Analysis of FIFA 18 Dataset and Predictions

**A PROJECT ON**

**“ANALYSIS OF FIFA 18 DATASET AND PREDICTIONS”**

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE COURSE OF

DIPLOMA IN

BIG DATA ANALYTICS FROM CDAC



**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY**

‘Plot no R/2’, Market yard road,

Behind hotel Fulera,Gultekdi

Pune – 411037.

MH-INDIA

**SUBMITTED BY:**

Nehal Rohidas Bandal

**UNDER THE GUIDENCE OF:**

Mr. Girish

Faculty Member

Sunbeam Institute of Information Technology, PUNE.

Analysis of FIFA 18 Dataset and Predictions



**Index**

**CERTIFICATE**

This is to certify that the project work under the title ‘Analysis of FIFA18 dataset and predictions’ is done by Nehal Rohidas Bandal in fulfilment of the requirement for award of Diploma in Big Data Analytics Course.

**Mr. Girish Gaikwad**

**Project Guide**

Date: 30/January/2019

Analysis of FIFA 18 Dataset and Predictions

**ACKNOWLEDGEMENT**

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudhale (Centre Coordinator, SIIT, Pune) and Mr. Nilesh Ghule and Project Guide Mr. Girish Gaikwad.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form. Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Nehal Rohidas Bandal

DBDA August 2018 Batch,

SIIT Pune

Analysis of FIFA 18 Dataset and Predictions

**TABLE OF CONTENTS**

**1. Introduction of Project**

1.1 Statement of the problem

1.2 Project Objectives

1.3 Overview of proposed solution approach

**2. Project Overview and Summary**

2.1 Analysis

2.2 Squad Management

2.3Technologies Used in Project

**3. System Design and Architecture**

3.1 System Flow diagram

3.2 Control Flow Diagram

**4.** **Exploratory Data Analysis**

4.1 Geographic Distribution of Players

4.2 Club Profiles

* 1. Scouting

4.4 Wage and Market Value

**5. Predicting Player Position and Overall Ratings**

5.1 Data Collection and Description

5.2 Data Preparation

5.3 Feature Selection

5.4 ML Algorithm Selection

5.5 Data Storage

**6. Building Best Squad**

7. **Application Overview**

**8. Conclusion and Future Work**

Analysis of FIFA 18 Dataset and Predictions

**1. Introduction**

Football is very popular global sport. Not just playing football, but certainly every country possesses that world class talent. Football is no just about kicking the ball. Lots of brains involved in build-up of even a single goal. All clubs do their homework before entering in the match. They study the pattern of opponent’s football. The tactics they've used in previous games. The study is now goes even deeper, now are examining the important players from opposition.

In an aim of becoming strong team, every club is focusing on recruiting talented players in their squad and nurture them by as they are the future prospects of their team. But since this game has spread all along, finding hidden talent in country where were class of football is not at good level is cumbersome task.

Clubs still follow traditional approach for recruiting new players by sending their scouts. Every club invests huge amount of money in scouting of young players. In this era of football, from broad to minute details of this game is available. Clubs are now moving from their old school techniques to modern approach.

Big clubs analyse every available data so that and find out useful insights about game, players. The analysis of data is going to play major role in recruiting players in club. Following modern day approach will play crucial role in any clubs revolution.

**1.1. Statement of the problem**

Clubs still invests lots of money in finding new talent. There are many talented footballer are not in eyes of the big clubs. There are some clubs who are not aware about not keeping track of their player’s growth. The purposed project will help clubs in finding best talent all around the globe. This project will be great help for the clubs in strengthening their squads and building youth academy. The analysis done in this project will play crucial role in deciding clubs strategy for upcoming transfer season. Analysis also helps club in keeping track of player’s development.

**1.2 Project objectives**

1. Perform exploratory analysis on FIFA 18 dataset.

2. Predict part of field where player is going to play.

3. Predict his exact position

4. Predict players overall ratings.

5. Best team for user given formation

Analysis of FIFA 18 Dataset and Predictions

**1.3 Overview of proposed solution approach**

To achieve all the objectives two main techniques are used. For predication of ratings and classification of field position of player machine learning is used.

