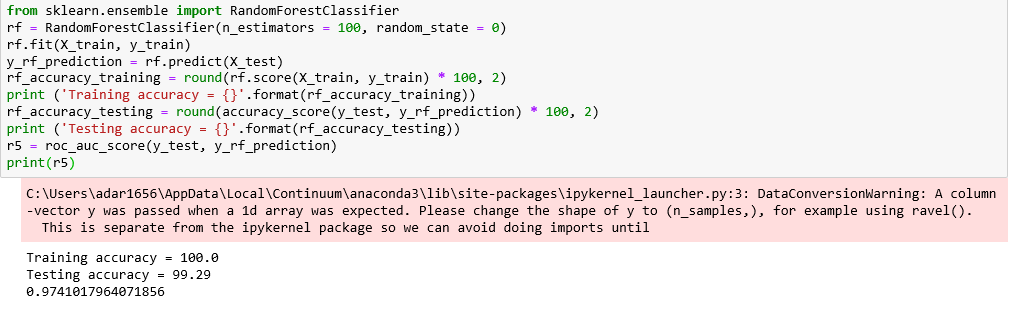
**Naive Bayes:**

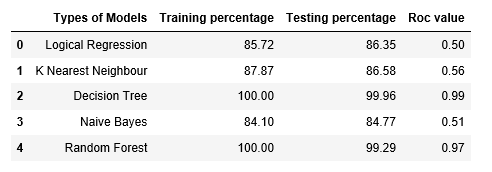
This technique is based on Bayesian theorem which is much suitable for inputs with high dimensionality.

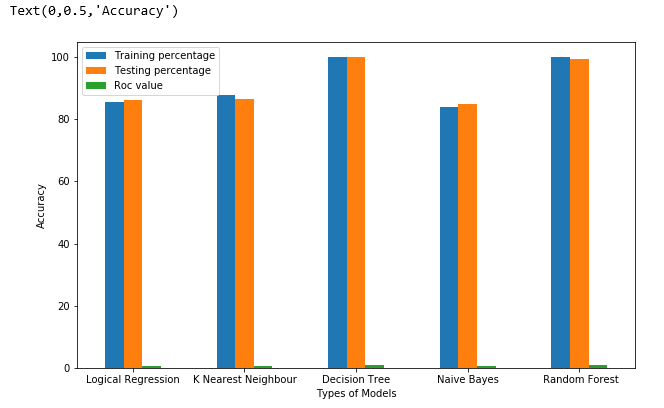


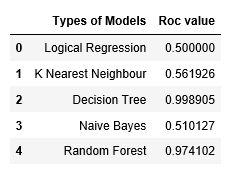
**Random Forest:**

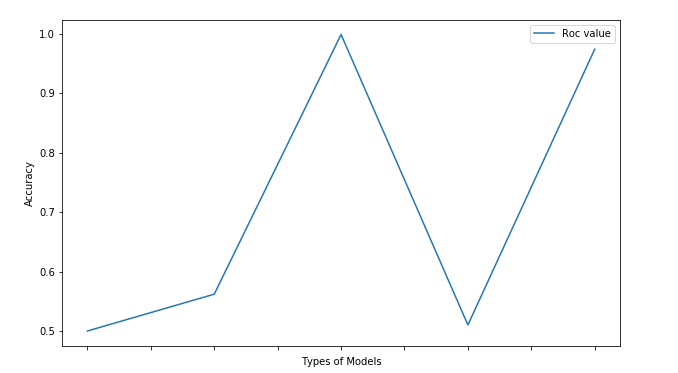
This technique can be used in both supervised as well as unsupervised learning. It constructs many decision trees and combines them together to obtain a good prediction value.







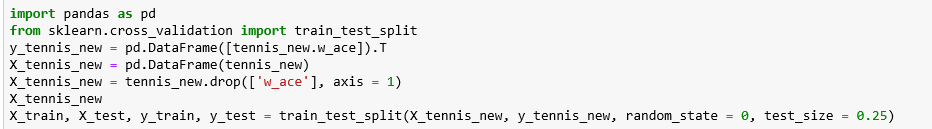




After performing the analysis, it is observed that, **Decision Tree** gives the best results with **training accuracy of 100** and **testing** **accuracy of 99.96. Decision Tree** also produces the best **roc value of 0.97** when compared to other models.

