NFL Player Archetypes Using Clustering

Emily Ahern
Indiana University
Bloomington, IN
enahern@iu.edu

Noah Bussell
Indiana University
Bloomington, IN
nabussel@iu.edu

Sydney Gargiulo
Indiana University
Bloominton, IN
skgargiu@iu.edu

Jack Schwartz
Indiana University
Bloomington, IN
schwajaw@iu.edu

***Abstract*—This project uses K-means clustering to identify performance-based archetypes of NFL players using NFL Combine data from 2008 to 2022. After preprocessing and standardizing key physical and athletic metrics, we applied clustering techniques without using position labels. Evaluation methods showed that $k = 2$ yielded the most distinct groupings. Visualizations such as PCA plots and radar charts revealed two clear athlete profiles—one emphasizing speed and agility, the other size and strength. The resulting clusters aligned closely with actual player positions, highlighting the effectiveness of unsupervised learning in profiling athletes.**

I. Introduction

The ability to analyze player performance data offers valuable insight into the factors that define different types of athletes and their positions on the football field. The critical intent of this project is to explore NFL players performance through the use of clustering techniques learnt throughout the span of this course, with the objective of uncovering patterns and grouping players into performance-based profiles, or archetypes. Clustering, a data mining technique, enables us to group players based on similar features such as physical attributes, game statistics, position-specific values, and other related attributes. By organizing NFL players into clusters according to their data, it is possible to observe the different attributes and their dynamics that contribute to a player's football archetype. This gives the opportunity to understand player types—such as wide receiver, running back, cornerback, linebacker, tight end, safety, or quarterback—and help inform strategic decisions in scouting and performance evaluation. THis project aims to demonstrate how clustering can be used to discover patterns in NFL data and offer a data-driven perspective for profiling athletes. As a result, the report below outlines the objectives, data description and statistics, methodology, implementation timeline, data preprocessing, model training insights, and key findings.

II. Data Description and Statistics

The dataset used in this analysis is derived from the Kaggle NFL Combine Dataset [1], which compiles data scraped from Pro Football Reference [2]. It includes performance and physical metrics for players who participated in the NFL Scouting Combine from the years 2000 through 2022. Each year's data was originally provided as a separate CSV file, which we have combined into a single comprehensive dataset to enable longitudinal analysis. Each row in the dataset represents an individual athlete and contains attributes such as player name, position, college team, height, weight, and results from various athletic drills.

The dataset focuses on players occupying key skill and defensive positions, including wide receiver (WR), running back (RB), cornerback (CB), tight end (TE), linebacker (LB), safety (S), and quarterback (QB). For each athlete, physical characteristics such as height (in inches) and weight (in pounds) are recorded, with observed ranges of 65 to

80 inches for height and 168 to 336 pounds for weight. Performance metrics from the combine include the 40-yard dash, vertical jump, bench press, broad jump, 3-cone drill, and shuttle run. The 40-yard dash measures straight-line sprint speed over a 40-yard distance, with recorded times ranging from 4.26 to 5.29 seconds. The vertical jump, which assesses lower-body explosiveness, ranges from 25 to 36.5 inches, while the broad jump, another measure of lower-body power, ranges from 98 to 140 inches.

The bench press measures upper-body strength by counting the number of repetitions an athlete can complete at 225 pounds, with observed values ranging from 3 to 34 reps. Agility and change-of-direction speed are evaluated using the 3-cone drill and shuttle run. The 3-cone drill involves a tight-turning pattern and recorded times range from 6.28 to 7.96 seconds, while the shuttle run (also called the 20-yard shuttle) consists of rapid lateral movements over a short distance, with times ranging from 3.75 to 4.96 seconds. All time-based variables are measured in seconds, length-based variables in inches, and weight in pounds.

This dataset enables a robust analysis of athletic performance across positions over time. It provides insight into how players' physical attributes correlate with their Combine results, and how these measurements vary by role on the field.

