# **PCA and Clustering on FIFA Data**

Report by

Bharat Nagaraju (A0178258N)

Edwin Tam (A0178396J)

Vigneshram Andiappan Selvaraj (A0178215A)

# **Index**

1. Introduction
2. Data Description
3. Data Understanding
4. Data Cleaning
   1. Missing Data
   2. Distribution and Outliers
   3. Variable Selection
5. Dimension Reduction
   1. Principal Component Analysis
   2. Eigen Values
   3. Scree Plots
   4. Component Selection
   5. Component Profiling
6. Clustering
   1. Description of clusters
7. Regression Analysis
8. Conclusion

# **Introduction**

The data which we picked for doing our analysis is the match statistics of the FIFA 2018. We used this public dataset available in Kaggle to discover interesting patterns in it. The data has been originally collected from 2018 FIFA World Cup Russia Official App and posted in Kaggle. We have first done a PCA to reduce the number of variables which was originally 23 for our analysis and built a model to cluster the data to find the patterns in it. Finally a logistic regression model was built to predict if there is a man of the match for individual games. The number of observations is around 126 which is the number of matches in the world cup.

# **Dataset Description**

The data set has 23 columns in it including the dependent variable. The data collected contains several statistics like the goals scored, penalties missed opportunities and accuracy. These will help us in identifying the matches were known for which characteristic of it. This can be further used to suggest matches based on the search criteria over internet.

# **Data Understanding**

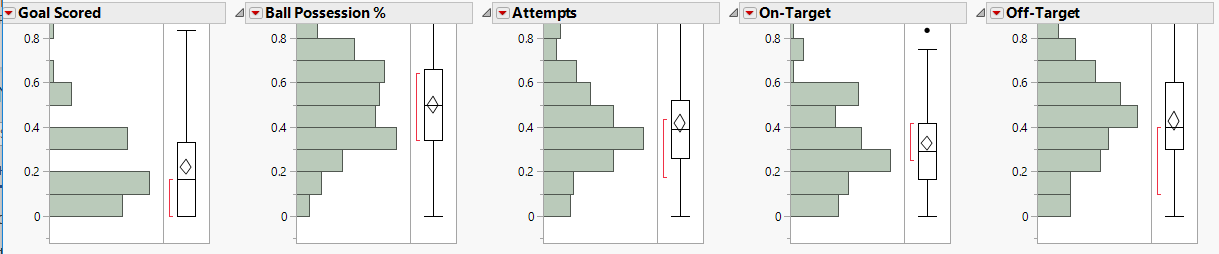
|  |  |
| --- | --- |
| Feature | Description |
| Goal Scored | Goals scored in the match |
| Ball Possession % | Ball Possessed by the team in the match |
| Attempts | Attempts to score goals |
| On-Target | Attempts that were on target |
| Off-Target | Attempts that were off target |
| Blocked | Blocks made to counter the shots |
| Corners | Corners earned |
| Offsides | A player is in an offside position if: he is nearer to his opponents' goal line than both the ball and the second last opponent |
| Free Kicks | Restart of play that is awarded to a team following most types of fouls |
| Saves | Goals saved |
| Pass Accuracy % | Passes that were on target |
| Passes | Passes made both accurate as well as inaccurate |
| Distance Covered (Kms) | Distance covered in whole match by both teams |
| Fouls Committed | Acts committed by players which are deemed by the referee to be unfair and are subsequently penalized |
| Yellow Card | Yellow cards are used to punish milder forms of misconduct by players |
| Yellow & Red | Players who got an automatic red card subsequently after he received two yellow cards. |
| Red | Number of direct red cards earned in the game |
| Man of the Match | Indicator to show if the game had a man of the match |
| 1st Goal | Time when first goal was scored |
| Round |  |
| PSO | Indicator to show if the game had a Penalty Shooutout. This is used to decide the winner if both teams are on a draw after extra time |
| Goals in PSO | Number of goals in PSO |
| Own goals | When a player scores a goal on their own side of the playing area rather than the one defended by the opponent |
| Own goal Time | Tie of own goal scored |

