# Machine Learning - Milestone 1 - Unsupervised Learning

Raúl Barba Rojas[1,1*], Diego Guerrero Del Pozo[1,1†] and Marvin Schmidt[1,1†]

[1*]Department of Computer Science, Universidad de Castilla-La Mancha, Paseo de la Universidad, 4., Ciudad Real, 13071, Castilla-La Mancha, Spain.

*Corresponding author(s). E-mail(s): Raul.Barba@alu.uclm.es;
Contributing authors: Diego.Guerrero@alu.uclm.es;
Marvin.Schmidt@alu.uclm.es;
†These authors contributed equally to this work.

## 1 Overview

The purpose of this document is to summarize the problem to be solved, as well as key aspects of the created solution during this first milestone. More detailed descriptions of each step can be found in the main Google Colab document[1].

### 1.1 Task Description

Within this assignment[2], Unsupervised Learning Techniques are applied onto the results of different football teams during the football world cup of 2018[3]. By applying Unsupervised Learning Techniques, patterns within the input data are revealed. Using these patterns, we can draw conclusions about the dataset, as further discussed where applicable. Moreover, these insights can be utilized within other Unsupervised Learning Techniques.

[1]Refers to document: "`Task01_UnsupervisedLearning_Barba_Guerrero_Schmidt.ipynb`"
[2]Refers to document: "`ML2022_Milestone_1_Task_Definition.pdf`"
[3]Refers to document: "`worldcup_2018_final_data.csv`"

## 1.2 Feature Selection and Dimensionality Reduction

The steps performed during this stage are essential to the usage of the dataset with other techniques. The dataset, initially having a shape of `(32, 69)` contains too many columns to be practically usable and visualizable. Having applied the aforementioned techniques, a reduction to three features was achieved using PCA explaining 88.21% of the original dataset. This also allows for a visualization of the contending teams by plotting the first two features (see Figure 1). Our approach as well as detailed explanations of this step can be found in chapters two and three of the main Google Colab document.

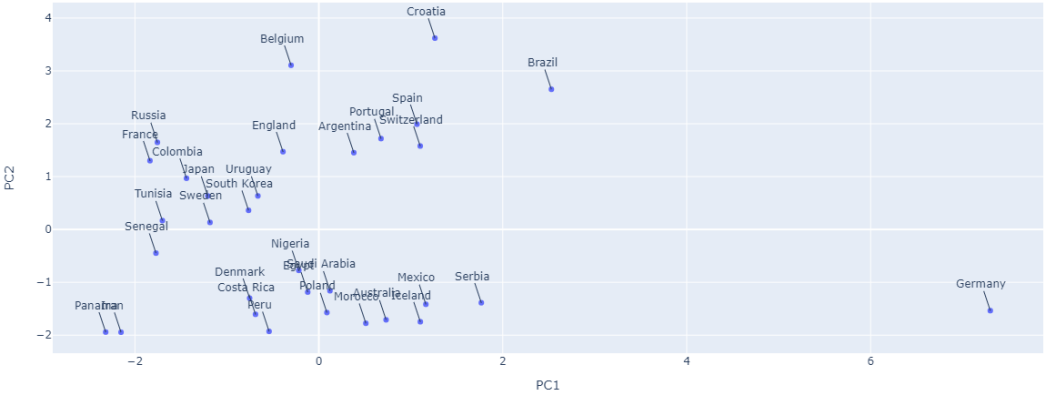


**Fig. 1** Visualization of dataset after application of PCA.

## 1.3 Outlier Detection

This step serves to detect anomalies within the data. By doing this, we reveal football teams that deviate the strongest from most other teams. In this stage, interpretations about these outliers are made, and, if necessary, filtered out in order to not affect the upcoming clustering steps. We use a parametrized instance of the DBSCAN technique to identify six teams as outliers, namely Brazil, Germany, Croatia, Egypt, Belgium and Serbia. The implementation of the DBSCAN technique, visualizations and interpretations of the outliers can be found in chapter four of the main Google Colab document.

## 1.4 Clustering

Clustering was performed using three distinct techniques in order to compare the results of each one after they've been implemented and parametrized. Namely, we've implemented clustering using HCA, K-Means and DBSCAN. Each of the three techniques uses the prepared dataset retrieved from the previous steps. The result of each technique, a categorization of teams into groups, is visualized using a similar plot as seen in figure 1. For instance, the resulting group assignment created by the HCA algorithm can be seen in figure 2. The detais regarding each implemented technique can be found in chapter five of the main Google Cloab document.

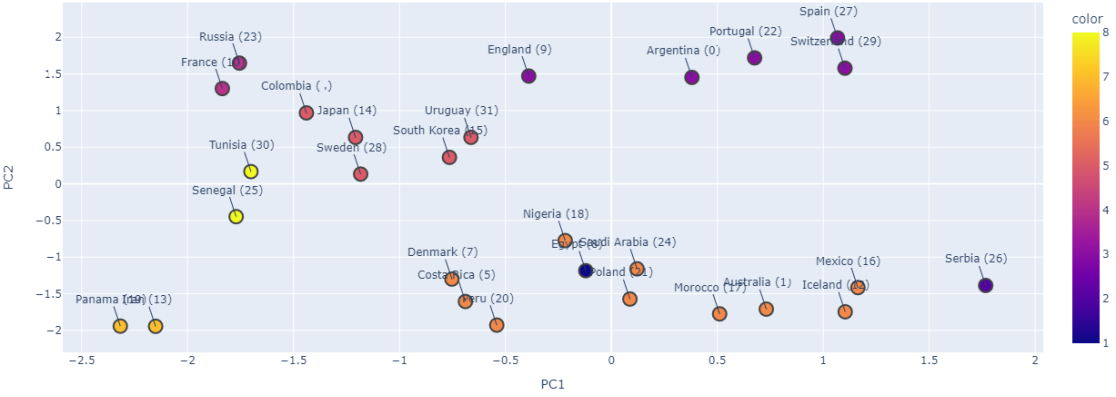


**Fig. 2** Visualization of groupings assigned to the dataset using HCA.

## 1.5 Conclusions

Using the most fitting results from the clustering algorithms, in our case the results of the HCA algorithm as seen in figure 2, we found interesting insights about the competing teams within this world cup. As we focused on the attack features of each team in the past steps, an interpretation of the shown visualization in combination with the world cup results was possible. For example, we concluded that both Russia and France are the teams with the best overall attack values of the whole tournament. The full list of drawn conclusions can be found in chapter six of the main Google Colab document.