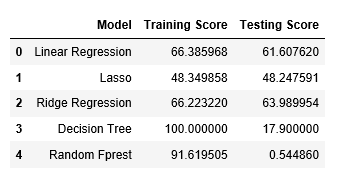
**Regression:**

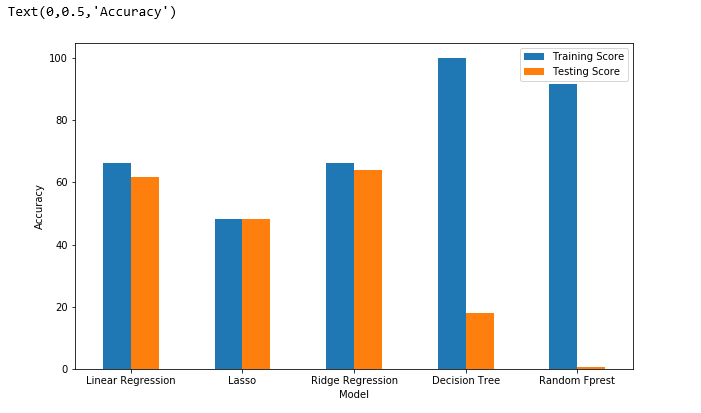
In regression analysis, I am using the column “w\_ace” (continuous variable) as my target variable. I analyzed the number of aces secured by the winner in that particular match.



For the experiment I took 25% of the data to be test set.

By using various regression models such as linear regression, lasso regression, ridge regression, decision tree and random forest, I obtained the following results.





From the obtained results, it is evident that**,** decision tree has best training score but ridge regression tops the testing score.

**\*Association:**

Association rule mining is mainly focused on getting the most co-occurring associations which exist between the multiple variables in a given set of data. To put it in a simple way it is the method of analyzing multiple activities which occur together. For example, a person buying a pair of shoes will also buy a pair of socks.

In my analysis, I have used the association rules to observe some interesting patterns and relationships among the variables. I have partitioned the dataset into 3 decade groups:

