End Semester Project
Analysis of FIFA 2019

UE17CS322 : Data Analytics

Team Name : DASK_19

Ketan Panwar PES1201700195
Subhranil Das PES1201700309
Section 5CSD
PES University

***Abstract* — This Project report includes analysis and prediction of several parameters of the dataset containing information of football players provided by FIFA 2019. This analysis includes an explanation of the approach used and observations found during the analysis.**

Keywords — Fifa 2019, Analysis, EDA Soccer, KNN, K-means clustering, Random Forest Regression, Linear and Polynomial Regression.

1. Problem Statement

We aim to assist the club managers to select the best possible team based on on-field abilities, analyse player's performances based on various fields , group similar players and predict how much to pay them and what their value is.

2. Introduction

FIFA 19 is a football simulation video game developed by EA Vancouver as part of Electronic Arts FIFA series. The club managers might face several issues such as finding a good team in a given budget, finding substitution for an injured player, recruiting new players and deciding beforehand what their initial salaries should be.

Our model aims to address these issues in the following manner :-

We are using the dataset to help club managers choose players based on potential and skill, grouping/clustering players having similar features. At the same time, we also compare existing teams on different levels. We find similarities between different players and suggest four substitutions for a player also considering their on field positions using KNN algorithm. We use polynomial regression model to predict wages of players, random forest regression model to predict the value of a particular player. We also suggest an entire football team ,All the on field position, based on statistics. We also group various players based on different parameters using K-means clustering.

3. Previous Work

Citation 1:

https://www.kaggle.com/applecider327/football-aNalytics-analysing-the-fifa-19-dataset

1. The author drops the column 'Wage' , which might be a deterministic factor for selection of players in clubs.

2. Our goal is to help the club managers select the best possible teams. Teams have to be diverse. Dropping the column 'Preferred Foot' might not be the best option as the club manager may end up choosing a team with most right footed players. Formulating a strategy to defeat such a team would be Easy.

3. The author is trying to measure the accuracy of the overall rating given for goalkeepers. For this he uses a self formulated field of 'clean sheet rate' , which counts the number of games where a goalkeeper concedes no goals. A goalkeeper conceding no goal may not necessarily be the goalkeeper's talent. It might also be the case that the other players of the same team rarely allowed the ball to touch the goal line of their side goal. Alternatively, we can use any

external data set having information of the goalkeeper's saves when he gets an opportunity to save.

4. The author is trying to measure the accuracy of overall scores of defenders. For this he uses a self formulated field of 'defence score' , which has the same problem as the previous point. The defender is alone responsible for the score.

5. The point worth mentioning is that the author performs the above calculations based on the performance of only 4 groups of 10 members. This statistic may not be enough to judge the accuracy of any field.

Citation 2:

https://www.kaggle.com/roshansharma/fifa-data-visualization/notebook1.

The author is trying to fill missing values of several fields inappropriately.

2. In 'Contract Valid Until' the values used is 2019 so, after 2019 the model will show those players available but actually they won't be.

3. The above problem pertains with 'joined'
and 'Club' field also.

4. Explanation of the Components of The System

A. ***Importing Data:*** We have downloaded a dataset named 'data.csv' and stored it in a local folder. This dataset contains information about football players on 89 different attributes.

B. ***Preprocessing:*** We remove unwanted variables that do not help make any predictions or analysis for our model. This leaves us with 49 attributes.

C. ***Cleaning:*** During analysis we find that most columns have missing values in 48 rows. Since this is a very small number compared to our dataset, we remove all rows with missing values.

D. ***Exploratory Data Analysis:*** We do the following things in this section:

No of unique clubs and no. of players in each such club.

1. Find the club which has maximum no. of players.
2. Finding Different Countries and No of Players in them.
3. Find Top Performer.
4. Plotting graph between Overall Performance and Age.
5. Plotting the share of each nation in terms of number of players.
6. Finding distribution of Mean Overall Rating among clubs.
7. Observing the effect of age on Wages.
8. Distribution of wages based on clubs.
9. Variation of rating with age.

Observations from all of the above are explained in the next section.

