

# Predicting Top Tiers of Forward Soccer Players

---

## FIFA 2019

By Josiane Pafeng

# Outline

---

- How can soccer players' performance statistics be used to identify top tiers best players?
- A look at the data
  - What is the overall score of the best players at each position?
  - Correlation between features and outcome
- How can we use unsupervised clustering techniques to identify the top tiers soccer players?
  - Dimensionality reduction
  - Clustering techniques
- Results
- Conclusion and the road ahead

## How can soccer players' performance statistics be used to identify top tiers best players?

- Improve clubs' transfer strategy and profit.
- Identify strengths and weaknesses in the opposite team.
- Help picks and bets for Fantasy Soccer fans.



<https://soccerlifestyle.com/what-is-the-weight-of-a-soccer-ball/>

### FIFA Positions

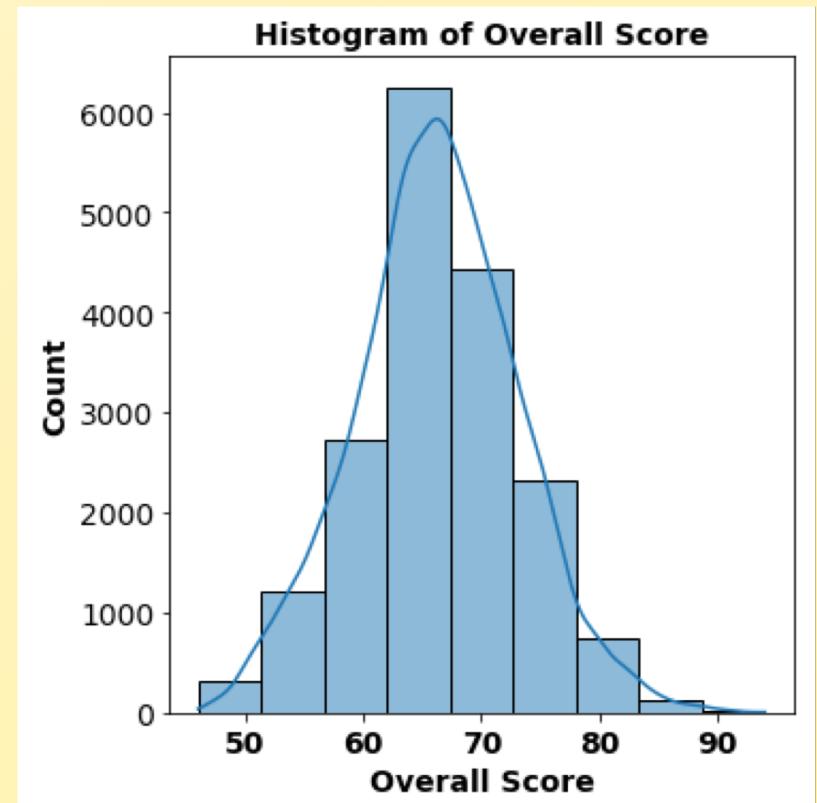
- Forward
- Midfielder
- Defender
- Goalkeeper



