INDIANA UNIVERSITY

SCHOOL OF INFORMATICS AND COMPUTING

# EUROPEAN SOCCER ANALYTICS REPORT

Abhinandan Sampathkumar

Prathik Rokhade

Venkat Sambandhan

## 1. INTRODUCTION:

Our project deals with an extremely large dataset called the European Soccer Database which is available as part of Kaggle Competition.

The soccer database includes 25,000+ matches, 10,000+ players, 11 European Countries with their most popular championships. It includes data for all these games from 2008 season to 2016 season. The database also consists of players and team attributes which is extracted from EA Sports FIFA games. Furthermore, it includes team line ups, betting odds from around 10 providers and details of the match events like ball possession, goal types, corners, fouls, cards etc. for 10,000 odd matches. The data is a collective set of SQLite database acquired from different data sources. The foreign keys for players and matches are same as the original data source. The Kaggle competition has already set up numerous kernels and we have gone through some of them which are interesting and challenging.

The main clients are Online Gambling Companies and websites which host Fantasy League and the major users are the bookies and sports analysts in respective companies. Our analysis would be helpful to determine most predictable European league and finding top n players overall. The project will be implemented using Python and R.

Some of the challenges that we face in this project are data pre-processing, data analysis and visualization. Fine tuning the data, predicting top N football players, developing attractive plots and graphs are some the important tasks in our project. With these results, it would help our clients attract and encourage more users to involve in match betting's and fantasy leagues. The bookies can use the predictions of league that we have for fixing the pre-game odds.

## 2. OBJECTIVE:

Below are the primary objectives of this project:

a. Identify the most predictable league in Europe using Python
b. Top N players of Europe
c. Visualizations using R Studio

The main tasks involved are data collection, data preprocessing, data cleaning, extracting the data from SQLite using Python, identifying top N players of Europe using entire dataset in R. We verify these results using the SQL queries for the same. We use the processed and extracted data to calculate entropy across leagues to identify the most predictable league.

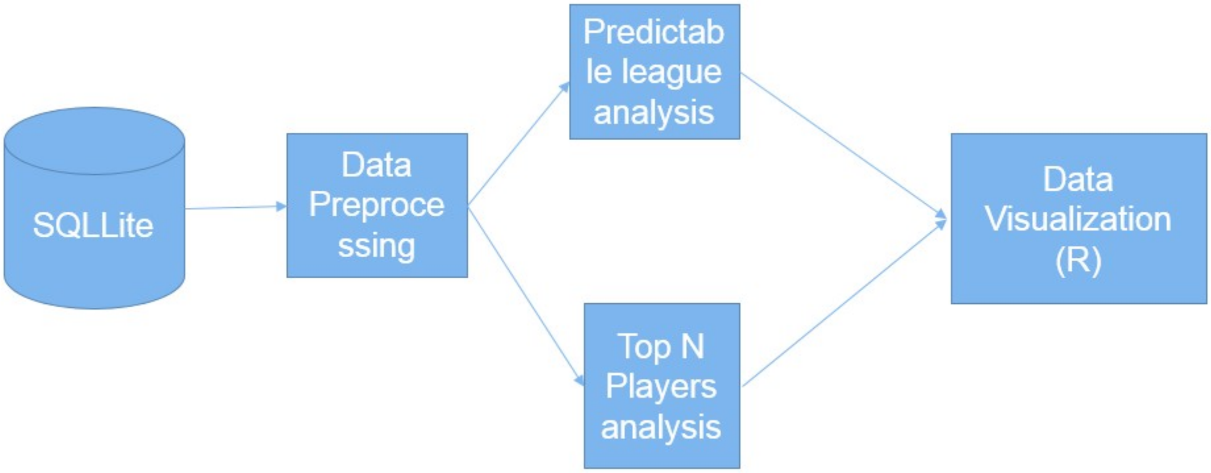
Then we create visualizations using modern age charts which are visually appealing and at the same time depict the desired results as well. The visualizations include league prediction, player comparison's based on attributes and the like.

# 3. DATABASE ENVIRONMENT:

## 3.1 CLIENT PROFILE:

The main clients are Online Gambling Companies and websites which host Fantasy League. Soccer is a major sport in Europe and there are at least 30-40 major games each week across leagues. Millions of people are involved in betting for these games and equal amount of people are excited to participate in fantasy leagues. Our data and the results we provide would be fantastic source for these clients to give some insight to their users. We provide result predictions with most predictable leagues and with your top N player’s insights can be produce for visualizations, which users of theses client websites might find helpful. With our results, it would be great for these clients to attract and encourage more users to involve in match betting’s and fantasy leagues.

