# “Valuing” Our Players : Market Value Prediction of Football Players using Machine Learning

Group 6
Niharika Patil | Marisa Yang | Madi Zhaksylyk

# Agenda

## DATA

Datasets and Exploratory Data Analysis

## CLUSTERING

What are some clusters of players?

## CLASSIFICATION

How do we classify players based on their performances?

## PREDICTION

How do we predict the market value of the players?

## CONCLUSION

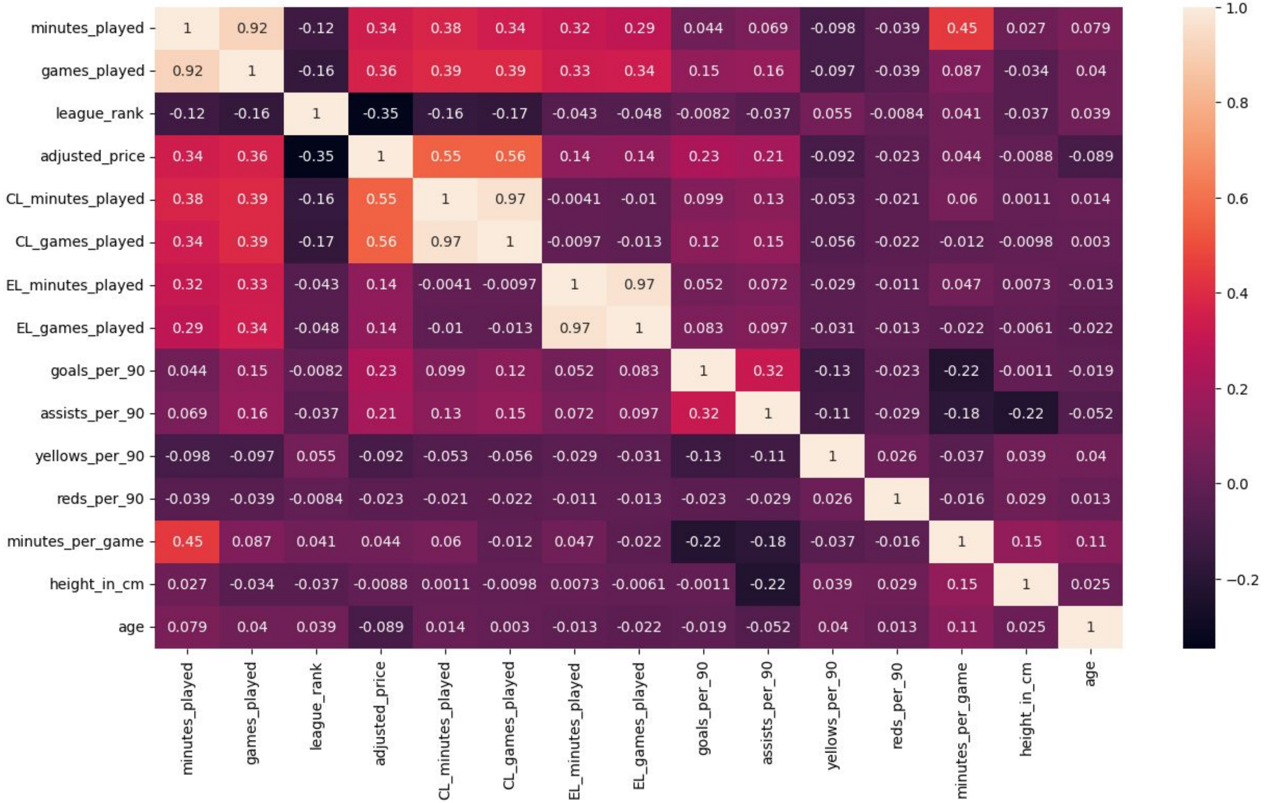
Key findings and takeaways, future work

# Datasets

| Competitions and Clubs | Players | Games | Appearances | Player Valuations |
|---|---|---|---|---|
| UEFA clubs and tournaments | Basic information on soccer players of major Europe leagues | Games played in major leagues in Europe | Individual player performance statistics (2014-2022) | Every player’s valuation |
| - Club Id, Competition Id, Dates, etc.<br>- League affiliation of club<br>- Merged on club_id | - Player Id, Height, DOB, age, etc.<br>- Goalkeepers not part of this study<br>- Merged using player_id | - Domestic leagues<br>- Champions league and Europa League<br>- Merged using club_id | - Player Id, minutes played, goals scored, red/yellow cards earned<br>- Performance metrics aggregated annually per club | - Played Id, Date of Valuation, Valuation<br>- 1 aggregate annual value per club per player<br>- Adjusted for inflation, merged with player_id and date |