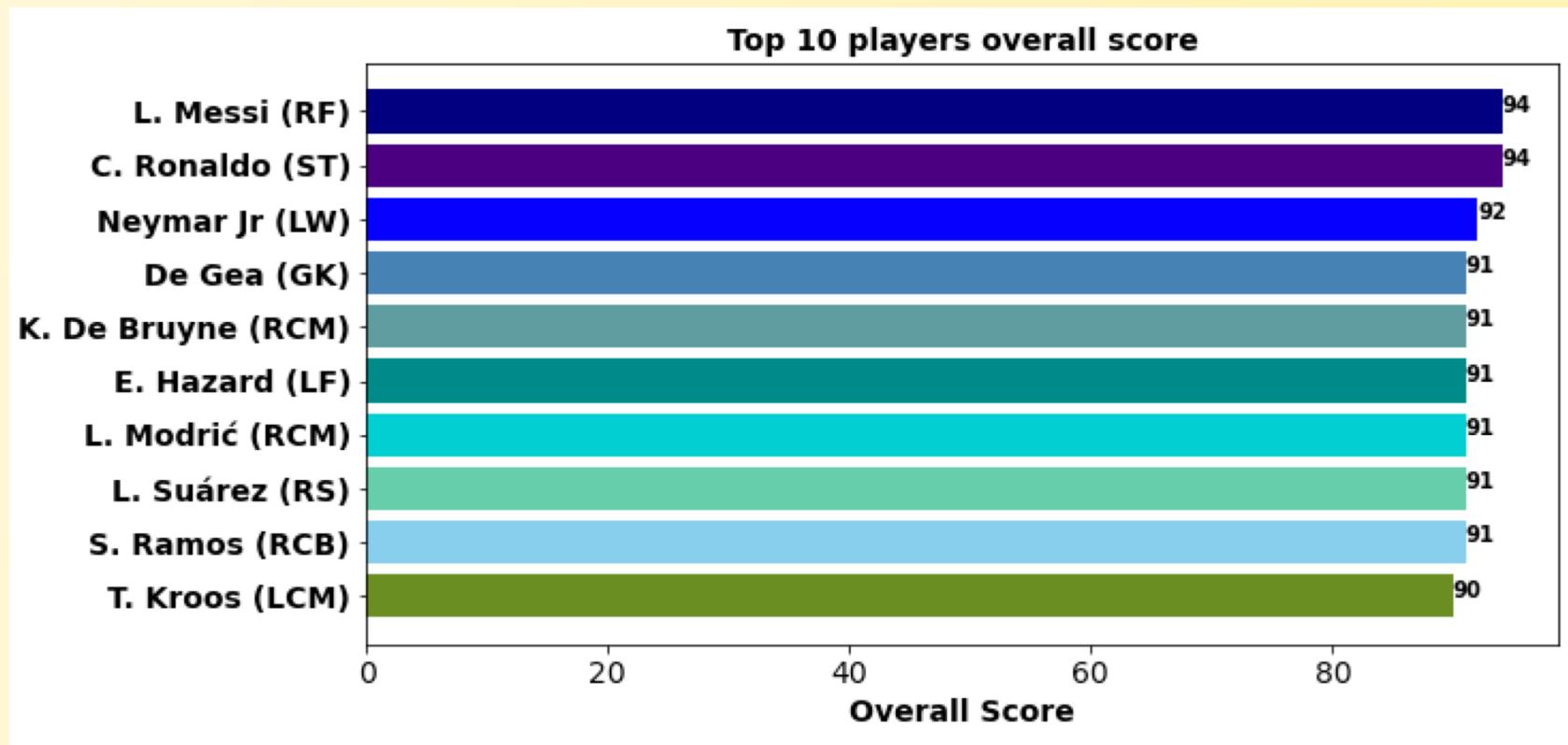
<https://fifauteam.com/fifa-ultimate-team-positions-and-tactics/>