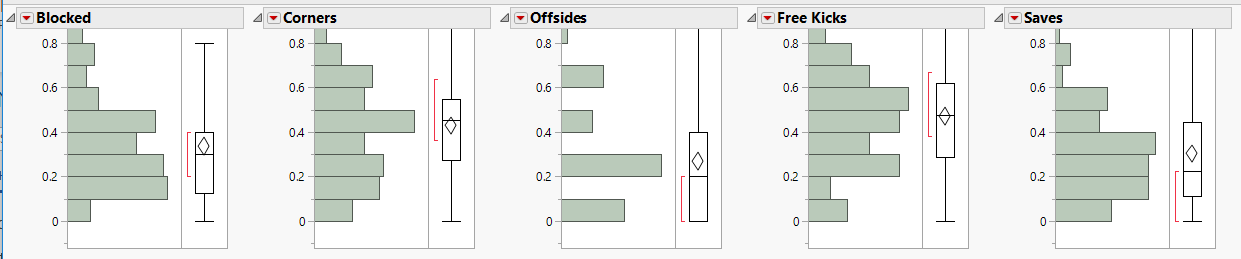
# **Data Cleaning**

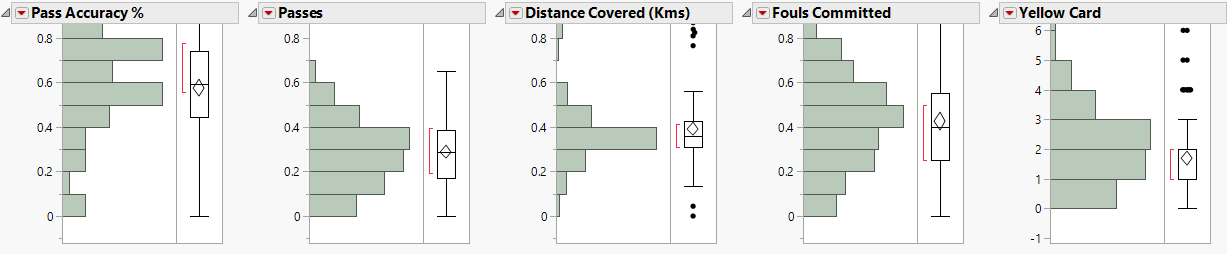
|  |  |
| --- | --- |
| ITEM | STATS |
| Number of variables | 23 |
| Number of samples | 124 |
| Rows with partial data | 0 |
| Number of Target Variables | 1 |