For providing various useful insights with the help of analysis, Data visualization tool tableau is used. The final application is developed using Python and Flask.

Analysis of FIFA 18 Dataset and Predictions

**2. Project Overview and Summary**

Project is divided into two parts:

1. Analysis

2. Squad Management

**2.1 Analysis**

Analysis of continent and countries: Their overall strengths, potential. How much and how quality young talent each country posses

Analysis of club: This analysis shows players list along with their value and wage, from which country club is recruiting players, their attack, midfield and defensive strengths. Using this analysis top player from club attribute-wise can be found out.

Analysis to help scouting: This analysis will help club in recruiting new talent in their squad. This analysis will be great help if club want to build their youth academy.

Analysis on transfer market: This analysis will help club to formulate their strategy for upcoming transfer-window.

**2.2 Squad Management**

At start there is login. Managers need to login first before having access to these components.

Squad Management includes list of all players playing in the club. Manager has given facility to see best team for different formations.

Simple form is provided where manager can upload details about newly available players. Using ML algorithms, field position and player rating of players are predicted and finally uploaded on database.

Analysis of FIFA 18 Dataset and Predictions

**2.3 Technologies used**

1. Python - Python (3.6.7) is used as programming language.
2. Flask - Flask (1.0.2) is a web application framework written in Python. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine.
3. Machine Learning - ML Algorithms are used for predicating players exact position and determining overall ratings.
4. MySQL - The data after cleaning and passing through machine learning algorithms is stored in MySQL databases which will useful for determining best squad for selected formation.
5. Tableau - For performing various analyses on given dataset Tableau is used as visualization tool. Tableau provides interactive stories and dashboards which are crucial part of the application.

Analysis of FIFA 18 Dataset and Predictions

**3. System Design and Architecture**

**3.1 System Flow diagram**

**MySQL**

**Local Data Store**

**Website**

**Flask Server**

Analysis of FIFA 18 Dataset and Predictions

**3.2 Control Flow Diagram**

**Already register-ed?**

**Register**

**Home Page**

**1. Analysis**

**2. Squad Management**

**1. Squad List**

**2. Best team for Selected Formation**

**1. World Football Analysis**

**2. Club Profile**

**3. Scouting**

**4. Player Worth**

**Logout**

**Upload new Players??**

Analysis of FIFA 18 Dataset and Predictions

**Save uploaded file on the disk**

**Read the uploaded file in appropriate data frame format**

**Remove NaN Values**

**Load the previously stored models.**

**Predict player Position**

**Predict player’s overall ratings**

**Upload complete data frame in MySQL database**

**Redirect to squad management page.**

Analysis of FIFA 18 Dataset and Predictions

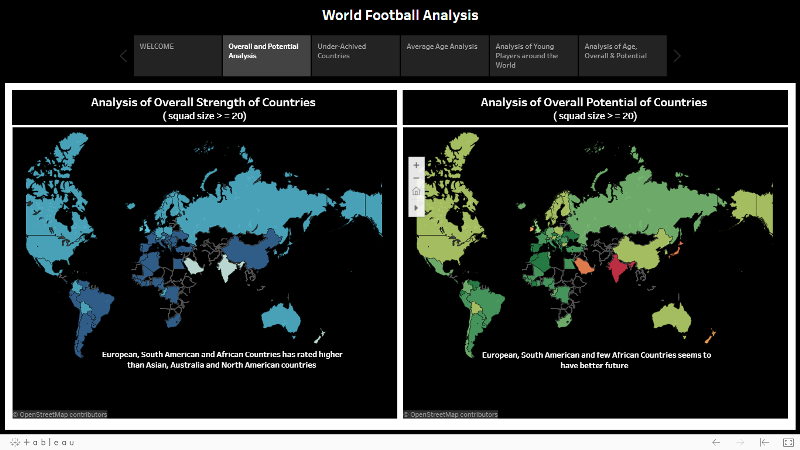
**4. Exploratory Data Analysis**

After preparing the dataset for analysis, explorations were made on a macro level like continent and player nationality level analysis as well as on player level. Through the process attempts were made to derive interesting correlations and trends by the use of visualizations.

