1. Introduction & Methods:

1.1 Introduction:

The project aims to predict football UEFA champion's league match outcome of taking two teams and in which year [2016-2023] match happening using machine learning, focusing on player ratings and team past performance with opponent in that year. The dataset used is derived from the FIFA video game series. It includes various player attributes such as overall rating, potential, age, and skill levels. The UEFA Champions League match data for the years 2016 to 2023 is also incorporated, allowing the model to consider team performance in its predictions. The motivation is to predict a future match result depending upon past data and enhance decision-making in the context of football management, leveraging data from the FIFA video game series and UEFA Championship league. This year-wise separation is due to the fact that a club team and player rating might changes from year to year.

1.2 Methods:

- The project utilizes **PySpark** for scalable data processing and employs machine learning algorithms.
- Goalkeepers typically have different attributes and play a distinct role on the field compared to outfield players. Including goalkeeper data when training a model for outfield positions may introduce noise or lead to suboptimal performance because the patterns.
- Feature importance is determined using **RandomForestClassifier**, identifying and removing the least important features. This results in feature reduction.
- **PCA** is applied for Principal component analysis as part of feature reduction.
- The **rank** function is used to find top-ranked player for each specified position based on the overall rating.
- **RandomForestRegressor** and **RandomForestClassifier** are used to predict player ratings and match outcomes while processing past match data respectively.
- There are some user defined functions defined for selecting best squad per club involves selecting players with the highest predicted ratings for each position, identifying and removing the least important features, pre-processing best squad obtained with least important features obtained prior.

2. Architecture:

2.1. Data Preprocessing of Players Rating Data:

Player Data:

- The dataset from the FIFA video game series (2016 version) is loaded, containing player attributes. we are considering FIFA version 16 game - data which is built on season 15 player's performance data.
- Features considered for analysis are

```
["overall", "potential", "age", "height_cm", "weight_kg",
 "weak_foot", "skill_moves",
 "attacking_crossing", "attacking_finishing", "attacking_heading_accuracy",
 "attacking_short_passing", "attacking_volleys", "skill_dribbling", "skill_curve",
 "skill_fk_accuracy", "skill_long_passing", "skill_ball_control", "movement_acceleration",
 "movement_sprint_speed", "movement_agility", "movement_reactions", "movement_balance",
 "power_shot_power", "power_jumping", "power_stamina", "power_strength", "power_long_shots",
 "mentality_aggression", "mentality_interceptions", "mentality_positioning", "mentality_vision",
 "mentality_penalties", "defending_marking_awareness", "defending_standing_tackle",
 "defending_sliding_tackle"]
```

- Goalkeepers are excluded from the analysis to focus on outfield player performance.
- The data is divided into three RDDs based on player positions: attack, midfield, and defense because the overall rating depends on different features/columns to different player based on these attack, midfield, and defense. We can see clearly about this in feature selection.

```
# All possible player positions
attack_positions = ['ST', 'RW', 'LW', 'CF']
midfield_positions = ['CAM', 'CM', 'CDM', 'RM', 'LM']
defense_positions = ['CB', 'LB', 'RB', 'RWB', 'LWB']
```

2.2 Feature Engineering for Players Data:

Feature Selection:

- RandomForestClassifier is employed to determine feature importance for each player position (attack, midfield, defense).
- The model assigns importance scores to each feature, helping identify the most influential factors in player '**overall'**.

```
overall: 0.42809957800138904
potential: 0.0753316295587487
age: 0.007958284283142408
height_cm: 0.001010531502104666
weight_kg: 0.0014325693954547213
weak_foot: 0.00010766760078229407
skill_moves: 0.0005284712246161653
attacking_crossing: 0.004777716911909683
attacking_finishing: 0.06712977513154324
attacking_heading_accuracy: 0.014866367656195235
attacking_short_passing: 0.023294570080184544
attacking_volleys: 0.02754531697836947
skill_dribbling: 0.025305597220512538
skill_curve: 0.004484080893339714
skill_fk_accuracy: 0.002501003558495606
skill_long_passing: 0.0026926736441458448
skill_ball_control: 0.0750023714463082
movement_acceleration: 0.0015857419368750523
movement_sprint_speed: 0.002495508870546176
movement_agility: 0.0012444083127007063
movement_reactions: 0.051824710142917195
movement_balance: 0.0009748270014885848
power_shot_power: 0.05467270701362513
power_jumping: 0.001954334343562746
power_stamina: 0.002373163931685464
power_strength: 0.005008883745056091
power_long_shots: 0.016870777851395197
mentality_aggression: 0.0032779593934172064
mentality_interceptions: 0.0017593125175371393
mentality_positioning: 0.07915733885459965
mentality_vision: 0.007470509162201403
mentality_penalties: 0.004298132802693403
defending_marking_awareness: 0.00091920780837798
defending_standing_tackle: 0.0012218789191630964
defending_sliding_tackle: 0.0008223923049157199
```

```
overall: 0.43991622159048344
potential: 0.10806252116193057
age: 0.015476730045176683
height_cm: 0.0006607026169758064
weight_kg: 0.0006074691348131796
weak_foot: 0.00010933699975867932
skill_moves: 2.7881947802923984e-05
attacking_crossing: 0.00512387364594369
attacking_finishing: 0.0006565080114886352
attacking_heading_accuracy: 0.0193514797344375
attacking_short_passing: 0.015548081588889996
attacking_volleys: 0.0007776844351764405
skill_dribbling: 0.002196797208743052
skill_curve: 0.0011695639876384302
skill_fk_accuracy: 0.0006836240917552829
skill_long_passing: 0.00895990601416382
skill_ball_control: 0.01274883447104176
movement_acceleration: 0.001136612877439211
movement_sprint_speed: 0.0023120438179871083
movement_agility: 0.0006351741284802379
movement_reactions: 0.040393985352918035
movement_balance: 0.0006184130739718443
power_shot_power: 0.0038126728623206632
power_jumping: 0.0010947322071510153
power_stamina: 0.004019146508597694
power_strength: 0.004597075225618811
power_long_shots: 0.001192840109741978
mentality_aggression: 0.018990971025399202
mentality_interceptions: 0.0631808988335738
mentality_positioning: 0.0005246638191618154
mentality_vision: 0.002343647193999652
mentality_penalties: 0.0008283857916692972
defending_marking_awareness: 0.08895084916694915
defending_standing_tackle: 0.08240882442269729
defending_sliding_tackle: 0.05088184689610318
```

The above two pictures are importances generated for attack players and defense players respectively. Important columns for attack players are different from defense players. This is the main reason in dividing dataset based on position.

- Least important features (with importance scores below a threshold) are identified and removed from the dataset.
- The following features are considered for attack players based on featureImportances generated by this model as others are <0.01 which makes them eliminated.
  [`'potential', 'attacking_finishing', 'attacking_heading_accuracy', 'attacking_short_passing', 'attacking_volleys', 'skill_dribbling', 'skill_ball_control', 'movement_reactions', 'power_shot_power', 'power_long_shots', 'mentality_positioning'`]
- Here are columns after elimination for defensive players
  `['potential', 'age', 'attacking_heading_accuracy', 'attacking_short_passing', 'skill_ball_control', 'movement_reactions', 'mentality_aggression', 'mentality_interceptions', 'defending_marking_awareness', 'defending_standing_tackle', 'defending_sliding_tackle', 'position', 'rank', 'features', 'pca_features']`
- The same process is implemented for other parts of data.

Principal Component Analysis (PCA):

- PCA is applied to reduce the dimensionality of the feature space.
- The transformed features, known as principal components, capture the most significant information while minimizing information loss.
- The number of principal components is set to 8 for feature engineering.

2.3 Model Training:

Data Splitting:

- The preprocessed data is split into training and testing sets for model evaluation.
- An 80-20 split is used, with a seed value (42) for reproducibility.

Random Forest Regressor:

- Three **RandomForestRegressor** models are trained separately for attack, midfield, and defense positions data portions.
- Each model predicts the future overall ratings of players in its respective data portion depending on **player_position**.