E. ***Finding Similar Players using KNN***: The Club Managers might come across situations where a certain player is unavailable due to some reasons. In these cases we help them by finding 4 most similar players using K-Nearest Neighbours Algorithm. We have used sklearn to implement KNN.

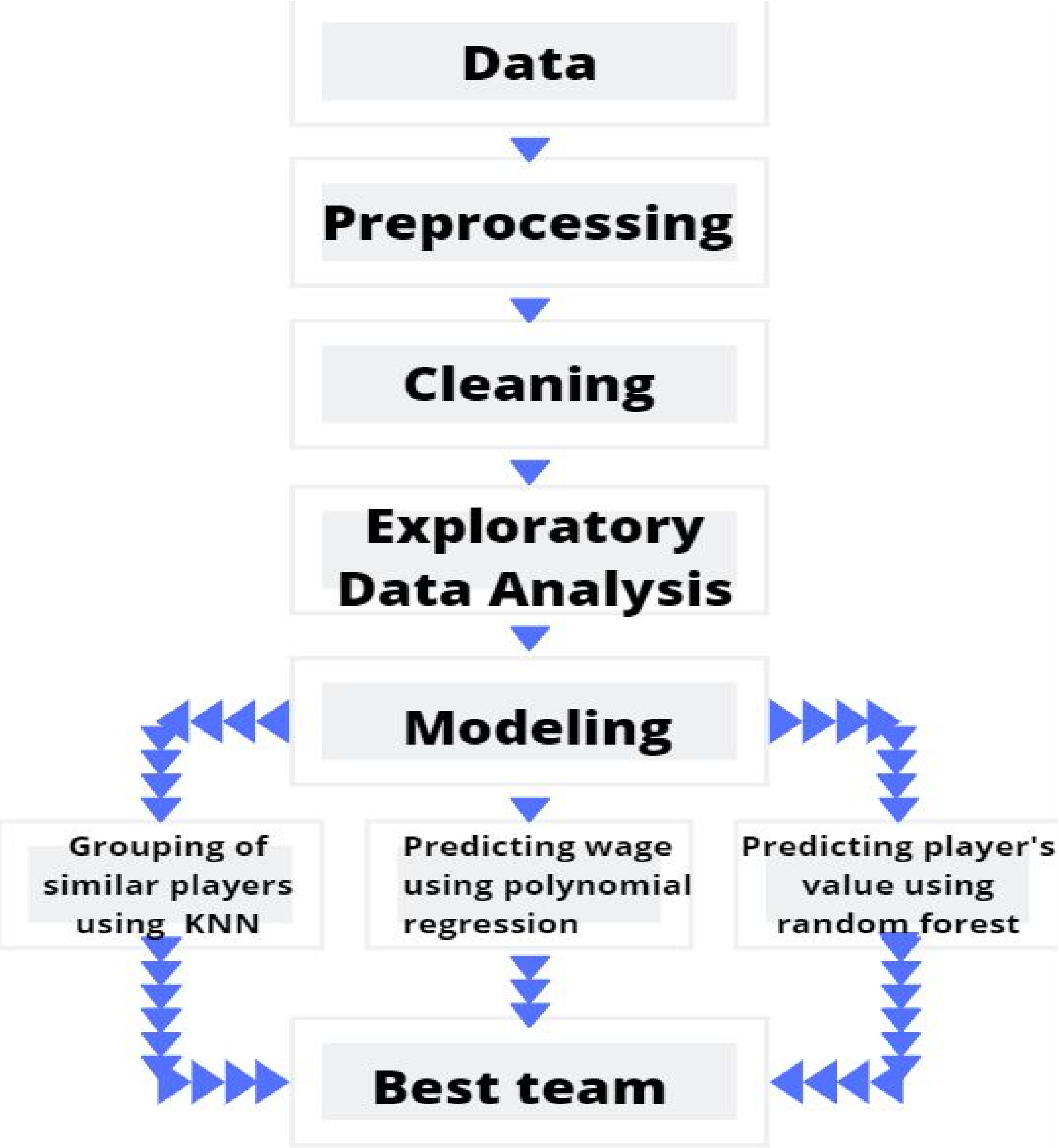
F. ***Predicting Wages using Polynomial Regression :*** When a new player is to be recruited by a club manager, he needs to know how much payment to offer to the player to be able to maintain a general budget. We do this using Polynomial Regression. We have used sklearn to implement Polynomial Regression.

