Unsupervised Learning on FIFA Player Skills

Erik Elias Mikael Maresia
963724
January 2022

Abstract

The purpose of this paper is to understand football player skills and if there are some patterns in them. Principal component analysis is used to reduce the dimension of the data and try to understand how and if it explains it. In the end k-means clustering is used to show support to the findings of the principal component analysis.

1. Introduction

FIFA (Fédération Internationale de Football Association) is an international non-profit organisation that is the governing body of football. FIFA Football is a video game series developed by Electronic Arts. The video game simulates real football teams and players, by portraying realistic information about the teams and players.

Player objects are divided into multiple individual attributes, which make up a player. In this paper these attributes are synonymous with player skills.

Examples of these attributes are acceleration, shot power, vision and player strength.

2. Data

The data analysed in this paper is player data from the FIFA Football video game of 2018. There are 75 variables in the dataset, but not all are of interest, including the player's age, nationality, name and the team it plays for.

Only measurable attributes are of interest, which might constitute a player's position. There are 19 of these attributes. These attributes are numerical variables that range from 0 to 100, with 0 being the worst and 100 the best.

In addition to dropping irrelevant variables, the variables of interest are of type *char* instead of numerical, although being numerical in value. Before beginning the analysis, this has to be changed.

3. Theory

This section briefly explains the theory used in this paper. This paper uses unsupervised learning methods tottery to find patterns in the data. Since there is no response variables involved, predictions are not of interest, rather to understand what makes up the data.

3.1 Principal Component Analysis

Principal component analysis (PCA) is a tool to summarise a large dataset with many variables, with fewer variables. These variables collectively explain most of the data. It is a classical dimensionality reduction tool where instead

of having p features, the features are summarised in m principal components in decreased order of importance.

The first principal component of the features $X_1, X_2, .. X_p$ is the linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

That has the largest variance. The second principal component has the second most variance, and is orthogonal to the first principal component, and so on.

ϕ_1 is the *loading* vector, which geometrically shows the direction in the feature space in which the data varies the most. z_1 is the score of the first principal component. This vector contains the *scores* of the first principal component. These scores are points that represent the original datapoint with respect its principal component axis.

3.2 k-Means Clustering

Often it is of interest if the data can be grouped into clusters. This chapter is going to discuss a popular method called k-means clustering.

The idea of k-means clustering is to group the data points into k distinct, non-overlapping clusters $(G_1, G_2, \ldots, G_k)$, where G_i is the set of n_i individuals. The way find the partitions of the n individuals into the k groups, is where the within-cluster variation is as small as possible. This is

$$WGSS = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \in G_l} (x_{ij} - \bar{x}_j^{(l)})$$

where $\bar{x}_j^{(l)} = \frac{1}{n_i} \sum_{i \in G_l} x_{ij}$ is the mean of the group G_l on variable j.

The k-means clustering algorithm follows rather simple instructions:

1. Find a starting value for each observation, from 1 to
2. Iterate until no more cluster assignments are made

a. For all clusters G, calculate the centroid, being the mean of the of the features for the observations in that cluster.

b. Assign each observation to a cluster whose centroid is the closest.

4. Analysis

This chapter will try to find patterns in the data. To begin with, PCA is used for dimensionality reduction and to explore what features might have an effect on the variance of the principal components. Secondly k-means clustering is performed and it is compared to the findings of the section on PCA.

4.1 PCA

To begin with, PCA is used on the dataset. Before PCA can be used, it must be ensured that the data points are all numerical and that they are scaled. The reason for this scaling is that PCA is very sensitive to variation. If one variable has a significantly larger variance than the other variables, PCA will capture this.

Computing the PCA, it is seen that there are 19 principal components. This is because generally there are always $min(n-1,p)$ principal components. Where n is the number of observations and p the number of features. The left side panel of Figure 1 shows the proportion of variance explained by each principal component, while the right panel shows the cumulative proportion of variance explained by each principal component. From these two plots it is evident that the first principal component explains the largest percentage of variance, $56.6\,\%$, while the second principal

