Database and Analytics Programming

Abstract— Football is a game of Emotions, Spirit as well as Tactics and Strategy. One of the Important aspects of Team Tactics is Team Formation. A series of football simulation video games is called Pro Evolution Soccer (abbreviated as PES). Analysis of its dataset was carried to find Various insights from player's attributes such as Age, Weight, Wages, Valuation, Work Rate, Nationality, Club. Each football player poses various distinguished characteristics which include their Skill Moves, Pace, shooting, strength, stamina, physics, Passing, Movement, Mentality, and other attributes related to Football. We have compared these attributes as it is a vital aspect for a perfect team selection. Using this analysis an algorithm is implemented to choose the best players for the user-defined formation. The results show that it leads to improved team structure through a systematic analysis of the PES football data sets. This kind of approach and analytical results can be useful to form a group of selected players as a team.

I. Introduction

A Statement of Project

Young players are scouted so that they could be trained, groomed at the early stage of the career and they are visioned to play for a longer time. An analysis is performed to find the age distribution of the football players in the dataset. Height and Weight analysis are one of the important aspects of scout selection as it denotes players fitness. For a club, it is necessary to have adequate money flow throughout the Football Season due to which they are bonded to a limited scouting budget. It becomes a tenuous task to keep track of fluctuating player release clauses and wages and select the player which fits within the specified budget. When it comes to selecting a player, analysis is done on several attributes based on player positions for example attackers and Midfielders are assessed based on Pace, Dribbling, Crossing, Finishing, Short Passing, accuracy, Volley, Vision, Reaction. Defender and Goalkeeper are analyzed based on Defending Marking, Standing Tackle, Sliding Tackle, aggression, interceptions, diving, handling, reflexes. This analysis aims to improve the various decision-making stages of football.

B Motivation and Relevance

There are 3.5 Billion estimated fans of Soccer all around the globe which is the highest in the Sports Category. As these numbers are increasing teams\clubs want to keep their fans happy by winning as many trophies as possible they can. As data analytics evolved vastly, its application in Sports Analytics increased heavily. Sports clubs have started investing in appointing Data Analysts to make a guided and supervised decision to scout (Buying New Players for the Club), select the best fit team for match day, choosing the appropriate playing strategy and tactics. In the traditional football era scouting was done by managers by physically examining the player on the football field but now that has become data-driven with help of analysis, statistics, visualizations.

C Elicitation of Research Question

Implement an algorithm to select the best players for the user-defined formation and improve squad structure through a comprehensive study of the PES soccer data sets and improve the various decision-making stages of football.

The technique used to achieve the stated goal is having different visualizations of the player’s attributes and extracting insights from it to apply to the implemented algorithm.

II. Related work

In the first research paper [1], The authors aim to minimize the time taken to pick a player for the squad, considering the expense and the strengths of the player as limitations. Using PowerBI and Python Pandas, the paper presents the statistical analysis of player success based on skills and abilities by minimizing the cost. This paper decreases the choice of risk factors for players by promoting different player characteristics based on market value, popularity, player effectiveness, and success on the nature

of competition from the national team. Despite this, different Machine learning solutions can be incorporated on different datasets having other aspects like player injuries, GPS data, player video performance for better decision making.

In the second research paper [2], The authors proposed a method to measure the occupation and production of spatial value during open play. Direct occupation of space here refers to space created for oneself while creating space refers to opening space for teammates by drawing opponents out of place. Authors have shown how it is possible to extract meaningful information relating to the occupation of spaces of value and the generation of spaces for teammates through Spatio-temporal (e.g., tackles, passes, fouls, shots, dribbles, etc.) data. The interpretation of off-ball movements includes the need for a more specialized per-match or even per-situation analysis, beyond the broader image that total output statistics of multiple matches can offer. The impact of various pressure tactics, the idea of potential space and how it could be used, the overall complex balance of space management between the two teams, and their success association could be more deeply studied and applied to better analyze the data.

In the third research paper [3], For the performance measurement and the ranking of soccer players, the authors are introducing a new-generation data-driven system. It allows the study of the statistical properties of soccer performance by providing a score that substantially synthesizes the performance level of a player in a match or a series of matches. Different data analysis approaches were applied in python to different features of data to extract the result and apply it in the model. To detect the position of a player during a match or a fraction of a match, more advanced algorithms could be designed using different machine learning models.