Evaluation:

- The performance of the RandomForestRegressor models for attack, midfield, and defense positions is assessed using the Root Mean Squared Error (RMSE) metric. RMSE provides a measure of the model's accuracy by quantifying the difference between predicted and actual overall ratings.
- Lower RMSE values indicate a closer alignment between predicted and actual player ratings, signifying higher model accuracy.

```
RMSE for Attack: 2.460411904871925
RMSE for Midfield: 2.6530998715858587
RMSE for Defense: 2.5249462324858842
```

2.3 Data Preprocessing of UEFA Champions League Match Data:

Best Squad Determination:

- For each participating club in the UEFA Champions League, the best squad of a customizable formation is determined based on predicted future overall ratings using **get_best_squad** method.
- For these obtained Squads, each Squad data is categorized into attack, midfield, and defense positions and used to predict their future overall rating using above attack, midfield, defense **RandomForestRegressor** model.

Average Overall Rating of Best Squad:

- The predicted future overall ratings RandomForestRegressor models are averaged for attack, midfield, and defense positions of each player in the best squad.
- Now, the average of these three predicted overall ratings is overall average rating of best squad and these values will be stored in a dictionary as follows.


```
{'AS Monaco': 67.66666666666667, 'Midtjylland': 69.96666666666667, 'Panathinaikos': 73.55555555555554, 'Club Brugge': 74.91666666666667, 'Arsenal': 82.6388888888889, 'Salzburg': 72.3888888888889,
'Sporting CP': 78.55555555555554, 'Zenit St. Petersburg': 75.5, 'Galatasaray': 74.33333333333333, 'Manchester City': 82.22222222222221, 'Basel': 74.6111111111111, 'SL Benfica': 79.33333333333333, 'Paris
Saint-Germain': 81.5, 'Shakhtar Donetsk': 77.58333333333333, 'Sevilla': 78.8888888888889, 'Valencia CF': 75.83333333333333, 'Lazio': 80.1111111111111, 'FC Porto': 71.33333333333333, 'VfL Wolfsburg':
77.0, 'Manchester United': 80.77777777777779, 'Chelsea': 83.44444444444444, 'CSKA Moskva': 78.55555555555554, 'Real Madrid': 85.72222222222223, 'Gent': 70.88888888888887, 'BSC Young Boys': 60.5,
'Celtic': 72.33333333333333, 'Dundalk': 61.833333333333336, 'Molde FK': 67.0, 'Juventus': 82.11111111111111}
```


- These are some of overall average rating of respective best squad of clubs in UEFA Champions League.

Joining Club Ratings with Match Data:

- A DataFrame (**club_ratings_df**) is generated to capture the average overall ratings of clubs participating in the UEFA Champions League. This DataFrame includes columns for the club name (**'club'**) and its corresponding average rating (**'rating'**).The input features are the predicted future overall ratings of home and away teams.
- The match data of UEFA Champions League is joined with **club_ratings_df** for home team rating ('home_rating') and away team rating ('away_rating') as new columns.
- A new column ('result') is created to label match outcomes. It assigns a value of 1 if the home team scored more goals than the away team and 0 otherwise. This binary labeling facilitates the training of the match outcome prediction model.
- After appending this data to match data of UEFA Champions League it will be like as follows,

