# Clustering Project

Data Analytics - I

September 12, 2020

## Dataset

This dataset contains features of around 18k football players. You are expected to complete the following tasks. Use only numerical attributes to cluster the data.

Language - **Python** (You can use python libraries except for the second task)

## 1. Data visualisation

You can use seaborn and matplotlib for visualisation. This is an open-ended task. For instance, some basic visualisation you can perform:

- Plot histograms of count of players on the basis of some attributes like height.
- Distribution of players in different clubs/country on the basis of some attribute.
- Features of players according to their position.
- Can you use visualisation techniques to find outliers? Like Ronaldo and Messi.

## 2. K-means

1. Implement k-means clustering algorithm from scratch.
2. Choose k = 3, 5, 7. Use only numerical attributes to cluster.
3. Use elbow method and Silhouette Score to get optimal number of clusters.
4. Analyse the results got in every case and try to mark each cluster.

## 3. Hierarchical Clustering

1. Cluster the data using any Agglomerative(bottom-up strategy) method of your choice.
2. Cluster the data using any divisive hierarchical clustering method (top-down strategy).
3. Plot a dendrogram for both. Compare the clusters formed by both and analyse the clusters formed.

## 4. DBSCAN

1. Use DBSCAN to cluster the data.
2. DBSCAN algorithm requires 2 parameters - epsilon and minPts. Show all experiments you did to arrive at the final eps and minPts.
3. Analyse the clusters formed.

Finally compare the clusters formed by each of the above technique. Which method is the best according to you for clustering the given dataset? Which clustering technique made the most meaningful clusters?

## Analysis of clusters

This is a naive example of how you can approach analysing clusters. Suppose you got 3 clusters.

- How good are the clusters? Use intra-class similarity and inter-class similarity to measure the goodness of clusters.
- Which attributes are the most similar in a cluster?
- Can these clusters be named according to mean of attributes present in each cluster like Forwords, Midfielders and Defenders.
- Are there any outliers? If any, then what do you interpret from those outliers. Which attributes were different in them?
- Find any hidden patterns if any.

## Submission details

- Make a zip file containing a detailed report and a folder named code containing all python code files.
- Naming convention: RollNumber1_RollNumber2_clustering.pdf,
  e.g., 20171117_20171185_clustering.pdf
- Only one team member needs to submit
- **Any kind of plagiarism will be severely punished.**