III. Methodology and Timeline

A. Timeline

Throughout the course of this project, we adhered to a structured timeline beginning in the week of April 21^{st}. During this week, we focused on data preprocessing. Which included cleaning the dataset and selecting, combining, and manipulating the data to observe and analyse the different attributes. We encoded categorical variables and removed unnecessary columns within the dataset. Initial plotting was also completed to gain an understanding of the data's structure and identify which attributes might be the most impactful for a deeper analysis. During the week of April 28^{th}, we implemented the K-means clustering algorithm and completed all associated coding tasks. Using the calculated Sum of Square Errors (SSE) and other related clustering methods such as Silhouette Score and Elbow Method, we analysed the clustering results. During this time, we began drafting our final report, identifying key findings, and finalizing the plots for the presentation. From May 3^{rd} to May 6^{th}, we concentrated on preparing the final presentation. This included compiling our findings into a PowerPoint, writing the final project report, and ensuring all files were formatted and ready for submission.

B. Methodology

Clustering methods are applied to the NFL player combined dataset [1] in this project to uncover athletic archetypes based on similar attributes. The methodology includes data preparation, performing K-means clustering, and analyzing the identified clusters. The dataset used contains NFL statistics from 2008 to 2022 [1]. The data was pre-processed by dropping NA values, converting units, and standardizing the data to ensure a balanced analysis. This process will be discussed in more detail later in the report.

a) Clustering Implementation

K-means clustering was applied to the NFL player combined dataset using scikit-learn once the data was pre-processed. K-means is a clustering algorithm that splits the dataset into k non-overlapping clusters, where each cluster is represented by a centroid. Each data point, representing an athlete, is assigned to the cluster whose centroid is closest. The K-means algorithm follows theses iterative steps:

1. Randomly initialize K centroids.
2. Assign each data point to the nearest centroid.
3. Recalculate the centroid of each cluster based on the assigned data points.
4. Repeat steps 2 through 3 until the centroids stabilize or minimal changes occur.

To enhance clustering consistency and improve the accuracy of the results, careful attention was given to the initialization of the centroids. To determine the optimal number of clusters *(k),* which balances intra-cluster compactness and inter-cluster separation, we used clustering methods such as the Silhouette Score and the Elbow Method. We computed clustering results for k values ranging from 2 to 5. Final clustering was used when $k = 2$, as this was found to be the optimal number of clusters according to both methods. However, alternative values were also explored for comparative analysis.

b) Cluster Evaluation and Interpretation

For dimensionality reduction, we used Principal Component Analysis (PCA) to reduce the data to two dimensions. With this, scatter plots of the data illustrated clear separation between clusters. To further visualize the clustering outputs for k values from 2 to 5, we used heat maps to illustrate the players' positions against the clusters, as well as

radar plots to display the average performance across different attributes. This methodology allowed us to cluster the NFL combined data and visually interpret the physical and performance features corresponding to the data-driven archetypes.

C. Data Preprocessing

We read our data from a Kaggle dataset containing NFL Combine statistics from 2000 to 2022 [1]. The original dataset was made up of twenty-three separate CSV files, each representing one year of Combine data. After combining all of the files, we ended up with a single DataFrame of size (3895, 13). Since some of the rows had incomplete information, we removed any row that had a "na" value. Additionally, certain columns were excluded because they wouldn't contribute meaningful insight to the analysis. Specifically, "Player," "School," and "Year" were dropped. "Pos" was also excluded since the purpose of clustering was to explore whether player position could be revealed through the data itself.

Next, we preprocessed the data to make it usable for modeling. The only feature that required type conversion was "Ht," which was originally formatted as "feet-inches." We split this into its components and converted it into a single integer representing the player's height in inches. Lastly, we scaled all numerical values using scikit-learn's StandardScalar class to ensure that features with larger numerical ranges didn't overpower the rest during modeling. At this point, the data was cleaned and ready for clustering and further analysis.