visitor	home	Date	Season	round	FT	HT	aet	pens	hgoal	vgoal	FTagg_home	FTagg_visitor	aethgoal	aetvgoal	tothgoal	totvgoal	totagg_home	totagg_visitor	tiewinner	hcountry	vcountry	home_rating	away_rating	result
Panathinaikos	Club Brugge	2015-08-05	2015	Q-3	3-0	NA	NA	NA	3	0	4	2	NA	NA	3	0	4	2	Club Brugge	BEL	GRE	74.91666666666667	73.55555555555554	1
AS Monaco	BSC Young Boys	2015-07-28	2015	Q-3	1-3	NA	NA	NA	1	3	1	7	NA	NA	1	3	1	7	AS Monaco	SUI	FRA	60.5	67.66666666666667	0
AS Monaco	Valencia CF	2015-08-19	2015	Q-PO	3-1	NA	NA	NA	3	1	4	3	NA	NA	3	1	4	3	Valencia CF	ESP	FRA	75.83333333333333	67.66666666666667	1
Club Brugge	Manchester United	2015-08-18	2015	Q-PO	3-1	NA	NA	NA	3	1	7	1	NA	NA	3	1	7	1	Manchester United	ENG	BEL	80.77777777777779	74.91666666666667	1
Club Brugge	Panathinaikos	2015-07-28	2015	Q-3	2-1	NA	NA	NA	2	1	2	4	NA	NA	2	1	2	4	Club Brugge	GRE	BEL	73.55555555555554	74.91666666666667	1
Zenit St. Petersburg	Gent	2015-12-09	2015	GroupH	2-1	1-0	NA	NA	2	1	3	3	NA	NA	2	1	3	3	NA	BEL	RUS	70.88888888888887	75.5	1
Zenit St. Petersburg	Valencia CF	2015-09-16	2015	GroupH	2-3	0-2	NA	NA	2	3	2	5	NA	NA	2	3	2	5	NA	ESP	RUS	75.83333333333333	75.5	0
Zenit St. Petersburg	SL Benfica	2016-02-16	2015	R16	1-0	0-0	NA	NA	1	0	3	1	NA	NA	1	0	3	1	SL Benfica	POR	RUS	79.33333333333333	75.5	1
Sporting CP	CSKA Moskva	2015-08-26	2015	Q-PO	3-1	NA	NA	NA	3	1	4	3	NA	NA	3	1	4	3	CSKA Moskva	RUS	POR	78.55555555555554	78.55555555555554	1
Galatasaray	SL Benfica	2015-11-03	2015	GroupC	2-1	0-0	NA	NA	2	1	3	3	NA	NA	2	1	3	3	NA	POR	TUR	79.33333333333333	74.33333333333333	1
Shakhtar Donetsk	Real Madrid	2015-09-15	2015	GroupA	4-0	1-0	NA	NA	4	0	8	3	NA	NA	4	0	8	3	NA	ESP	UKR	85.72222222222223	77.58333333333333	1

RandomForestClassifier Prediction:

- A RandomForestClassifier model is trained to predict match outcomes (home team win or away team win).
- The input features are the predicted future overall ratings of home (**home_rating**) and away teams(**away_rating**) are assembled using **VectorAssembler** passed as feature columns to RandomForestClassifier and label is result of match outcome(**result**).
- After the training the model, predictions were made for test data as follows

home_rating	away_rating	result	prediction
67.66666666666667	75.83333333333333	1	1.0
71.33333333333333	83.44444444444444	1	0.0
74.33333333333333	79.33333333333333	1	1.0
75.5	79.33333333333333	0	1.0
77.0	80.77777777777779	1	0.0
78.55555555555554	77.0	0	1.0
79.33333333333333	74.33333333333333	1	1.0
81.5	82.22222222222221	0	1.0
83.44444444444444	81.5	0	1.0
85.72222222222223	77.0	1	1.0
85.72222222222223	77.58333333333333	1	1.0
85.72222222222223	82.22222222222221	1	1.0

Evaluation:

- The performance of the **RandomForestClassifier** is assessed using the Root Mean Squared Error (RMSE) metric. RMSE provides a measure of the model's accuracy by quantifying the difference between predicted and actual result.
- Lower RMSE values indicate a closer alignment between predicted and actual player ratings, signifying higher model accuracy.

```
RMSE for prediction: 0.7071067811865475
```

2.4 Match Outcome Prediction for a Specific Match:

Team and Best Squad Selection:

- Two teams are chosen as the home and away teams, respectively.
- The best squads for both teams are selected based on a predefined formation (4-3-3)(customizable) and considering the overall player ratings. This is acheived through the **get_best_squad** function, which identifies the highest-rated players for each position.
- After the training the model, predictions were made for test data as follows

Feature Engineering and Preprocessing:

- For both the home and away teams, the selected squads are further processed:
    - Position-specific data (attack, midfield, defense) is filtered from the overall squad data.
    - The least important features, as determined by the previously trained RandomForestClassifier models (**model_attack**, **model_midfield**, **model_defense**), are removed from the datasets.