Decade 1 = 1989 – 1998

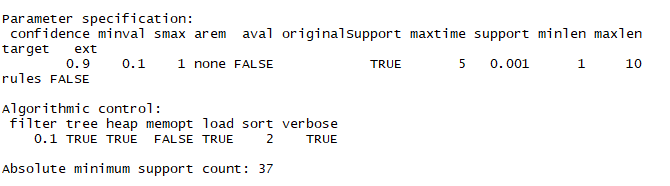
Decade 2 = 1999 – 2008

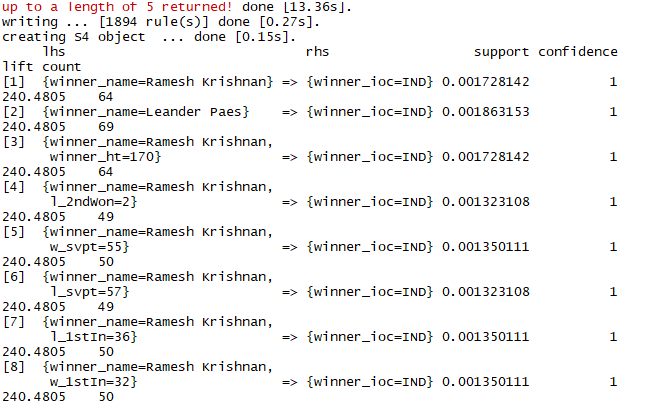
Decade 3 = 2009 – 2018

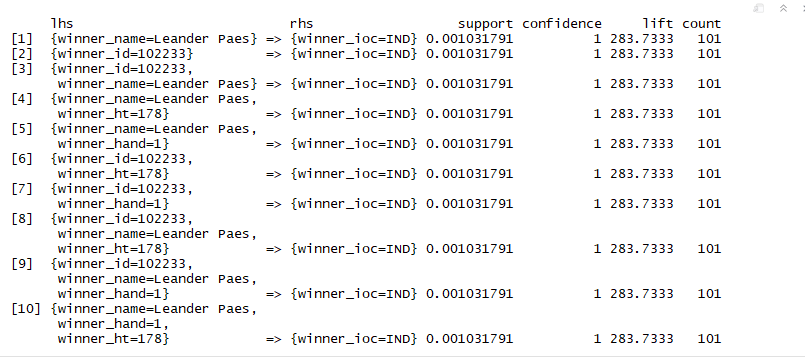
Considering every individual decade, the analysis was done.

**Decade 1:**

In decade 1, I have taken the “rhs” value as “winner\_ioc” and in particular I selected the country as India (IND) to observe interesting results.







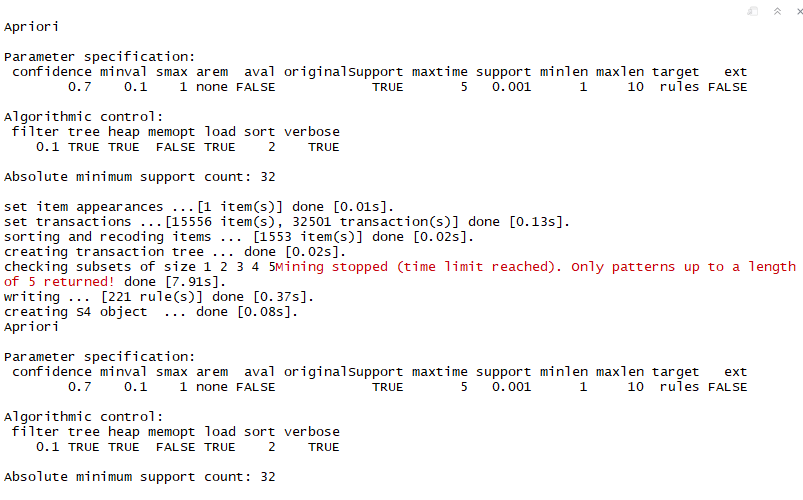
After executing the code, I got the relationships which are related to the country “IND” and I also obtained the relationship between various columns which are related to the Indian player Ramesh Krishnan.

I also obtained the relationship between the columns for the player Leander Paes.

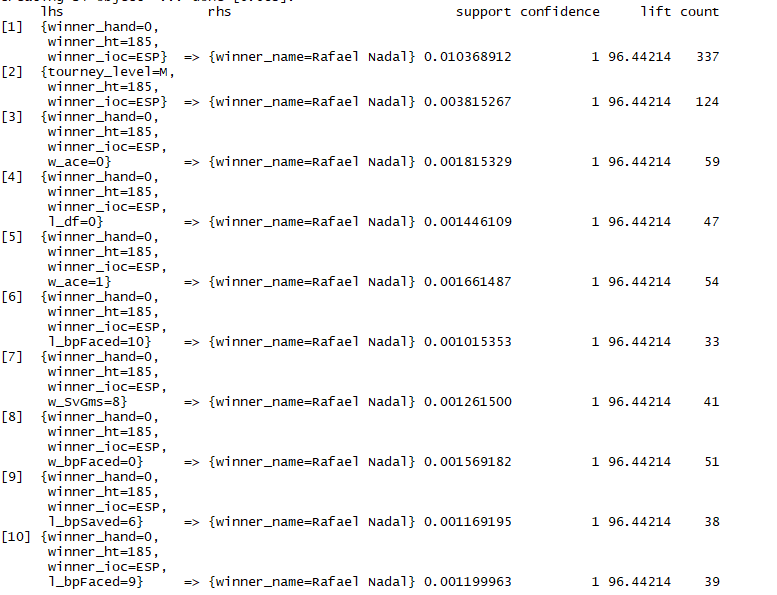
The obtained rules have good confidence rate of 1.