## 3.2 PRODUCT/INFORMATION FLOW DIAGRAM:



## 3.3 USER PROFILE:

The major users are the bookies and sports analysts in respective companies and also, lot of users who play betting and take part in fantasy leagues.

The bookies can use the predictions of league that we have for fixing the pre-game odds and the Top N Players analysis can be used for the impact these players could have. Betting’s on goals scored, impact player, man of match and in impacting the final result of the game which could change the odds are some statistics bookies can use.

Sport analysts could use this to perform to the tasks that we are doing. Tasks like predicting each game, impact of home games can be performed using the data. General audience who use the online gambling websites and play fantasy leagues actively are our primary users who could use the results and make smarter decisions while placing bets and making team changes.

## 4. GOALS AND OBJECTIVES:

The main clients are Online Gambling Companies and websites which host Fantasy League and the major users are the bookies and sports analysts in respective companies. Bookies use it to predict win/loss/draw for betting. Data analysts use this data to do future analysis to make the right buys during transfer window, understand a player’s body condition including his stamina, fatigue etc.

### 4.1 CLIENT/USER GOALS:

One can use the database to enter, store and access information about each league, team, player and matches:

- Country and League details including unique ids for all.
- Player details including name, height, weight and birthday.
- Player attributes including rating, potential, preferred foot, shooting, dribbling, defending etc.
- Match details including goals, date, season etc.
- Team attributes with team rating and team strengths and weakness
- Different leagues details.
- Predict game results, man of the match based on opponents and external factors.
- Predict player’s form.
- Number of teams, players, countries, leagues.

### 4.2 CLIENT/USER REQUIREMENTS:

- An id to indicate country, league, match, players.
- A field to indicate overall rating of the player based on his attributes.
- A field to indicate overall rating for the team based on the players.
- A field to predict the player’s value in the future.
- A field to predict a player’s current form.

### 4.3 DATABASE GOALS:

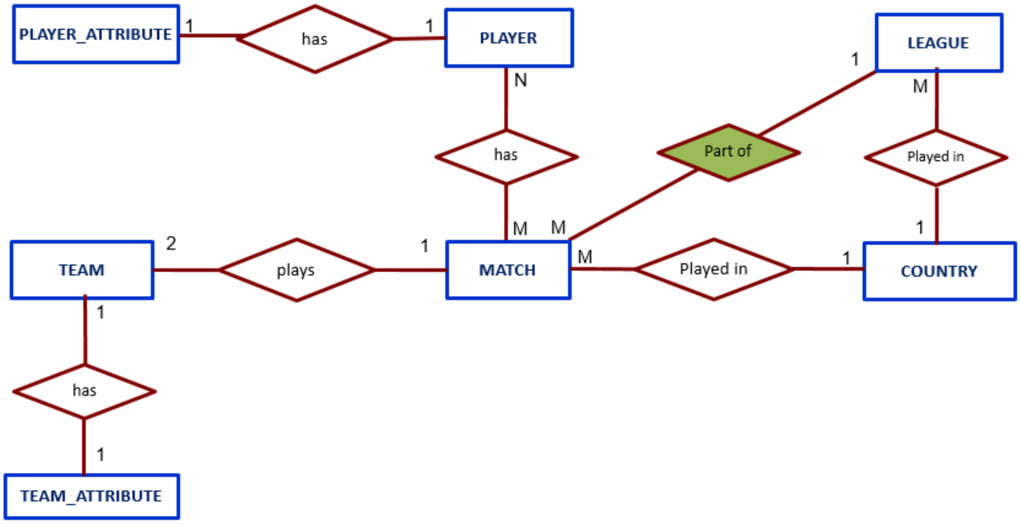
- A thorough data collection and processing of data so that the clients get easy access to the data.
- Eliminate the need for a separate spreadsheet to track attribute rankings from the recommendation forms.
- Create a more efficient and uniform method of storing and accessing information about the scholarship applicants.
- Create a database that is easy/intuitive for users with various levels of computers skills since data will be entered by various members associated with FIFA.
- Easily generate reports from the given data.
- Identify the most predictable league of Europe.
- Identify top players of the Europe.