# Exploratory Data Analysis

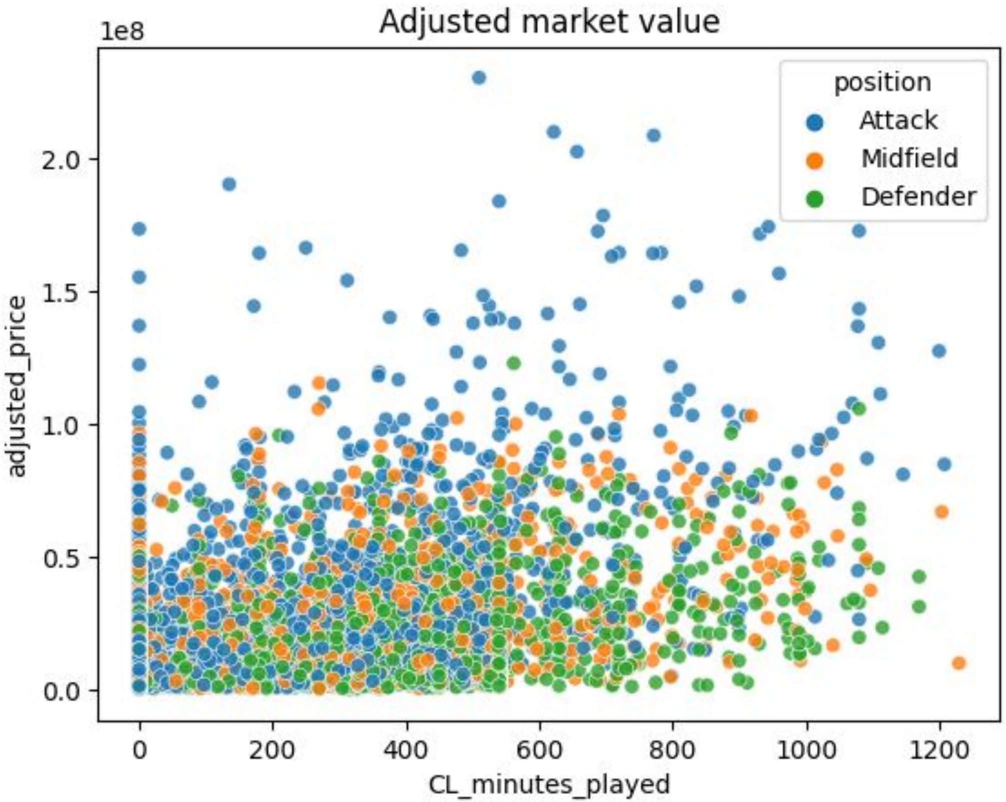
## Correlation Matrix



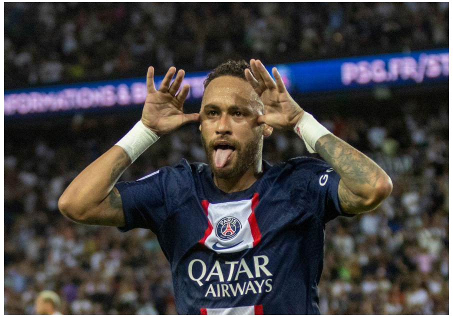

- Participation in the Champions League highly correlates with market value (**0.56 and 0.55**)
- Games and minutes in the domestic competition (**0.36 and 0.34**), and goals and assists per 90 minutes (**0.23 and 0.21**)
- League rank is **negatively** correlated with the price

# Exploratory Data Analysis

## Minutes played in the Champions League vs Market Value




- Players in **Attacking positions** tend to have a higher values.
- Price does not seem to be affected by the duration played.




**Fun Fact**: NJR had one of the most expensive transfers in football history!

# Clustering

## KMeans

separated by their
distance to each other

## Agglomerative Clustering

starts with individual
points, iteratively merges
the closest clusters

## DBSCAN

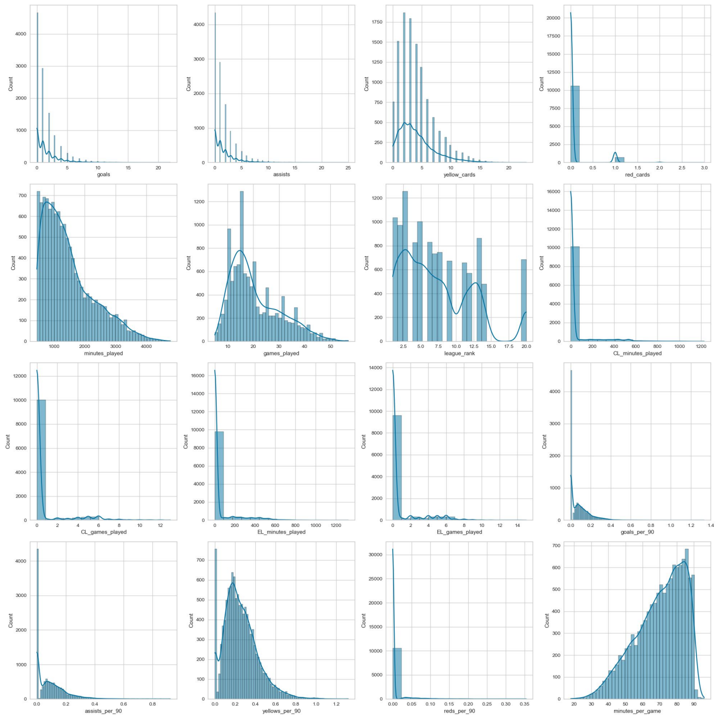
density-based, clusters of
varying shape and size,
identifies outliers as noise

## Gaussian Mixture

probabilistic version of
KMeans, dataset is made
up of multiple Gaussians

# Data Preparation

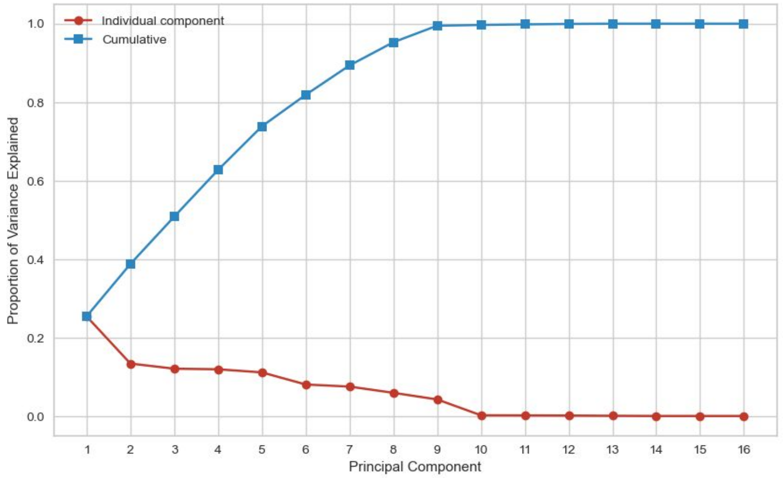
## PowerTransformer



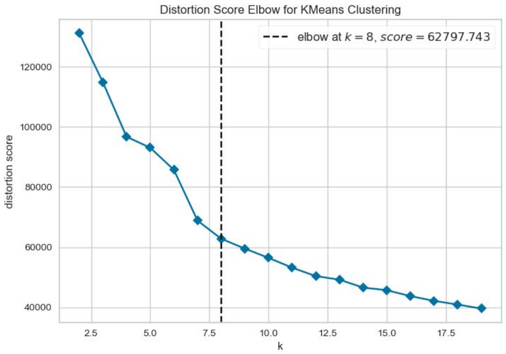

## PCA

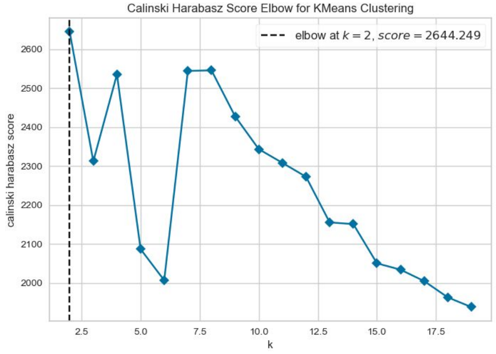
```
Ratio of Variance Explained:
___________________________

[2.53925984e-01 1.33644396e-01 1.20561920e-01 1.19069409e-01
 1.11180957e-01 8.02940907e-02 7.49434834e-02 5.90803093e-02
 4.22277525e-02 1.79426821e-03 1.53381082e-03 1.19503850e-03
 5.33295512e-04 1.21355531e-05 3.07921770e-06 7.02267801e-08]
```

The first principal component 25.4% of the variance in the data, the second - 13.3%, the third - 12%, and so on.