III. METHODOLOGY

A *Description of the Datasets*

The chosen datasets consist of very large data on different players of PES. 4 different datasets are related to each other on the Player Id attribute which is unique for each Player. The player's data is in different datasets such that one dataset has player's details, the second consists of player's position data, the third has data of goalkeepers and defenders, and the fourth has attackers and midfielder's data. The main reason to choose these datasets is that the datasets consist of enumerable attributes of the players which helped to analyze the player's data with different aspects using many data visualization plots and graphs.

1. *Player Details Dataset:*

In this dataset, the player's details such as age, value, wages, birth date, height, weight, etc. are provided. This dataset gives all the major attributes of the player which defines an individual player. The nation and club for which the player plays in the game are also given in this dataset.

2 *Player Positions Dataset:*

Different players have different best positions in the game, which are expressed on a scale of 100. With this position data, one can easily understand which player plays well in which position.

3 *Defender and Goalkeepers Dataset:*

This dataset is for the defenders and goalkeepers in the game. The Player's strength, stamina, and physic are provided in this dataset. The scale of different defending techniques of a player is also given along with the goalkeeping styles of a goalkeeper.

4 *Attacker and Midfielders Dataset:*

In this dataset, the attributes for the attackers and midfielders are explained. Various techniques of attackers like crossing, finishing, heading accuracy are provided on a scale of 100. Similarly, the skills of an attacker and mid-fielders like dribbling, ball control, curve, etc. are given.

B *Data Gathering*

Kaggle is one of the largest data sources for achieving your data science goals. The four different semi-structured datasets on football are chosen from the repository and are programmatically stored in MongoDB using Python.

C *Data Pre-Processing*

1. *Player Details Dataset:*

In the Player Details dataset, there were many columns with missing data. For example- "nation_jersey_number", many players just play for the clubs and do not play for their respective countries. As a result, their nation_jersey_number value is missing and we cannot fill this column by any processing technique, hence removing the column. Several other less important columns such as "player_url", "body_type", "real_face", "player_tags", "loaned_from", "joined", "team_jersey_number", "contract_valid_until" were removed from the analysis as the percentage contribution to the analysis was very low. Missing values in "release_clause_eur" were replaced by the mean value of the column.

2 *Player Positions Dataset:*

As a Goalkeeper plays only in one position and this dataset comprises different player positions. There are many null values for all other positions for Goalkeepers. As there are 2036 Goalkeepers in the dataset, all other Position Columns are having Null values. Hence, replaced all null values with zero.

3 *Defender and Goalkeepers Dataset:*

The processing of the various aspects of this dataset was carried out. This is the dataset for the defenders and goalkeepers, so the various characteristics of the attackers were 0. Hence, Replaced all null values with 0.

4 *Attacker and Midfielders Dataset:*

In this dataset, processing has been performed using various player attributes. Since this dataset was for attackers and midfielders, the various characteristics of the Goalkeepers were 0. Replace all null values with 0.

D ***Technologies Used- Python, MongoDB, PostgreSQL***

1. *Python*

We chose Python because it functions on various platforms (Windows, Mac, Linux, Raspberry Pi, etc.). It has a basic syntax like the English language and allows developers to write fewer lines of programs than certain other programming languages. It operates on an interpreter framework, meaning that as soon as it is written, code can be executed, which means that prototyping can be very easy. Python can be handled procedurally, in an object-oriented way, or in a functional way.

All these language features mentioned above helped us to examine the dataset and to identify different trends in it.

2 *MongoDB:*

When designing a database schema, we could not know in advance all the queries that will be made to analyze the dataset. So, we used the ad hoc query in MongoDB, which is a short-lived command whose value depends on a variable. Each time an ad hoc query is executed, the result may vary depending on the variables in question. It has proper indexing for better query executions. It allows the "Sharding" of Data. It is the process of dividing larger datasets across multiple distributed collections that helps the database to distribute and better execute what might otherwise be problematic and cumbersome queries.

3 *PostgreSQL*

PostgreSQL is an extensible database that enables you to specify your data styles, index types, functional languages, etc. If you do not like some aspect of the framework, you can still create a custom plugin that can be optimized to suit the needs, e.g., by installing a new optimizer. This feature was required in this project and to have the structured dataset stored in a database.