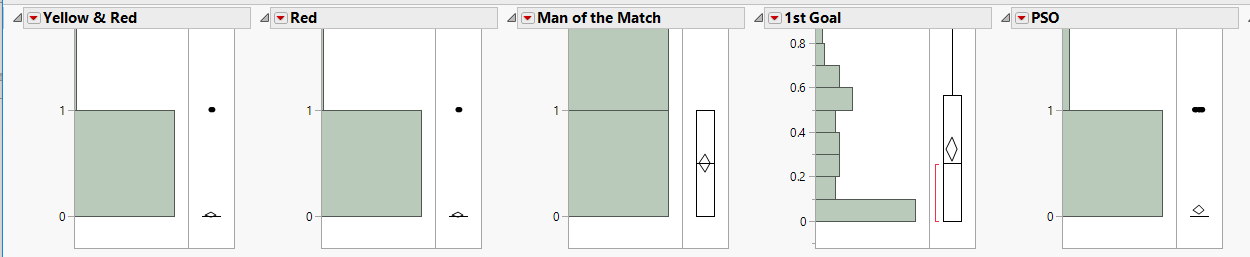
Data Normalization

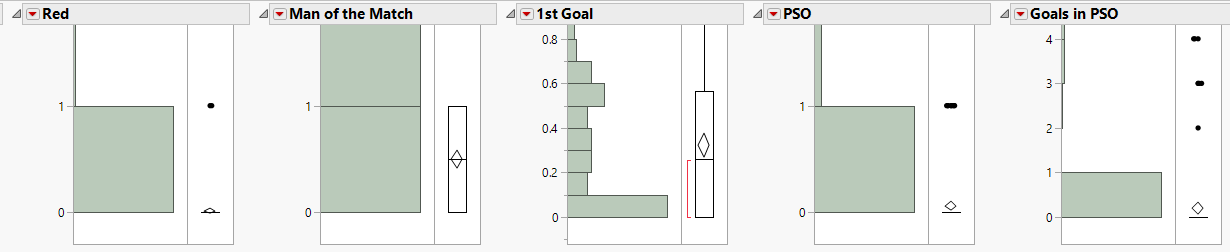
Since there is no missing data in the whole data set we can move ahead to the normalization section.











Although there are a few outliers, they actually are not due to dirty data. It is due to the fact that in football during the league stage we do not have extra time and penalty shoot out. Hence these features show up some outliers for matches in the knockout phase. We might end up finding some interesting patterns because of these outliers too and hence we have included these data too in clustering exercise.

# **Dimension Reduction**

## **Principal Component Analysis**

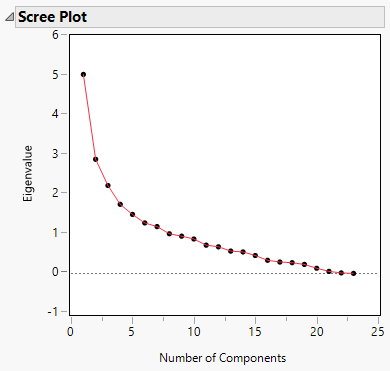
We use principal component analysis to get reduced number of independent linear combinations (principal components) that can be used to project as much of the variability in the original variables as possible. These are the summary plots generated after applying the appropriate settings and carrying out PCA. After performing PCA, we were able to analyze the total variance between the dimensions within our dataset. The Principal Component Analysis also helps us decide the number of principal components based on the largest eigenvalues obtained. This gives us insight about the direction in which the largest variance is present. To decide the number of total principal components, we searched the Eigenvalues table for the components with eigenvalue greater around 1.

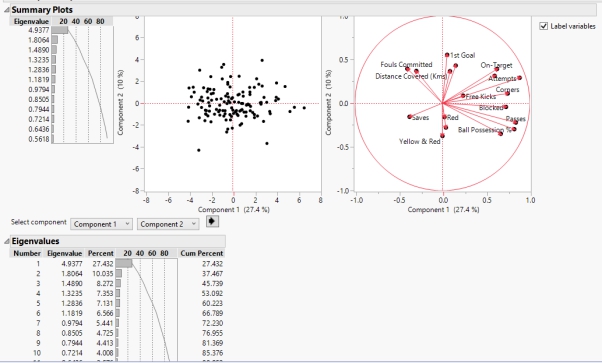
## **Eigenvalues**

From the Eigenvalues table, we observed that the first 7 principal components were adequate in explaining 72.30% of the cumulative variance. Hence, from the Eigenvalues Table, we have selected the total number of Principal Components to be 7. We can also double check with the findings from the corresponding Scree Plot generated.

## **Scree plots**

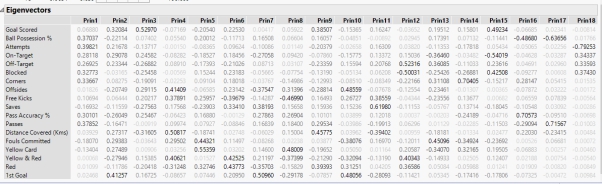
We analyzed the Scree Plot to figure out the value at the elbow of the graph to verify the number of Principal Components required. The value at the elbow point coincides with the selected number from the Eigenvalues Table. We observed an elbow at the point corresponding to 8 principal components.





## **Component Selection**

Out of the resulting 26 principle components, we chose the top 7 principal components with Eigen values greater than 1 and this corresponds to 72.30 % coverage.