**4.1** **Geographic Distribution of Players**

This analysis shows the spread of football around the world. How many players from each country are currently playing at international level of football.

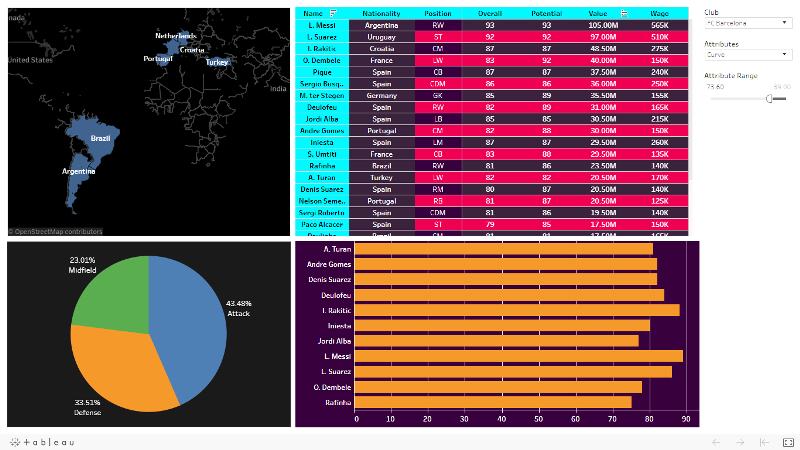
Geographical analysis helps to identify overall strengths of players from different part of the world. This analysis useful to compare strengths of countries. Find out which countries are stronger, which countries have very bright future. This analysis is useful to take a look at the countries which are not playing at standard level. Which are those countries where clubs can look for the best young talent can be find out from these analyses.



Analysis of FIFA 18 Dataset and Predictions

**4.2** **Club Profiles**

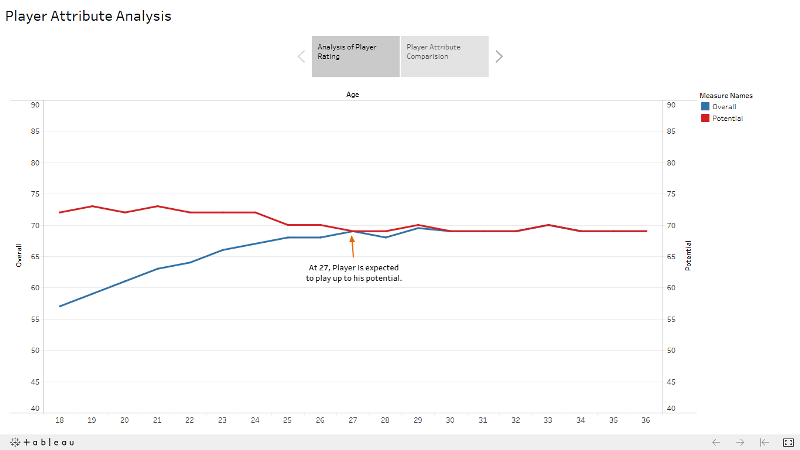
Club profiles are very useful to understand the teams recruiting strategy. Find out highest purchase players, Club's overall strengths at attack, midfield and defence. Attribute-wise top players can be found out, which is very helpful for club to target certain players in next transfer window.