# 5. DATABASE PROFILE:

## 5.1 BUSINESS RULES:

- Each player plays in in one or more matches. Each match is associated with 22 players.
- Each Player is associated with one Player Attribute. Each player attribute is associated with one player.
- Each match will be played in one country. A country can host many matches.
- Each match will be part of one league. A league can have many matches.
- Each match has 2 teams and each team plays 1 match.
- Each league can be part of single country. A country can have many leagues.
- Each Team is associated with one Team Attribute. Each Team Attribute is associated with one Team.

## 5.2 ER Diagram:

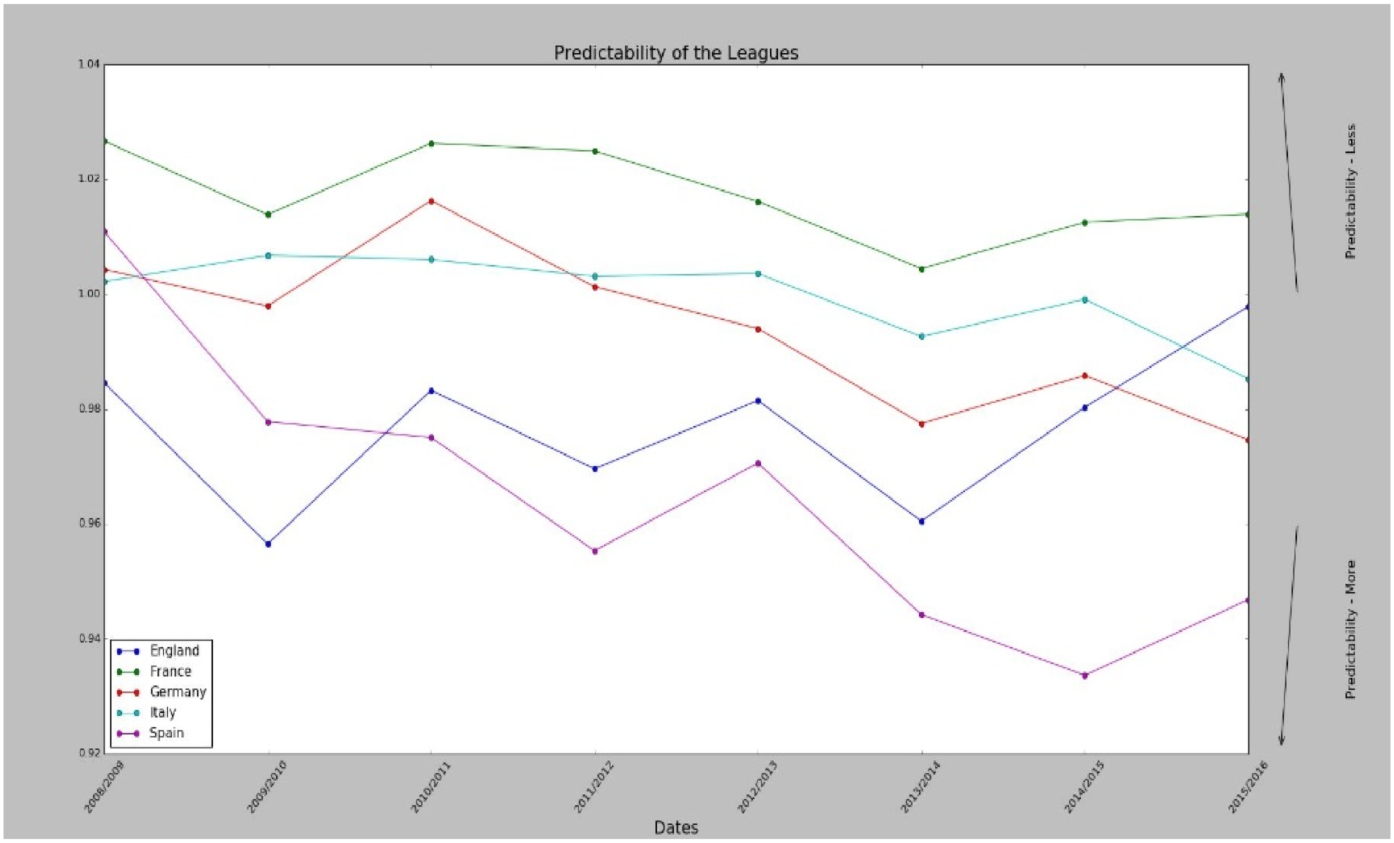


