PSG COLLEGE OF TECHNOLOGY, COIMBATORE 641004

Department of Computer Science and Engineering

19ZO02 - SOCIAL AND ECONOMIC NETWORK ANALYSIS

ASSIGNMENT PRESENTATION - TEAM 5

By

HARINI S – 19Z317

JAYASREE B S– 19Z322

NIVEDHA K – 19Z336

SELVA KEERTHANA B G – 19Z346

SUCHITHA MALISETTY - 19Z350

BE CSE G2 (2019 - 2023)

Analyzing and Visualizing the graph of football players

Project Description

- Constructing a football graph and analysing various centrality measures and average path distance among footballers.[1]
- Performing an exploratory data analysis (EDA) on the cleaned dataset.
- Predicting the overall ratings/brand value of players of FIFA 2020 by using OLS regression models.[9]
- Classifying the footballers as defenders, midfielders, goalkeepers and forwarders using PCA and K-means clustering.[10]

Dataset Description

FIFA 20 complete player dataset[2]

FIFA is a game released annually by Electronic Arts under the EA Sports label.As of 2011, 51 different countries could access the FIFA franchise, which has been localised into 18 foreign accents. By 2019, the FIFA series had sold more than 282.4 million copies, making it the best-selling sports video game franchise worldwide according to Guinness World Records. One of the most popular video game franchises, as well.

The FIFA dataset comprises of

- Every player available in FIFA 15, 16, 17, 18, 19, and also FIFA 20
- 104 attributes or features
- Player positions, with the role in the club and in the national team
- Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc.
- Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, etc.

Tools used:

Gephi - Gephi is a Java-based open-source network analysis and visualising software suite built on the NetBeans platform. It is a tool used by data analysts and researchers that want to investigate and comprehend graphics. The user's interaction with graph data by modifying the structures, forms, and colours to expose underlying patterns.
Import the dataset into Gephi[4] and the Layout algorithms give the shape to the graph. The statistics and metrics framework offer the most common metrics for social network analysis (SNA) and scale-free networks. Some of the metrics we used is Degree centrality, Betweenness centrality, Closeness centrality, Eigenvector centrality, Clustering coefficient, average path length and eccentricity

Google Colab - Google Colab is an online IDE for Python that was published in 2017. Colab is a great platform for data analysts to use to run Machine Learning and Deep Learning projects that require cloud storage.

Import the dataset to Colab and visualise the Exploratory Data Analysis(EDA) for Overall rating distribution by year, Comparison of the top 5 players from 2015 to 2020 and Best overall rating for each position by year. Predict overall rating of a player by linear regression model. Also, a community detection was performed using PCA for dimensionality reduction and K-means clustering to obtain the communities.

Libraries used :

- **Standard Java Libraries** - The Java Class Library is a collection of dynamically loadable libraries that Java Virtual Machine languages can use during execution. Some of the basic java libraries are Java. util. , Java. lang. ,Java. Math. etc.,
- **Numpy -** NumPy is the fundamental package for scientific computing in Python. A multidimensional array object, numerous derived objects (such as masked arrays and matrices), and a selection of procedures for quick operations on arrays are all provided by this Python package.The ndarray object is the basis of the NumPy package.
- **Pandas -** Pandas data processing and analysis of libraries It provides data structures and procedures for attempting to manipulate numerical tables and time - series data in particular.
- **Seaborn -** It is a matplotlib-based Python data visualization library. It offers a sophisticated drawing tool for creating eye-catching and educational statistical visuals.
- **Matplotlib -** For the Python programming language and its NumPy numerical mathematics extension, Matplotlib is a graphing library. For integrating charts into programmes utilising all-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK, it offers an object-oriented API.
- **matplotlib. pyplot -** matplotlib. pyplot is a set of routines that makes matplotlib behave like MATLAB. Each pyplot function alters a graph in a certain way, such as creating a graph, creating a plotting area in a figure, plotting certain points in a plotting area, decorating the graph with labels, and so on.
- **gridspec -** The gridspec class is used to describe the geometric features of the grid on which a subplot will be placed.
- **Math** - To do different complex computations, such as arithmetic, trigonometric, logarithmic, and exponential computations.
- **Scikit-learn (Sklearn)** - The most effective and reliable Python machine learning library is named Skearn (Skit-Learn). Through a Python consistency interface, it offers a variety of effective tools for statistical modelling and machine learning, including classification, regression, clustering, and dimensionality reduction.