# A Look At The Data

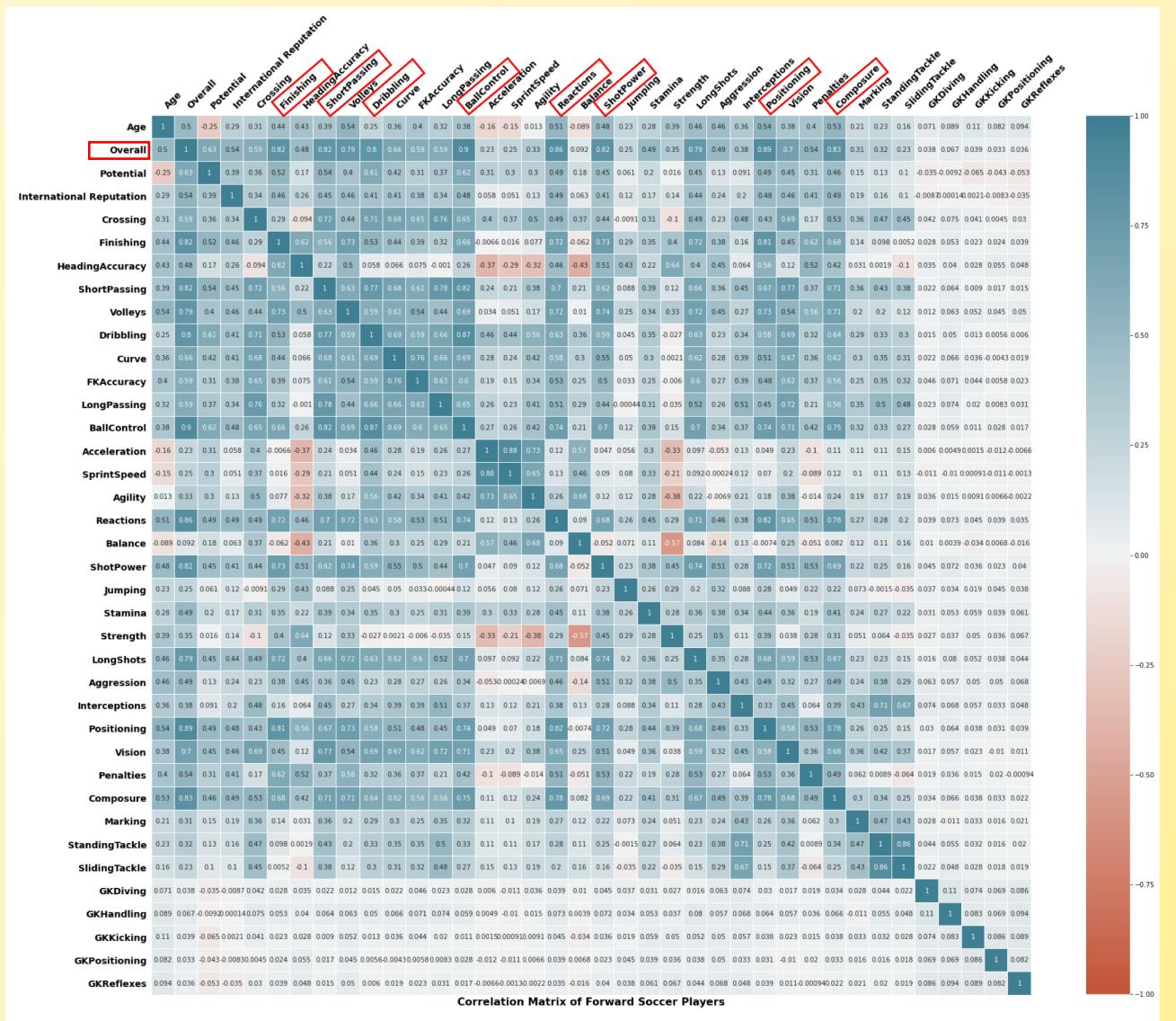
- FIFA-19 dataset
  - Performance statistics from 18k+ international soccer players.
- 35 features and 1 outcome (overall score) variable.
  - No labelled data as we don't know the number of clusters or tiers.



## What is the overall score of the best players at each position?

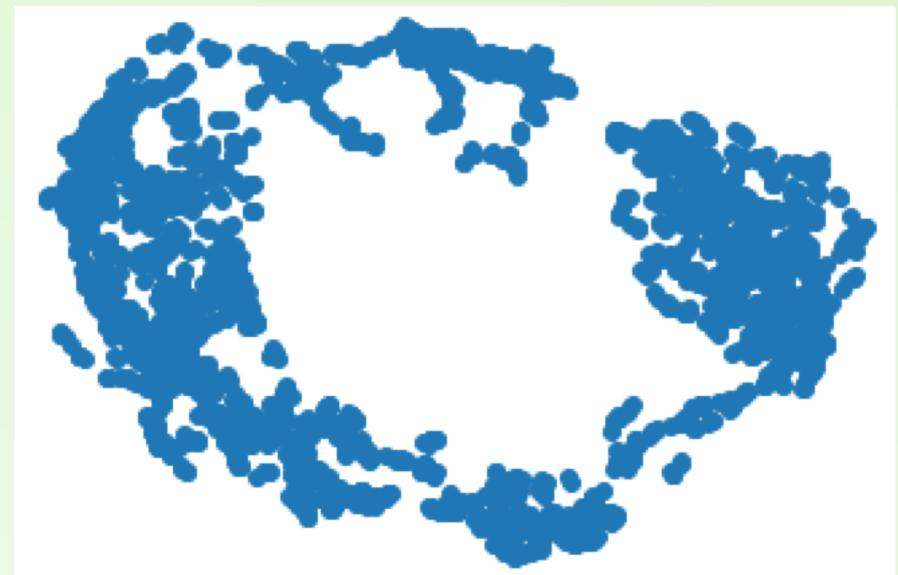
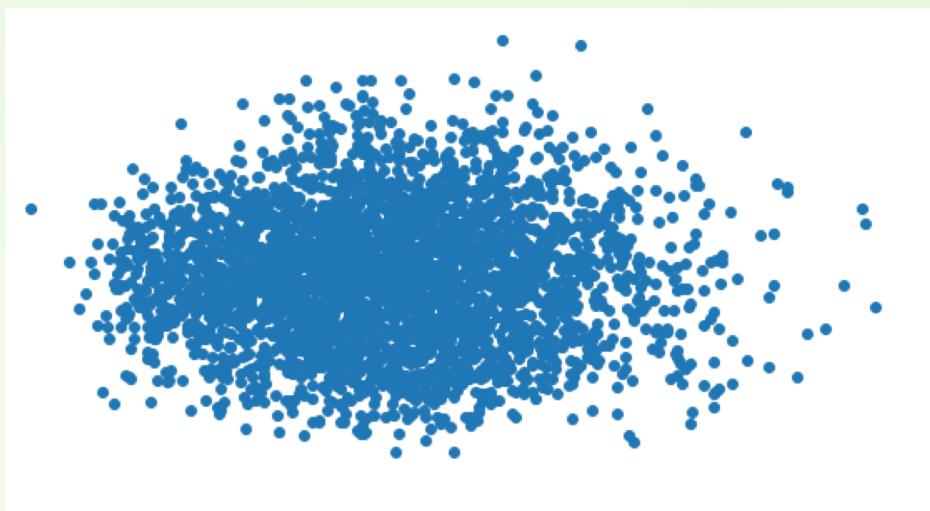