- Principal Component Analysis (PCA) is applied for feature engineering, reducing dimensionality and capturing essential information.

- The RandomForestRegressor models (model_attack_r, model_midfield_r, model_defense_r) previously trained are used to predict the future overall ratings of the selected squads**.**
- The average predicted future overall ratings are calculated for both the home and away teams based on their respective attack, midfield, and defense components. A DataFrame is created, containing the calculated home and away team ratings.

Prediction and Probability:

- The previously trained **RandomForestClassifier** model is then used to predict the match outcome and provide the associated probability.
- A team which is more likely to win with certain probability is printed.

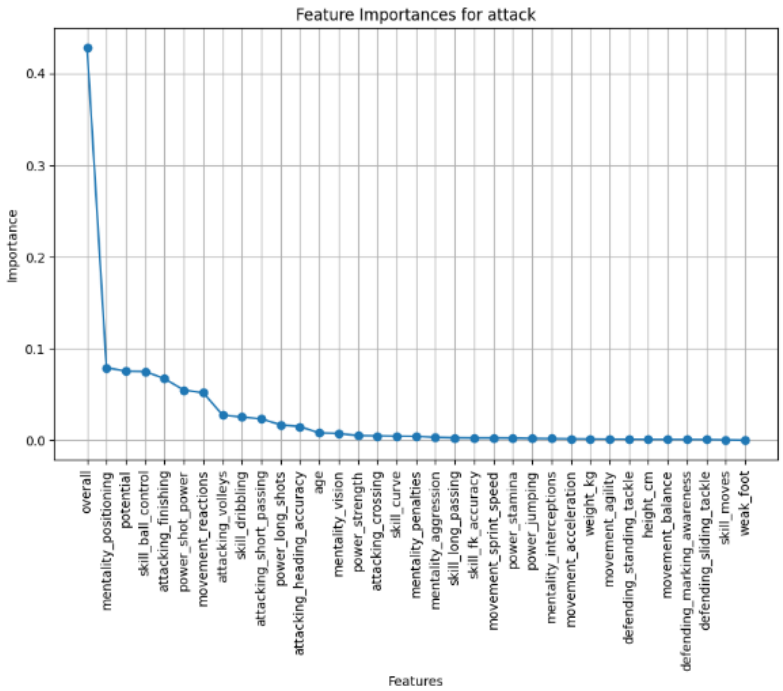
2.5 Match facts for a Specific Match:

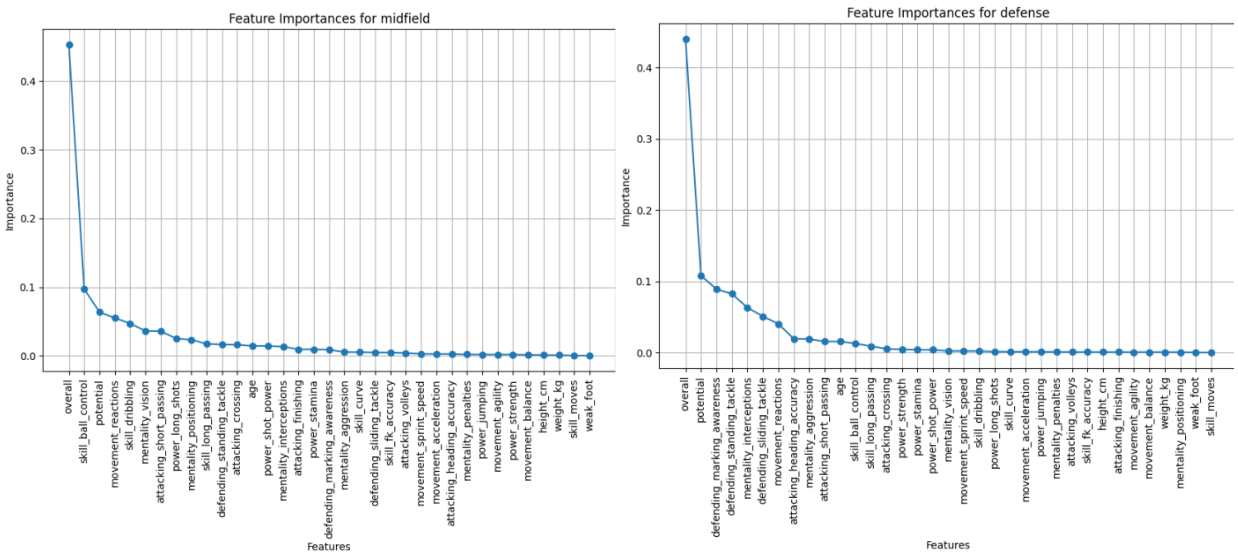
- Comparing two football teams based on their average ratings in attack, midfield, and defense, analyzing which team is superior in each aspect or if they are equal.
- Analyzing who is most rated player among two teams.