Challenges faced

- Feature scaling and transformation of the dataset and deriving the necessary features to perform community detection and prediction.
- Identifying relevant features in the dataset to perform dimensionality reduction for PCA.
- Tuning the accuracy for overall rating prediction of each individual player in the dataset.
- Understanding the detailed working of PCA
- Explored several scaling methods such as log, standard, min-max and log-normal scaling before arriving at the optimal method to carry the analysis forward.

Contribution Of Team Members

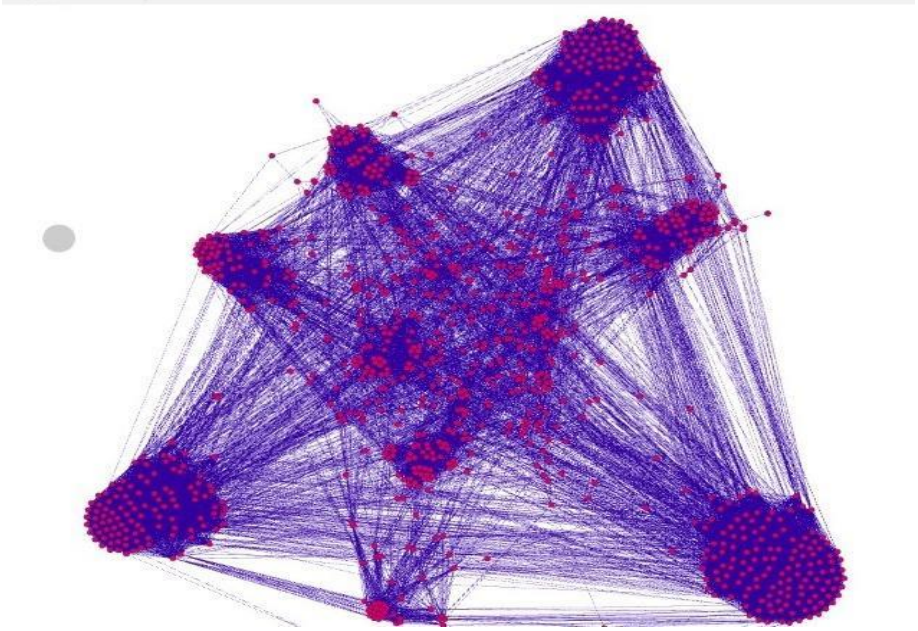
Roll No	Name	Contribution
19z317	Harini S	• A K-means clustering model was built and the data from PCA was fed to predict the clusters(positions of players) in the dataset. • Visualising the generic clusters derived with respect to the specific player positions.
19z322	Jayasree B S	• Performed Exploratory data analysis on FIFA dataset. • Predicting the overall ratings/brand value of players of FIFA 2020 by using OLS regression models. • Worked on the project report
19z336	Nivedha K	
19z346	Selva Keerthana B G	
19z350	Suchitha Malisetty	• Cleaning and understanding the FIFA dataset • Feature scaling and transformation of dataset to perform PCA • Constructing a model to perform PCA and visualising using a Biplot.