G. ***Grouping Similar Players using K-Means:*** During Auctions, the club managers need to select various players based on various parameters. We provide clustered visualisation of players based on age, potential and overall in one case and the playing abilities in the 2nd case. For both these cases, we use K-Means Clustering. The implementation was done using sci-kit learn.

H. ***Predicting Value using Random Forest Regression:*** In football, value of a player affects several decisions within a club. So, here we predict the value of a players given their potential and overall using random forest regression with 100 estimators. We choose potential and overall because both of them show the same type of trend when plotted against value (shown in EDA). The implementation was done using sci-kit learn.

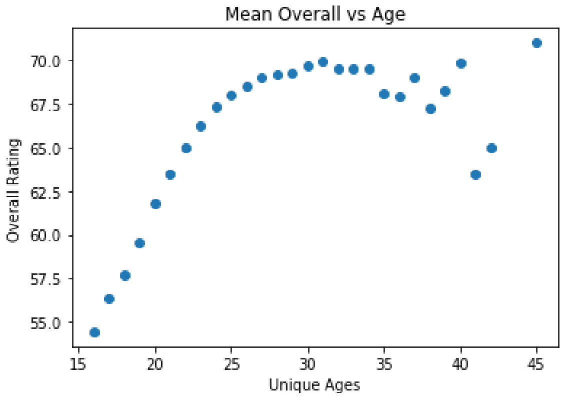
I. ***Finding Best Team:*** Here we tried to step out of the box a bit and formulated a dream team based on most optimal player abilities for all the positions. We did not use any machine learning model for it, rather we depended on statistics.

5. Proposed System - Block Diagram


Data
Preprocessing
Cleaning
Exploratory Data Analysis
Modeling
Grouping of similar players using KNN
Predicting wage using polynomial regression
Predicting player's value using random forest
Best team

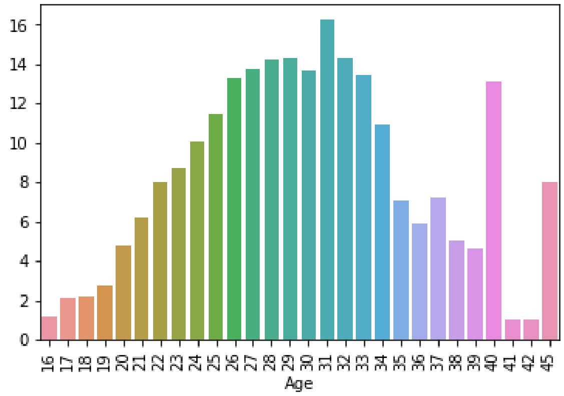
6. Experiments and Results

Mean Overall vs Age:



From the above plot we observe that the mean overall rating increases as the age increases upto 30. Mean Overall Rating remains constant till 35 beyond which it decreases. We infer that young players gain experience as the play over the years, then reach a saturation level, beyond which with age their performance decreases.

Effect of Age on Wages:



Age is maximum for the ages 27-32. It decreases on both sides.

Similar Players using KNN:

These are 4 players similar to L. Messi :
Name: E. Hazard
Position: LF

Name: Neymar Jr
Position: LW

Name: P. Dybala
Position: LF

Name: M. Reus
Position: LM

Wage Prediction:

Lionel Messi has Age = 31, Overall = 94, Potential = 94, Value = 110, Int. Rep = 5

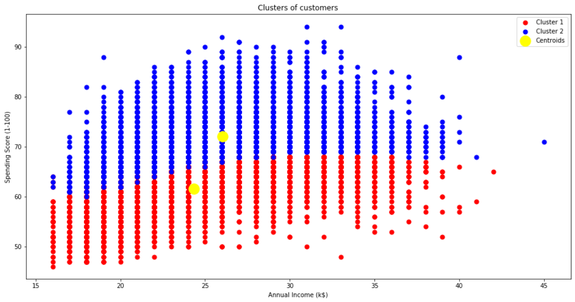
His predicted Wage is 527.37 Million and his actual Wage is 565 Million