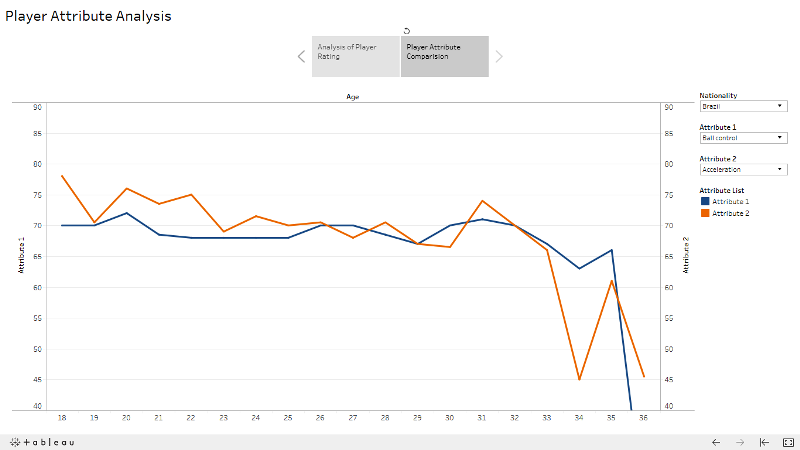


Analysis of FIFA 18 Dataset and Predictions

**4.3** **Scouting**

For selected features which country contains players that are good at these features, this is the analysis clubs has to look for. E.g. Analysis shows that Brazil produces better forwards than any other country. Overall v Potential for all the ages shows that player plays at his full potential at the age near to 27.

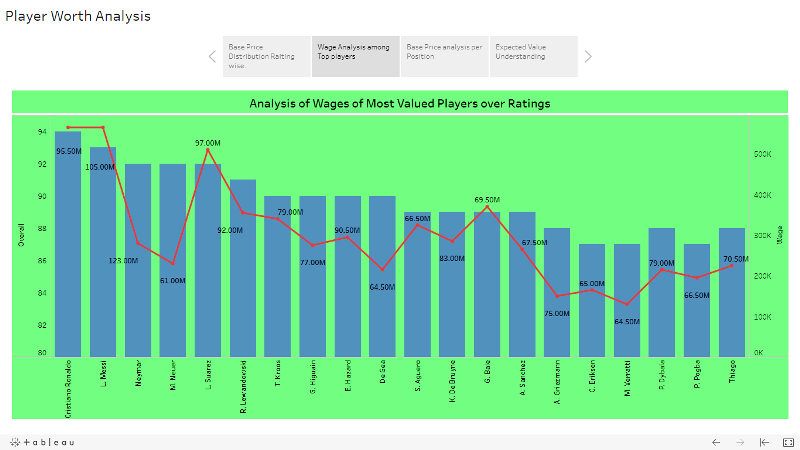


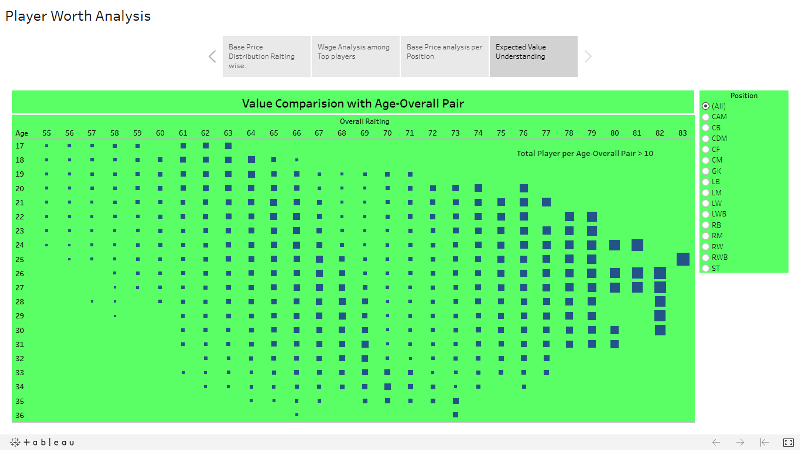


Analysis of FIFA 18 Dataset and Predictions

**4.4** **Wage and Market Value**

Here analysis is done on distribution value over the age. While comparing value, wage for overall and potential of 20 most valued players, it was observe that market value is more dependent on potential of the player whereas the player wages are more affected by the experience of the player and their seniority in teams. The increase in player weekly wage is also much more gradual with age as compared to player value, which was prone to much variation.





Analysis of FIFA 18 Dataset and Predictions

**5. Predicting Player Position and Overall Ratings**

The first objective consist of 4 classes Attack, Midfield, Defence and GK which indicate that where the player is going to play. Second objective focuses on predicting player’s exact position. There exist total 19 positions i.e. 19 classes.

Forward: ST, CF, RF, LF, RW, LW

Midfield: CAM, CM, LM, RM, CDM

Defence: CB, RB, LB, RCB, LCB, RWB, LWB

Goalkeeper: GK

Overall rating of player is on scale of 0 to 100.