E ***Process Flow:***

Figure 1: Process Flow Diagram

Four JSON dataset of PES (Pre-Evolution Soccer) 2020 were taken from Kaggle, the semi-structured JSON was imported into the python environment via a script file (Jupyter Notebook). Now, programmatically, the JSON data in python was stored in the MongoDB database (PES Players 20) with four collections storing one dataset each for further study and for identifying trends in the data.

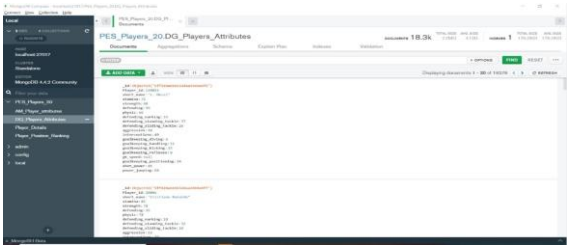


Figure 2: Data Structure In MongoDB

To continue with the study, first, the processing of data must be performed, and a structured database must be created. All four collections have been extracted from MongoDb in python in four separate data frames. Dataset analysis has been performed and maintained in a standardized PostgreSQL database. Later, all four collections were taken from PostgreSQL and merged into one data frame with the player ID attribute. Visualization was performed on a single data frame of 89 columns and data trends that shaped a community of chosen players as a team.

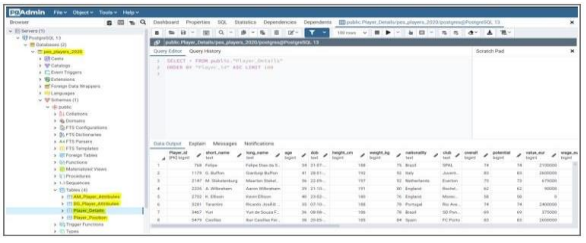


Figure 3: Data Structure In PostgreSQL

F. Implemented Algorithm:

Start

Taking User Input which is a short form of 11 playing positions

Store this input in a list.

Create a function team selection

Pass the list as a function call.

Inside function

Create a copy of the data frame

Filter the data frame by a team

Create the empty data frame name (Team)

For loop to iterate over player position list passes as a function argument

Filter the data frame on bases position given on the respective index of the list

then select the id of the first occurrence of a player who has a max rating for the filtered position.

After this select the player name overall value of the player and append it to the data frame.

Now using the drop function remove the player from the data frame as it should not get a repeat.

This loop will run until the last position player is added to the data frame.

Then return statement returns the data frame TEAM () by converting it into the array of 11*3

Print Team

End

IV. RESULTS

Below are the results obtained from the analysis and visualizations performed.

The four tables from PostgreSQL were merged in a single data frame on Player ID to ease the process of visualization.

Age Distribution and overall rating of all the PES players:

Figure [4] shows the age distribution of the players in the PES game and Figure [5] compares the age of the players against their overall rating. It is visible from the distribution that a maximum number of players fall in the age group of 20 to 25 years. The distribution shows a decreasing trend beyond this point. The scatter plot clearly shows that players whose age is below 30years have a higher overall rating. But we can see higher overall rating points for players that have age above 30 years. Thus, we can say that in general, the overall rating is high for young players but at the same time, certain players have higher overall ratings irrespective of their age as they have more experience. Thus, age cannot be considered as a very important factor in building the best team.

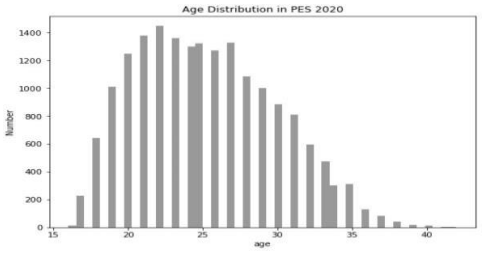


Figure 4: Age Distribution

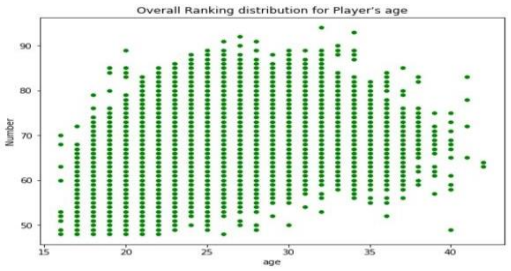
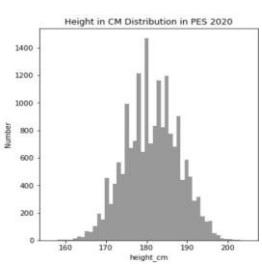