## **Component Profiling**

Based on the type of columns involved in each type of the clustering done, we have assigned names for all the selected components, and for the type of care that the diabetic patient received. We have:

|  |  |  |
| --- | --- | --- |
| No | Principal Component | Description |
| 1 | Ball Control | Reflects how the ball was controlled well in the game |
| 2 | Factors influencing Goal | Shows the critical factors which influence goal |
| 3 | Goals Scored | Gives an overview of the goals scored in the match |
| 4 | Missed Opportunities and Distance covered | Highlights the opportunities missed when trying to score a goal |
| 5 | Fouls turned to yellow card | Covers how the fouls eventually turned into a Yellow card |
| 6 | Number of Foul Cards earned | Gives an overview of number of Foul cards earned |
| 7 | Good Performance | Highlights the standout performance in the match |

# **Clustering**

# **Implementation Code Approach**

K-means clustering aims to converge on an optimal set of cluster centers (centroids) and cluster membership based on distance from these centroids via successive iterations. It is intuitive that the more optimal the positioning of these initial centroids, the fewer records will go into wrong clusters or we end up forming a new cluster altogether. Therefore, thinking about ways to find a good implementation code is to fix a better set of initial centroid positions.

What we will do differently, specifically, is to draw a sample of data from our full dataset, and run short runs of the k-means clustering algorithm on it (not to convergence), short runs which will include, out of necessity, the centroid initialization process. We will repeat these short runs with a number of randomly initialized centroids, and will track the improvement to the measurement metric -- within-cluster sum-of-squares -- for determining goodness of cluster membership (or, at least, one of the valid metrics for measuring this). The final centroids associated with the random centroid initialization iteration process which provide the lowest inertia is the set of centroids which we will carry forward to our full dataset clustering process.

The hope is that this up-front work will lead to a better set of initial centroids for our full clustering process, and, hence, a lesser number of k-means clustering iterations and, ultimately, less time required to fully cluster a dataset.

This would obviously not be the only method of optimizing centroid initialization. In the past we have discussed the naive sharding centroid initialization method, a deterministic method for optimal centroid initialization. Other forms of modifications to the k-means clustering algorithm take different approaches to this problem as well (see [k-means++](https://en.wikipedia.org/wiki/K-means%2B%2B) for comparison).

The approach for our task will be as follows:

* Prepare the data
* Prepare our sample
* Perform centroid initialization search iterations to determine our "best" collection of initial.
* Use results to perform clustering on full dataset

## **Description of clusters**

Based on the clusters obtained and the summary using the principal components, the key characteristics of each were examined



Poor Ball Control Matches

This cluster shows clearly the group of matches where the ball control has been very poor. This can be used to analyse how the ball control can be improved based on the errors found.

Matches with Clinical Performance

It can be seen that this cluster shows a high value for the Ball control which obviously means this is a set f matches where technically the things have been done in a good way. They can be used to analyse and see how and what tactics have been used under different circumstances to maintain good ball control.

Chances turned into Goals

This cluster highlights all the matches where the chances have been eventually turned into goals. Hence this can be used to analyse and find those factors which are critical for goal scoring

Missed Opportunities

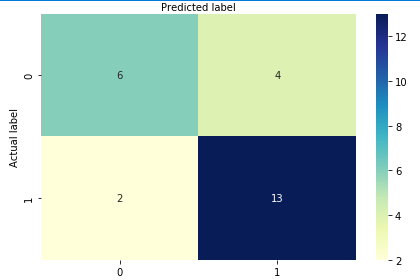
Here we can understand the factors which lead to missed chances. Primarily we can analyse this cluster to understand the factors which influence a goal and how it has been squandered. This can be used to understand how goal scoring opportunities can be turned successfully to goals by reducing the mistakes.

# **Regression Analysis**

The regression analysis was done with Man of the match as the target variable. Using the derived factors and linear regression, we will now test our model for its performance. The Accuracy, Precision and the Recall was taken while taking Principal Components 1 to 7.

* Below is the model performance variables for the model built with original variables and one with principal components. It can be seen the performance is almost the same despite taking only the principal components for logistic regression.

Model Performance on Principal Components



Model Performance on Cleansed Original Data