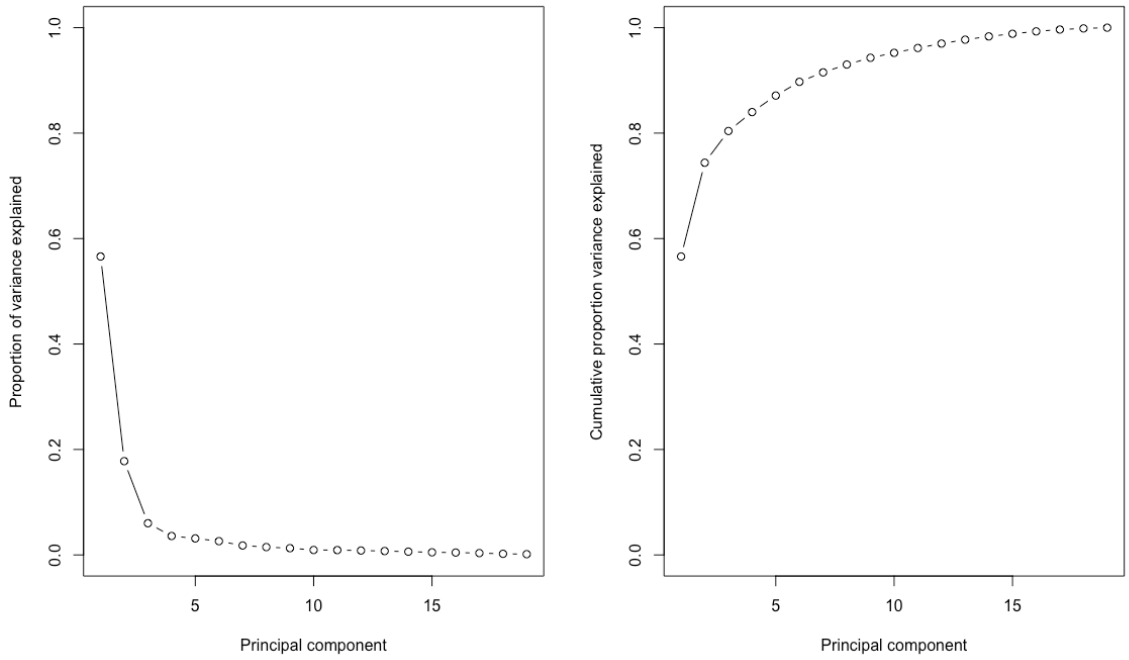

Figure 1

component explains $\approx 17.8\,\%$. The two principal components explain roughly $74.4\,\%$ of the of the variance of the data. The first three principal components explain more than $80\,\%$ of the of variance.

Plotting the scores and loadings of the first two principal components will give a biplot as seen in Figure 2. This biplot reveals many things about the data. The first principal component (PC1) explains the change on the X axis, while the second principal component (PC2) the y axis. The black dots on the plot are the principal component scores and the red arrows the loadings.

Some arrows seem to cluster very closely together. The most evident ones are the ones that point to larger values on the x axis, with labels such as *Marking, Interceptions, Sliding.tackle* etc. Arrows that are very close together mean that they are highly correlated positively - the higher the Marking skill is, the higher the Interceptions skill tends to go etc. The same is also apparent as the opposite. The higher the Marking skill tends to go, the

smaller the Finishing skill tends to go - thus they are highly negatively correlated.