Neymar Junior has Age = 26, Overall = 92, Potential = 93, Value = 118, Int. Rep = 5

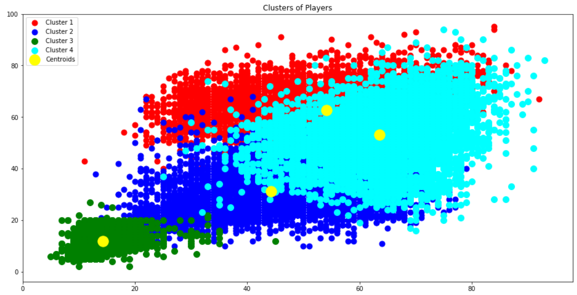
Predicted Wage is 293.44223451170944

Grouping Players using KMeans Clustering:

Age, Overall ,Potential



Player’s Features



Value Prediction:

We are using our model to predict wage for J. Oblak [overall=90 and Potential=93].

Predicted Value is 70.72425
His Actual Value was 68

We are using our model to predict wage for J. Oblak [overall=90 and Potential=93]

Predicted Value is 32.04613730713731
His Actual Value was 34

Recommending a Dream Team:

GoalKeeper :
De Gea [Club : Manchester United , Position : GK , Age : 27]

Forward :
S. Giovinco [Club : Toronto FC , Position : CF , Age : 31]
E. Hazard [Club : Chelsea , Position : LF , Age : 27]
J. Martínez [Club : Atlanta United , Position : LS , Age : 25]
L. Messi [Club : FC Barcelona , Position : RF , Age : 31]
A. Saint-Maximin [Club : OGC Nice , Position : RS , Age : 21]
Cristiano Ronaldo [Club : Juventus , Position : ST , Age : 33]

Midfielder :
H. Nakagawa [Club : Kashiwa Reysol , Position : CAM , Age : 23]
Casemiro [Club : Real Madrid , Position : CDM , Age : 26]
N. Keïta [Club : Liverpool , Position : CM , Age : 23]
Paulo Daineiro [Club : Ceará Sporting Club , Position : LAM , Age : 34]
David Silva [Club : Manchester City , Position : LCM , Age : 32]
N. Kanté [Club : Chelsea , Position : LDM , Age : 27]
Douglas Costa [Club : Juventus , Position : LM , Age : 27]
Neymar Jr [Club : Paris Saint-Germain , Position : LW , Age : 26]
L. Modrić [Club : Real Madrid , Position : RCM , Age : 32]
P. Pogba [Club : Manchester United , Position : RDM , Age : 25]
Gelson Martins [Club : Atlético Madrid , Position : RM , Age : 23]
R. Sterling [Club : Manchester City , Position : RW , Age : 23]

Defender :
D. Godín [Club : Atlético Madrid , Position : CB , Age : 32]
Jordi Alba [Club : FC Barcelona , Position : LB , Age : 29]
G. Chiellini [Club : Juventus , Position : LCB , Age : 33]
M. Pedersen [Club : Strømsgodset IF , Position : LWB , Age : 28]
M. Millar [Club : Central Coast Mariners , Position : RWB , Age : 21]

7. Conclusion

We compared the players using various fields. The main concern we served was to help the club managers choose the best possible team. We also formed clusters based on various parameters. We also have models which predict some unknown parameters using other known variables which helps club managers in recruiting new players.

8. Contribution

Subhranil Das - 50%
Ketan Panwar - 50%
(Everything was done together every time)

9. References

[1] https://en.wikipedia.org/wiki/FIFA_19.
[2] https://www.kaggle.com/applecider327/football-analytics-analysing-the-fifa-19-dataset
[3]https://www.google.com/
[4]https://www.guru99.com/scikit-learn-tutorial.html
[5]https://github.com/sharmaroshan/FIFA-2019-Analysis
[6]https://ieeexplore.ieee.org/document/8697111
[7] https://scholar.google.com/
[8]https://towardsdatascience.com/9-data-visualization-tools-that-you-cannot-miss-in-2019-3ff23222a927
[9]https://stackoverflow.com/questions/4212145/getting-started-with-data-visualization
[10]
https://datascience.stackexchange.com/