**5.1 Data Collection and Description**

Data collected from Kaggle. Data developed by Electronic Arts for the latest edition of their FIFA game franchise. Through several research projects done on soccer analytics, it has been established in the field of academia that the use of data from the FIFA franchise has several merits that traditional datasets based on historical data do not offer. Since 1995 the FIFA Soccer games provide an extensive and coherent scout of players worldwide.

For each attribute, we have an integer from 0 to 100 that measures how good a player is at that attribute. Examples of attributes are: dribbling, aggression, vision, marking and ball control. Observe that it seems to be unfeasible to accurately characterize players in these attributes automatically. Thus, all of those are gathered and curated by the company whose job is to bring the gameplay closer to reality as possible, hence preserving coherence and representativeness across the dataset.

The FIFA 18 dataset that has been used for this analysis provides statistics of about 17930 players on over 78 different attributes. These attributes are optimal indicators to determine the performance of a player at a particular playing position.

**5.2 Data Prepration**

Since all columns in dataset are not required for completion of objectives, these unnecessary columns such as Sr.No, Photo, Flag, Club Logo, and Special are removed.

Certain player names contain accented characters. These are converted into normal English words.

The Wage and Value of all the players were in form of strings with the symbol of the currency in front of the values. These values are converted to numeric values.

Analysis of FIFA 18 Dataset and Predictions

Some records contain missing values. These missing values are replaced by appropriate substitutes.

There are values for physical attributes in the form like '55+2' so need to compute this values.

There is need of introducing some new column such as 'continent', field position, position. These columns are added into the dataset.

**5.3 Feature Selection**

It is very important step in any machine learning project because it provides following advantages:

1. Reduce over fitting and complexity of model

2. Improve accuracy

3. Reduce training time

By using knowledge about the game, un-important features such as name, age, club, continent, value, wage are dropped. The only features that are important for selected objective are players attribute. So the player attributes as used as independent columns.

Next observe the correlation plot between selected features. There exists some amount of collinearity between some features. The GK attributes such as GK Diving, reflexes, kicking, handling, positioning are highly correlated with each other. So only kept GK diving attribute, which is sufficient for identifying GK.

Another pair where collinearity is found between sprint speed and acceleration. Removing anyone of them is resulting in decreasing accuracy. Removing them is will result in loss of important information which is responsible for classifying players at different position, So decided to keep these features.

To remove this collinearity without removing features there are two ways to achieve this. First is Principle Component Analysis (PCA) and second is Linear Discriminant Analysis (LDA). Both Linear Discriminant Analysis and Principal Component Analysis are linear transformation techniques that are commonly used for dimensionality reduction. PCA can be described as an “unsupervised” algorithm, since it “ignores” class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is “supervised” and computes the directions that will represent the axes that that maximize the separation between multiple classes.

Both techniques focuses on removing collinearity between existing independent variables by transforming original feature set into new low dimensional feature set. PCA feature set contains vector ordered in descending order of variance. 1st vector in feature set captures maximum features. next vector captures 2nd most variation and it is orthogonal to 1st principle component indicating that there is zero correlation.

Analysis of FIFA 18 Dataset and Predictions

The general LDA approach is very similar to a Principal Component Analysis (for more information about the PCA, but in addition to finding the component axes that maximize the variance of our data, we are additionally interested in the axes that maximize the separation between multiple classes.

**5.4 ML Algorithm Selection**

In ML Algorithm selection process, complete emphasis is on accuracy and performance both.

Looking at dependent variable, there exist more than two classes. So Naive Bayes is not suitable algorithms for completing objectives.

Main assumption in Logistic regression algorithm is that there is no collinearity in the independent variables but in selected features there exist collinearity. After performing PCA and observing scree plot it is decided that select top 4 features as they contains maximum variation. So using PCA with Logistic regression gives accuracy of 55%. Alternate approach was using Linear Discriminant Analysis. LDA focuses of maximizing the between class distance and minimizing within class distance. Looking at classes there is very less difference in two classes as attribute values required to distinguish between two classes are very close. So LDA was best choice for this.