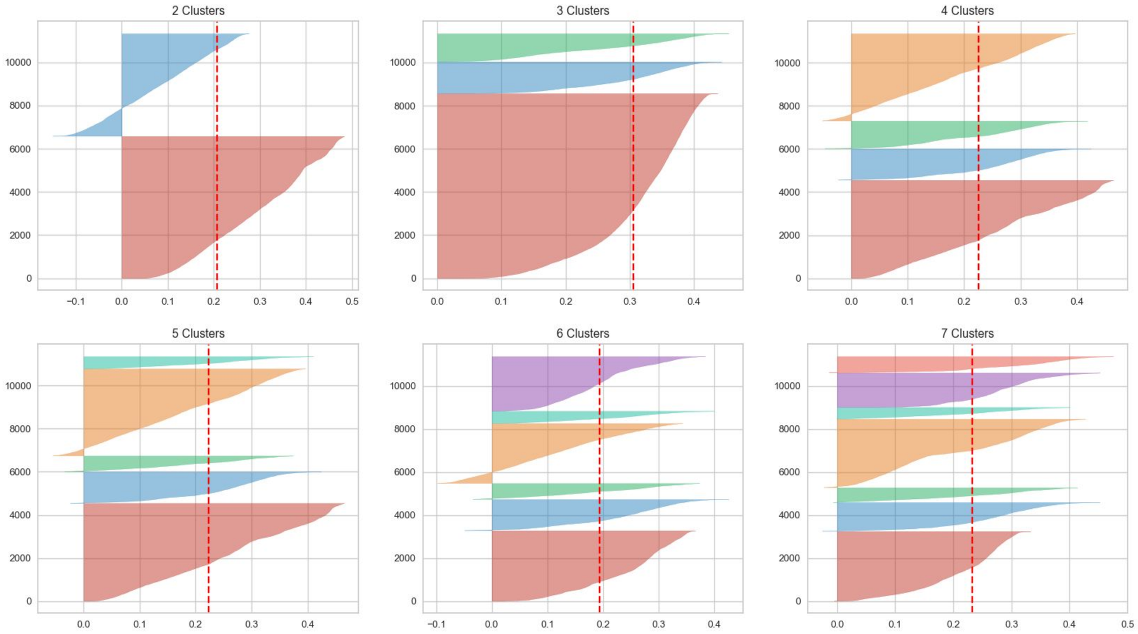


# KMeans Clustering



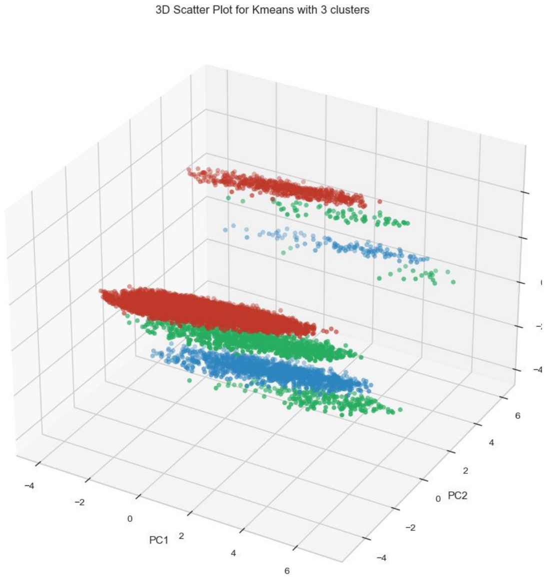



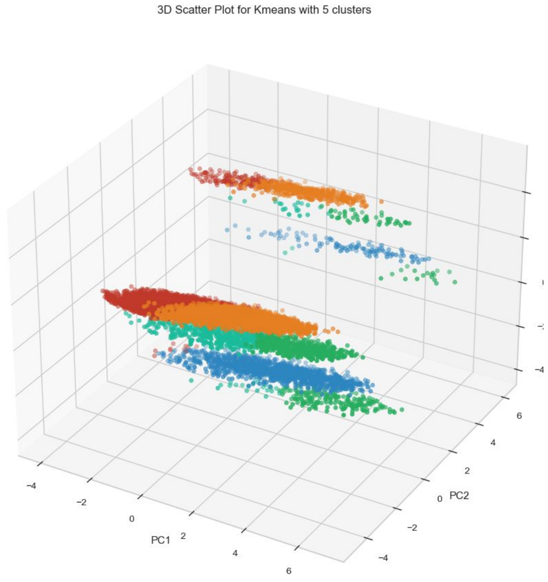

## Silhouette Visualizer

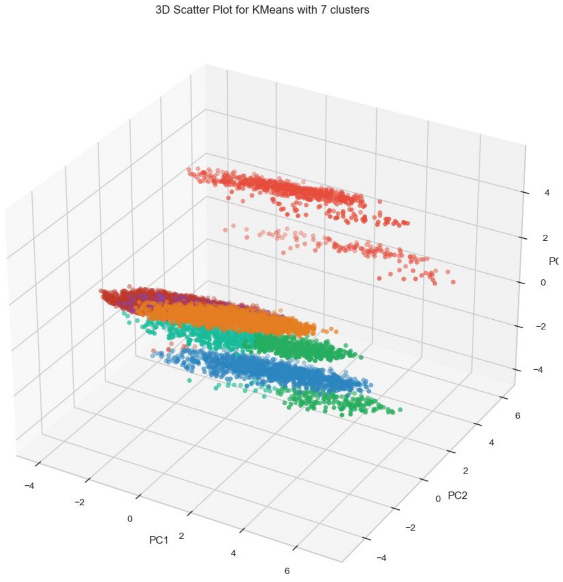


# KMeans Clustering

# KMeans Clustering

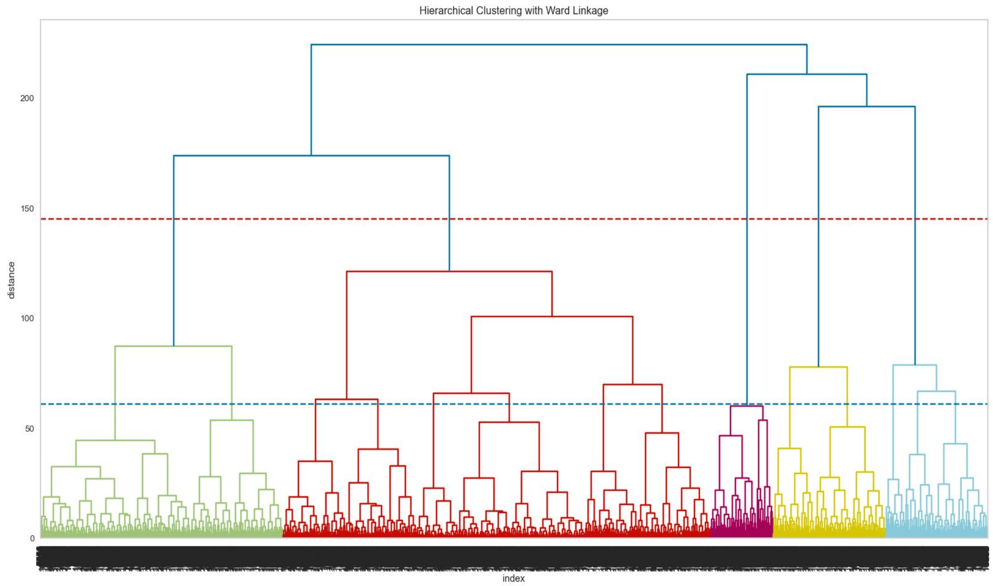


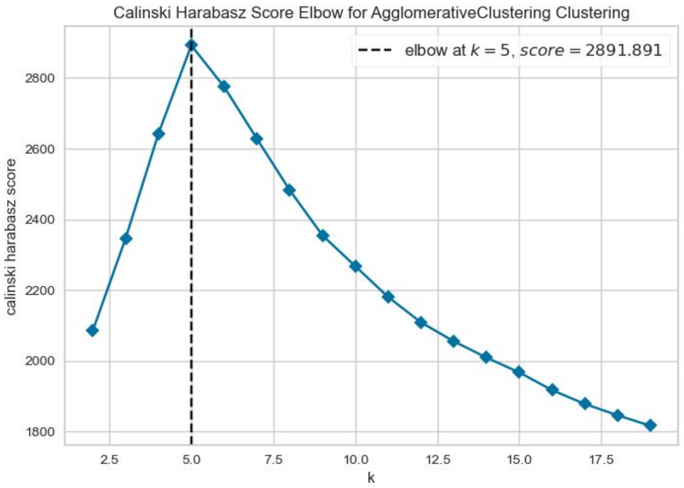




## 3D Scatter Plots

# Agglomerative Clustering



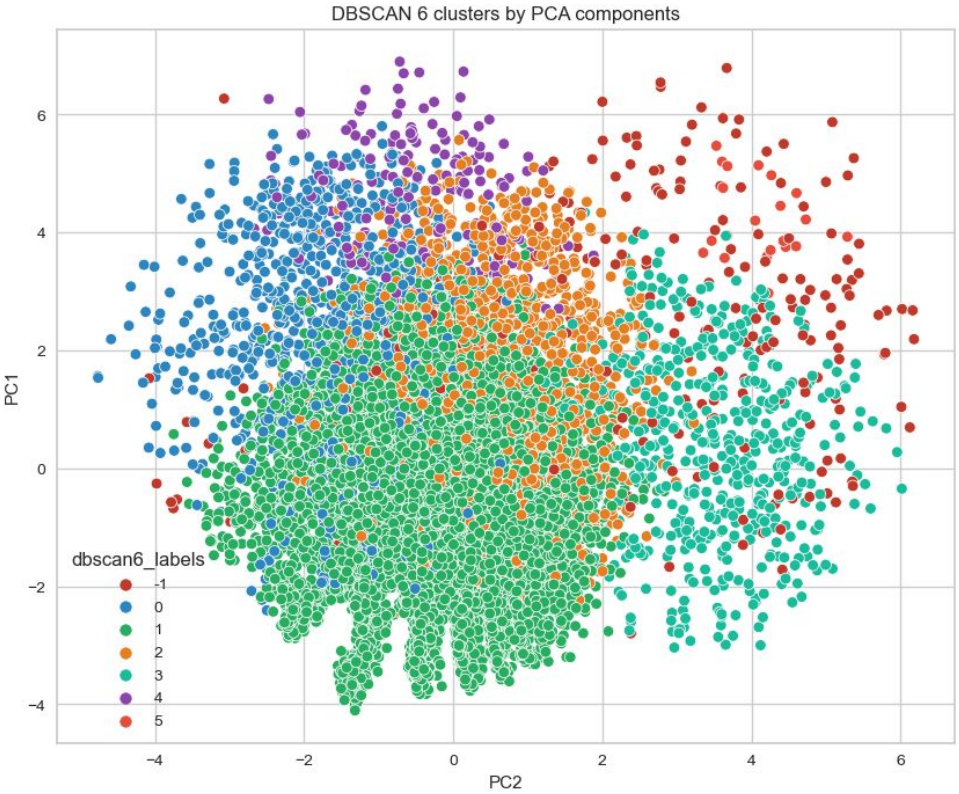




## Dendrogram using Ward Linkage

## Calinski-Harabasz Score

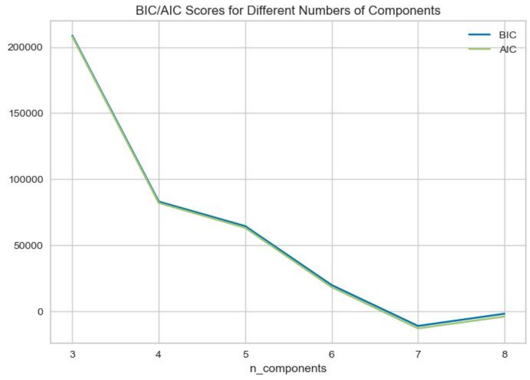
# Agglomerative Clustering

# DBSCAN



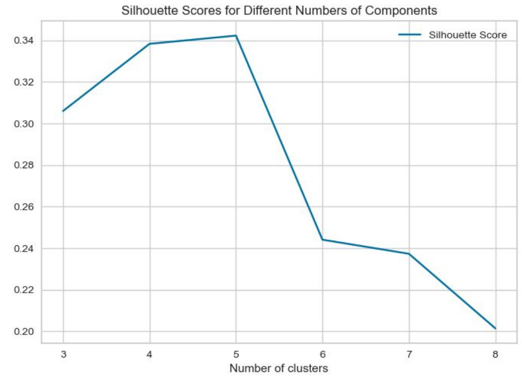

| | no_of_clusters | silhouette_score | epsilon_values | minimum_points |
|---|---|---|---|---|
| **0** | 21 | 0.123754 | 1.0 | 12 |