Figure 5: Overall Ranking Distribution

Fitness of the PES Players:

The three figures [6], [7], and [8] show the height and weight wise distribution of players and the scatter plot indicates the proportionality in them. These three visualizations help us understand the level of fitness of the players. To build the best team, one most important factor is the fitness of the player which can be determined by their Body Mass Index (BMI). BMI of a fit person should lie within a specific range and needs the height and weight of the person to be proportional. The scatter plot in figure [8] shows exactly this. We can see that weight and height are directly proportional to each other. And thus, it helps in building a team with the best players.



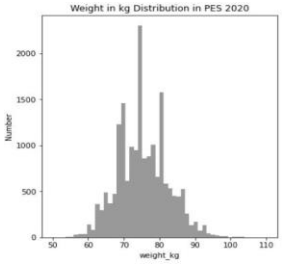


Figure 6: Height Distribution **Figure 7: Weight Distribution**

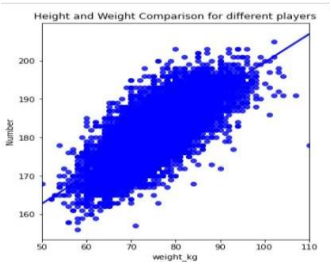


Figure 8: Scatter Plot

Wage and Value of Player:

Figure [9] represents a scatter plot of the weekly wage of a player in euro against the value of that player i.e., the amount in which that player was signed. It can be noted that the wage is higher for a player who has a higher value of signing. In most cases, a player who has a high value is the one who can get the best results for the team as his performance is very good.

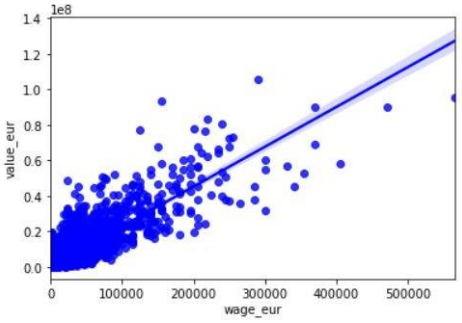


Figure 9: Weekly Wage against the value of the player

Attributes of a player against Overall Rating:

The six scatter plots in figure [10] show the attribute wise rating in comparison with the overall rating of the players. The first plot shows the potential of the player against their overall rating. In the remaining 5 scatter plots a single straight line at point zero for physic, shooting, passing, dribbling, and defending is for goalkeepers. As the goalkeepers are not judged on these attributes, they do not have a rating for these five attributes. Thus, these five graphs can be used to analyze the attackers, midfielders, and defenders.

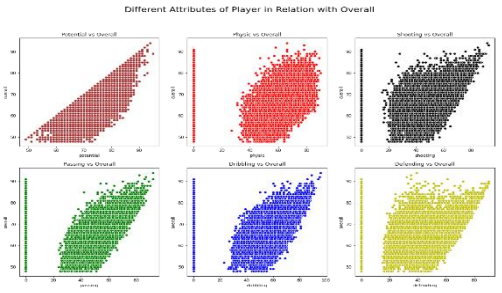


Figure 10: Different Scatter Plots

Player distribution according to overall rating:

Figure [11] shows the frequency distribution of players depending on their overall rating. It is visible that the maximum number of players have an overall rating ranging between 60 and 70. There are a few players who have a rating of more than 70 and even less having a rating above 85.