# Correlation between features and outcome



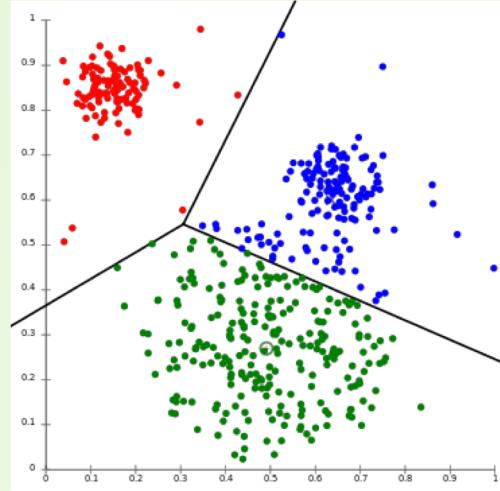
# Dimensionality reduction

- PCA (Principal Component Analysis)
  - Finds low-dimensional representation of high-dimensional data by retaining as much information in the data as possible.
- UMAP (Uniform Manifold Approximation and Projection)
  - Computationally fast, useful for visualization and feature engineering.
  - Preserves local and global structure of the data.



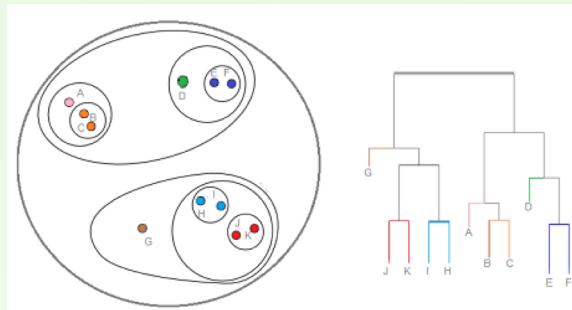
# Clustering techniques

## ❖ K-Means Clustering



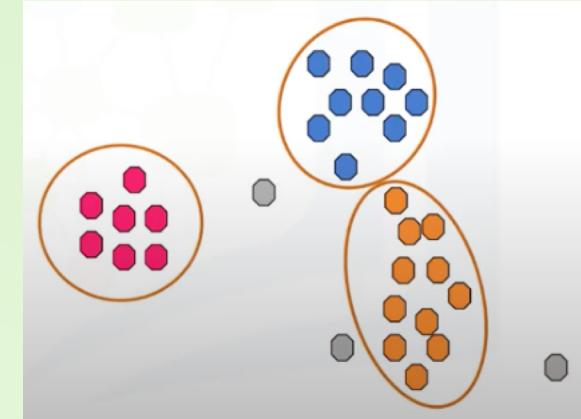
<https://aws.amazon.com/blogs/machine-learning/k-means-clustering-with-amazon-sagemaker/>