## 6. EXPERIMENTS AND RESULTS:

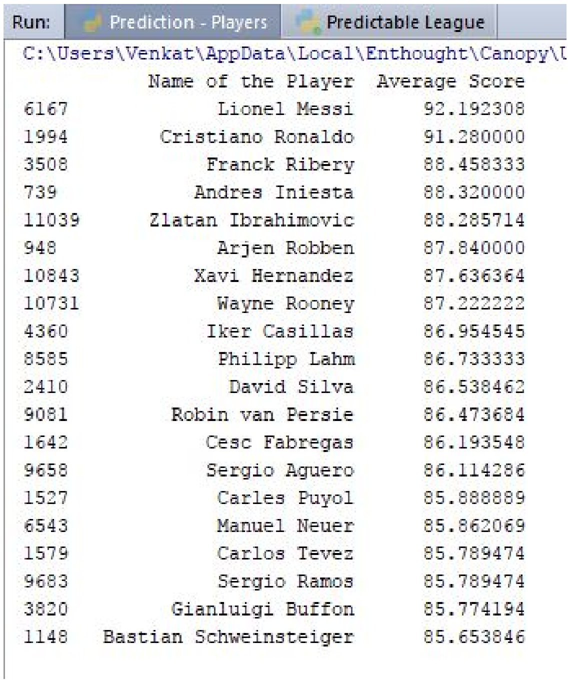
Below are the final visualization results obtained:

1. To build the ‘Most predictable league model’, we made use of probability and mean entropy which was implemented in Python using Pycharm. The result was plotted as below.

$$H = -\sum p(x) \log p(x)$$

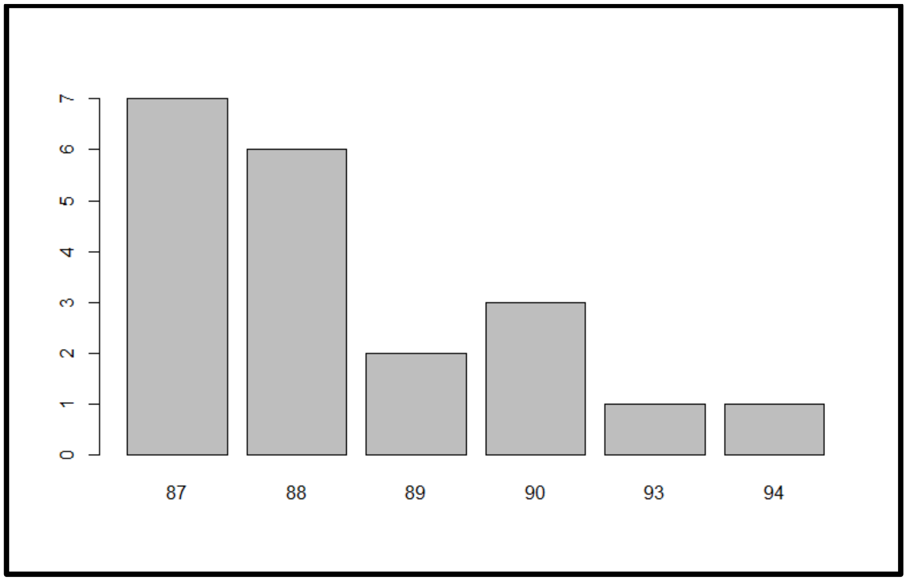


2. To estimate the top 20 players based on league/country/overall, We calculated the average of any specific parameter over all the matches recorded, in the ‘player_attributes’ table and sorting it in descending order with respect to country/league/overall.