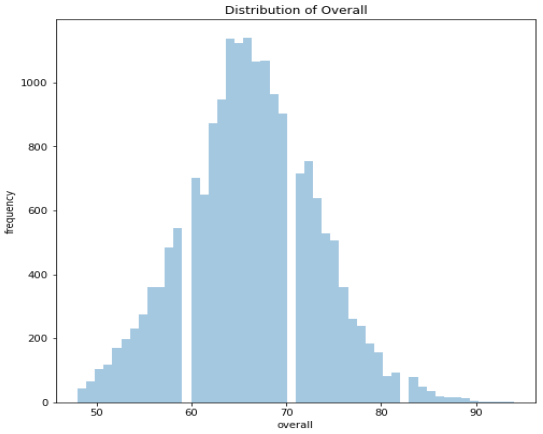


Figure 11: Different Scatter Plots

Attackers, Midfielder, Defender Distribution:

PES Dataset contains players from various playing positions, and which is drilled down into three subcategories Attacker's, Midfielder's, and Defender's Playing Positions. This helps the club\national team to understand several formations in which they can form their players, Number of backup players must replace a player of the same position in case of injury.

Figure [12] explains the distribution of players according to the Attacking Positions such as ST, LS, RS, RW. LW, RF, LF, CF.

Figure [13] explains the distribution of players according to the Midfielder Positions such as ST, LS, RS, RW. LW, RF, LF, CF.
Figure [14] explains the distribution of players according to the Defender Positions such as ST, LS, RS, RW. LW, RF, LF, CF.

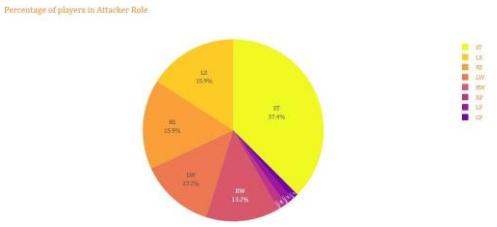


Figure 12: Pie Chart of Player % in Attacker Role

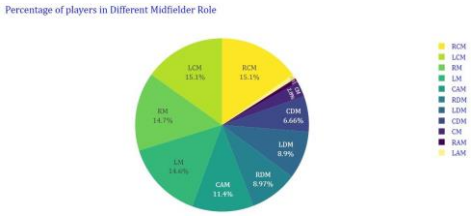


Figure 13: Pie Chart of Player % in Midfielder Role

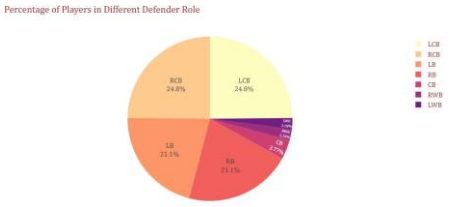


Figure 14: Pie Chart of Player % in Defender Role

International Reputation:

International Reputation is one of the important aspects when selecting a player for their national team. Reputation for a particular player ranges from 1-5. Below figure [15] provides us the list of players with the Highest International Reputation.

	short_name	age	club	nationality	overall
0	L. Messi	32	FC Barcelona	Argentina	94
1	Cristiano Ronaldo	34	Juventus	Portugal	93
2	Neymar Jr	27	Paris Saint-Germain	Brazil	92
19	L. SuÃ¡rez	32	FC Barcelona	Uruguay	89
31	M. Neuer	33	FC Bayern MÃ¼nchen	Germany	88
96	Z. IbrahimoviÄ‡	37	LA Galaxy	Sweden	85

Figure 15: Players with the Highest International Reputation

Physics, Strength, Stamina:

Physic, Strength & Stamina are such factors which are monitored daily and sometimes hourly during training sessions that is why it is necessary to analyze them in such a way that it is efficiently displayed, easily understandable and comparable to the coach and medical staff. Below the figure is a scattered 3D plot of Players based on their strength, stamina and physique, where each circle denotes an individual player. This can be done for other national and club teams.

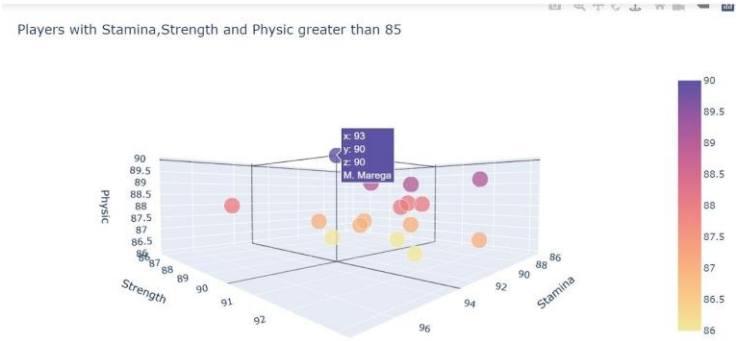


Figure 16: Scatter 3D Plot

Wage by WorkLoad & Player Position:

There are two transfer windows in which players are brought by club teams, Summer and Winter Transfer windows. A list of players gets shortlisted to put a bid for those players and decide their weekly wages and bonuses. Wages and Bonuses vary from Player to Player and it is dependent on their work rate and the player playing position. Below figure [17] indicates average wage by Player Playing Position. The below figure [18] indicates the average wage by work rate..

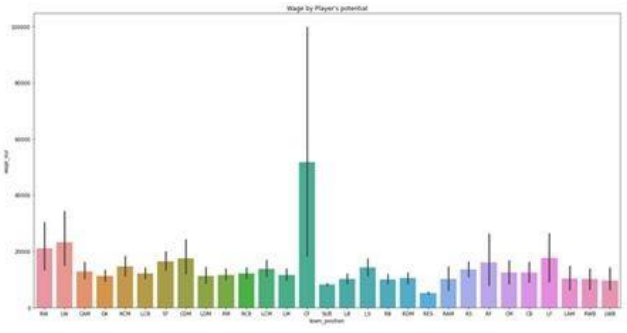


Figure 17: Average Wage By Player Playing Position

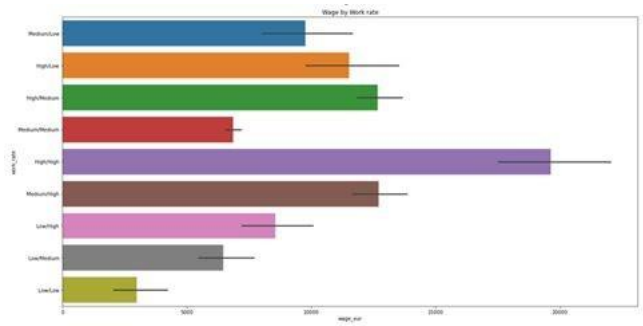


Figure 18: Average Wage By Work Rate

Top 10 Nations:

The pie chart in Figure [19] shows the top 10 nations to which the maximum number of PES players belong to.

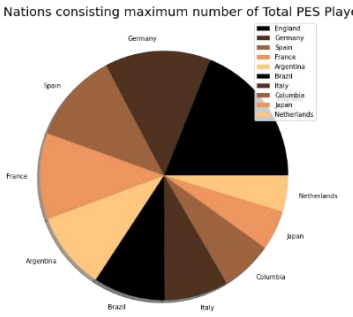


Figure 19: Top 10 Nations

Comparison of Defending rating for different player positions:

The scatter plot in Figure [20] shows a comparison of the importance of player defending rating for different player positions. A player playing at the center back position has a linear relationship with the player attribute rating for defending. This is correct because a player playing at the center back position should have a good defending rating. Whereas a striker(graph1) and a center attacker mid need not be good at defending and hence, would have no relation with the defending rating.

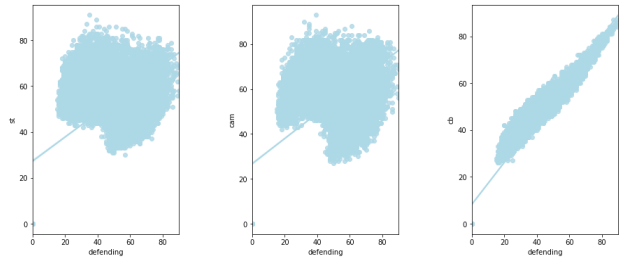


Figure 20: Comparison of Defending rating

Wage Distribution of top 10 rated Players:

The bar plot below (Figure [21]) visualizes the wages of the top 10 players that have the highest overall rating. This can help in understanding which players can be included in the team depending on the club's or National team's budget. Practically, a team's budget is overpowered by the overall rating as the team would always want to win by having the best players, which is the goal of this project.