| **1** | 22 | 0.097304 | 1.0 | 14 |
| **2** | 18 | 0.080026 | 1.0 | 16 |
| **3** | 22 | 0.171897 | 1.1 | 12 |
| **4** | 21 | 0.162880 | 1.1 | 14 |
| **5** | 19 | 0.154562 | 1.1 | 16 |
| **6** | 22 | 0.191877 | 1.2 | 12 |
| **7** | 20 | 0.187766 | 1.2 | 14 |
| **8** | 17 | 0.188554 | 1.2 | 16 |
| **9** | 17 | 0.192461 | 1.3 | 12 |
| **10** | 16 | 0.182712 | 1.3 | 14 |
| **11** | 15 | 0.195867 | 1.3 | 16 |
| **12** | 7 | 0.287239 | 1.4 | 12 |
| **13** | 8 | 0.268507 | 1.4 | 14 |
| **14** | 9 | 0.234230 | 1.4 | 16 |
| **15** | 7 | 0.315194 | 1.5 | 12 |
| **16** | 8 | 0.303749 | 1.5 | 14 |
| **17** | 8 | 0.274046 | 1.5 | 16 |
| **18** | 8 | 0.318147 | 1.6 | 12 |
| **19** | 7 | 0.317981 | 1.6 | 14 |
| **20** | 7 | 0.315938 | 1.6 | 16 |