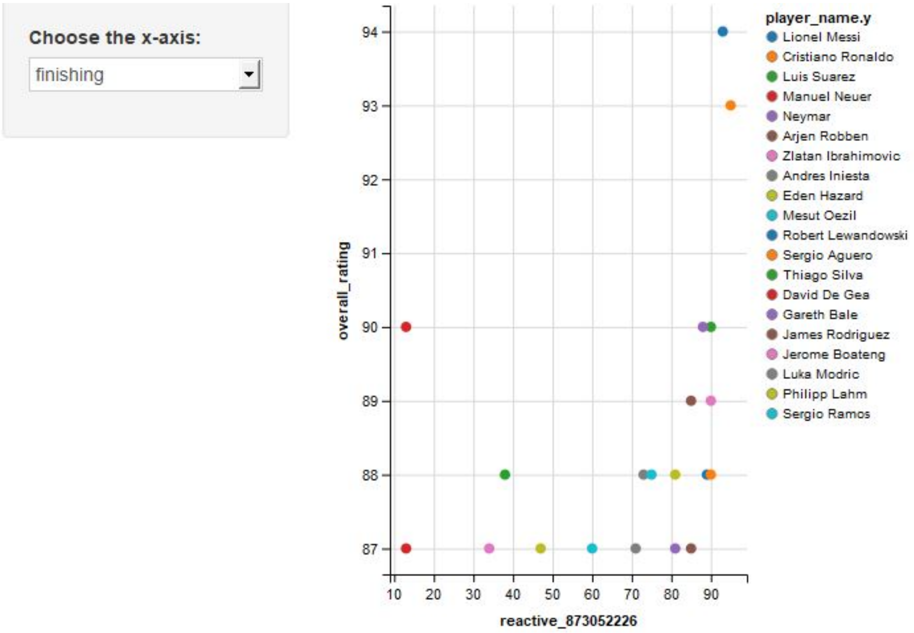


```
C:\Users\Venkat\AppData\Local\Enthought\Canopy\l
           Name of the Player  Average Score
6167             Lionel Messi      92.192308
1994        Cristiano Ronaldo      91.280000
3508            Franck Ribery      88.458333
739            Andres Iniesta      88.320000
11039       Zlatan Ibrahimovic      88.285714
948              Arjen Robben      87.840000
10843         Xavi Hernandez      87.636364
10731            Wayne Rooney      87.222222
4360            Iker Casillas      86.954545
8585             Philipp Lahm      86.733333
2410              David Silva      86.538462
9081         Robin van Persie      86.473684
1642            Cesc Fabregas      86.193548
9658            Sergio Aguero      86.114286
1527             Carles Puyol      85.888889
6543             Manuel Neuer      85.862069
1579             Carlos Tevez      85.789474
9683             Sergio Ramos      85.789474
3820         Gianluigi Buffon      85.774194
1148   Bastian Schweinsteiger      85.653846
```

3. The below graph gives the range of overall rating of Top 20 players along its frequency. This is helpful to check if given a players rating, wether he falls in top 20 players and also predict in what percentage.

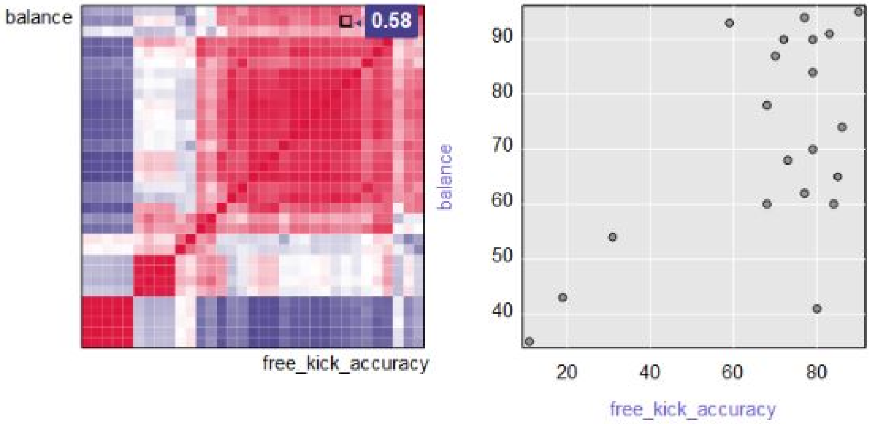


4. The relationship between the overall_score and other numeric variables is plotted using an interactive scatter plot using R.