**Decade 2:**

In decade 2, I thought of viewing the association rules of my childhood favorite and all time champion Rafael Nadal.

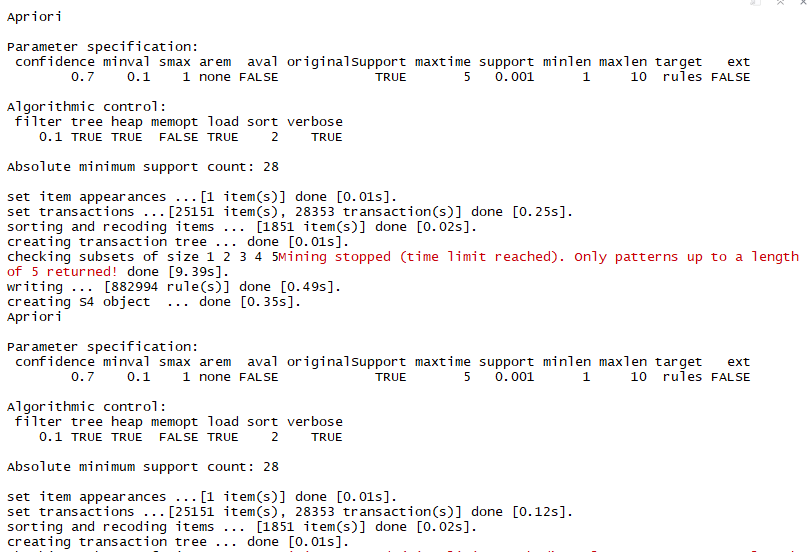


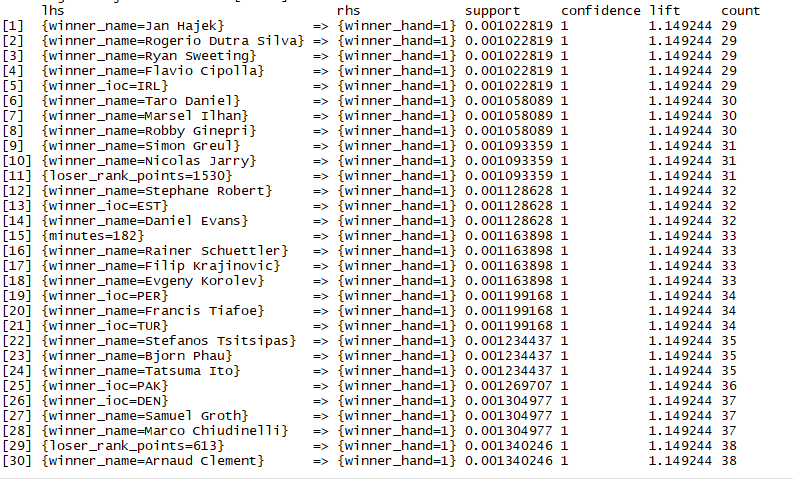
With the confidence of 0.7, I got 221 rules. When the confidence level is increased to 0.9, in total 157 rules were generated.



**Decade 3:**

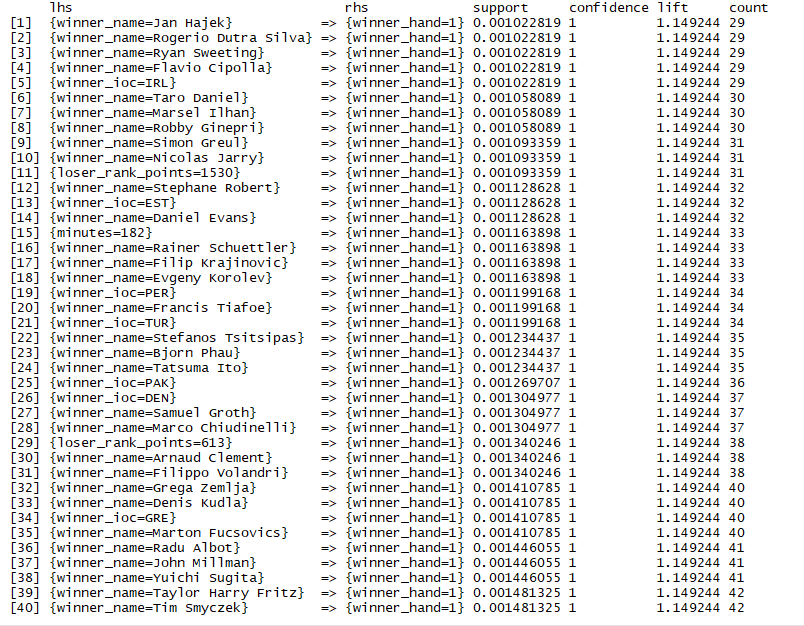
In decade 3, I analyzed the players with winner hand as right. But I didn’t get better results.