# Gaussian Mixture

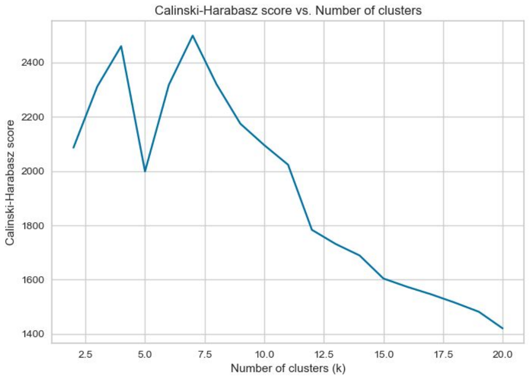



## AIC/BIC

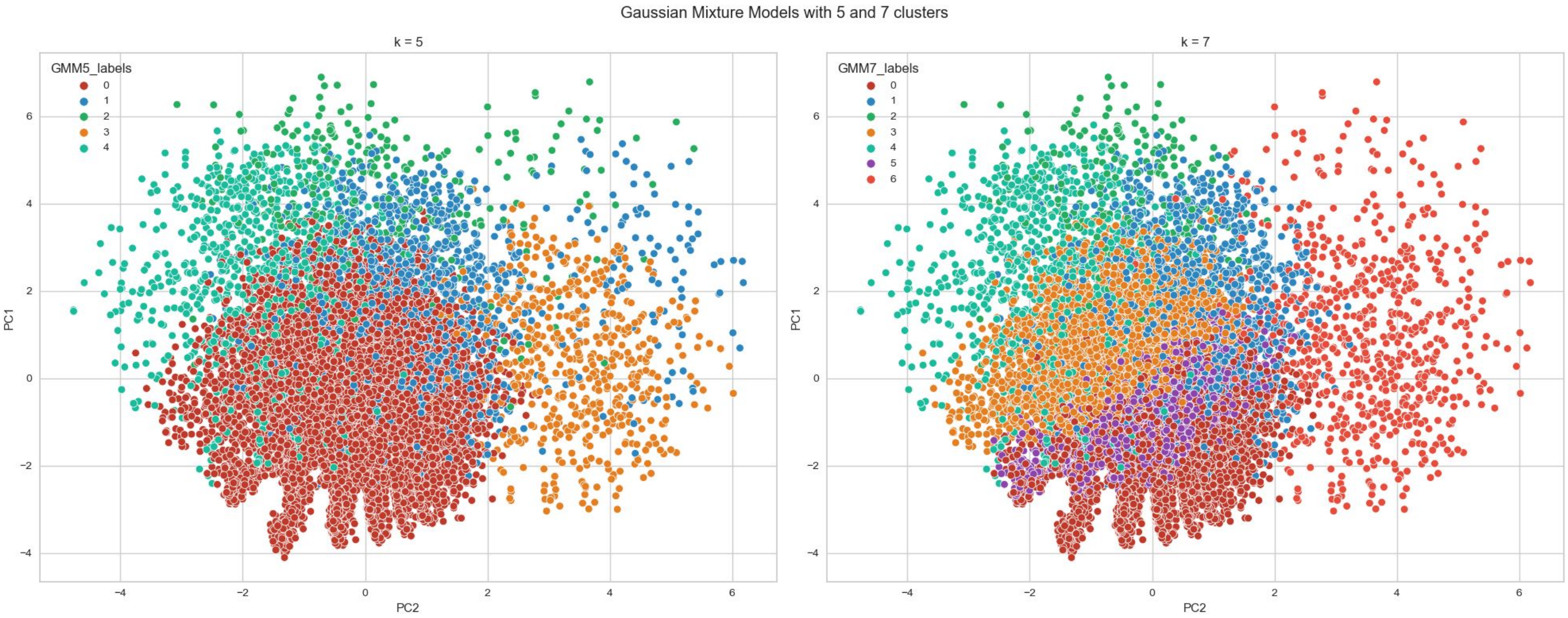
## Silhouette Score



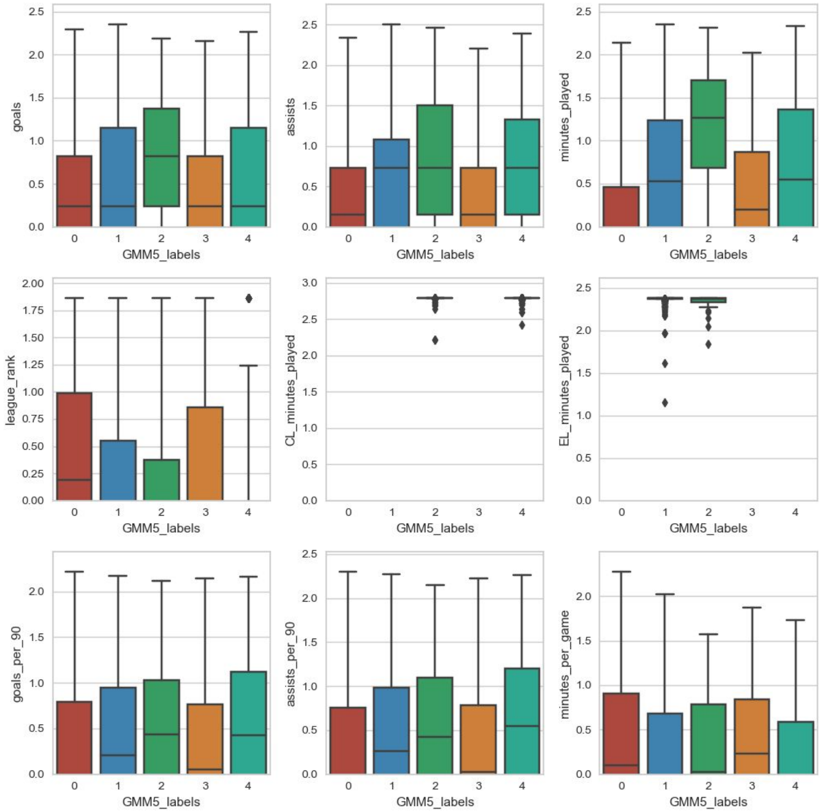

## Calinski-Harabasz



# Gaussian Mixture



# Model Selection: Gaussian Mixture




## Boxplots of Features

- There are observable differences across clusters for almost all features
- Goals, assists, minutes played, CL and EL minutes played are the features where clusters have significant differences
- Only clusters 2 and 4 have players who participated in the Champions League

# Model Selection: Gaussian Mixture

| cluster | goals | assists | yellow_cards | red_cards | minutes_played | games_played | league_rank | adjusted_price | CL_minutes_played | CL_games_played |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.158506 | 1.199850 | 3.603797 | 0.000000 | 1303.204222 | 18.135398 | 7.806395 | 3.264632e+06 | 0.000000 | 0.000000 |
| 1 | 2.043151 | 2.252740 | 4.821233 | 0.078767 | 1848.373288 | 26.180137 | 6.897945 | 9.341918e+06 | 0.000000 | 0.000000 |
| 2 | 3.088328 | 3.369085 | 6.722397 | 0.271293 | 2529.246057 | 35.763407 | 5.460568 | 2.089834e+07 | 313.776025 | 4.378549 |
| 3 | 1.410488 | 1.566004 | 4.428571 | 1.043400 | 1547.249548 | 21.258590 | 7.285714 | 3.776795e+06 | 0.000000 | 0.000000 |
| 4 | 2.408998 | 2.852761 | 4.753579 | 0.000000 | 1907.319018 | 27.693252 | 4.534765 | 2.376384e+07 | 360.126789 | 5.224949 |

| cluster | EL_minutes_played | EL_games_played | goals_per_90 | assists_per_90 | yellows_per_90 | reds_per_90 | minutes_per_game | height_in_cm | age |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.000000 | 0.077073 | 0.079369 | 0.259744 | 0.000000 | 70.919790 | 180.440391 | 26.560580 |