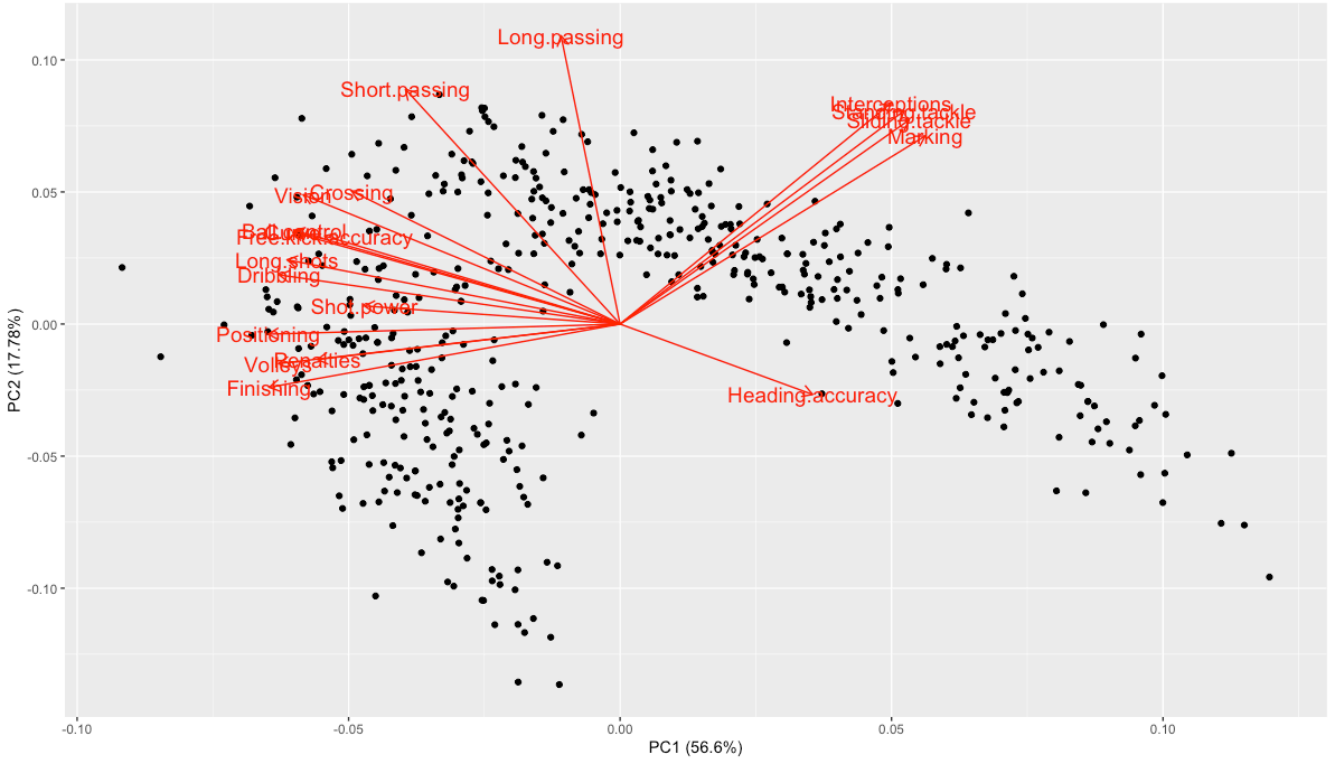

Figure 2

Since it is known that certain skills are associated with certain player positions on the field, this might shed light into what the first two principal components might be explaining. As mentioned above, the first principal component seems to be explaining if a player is more an offensive or defensive player, but that leaves the seconds principal component unexplained. The second principal component explains the change on the y axis of the biplot. One feature stands out the most, *Long.passing*.

To understand more on to what the principal components might be explaining, it might be useful to look at a correlation matrix of the principal components and the original data. This correlation matrix can be seen in Figure 3. Immediately it is clear that the first principal component is strongly correlated with almost all of the features. The second principal component is

strongly correlated with much fewer features and the third one with even fewer.