Using LDA with Logistic regression yields accuracy of 65%. Other algorithms which are used for classification are Support Vector Machine (SVC) and K-Neigherst-Neighbours (KNN) which takes to much time to process and give poor accuracy, around 38%. Decision tree has given 67% accuracy on predicting where player is going play.

Looking model complexity, using ensemble is best approach. Ensemble model always provide best result as they focused on making predications accurate. Ensembles internally use decision trees. Tree base models are immune to any collinearity between features. Because they use entropy calculations for building tree and entropy calculations involve probabilities. So there is no effect of collinearity.

RandomForest is homogenous bagging method which provides 94% accuracy after fine parameter tuning. RandomForest builds many decision trees parallely and select final result based on output of all the decision trees. For classification problem it takes the mode of outputs of the decision trees and for predication problem it takes average of outputs of the decision trees. The drawback of RandomForest is it takes lot of time to give final result.

Gradient Boosting is boosting technique which gives 96% of accuracy. Gradient Boosting builds decision trees sequential. It focuses of correcting those observations which are previously classified wrong. This algorithm also takes too much time to yield result.

XGBoost is advanced version of gradient boosting. XGBoost gives accuracy around 97. XGBoost internally implements gradient boosting algorithm. XGBoost is mainly focused on making computations faster. It uses all cores of CPU.

Analysis of FIFA 18 Dataset and Predictions

So the result obtained by XGBoost was very quick compared with other algorithm. XGBoost has an option to penalize complex models through both L1 and L2 regularization. Regularization helps in preventing overfitting.

Parameter need to tune while using XGBoost are n\_estimators and NThread. n\_estimators indicate number of decision trees need to be built to get maximum accuracy. After checking

Performance and accuracy for different values of n\_estimators, 125 is the value which gives better accuracy and take lesser time compare to other algorithms.

**5.5 Storing data in database**

After cleaning and predictions entire data is stored in MySQL table. So this data can be retrieved on web page. The data is stored in 4 different tables. First table contains players’ personal information. Second table contains data related to physical attributes. Another table contains overall ratings for each field position for each player and finally one table contains abstract view of player which includes player's wage, value, position, overall ratings.

Analysis of FIFA 18 Dataset and Predictions

**6. Building Best Squad**

In football there exist various formations like 4-4-2, 4-2-3-1, 5-3-2 etc.

Using players’ statistics, Best squad for user selected formation is obtained. For this section data stored in MySQL is used.

Consider 4-4-2 formation.

4-4-2: GK, LB, CB, CB, RB, LM, CM, CM, RM, ST, ST

This is normal formation but teams always look for flexibility in their formation. In ST position manager might prefer CF, RF or LF who are playing better than player at ST position.

The same case is for other positions too. So program is built like that it will consider all alternate options existing for position before displaying best squad for selected formation.

Analysis of FIFA 18 Dataset and Predictions

**7. Application Overview**

The web application is developed using Flask server. Flask is a web application framework written in Python. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. To developed web pages, HTML and CSS are used. To access python lists and tuple on web page Jinja language is used.

The application flow is as following:

1. Manager needs to login initially. if not registered then register first.

2. Manager will directed towards home page where he will see two options

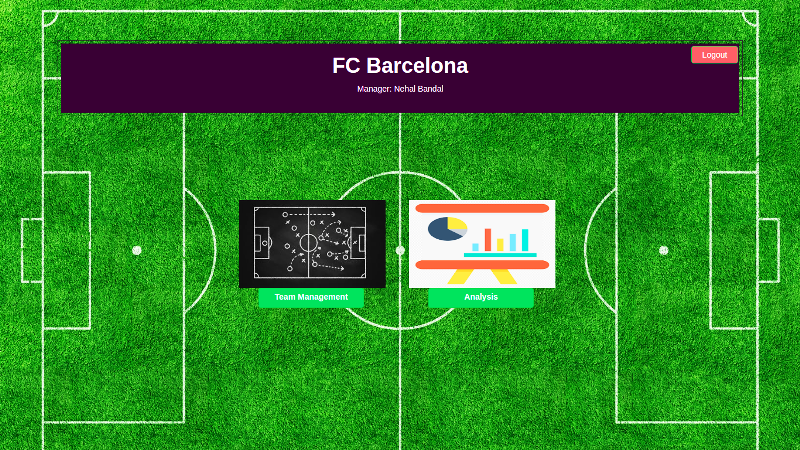
A. Team Management

B. Analysis

3. In team Management tab, manager will able to see his current squad and best playing 11 as per selected formation. One small form is also provided so that he can upload file of newly available players. After submitting file, each player in list will be assigned to his exact position and his overall ratings and then this complete data will be uploaded in database.