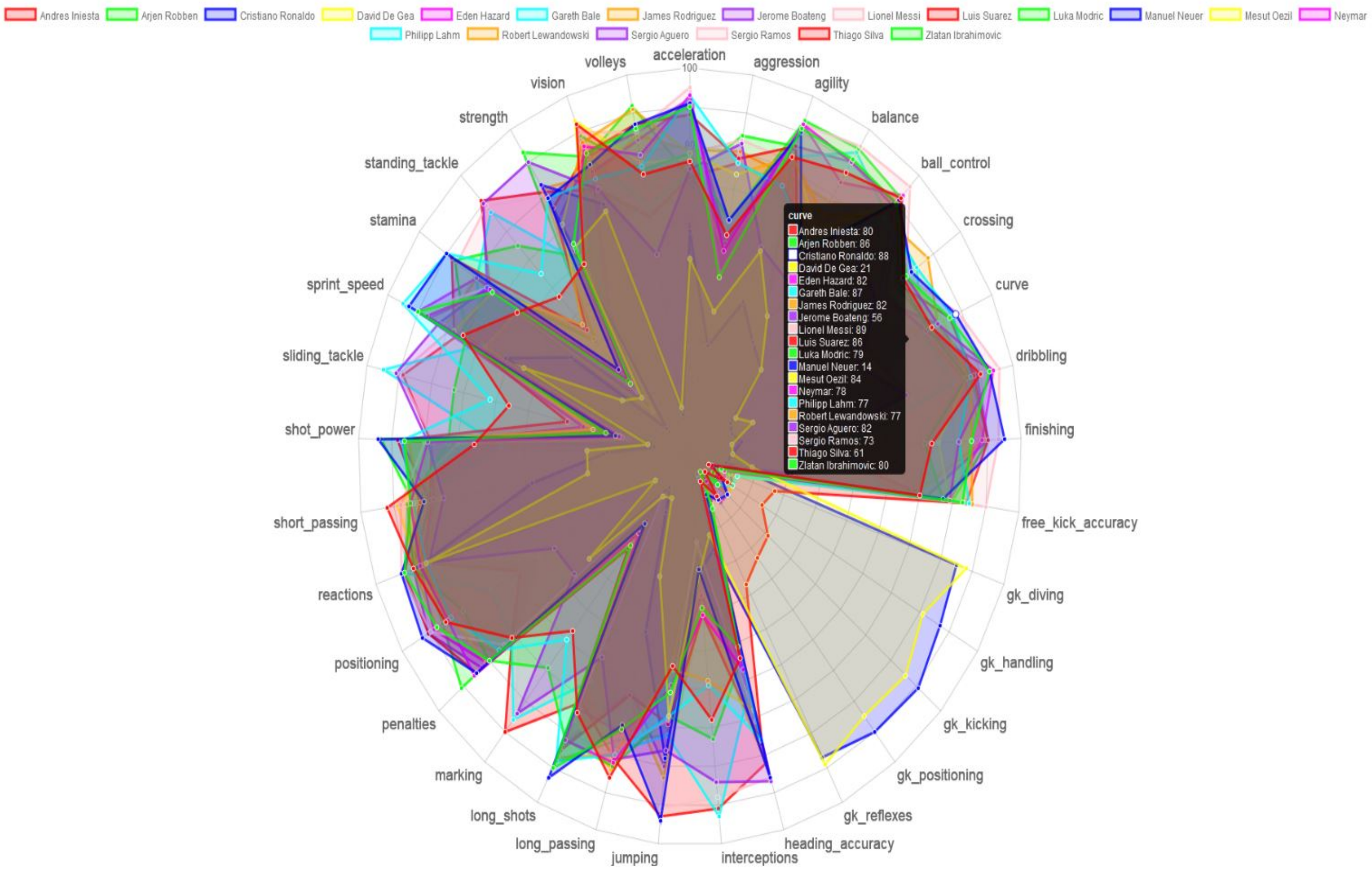


Choose the x-axis: finishing

5. We have also implemented an interactive correlation matrix amongst all attributes. If you hover at a point it will show the correlation value between two attributes at a time. As seen in graph below, the correlation value between 'balance' and 'free_kick_accuracy' is 0.58. This may be used in future for attribute selection when we use a different model like regression.



6. We have implemented a visualization to view complete profile of individual players using a radar chart. This is currently being used in most of the Soccer online games and betting companies.



# 7. PROJECT ASSESSMENT:

## 7.1 CHALLENGES/LIMITATIONS

- The dataset is too large, it needs to be optimized to run faster queries.
- Need to integrate data and technologies under one platform.

## 7.2 FUTURE WORK

- Optimize the dataset by the process of normalization without any loss of data.
- Provide admin rights using sophisticated encryption keys or RSA.
- Develop a web-app using duo authentication.
- Implement an interactive UI using HTML5, CSS and jQuery.
- Embed high charts and advanced plugins to showcase the visualizations in the web-app.