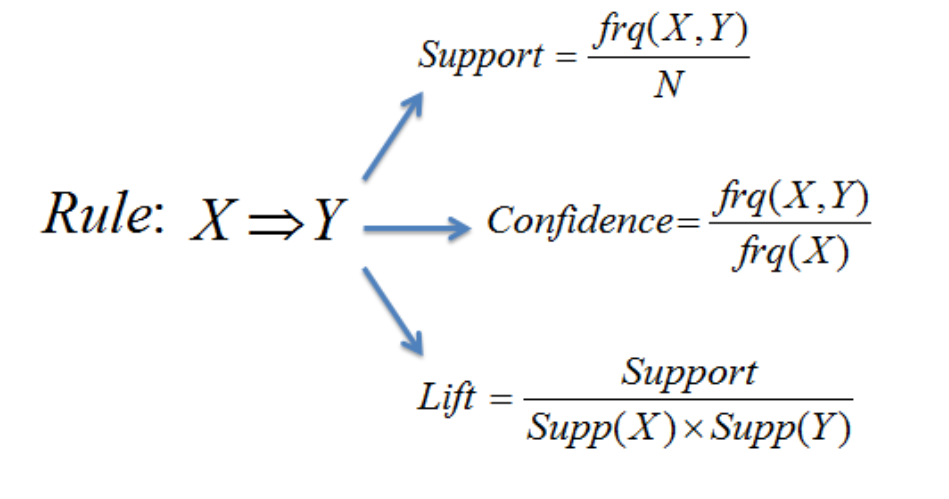


From the output of decade 3, it is much evident that the relationships were not generated.

I tried increasing the confidence rate from 0.7 to 0.9, but still there wasn’t any change in the outcome.



The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.



**Support:**

Support is the number of transactions that include items in the variable X and Y parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

**Confidence:**

Confidence is an indication of how often the rule has been found to be true.

**Lift:**

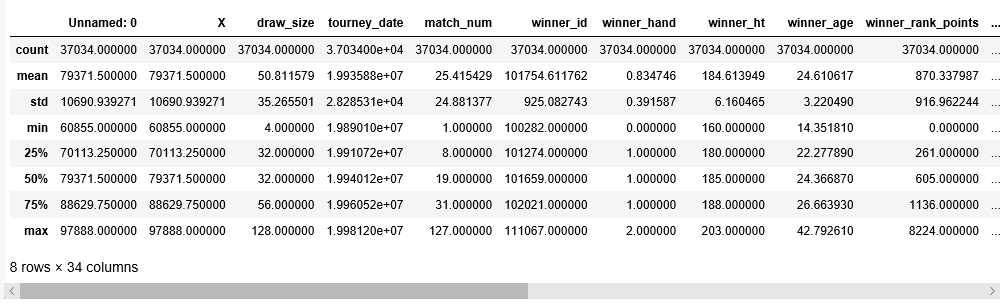
The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule.

**\*Unsupervised Learning:**

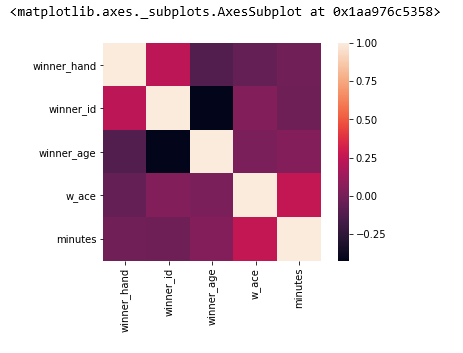
**Clustering Analysis:**

Clustering is a method of collecting a group of similar data from a dataset. A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. Hierarchical, density based, model based, k-means, grid based are some of the commonly used clustering algorithms.