| 1 | 314.615753 | 4.503425 | 0.094093 | 0.104958 | 0.243763 | 0.004427 | 68.738613 | 181.015394 | 26.260959 |
| 2 | 215.034700 | 3.075710 | 0.106576 | 0.119277 | 0.240007 | 0.012357 | 69.306350 | 181.214511 | 26.971609 |
| 3 | 0.000000 | 0.000000 | 0.079424 | 0.088072 | 0.273078 | 0.076955 | 71.859753 | 180.658228 | 26.721519 |
| 4 | 0.000000 | 0.000000 | 0.109625 | 0.129396 | 0.234401 | 0.000000 | 66.536462 | 181.056765 | 26.456033 |

# Cluster Characteristics

## Important

1

1460 players. score around 2 goals and give 2 assists in a season on average. important players in their teams, but may miss a few rounds. Their teams are usually not from top-3 leagues, but participate in the Europa League.

## Key

2

317 players. Prefer attacking style. As compared to other clusters,, score more goals and give more assists, but also get yellow and red cards. Participate in most of the games, often play in the Champions League or the Europa League.

## Enforcer

3

553 players. Sometimes score goals and give assists (but not very often). May miss some games during the season. Get yellow and red cards too often. Their teams do not usually qualify to European international tournaments and play in medium-level leagues.

## Top

4

978 players. Score more goals and gives more assists than other clusters, get around 5 yellow cards but never get red cards. Pla 28 games on average in higher-level leagues. They participate in the Champions League but not in the Europa League.

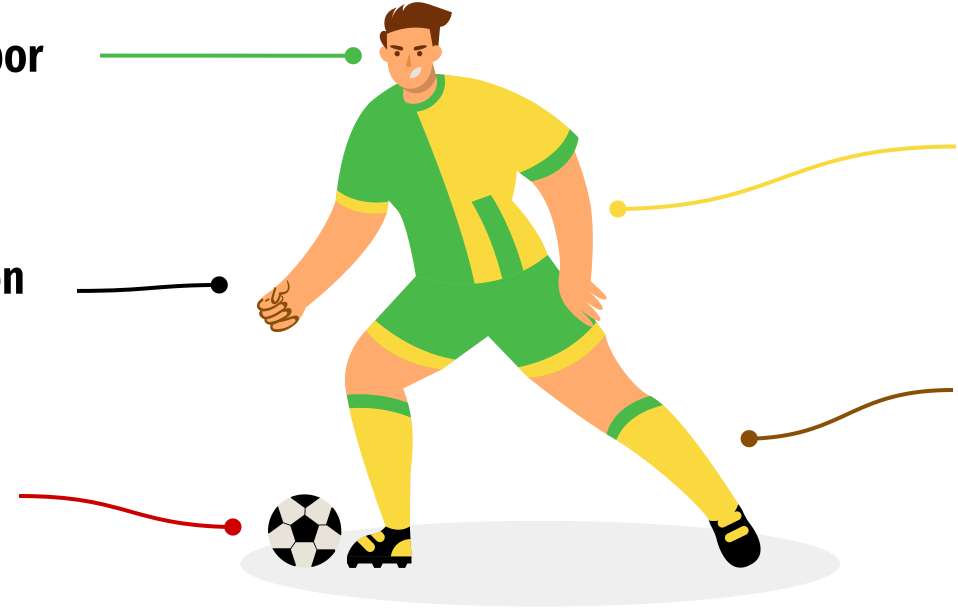
## Mid-Tier

0

8006 players. Most likely 'unremarkable'. Do not score a lot of goals or give assists, participate in some of the games, but not in most of them. Their teams play in lower-level leagues and do not usually participate in international club competitions.