3. Results:

3.1 Feature selection Based on Importances:

- We are aware of the process by now. Least important are eliminated from data.
- Below graphs depicts how importance for various features influence overall rating of player based on his position.





3.2 Results after a Specific team selection:

- Two teams are selected home_team = "Arsenal" and away_team = "Chelsea"

overall	potential	age	height_cm	weight_kg	weak_foot	skill_moves	club_name	long_name
87	88	26	183	76	2	4	Arsenal	Mesut Özil
83	83	29	186	75	3	2	Arsenal	Laurent Koscielny
83	83	30	198	90	3	2	Arsenal	Per Mertesacker
79	82	24	178	74	3	3	Arsenal	Francis Coquelin
82	85	24	177	70	3	3	Arsenal	Aaron James Ramsey
80	80	29	178	72	3	2	Arsenal	Ignacio Monreal E...
86	87	26	169	62	3	4	Arsenal	Alexis Alejandro ...
79	79	29	177	76	3	3	Arsenal	Mathieu Debuchy
82	82	28	192	88	3	2	Arsenal	Olivier Giroud

position	rank
CAM	1
CB	1
CB	1
CDM	1
CM	1
LB	1
LW	1
RB	1
ST	1

overall	potential	age	height_cm	weight_kg	weak_foot	skill_moves	club_name	long_name
83	87	23	179	67	3	4	Chelsea	Oscar dos Santos ...
84	84	34	187	90	4	2	Chelsea	John Terry
84	85	26	194	84	3	3	Chelsea	Nemanja Matić
86	86	28	175	74	3	3	Chelsea	Francesc Fàbregas...
82	84	25	178	75	3	3	Chelsea	César Azpilicueta...
83	83	27	167	62	5	4	Chelsea	Pedro Eliezer Rod...
80	80	31	185	91	3	2	Chelsea	Branislav Ivanović
85	86	26	188	85	4	3	Chelsea	Diego da Silva Costa

position	rank
CAM	1
CB	1
CDM	1
CM	1
LB	1
LW	1
RB	1
ST	1

This is truncated output of showing only understandable columns

- From this we will be predicting each player overall rating from which average overall rating of best squad of whole team and those values are passed to already trained **RandomForestClassifier** model.
- Thes results predicted will be as follows,

home_rating	away_rating	new_features	rawPrediction	probability	prediction
82.638885	83.44444	[82.6388854980468...	[3142.37224302937...	[0.62847444860587...	0.0

Chelsea got more average overall rating for their best squad compared to **Arsenal**

- Various other match facts from this prediction is

```
Both are equal in attack
Chelsea is most midfield team
Chelsea is most defensive team
```

4. Conclusions:

- The project successfully predicts player ratings and match outcomes based on FIFA player attributes.
- Feature importance analysis provides insights into the significance of different attributes in predicting player ratings.
- Club ratings provide a comprehensive assessment of a team's strength, which affects forecasts of match results.