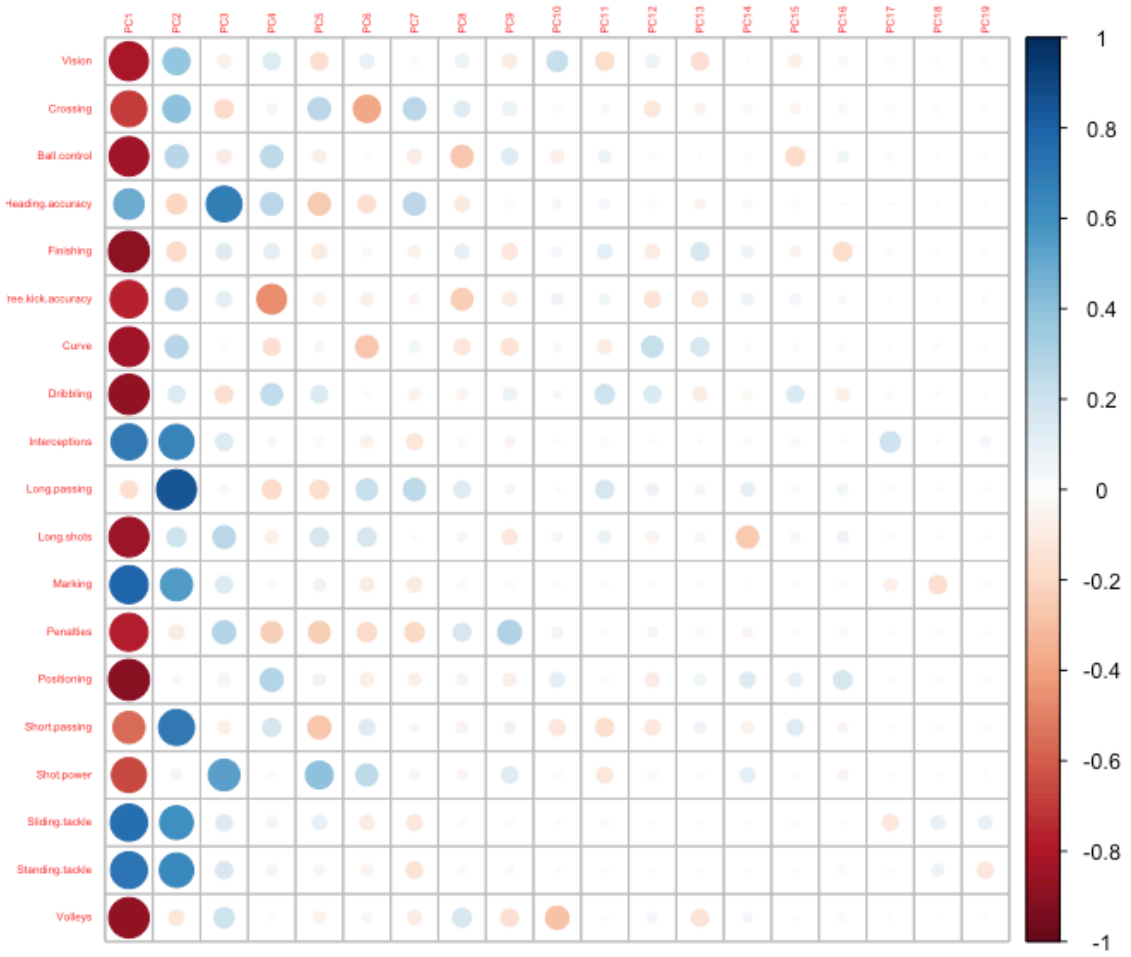

Figure 3

The observation made above on the first principal component possibly explaining defensive or offensive traits is more evident in the correlation matrix. Highly offensive skills are highly negatively correlated (e.g. *Long.shots*, *Shot.power*, *Finishing*), while highly defensive skills are highly positively correlated (*Interceptions*, *Sliding.tackle* etc.). The question still remains on what the second principal component is explaining. This is question is perhaps easier to answer when first looking at the correlation matrix. The highly positively correlated features are features that might be associated to defenders or more namely, midfielders. The negative

correlations show more offensive features, such as *Finishing* and *Volley*. So it might be, that the second principal component is explaining the differences between attacking and midfield skills.

4.2 k-Means Clustering

Taking what was discussed in the previous chapter, the next question would be if it would be possible to cluster the data in a sensible way into three categories. k-Means clustering is a popular unsupervised statistical learning method suited for this task.

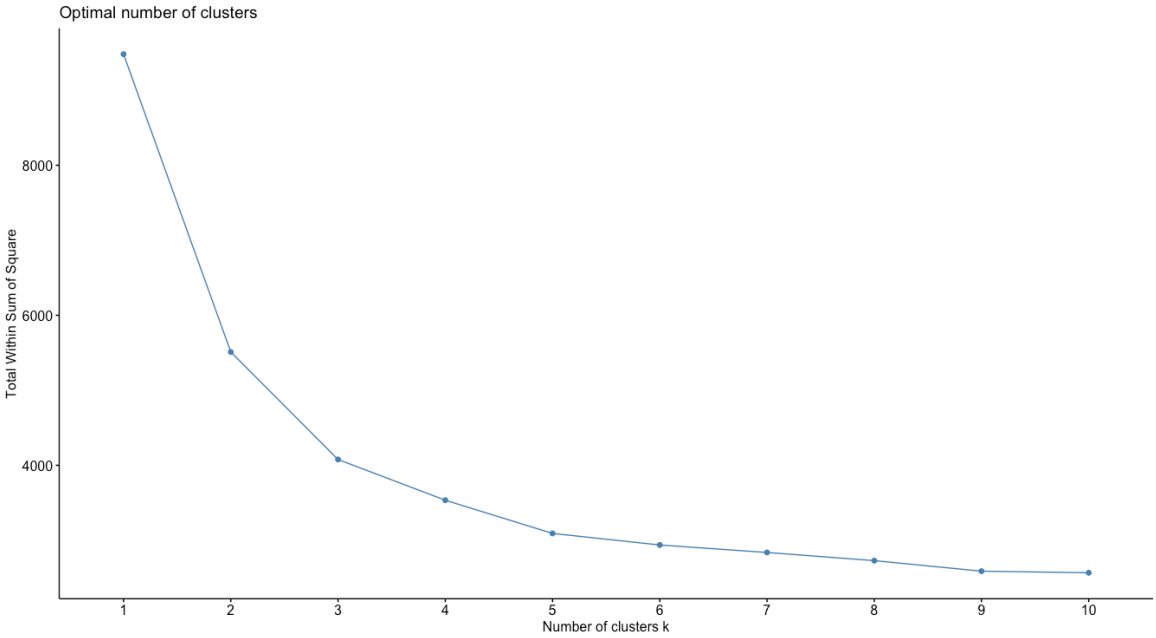

Figure 4

One practical issue with k-means clustering is to decide on what value to choose for k. A way for deciding the value of k, is to compute many k-means models with a different value for k and compute the total within-sum of squares for each model.