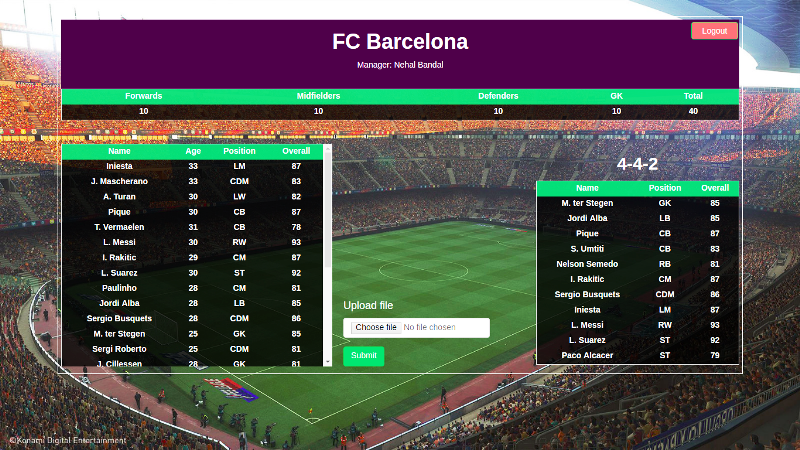
4. Analysis Tab, displays various analysis done in tableau. which include analysis of world football country and continent wise, Analysis of club, Analysis of value and wages, analysis which will help in recruiting new players.

5. To end the session manager needs to click Logout button.

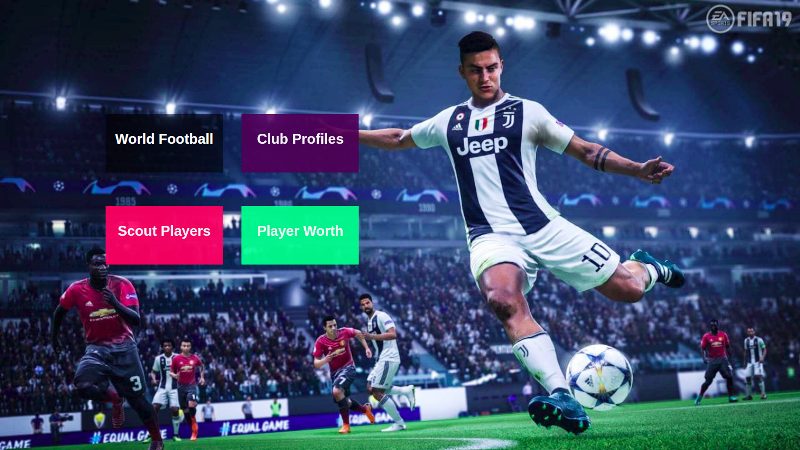


Analysis of FIFA 18 Dataset and Predictions

**Team Management Page**



**Analysis Page**



Analysis of FIFA 18 Dataset and Predictions

**8. Conclusion and Future Scope**

**8.1 Conclusion**

Deciding suitable position for player and analysing flexibility of player is really difficult job for managers. Thus, I've attempted to make use players physical attribute to predict suitable position for players and also predicting players overall ratings for all the positions which indicate players flexibility.

Using analysis, scouting of players around world is been made easy. Providing overview of recent trends in transfer market, help clubs to perform better in upcoming transfer season.

**8.2 Future Scope**

1. Assigning clubs to the country, league in which they are playing. This will help to focus on particular club in particular country which will be great to perform analysis on particular league.

2. Creating player's profile page which includes players game stats such as goals, assists, shot on target, passing accuracy, etc. along with physical attributes, this will improve scouting and help immensely in recruiting of players.

3. List of club’s fixtures can be added to squad management page.