IV. Model Training Insights

When performing modeling and analysis on the DataFrame, K-means clustering was the primary method used. We experimented with different values of *k*, starting with two clusters and increasing up to five, in order to evaluate which *k*-value would produce the most meaningful and accurate grouping.

A. Visualizations of Clustering with Different Values of k

a.) PCA Clusters

The PCA cluster visualization is a two-dimensional plot that reduces the dimensionality of the data while preserving as much of the original variance as possible. Each point on the plot represents a row from our dataset, positioned based on a combination of the original attributes as determined by the PCA transformation. The advantage of using PCA before clustering is that it helps remove noise and irrelevant variance, which can improve the clarity of the clusters. It also speeds up the clustering process by reducing the number of dimensions the model needs to work with.

b.) Heatmap of the Player's Position against the Clusters

The heatmap visualization shows how each football player's position was distributed across the clusters. Since the position attribute was intentionally excluded from the clustering analysis, the heatmap helps reveal whether the resulting clusters align with any real-world groupings based on position. This can offer insight into whether clustering provides meaningful structure in the data. To account for the fact that some positions had significantly more entries than others, sometimes over four times as many, we recorded the percentage of each position within each cluster instead of raw counts.

c.) Radar Plot

The radar plot illustrates the average performance across eight key combine attributes—40-yard dash, vertical jump, bench press, broad jump, 3-cone drill, shuttle run, height (in inches), and weight—for each cluster. The unique shape of each cluster's polygon reflects differing athletic profiles, such as speed-focused athletes with high agility and low weight versus strength-oriented players with greater size and upper-body performance. This visualization helps reveal how the clustering may correlate with different football positions, even though position data was not used in training.

B. Model Evaluations

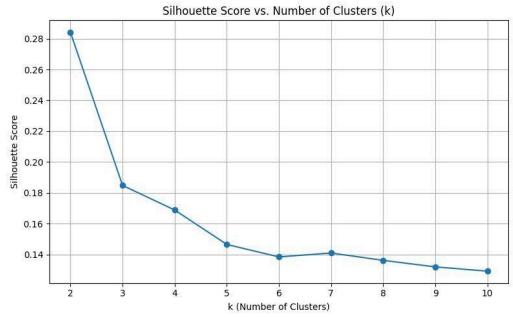
When performing k-means clustering, we used two main methods to determine which value of *k* would result in the most effective clustering. The first method was the silhouette score, which measures how similar each point is to its own cluster compared to other clusters. Higher silhouette scores indicate well-defined and separated clusters. The second method was the elbow curve, which plots the within-cluster sum of squares (WCSS) for various values of *k*. The "elbow" point in the curve, where the rate of improvement sharply decreases, can suggest an optimal number of clusters. By combining insights from both methods, we were able to select a *k* that balanced accuracy and interpretability for our analysis.

After performing K-Means clustering with values of *k* ranging from two to five, both evaluation methods—the silhouette score and the elbow curve—strongly suggested that *k = 2* provided the best clustering performance. Specifically, the silhouette score reached its highest value at *k = 2*, with a score of 0.2842 (Table I.), indicating that the data points were well matched to their assigned clusters and distinct from other clusters. Similarly, the elbow plot showed a noticeable bend at *k = 2*, where the rate of improvement in reducing the within-cluster sum of squares began to level off. This convergence of evidence from both methods supports the idea that dividing the data into two clusters results in the most compact, well-separated groupings, making it the most effective choice for our analysis.

TABLE I.

Silhouette Scores	
K Value	***Score***
2	***0.2842***
3	***0.1848***
4	***0.1688***
5	***0.1465***

FIGURE I.



V. Key Findings

A. Clustering was most efficient at k = 2

After evaluating several clustering models, both the silhouette score and the elbow method confirmed that k = 2 provided the most well-defined clusters. This meant that players could be most naturally divided into two distinct athletic groups based on their combine results. The silhouette score peaked at 0.2842 for k = 2, and the inertia curve showed a noticeable bend at this point, confirming that this value of k offered the clearest and most compact separation of data for clustering.