In Figure 4 the total within-sum of squares is seen for 10 different values of k. A possible candidate for choosing k could be 3, since it seems to be an

elbow point on the graph - the proportion of changes in the total within sum of squares gets smaller as $k > 3$.

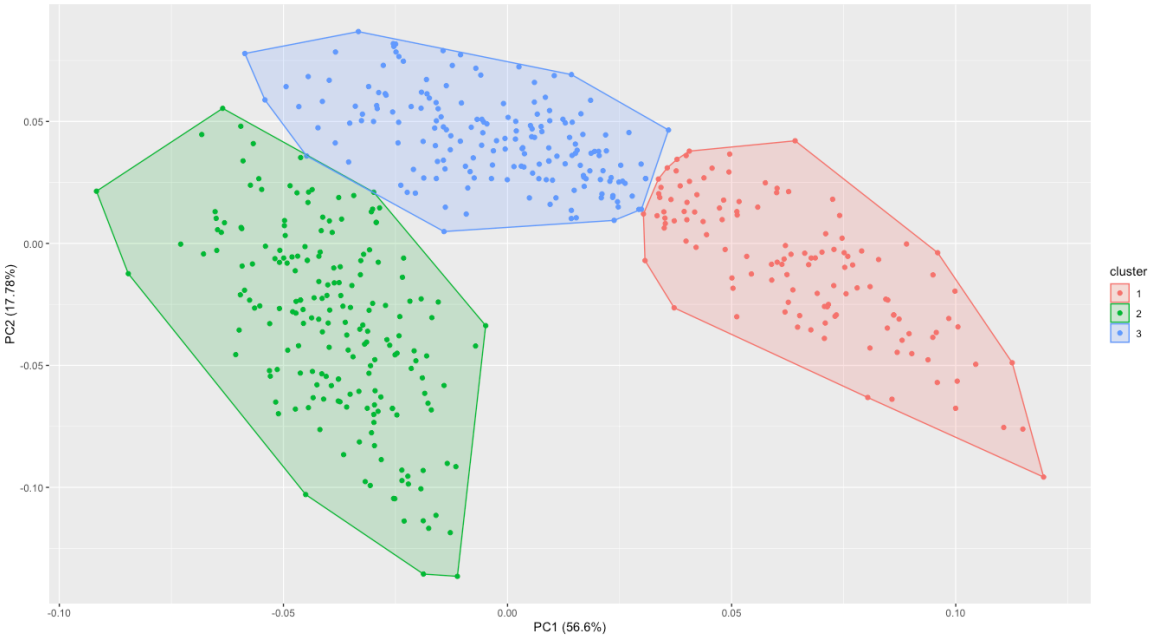

Figure 5

The results of setting $k = 3$ can be seen in Figure 5. The results of the three clusters are somewhat equal to what was analysed with PCA. There are three separate clusters, which might be explained by player positions as was discussed in the previous section.

5. Conclusions

This paper discussed unsupervised statistical learning methods, which can be applied to data without a response variable. This paper focused more on understanding what patterns there might be in the data, which cannot be immediately seen due to the amount of features.

When analysing the data with PCA, there were a few things which rose to importance. The first one being that the data, which was a dataset of 19 features, could be represented with a much smaller dimension - without

loosing too much of the information. Namely $80\,\%$ of the variability of the data was explained with just three principal components.

The principal component analysis showed that the data had some underlying patterns. This was shown by, for example some features grouping together showing positive and negative correlations between the data.

The results of the analysis of the k-means clustering support the finding of the principal component analysis. The possibility that the data is showing player positions on the fields was shown by the fact that the optimal choice for k was 3 (Figure 4). This might give support to the idea of the data being in three different categories: defensive, attacking and midfield players.

6. References

- Gareth, James et al. An Introduction to Statistical Learning with Application in R, Springer, New York, USA, 2013, pp. 373–413.
- Everett, Brian et al. An Introduction to Applied Multivariate Analysis with R, Springer, New York, USA, 2011, pp. 61-102;163-200
- Härdle, Wolfgang Karl et al. Applied Multivariate Statistical Analysis, Springer, New York, USA, pp. 320-325

Appendix

The R code that was used in this paper and the dataset can be found from: https://github.com/maresiaerik/statistical_learning_course