Annexure I: Code - FileName: FIFA_Analysis.ipynb

https://github.com/jayasree012/Football_sena_project

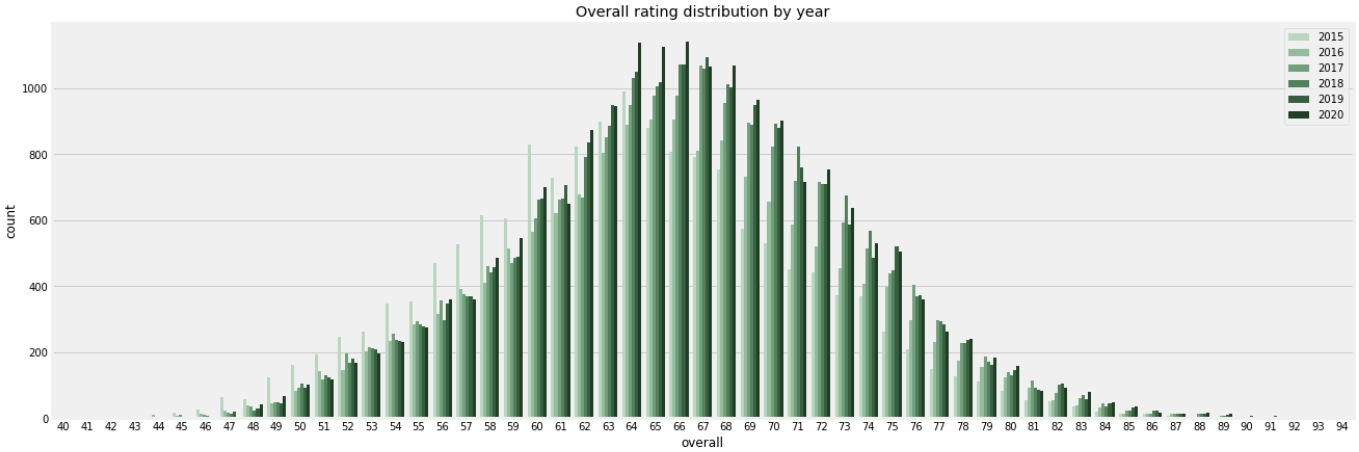
Annexure II: Snapshots of the Output

Visualisation of dataset using Gephi

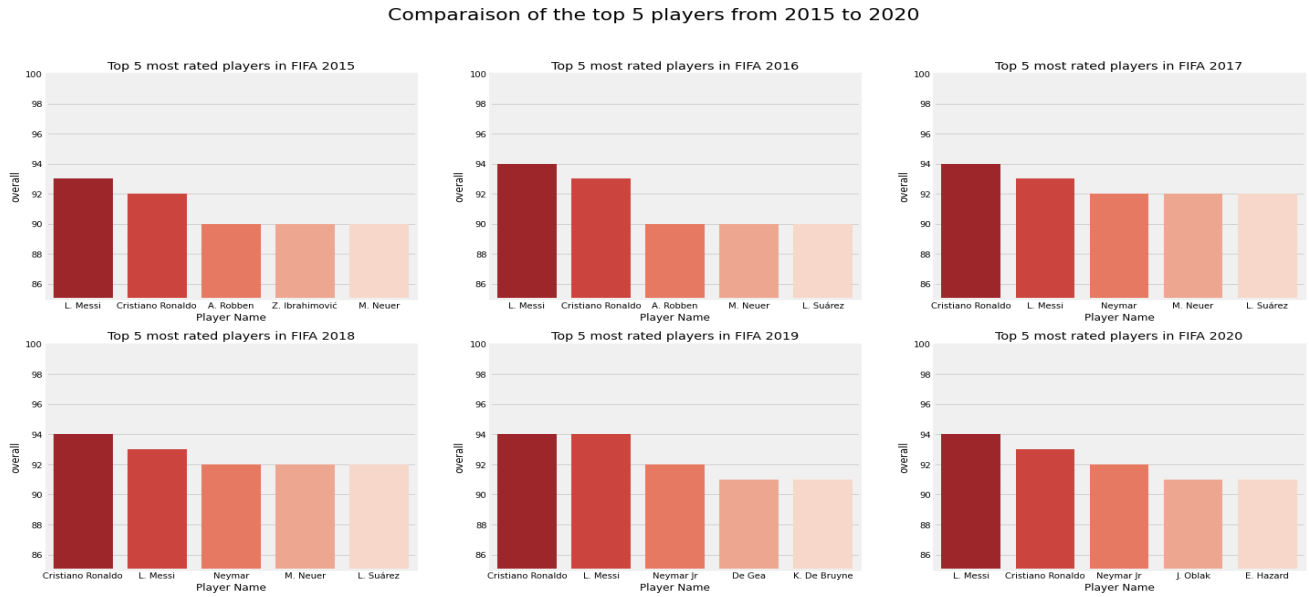


EDA results performed on the dataset -

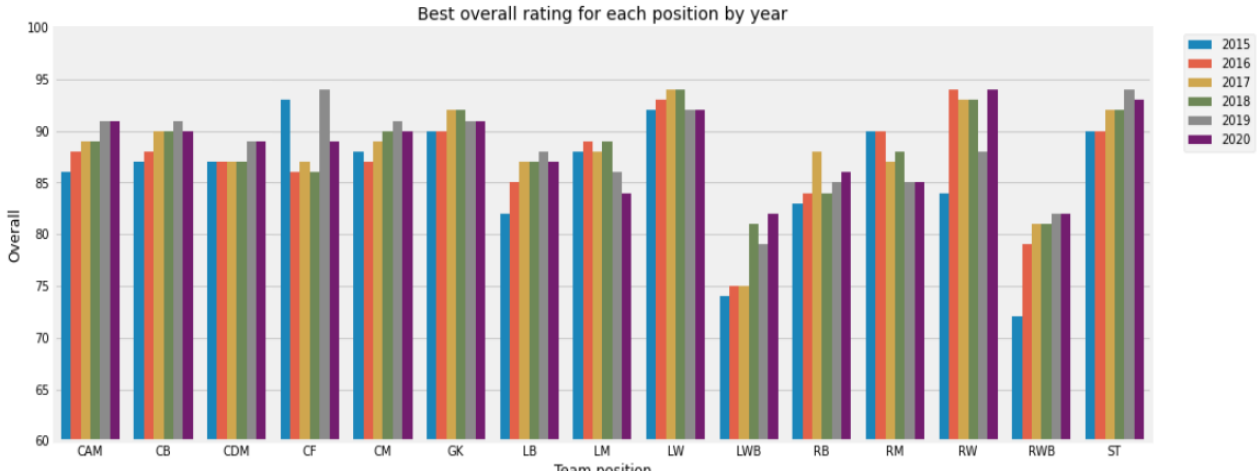
1.



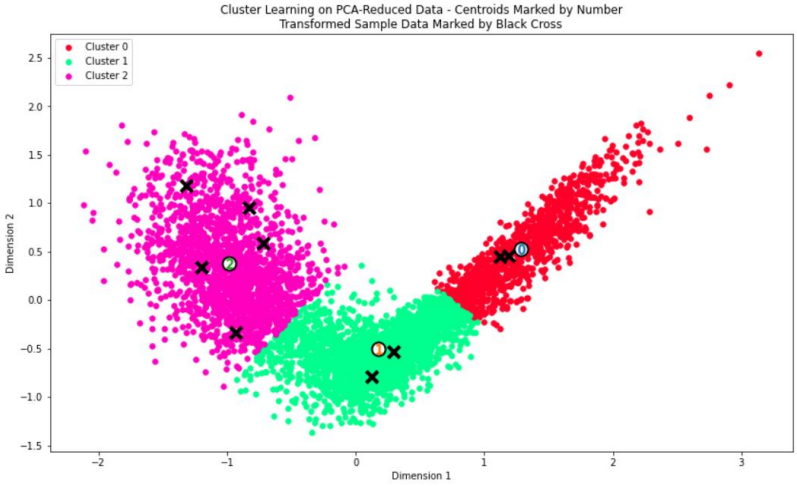
2.



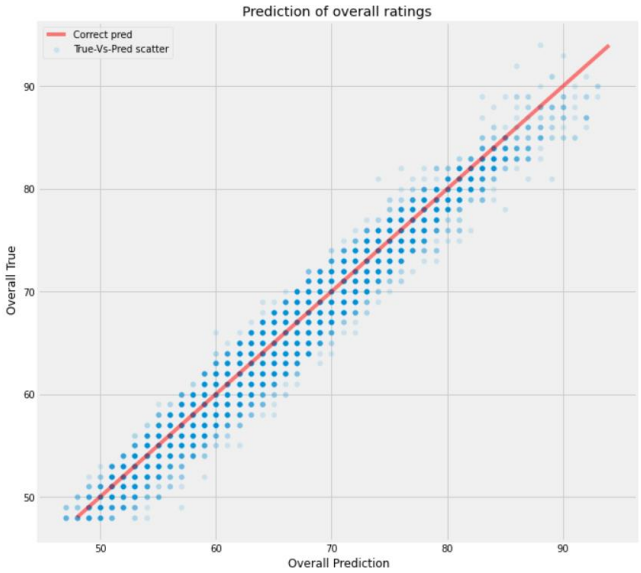
3.



Clustering



Overall Rating Prediction



References

1. https://towardsdatascience.com/degrees-of-separation-amongst-footballers-4163b86f88d
2. https://www.kaggle.com/code/hamzaboulahia/fifa-data-analysis/data
3. https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset
4. https://gephi.org/users/quick-start/
5. https://www.kaggle.com/code/vtaquet/fifa19-dataset-eda-and-clustering-analysis
6. https://www.kaggle.com/code/hamzaboulahia/fifa-data-analysis/notebook
7. https://anvil.works/blog/plotting-in-python
8. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
9. https://www.w3schools.com/python/python_ml_k-means.asp
10. https://exploratory.io/note/kanaugust/Introduction-to-PCA-Principal-Component-Analysis-with-FIFA-Soccer-Data-POb2BMx6ap#:~:text=PCA%20is%20an%20algorithm%20that,as%20possible%20with%20fewer%20dimensions