B. Clustering revealed distinct athletic archetypes

The clusters clearly represented different types of athletes. One cluster was composed of faster, more agile players who scored well in the 40-yard dash, shuttle, and 3-cone drill typically associated with positions like WR, CB, and S. The other cluster included stronger, more physically dominant players who excelled in bench press, height, and

weight aligning more with positions like TE, LB, and QB. These differences illustrate that clustering effectively surfaced performance-based archetypes from the data.

C. Position distribution aligned closely with cluster traits

Although position data was not used during clustering, a post-analysis heatmap showed that player positions aligned strongly with their assigned clusters. This suggests that the combine drills alone without knowing a player's listed position are sufficient to reveal meaningful groupings that correspond to real-world roles on the football field.

D. Visual tools such as PCA and radar plots enhanced interpretation

PCA projections allowed us to reduce eight-dimensional combine data to two dimensions, making it easy to visually confirm cluster separation. Radar charts helped us interpret each cluster's average performance across all metrics, highlighting key differences in athletic profiles. These visualizations made the clustering results more interpretable and actionable.

FIGURE II.

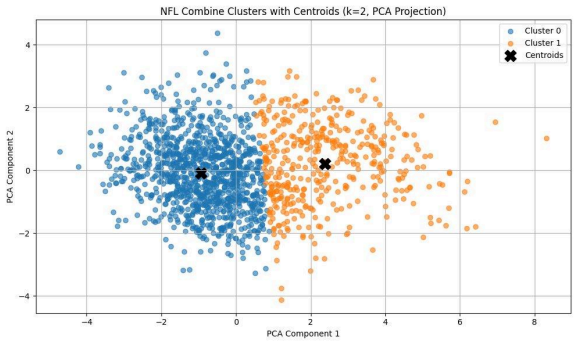


Fig. 2. PCA Plot of k=2

FIGURE III.

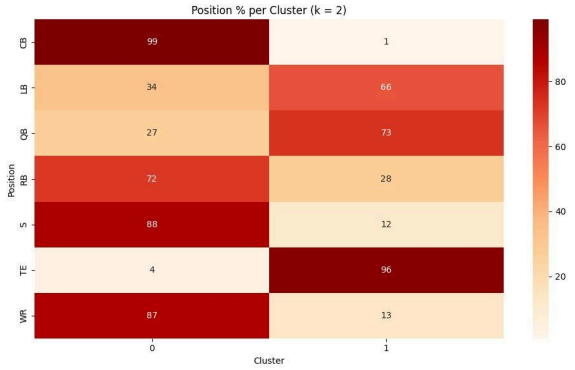


Fig. 3. Heatmap of k=2

FIGURE IV.

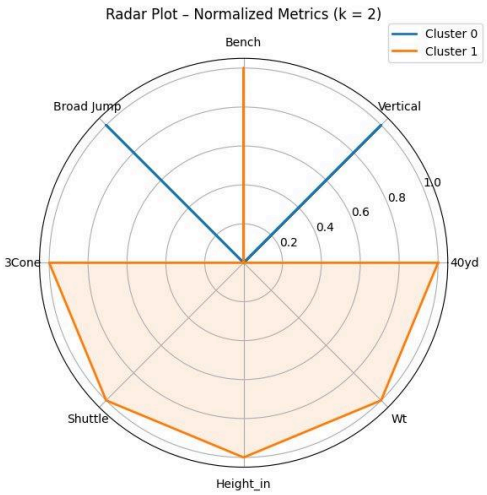


Fig. 4. Radar Map of k=2

[1] Weg, Mitchell. "NFL Combine Results Dataset 2000-2022." *Kaggle*, 10 Apr. 2023, www.kaggle.com/datasets/mitchellweg1/nfl-combine-results-dataset-2000-2022.