# Classification of Players' Playing Positions

Attackers, Midfielders, Defenders



# Methodology

1

2

3

4

### Scaling Data

Nominate models!

### Balance

Select which feature to use!

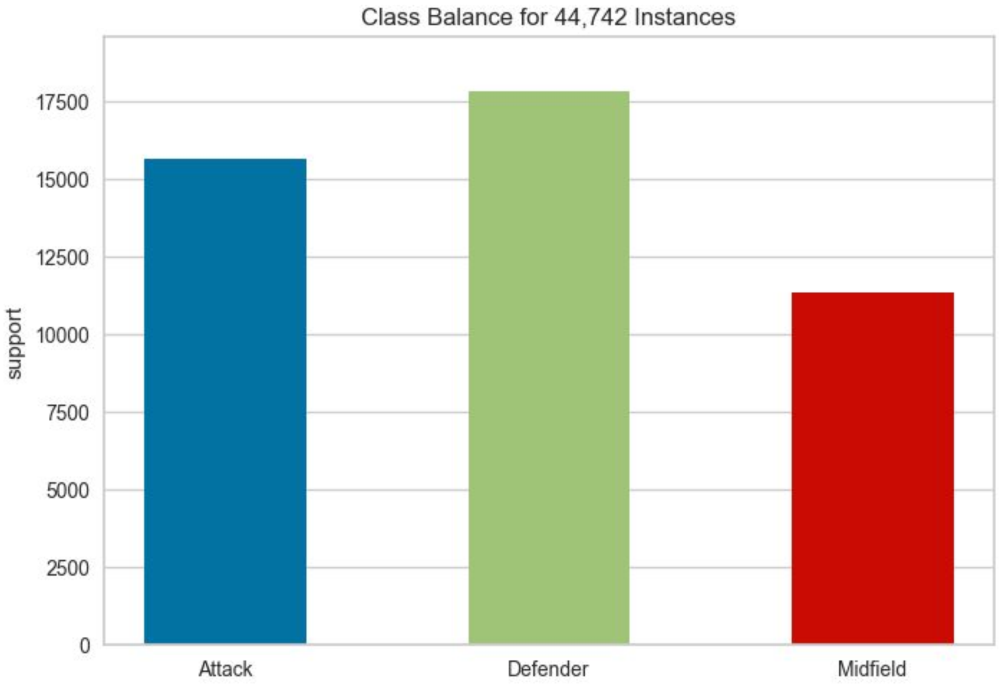
### Train Model

Scaling! (if required!)

### Parameter Tuning

Find best Parameter!

# Class Imbalance

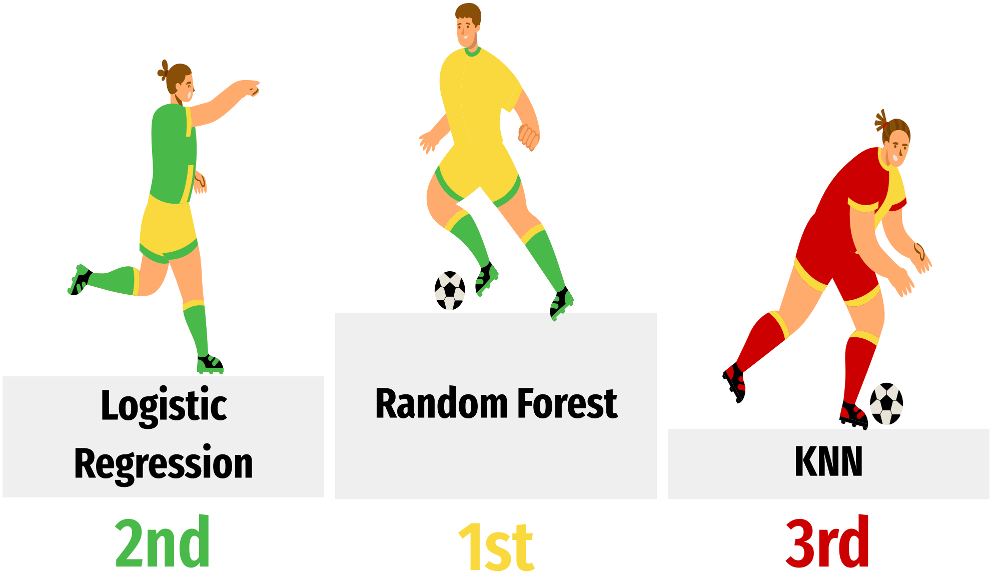



- There is a clear difference in the number of instances for each class. This could be due to difference in the number of players in each category
- However, there is no significant domination of one class over the other
- Include class balancing methods in the pipeline for each model during parameter tuning

# Results

| Model | Accuracy | Precision | | | Recall | | | F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **KNN** | 68.0% | 75% | 68% | 49% | 82% | 82% | 28% | 78% | 75% | 36% |
| **Logistic Regression** | 70.1% | 78% | 69% | 53% | 83% | 86% | 28% | 80% | 77% | 37% |
| **Decision Tree** | 66.2% | 74% | 67% | 47% | 80% | 79% | 28% | 77% | 72% | 35% |
| **Random Forest** | 71.79% | 79% | 70% | 51% | 84% | 87% | 28% | 81% | 78% | 37% |

# Findings




1. The classification model in general perform decently
2. They perform especially well for the ‘attackers’ but poorly for ‘midfielders’
3. Balancing classes doesn’t seem to fix the issue
4. This could be due to the fact the there are minimal performance indicators that quantify the performance of a midfielder

# Prediction of player's market value

Age: 18
Height: 5 ft 8 in
Position: Midfielder
Current team: Barcelona
Goals: 3
Games_played: 13

How much does he worth?

# Methodology

1

**Feature Engineering**

Encode when necessary!

2

**Feature Selection**

Select which feature to use!

3

**Scaling**

Scaling data! (if required!)

4

**Parameter Tuning**

Find best Parameter!

