Weekly Report (15/03/25)

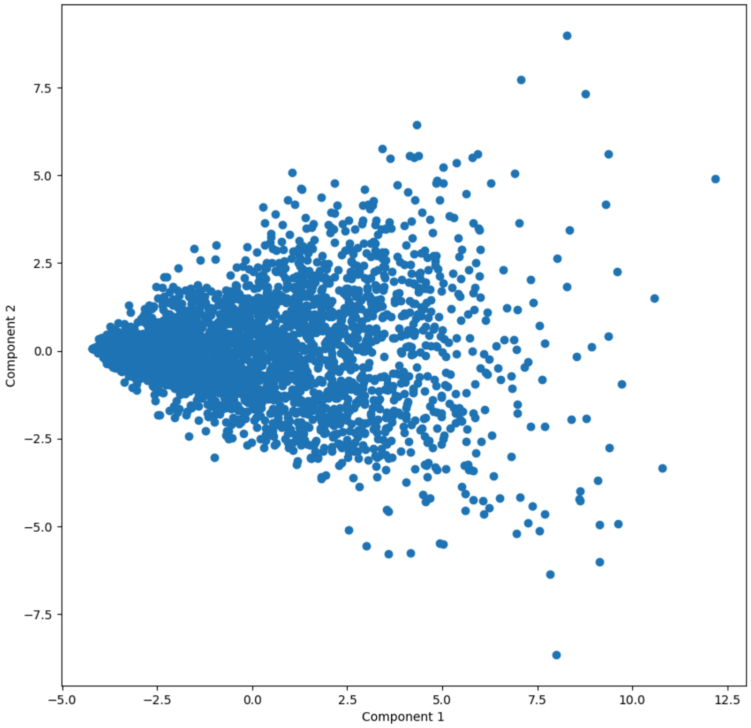
We have implemented some preliminary approaches this week and identified their problems along with the reasons for said problems. We will use these to guide future approaches.

K-Means Clustering

We considered the problem of clustering individual athletes into clusters based on their average statistics over all their appearances. One approach was to use K-Means clustering, a standard algorithm for unsupervised clustering problems.

We investigated the Within Cluster Sum of Squares (WCSS) as a metric for comparing different k values. Via the "Elbow Method", we concluded that 11 clusters was optimal for this metric. Upon investigating the $k = 11$ model with other unsupervised metrics (Silhouette Coefficient, Davies-Bouldin Index), we found the approach to be unsatisfactory.

We believe that the primary reason for the model's failure is the Curse of Dimensionality. Since our dataset is multidimensional, the usage of distance by K-Means becomes inefficient. We then investigated the dataset after dimensionality reduction via PCA to 2 parameters, however K-Means still resulted in suboptimal results.



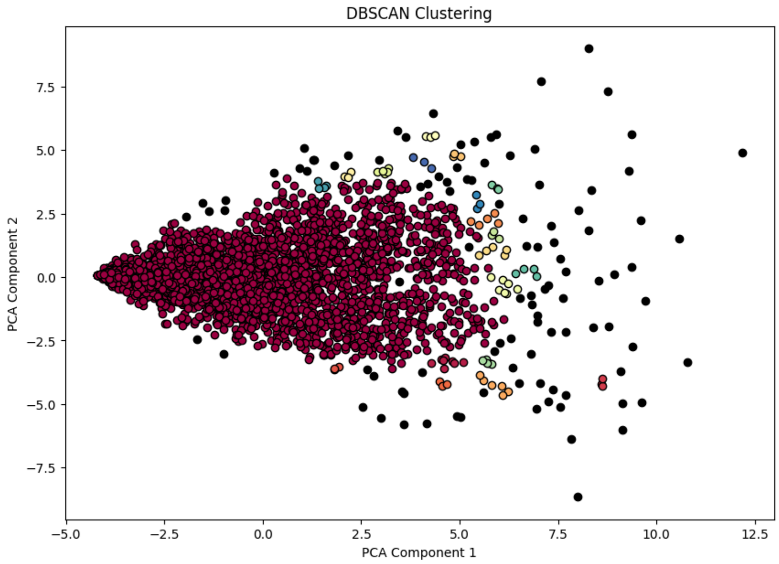
Dataset scatter plot after PCA reduction to 2 components/parameters

On a cursory glance of the PCA data, we see that there does not seem to be clear clusters. It may be the case that we would be more successful in considering the entire dataset instead. We will continue to implement clustering approaches that do not require distance as a metric as well as algorithms optimised for multidimensional data (such as Shared Nearest Neighbour, Biclustering and Subspace Clustering).

DBSCAN

We implemented DBSCAN (Density-Based Spatial Clustering of Applications with Noise) as an alternative approach for clustering to K-Means. Directly applying DBSCAN to the dataset was unsatisfactory as the clustering algorithm suffered due to the Curse of Dimensionality.

We integrated DBSCAN with PCA (Principal Component Analysis) to mitigate this issue. PCA reduces the dataset's dimensionality by identifying principal components that capture the most variance. Additionally, PCA helps reduce noise by focusing on the most significant features, leading to better clustering results. However, even with the integration of PCA, the results were far from desirable. Another issue is the lack of explainability of the clusters created by DBSCAN.



ANFIS

We worked with the package of ANFIS (Adaptive Neuro Fuzzy Inference System) from github. The package was written in python 2 and was converted to python 3. As the code had no documentation, we studied the package and reworked it according to our use case. Implementation on test dataset was done.

We started using the package on given dataset but we encountered some problems. We think these problems are in the logic of the package as it is not very refined. After some edits we will try to use the package on athlete dataset.

We will use ANFIS to get:

- Best player combinations against specific opponents.
- Individual player impact on the team's success.
- Team strength

Random Forest Regressor

The game score formula highlights that the variable Points (PTS) is the most dominant variable in determining a player's game score. Unlike other factors that are multiplied by coefficients (such as 0.4 for Field Goals Made or 0.7 for Assists), Points are directly added, making them the primary driver of score variations. This means that regardless of a player's contributions in other areas, their ability to score remains the strongest determinant of overall performance. Feature importance analysis from our model confirmed this, as Points consistently had the highest weight in predicting game scores.

To analyze and predict game scores based on player statistics, we implemented a Random Forest Regressor on the dataset. The R^2 value of our Random Forest Regressor was 0.93, indicating that the model effectively captures the relationship between game statistics and overall performance. However, after removing Points (PTS) from the dataset, the R^2 dropped to 0.91, showing a decline in predictive accuracy. This aligns with the Game Score Formula, where PTS is the dominant factor, significantly influencing the final score. Random Forest helps in this scenario by aggregating multiple decision trees, reducing variance, and improving overall prediction reliability.