## ❖ Hierarchical Clustering



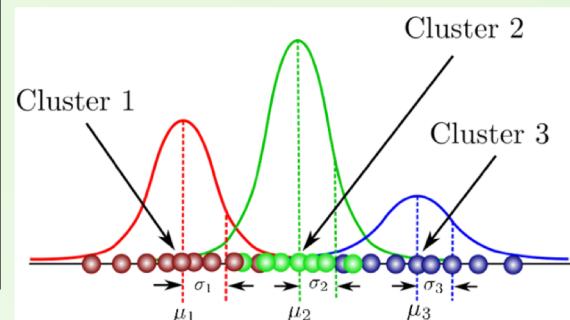
<https://www.statisticshowto.com/hierarchical-clustering/>

## ❖ DBSCAN Clustering (Density-Based Spatial Clustering of Application with Noise)

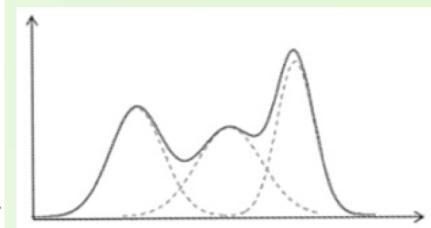


<https://www.youtube.com/watch?v=6jl9KkmgDIw>

## ❖ Gaussian Mixture Models



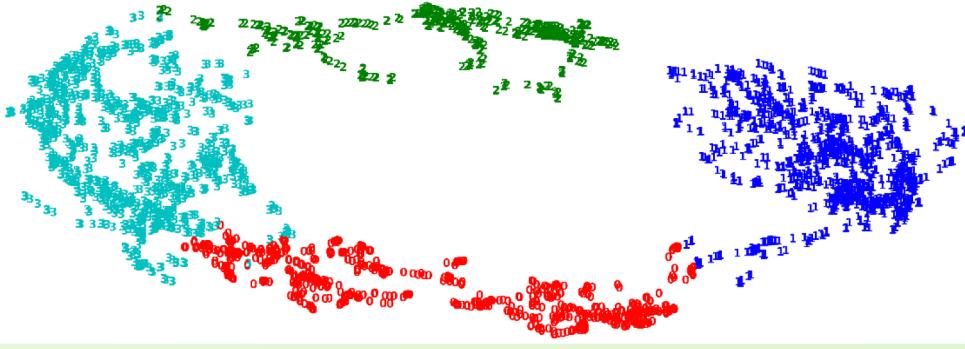
<https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>



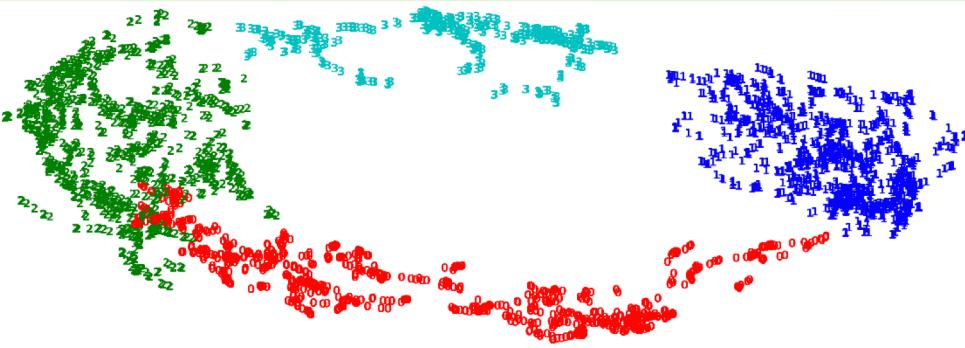
<https://www.slideshare.net/dulalsaurab/sample-project-38012289>