In my analysis, I have used k-means clustering and Agglomerative clustering to obtain the results. I also tried nb and hierarchical clustering, but due to huge size of the data I didn’t get the desired results.

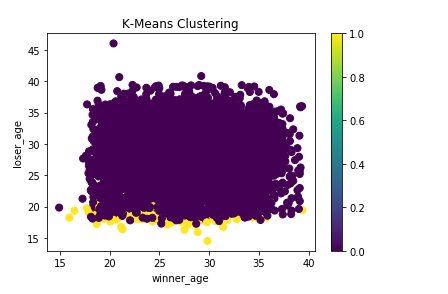


The above table is a description of the dataset which I am using for the analysis. It describes the count, mean, standard deviation of the values.



The above diagram represents the correlation among the variable winner hand, winner id, winner age, winner ace and the no. of minutes. It is clear that winner hand has comparatively good correlation with winner id than any other values.

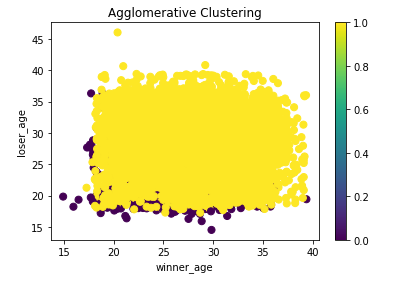
**K-Means Clustering:**



I used the winner age and loser age to plot in the clustering. Most of the players belong to the mid-20s group.

**Agglomerative Clustering:**

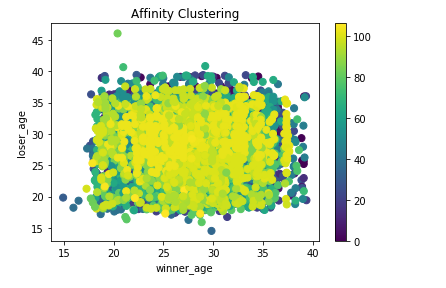
Agglomerative clustering is a "bottom-up" method for creating hierarchical clusters.



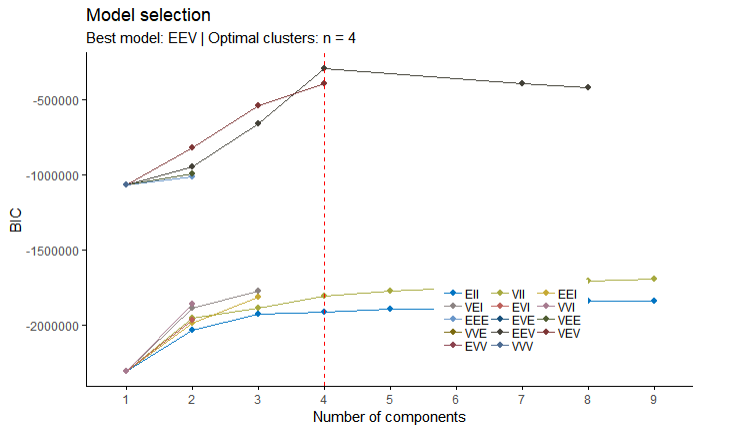
Winner age and loser age has been clustered into two groups.

**Affinity Clustering:**

Affinity clustering method generally converts the data into high dimensional values, performs the operation and gives out the lower dimensional values.



The winner age and loser age values are plotted into different clusters.

**Model Based Clustering:**

I tried running h-clust, nb-clust and m-clust using R program, but I got outputs only for the m-clust.

The m-clust suggests the optimal number of clusters to be 4 and the best model as EEV.

**Conclusion:**

To summarize, I performed analysis on Classification, Regression, Association and Clustering. I got the desired results for most of the methods except clustering analysis.