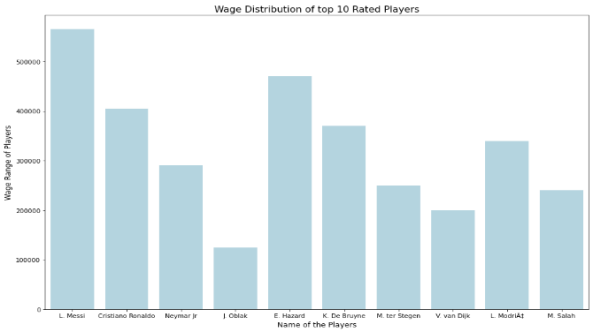


Figure 21: Wage Distribution of top 10 rated players

Algorithm Implementation Result

Team Selection:

One of the Important decision-making process is selecting the Playing XI for the next match from available Players. Each club\national team contains a squad of 30-40 Players. These players can be formed in multiple different formations as per the coach's analysis which often depends on Opposition Teams Playing Formation. To make this process data-driven and efficient we have implemented an algorithm to select the best team from specified club\nation team name and formation.

1. Best team selection - for a specified position.

4-3-3

	Position	Player	Overall
0	GK	J. Oblak	91
1	LB	Jordi Alba	87
2	CB	S. de Vrij	84
3	CB	F. Acerbi	83
4	RB	J. Kimmich	86
5	LM	P. Aubameyang	88
6	CDM	Sergio Busquets	89
7	RM	C. Eriksen	88
8	LW	Cristiano Ronaldo	93
9	ST	H. Kane	89
10	RW	L. Messi	94

Figure 22: Best Team - Specified Position

2. Best team Selection - Based on Club and specified position.

```
4-3-3
Team Selection Based on Club :  Real Madrid
   Position          Player Overall
0        GK     T. Courtois      88
1        LB         Marcelo      85
2        RB        Carvajal      85
3       LCB    Sergio Ramos      89
4       RCB       R. Varane      85
5       LCM        T. Kroos      88
6       CDM        Casemiro      87
7       RCM      L. ModriÄ‡      90
8        LW       E. Hazard      91
9        CF      K. Benzema      87
10       RW   VinÃcius Jr.      79
```

Figure 23 : Best Team - Club and Specified Position

3. Best team Selection - Based on Nation and specified position.

```
Team Selection Based on Nation :  Spain
   Position            Player Overall
0        GK             De Gea      89
1        LB         Jordi Alba      87
2        RB            Carvajal     85
3       LCB       Sergio Ramos      89
4       RCB             PiquÃ©      88
5       LCM        David Silva      88
6       CDM   Sergio Busquets       89
7       RCM             Parejo      86
8        LW          Oyarzabal      82
9        ST    Borja Iglesias       83
10       RW   IÃ±aki Williams       82
```

Figure 24 : Best Team - Nation and Specified Position

V. Conclusion

This paper introduced a data Science approach and an Algorithm to pick players and form a dream team from various nations and clubs considering the features such as Skill Moves, Pace, shooting, strength, stamina, physics, Passing, Movement, Mentality, and other football-related attributes. Such methods and empirical findings can help shape a squad of chosen players as a dream team.

VI. *Future Work*

1. One of the future scopes of this project is to analyze player match data from several seasons and years, which includes attributes such as the number of goals scored, assists provided, total shots on target, accurate passes, pass success, chances created, etc.

2. Match statistics of different clubs and teams such as ball possession, yellow cards, scores, etc. can be analyzed and included in the analysis while creating a dream team.

3. Match Results and odds of a win, loss, and draw can be predicted from the past data of the matches, also injuries, performances can be forecasted by using match statistics.

4. Real-time tracking of the live match to measure match statistics.

REFERENCES

[1] P. Rajesh, Bharadwaj, M. Alam and M. Tahernezhadi, "A Data Science Approach to Football Team Player Selection," 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA, 2020, pp. 175-183, doi: 10.1109/EIT48999.2020.9208331.

[2] Fernández, Javier & Bornn, Luke. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer.

[3] Pappalardo, Luca & Cintia, Paolo & Ferragina, Paolo & Massucco, Emanuele & Pedreschi, Dino & Giannotti, Fosca. (2019). PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Transactions on Intelligent Systems and Technology. 10. 1-27. 10.1145/3343172.

[4] https://pandas.pydata.org/

[5] https://www.psycopg.org/docs/usage.html