# Results

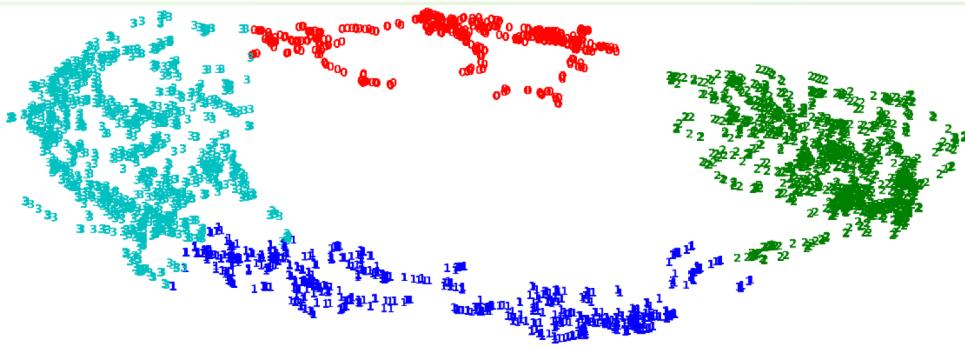
K-Means – 4 clusters



Hierarchical Clustering  
Linkage: ward

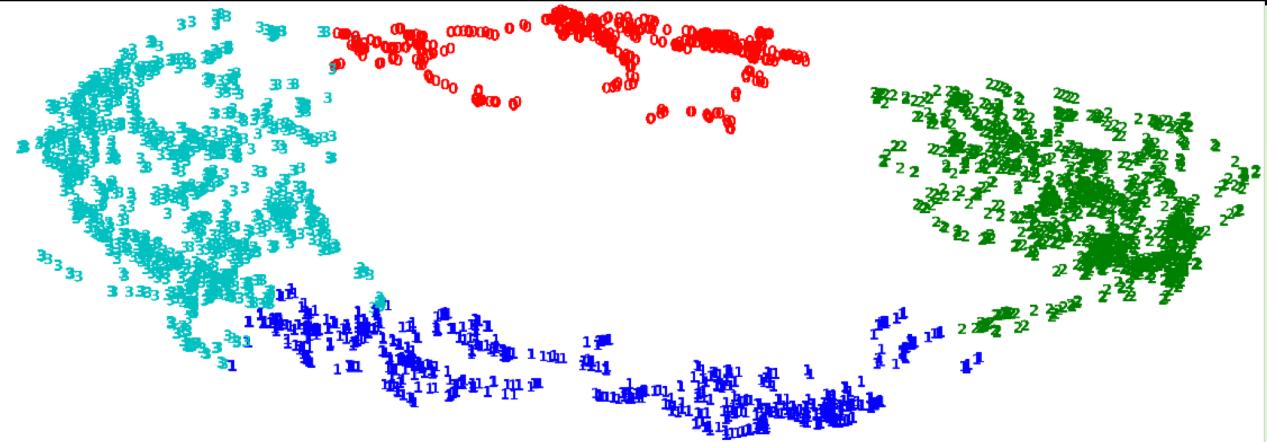


Gaussian Mixture Models  
n\_components=4,  
covariance\_type = spherical



# Results

## K-Means – 4 clusters



Cluster	Player (forward)	Overall Score									
0	D. Candeias	74	1	J. Tagueu	72	2	B. Dost	83	3	L. Messi	94
0	R. Quaison	74	1	S. Bahoken	72	2	N. Petersen	79	3	C. Ronaldo	94
0	M. Bolaños	73	1	E. Gallego	71	2	D. Sousa	78	3	Neymar Jr	92
0	Cristiano	73	1	S. Abdullahi	71	2	W. Weghorst	78	3	E. Hazard	91
0	T. Villa	73	1	J. Marriott	71	2	G. Hoarau	77	3	L. Suárez	91

# Results

Model	Silhouette Score	Cluster Visualization
K-Means (4 clusters)	0.52 +/- 0.17	Very Good
Hierarchical (Agglomerative): Linkage= complete	0.49 +/- 0.23	Good
Hierarchical (Agglomerative): Linkage= average	0.48 +/- 0.23	Poor
Hierarchical (Agglomerative): Linkage= ward	0.50 +/- 0.23	Good
DBSCAN	0.48 +/- 0.31	Poor
Gaussian Mixture Models (n_components = 3, covariance = full)	0.47 +/- 0.17	Very Poor
Gaussian Mixture Models (n_components = 4, covariance = spherical)	0.52 +/- 0.18	Very Good

## Conclusion and The Road Ahead

- UMAP is the best dimensionality reduction technique on **forward** soccer players.
    - Reduced 35 features into 6 main components.
  - Best clustering techniques are K-Means with 4 clusters and Gaussian Mixture Models with 4 components and spherical covariance.
    - Best silhouette score, consistent standard deviation of silhouette samples and best cluster visualization.
  - Best players to pick by clubs and fantasy soccer fans, or to analyze by the opposition teams are those with highest Overall score (L. Messi and C. Ronaldo best 2 **forward** players).
- Could exclude goalkeeping features when applying clustering techniques on **forward** players.