# Models

Polynomial Regression Model

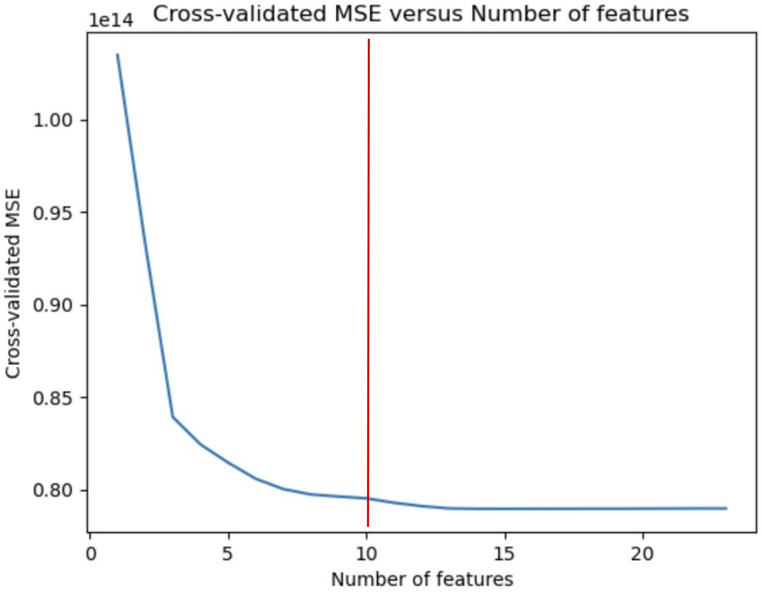
Ridge

Random Forest Regressor

Lasso

XGBoost

# Feature Selection




**Features selected by SFS:**

- 'goals'
- 'assists'
- 'games_played'
- 'league_rank'
- 'CL_minutes_played'
- 'CL_games_played'
- 'EL_minutes_played'
- 'minutes_per_game'
- 'age'
- 'Midfield'

**Targeted variable:**

- 'adjusted_price'

# Results

| | Polynomial Regression Model (degree=3) | Random Forest Regressor | XGBoost | Ridge | Lasso |
|---|---|---|---|---|---|
| **Training MSE** | 49111194845065.695 | 39702721998524.586 | 21345341079028.234 | 79337967276423.03 | 79337078419326.61 |
| **Testing MSE** | 52749101598593.016 | 51278113964524.7 | 50691270912746.78 | 77114038180465.47 | 77118708271844.56 |
| **Training R-squared** | **0.6717** | **0.7346** | **0.8573** | **0.4696** | **0.4696** |
| **Testing R-squared** | **0.6484** | **0.6582** | **0.6621** | **0.4860** | **0.4859** |

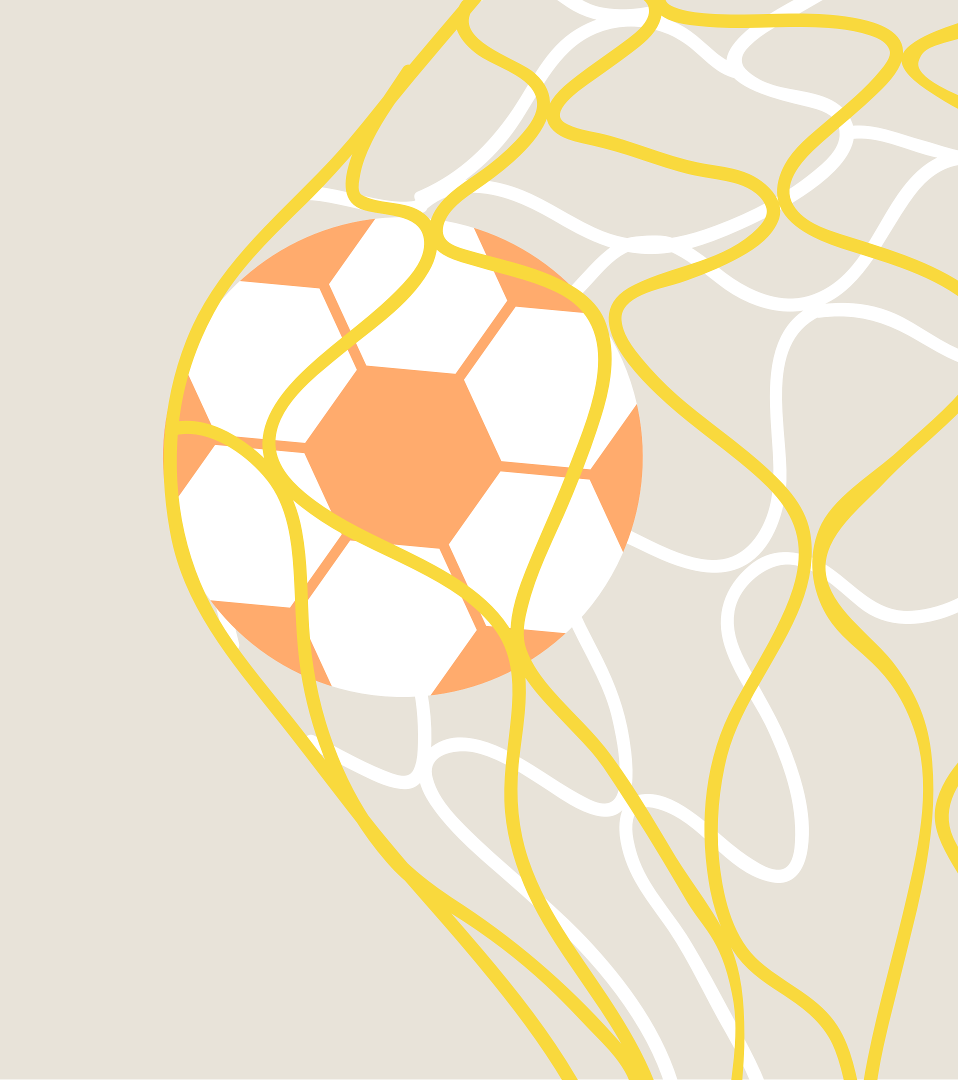
# Key Findings & Takeaways

- There are 5 identifiable clusters of midfielders, basically distinguished by number of goals & assists, participation in the international tournaments. The clustering logic also fits the market value differences
- Most of the classifiers perform well, but decision tree performs the best with 72% accuracy. All the classifiers seem to have trouble classifying the minority class ‘Midfielders’, which was improved by using class balancing techniques. This might be due to lack of quantitative performance metrics for that class, and including this data into the data would result in more precise and accurate classifications.
- The selected features in prediction model demonstrated ideal R-squared value of approximately 65%. While features including goals, assist, time played are straightforward, adding additional  performance metrics can potentially help even more precise prediction

# Future Work

- Include more performance metrics (successful tackles, key passes, pass accuracy, distance covered, clearances, blocks, clean sheets, etc.)

# Q & A



# Thank You!

# Resources

1. https://www.flaticon.com/free-icon/crown_2385865
2. https://www.flaticon.com/free-icon/slider_983738?term=parameter&page=1&position=9&origin=search&related_id=983738
3. Slides.go, https://slidesgo.com/faqs