P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 7 Issue 1| April 2025
10.30598/variancevol7iss1page113-122

CLUSTERING AND VISUALIZATION OF OLYMPIC ATHLETE DATA BASED ON PHYSICAL AND DISCIPLINARY ATTRIBUTES

Hilwin Nisa[1*], Abela Chairunissa[2]
[1,2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University
MIPA Center Building, Veteran Street, Lowokwaru, Malang, 65145, Indonesia

Corresponding Author's Email: *nisahilwin@ub.ac.id*

Abstract: *This study aims to identify hidden patterns in international athlete data through clustering and data visualization approaches. The goal is to group athletes based on physical characteristics and sports disciplines to uncover meaningful trends. Utilizing a dataset of over 200,000 entries from 1896 to 2016, the study applies K-Means, Agglomerative and DBSCAN clustering methods. Preprocessing steps include handling missing data, selecting relevant variables (Height, Weight, Age, Sex, Sport, and Medal), and data normalization. The Silhouette score for K-Means (0.273647136516163645), Agglomerative (0.26134664130023655), and DBSCAN (-0.23920792207945957) indicates suboptimal clustering with overlapping clusters. K-Means clustering performs slightly better among the three methods. The findings are visualized through cluster plots and an interactive map showing medal distribution. This study highlights the limitations of traditional clustering methods for large datasets and suggests future exploration with advanced techniques.*

Keywords: *Agglomerative, Cluster Analysis, DBSCAN, K-Means, Sport Analytics*

1. INTRODUCTION

Clustering, as one of the unsupervised learning techniques, helps uncover hidden patterns and trends within complex datasets. It is a data analysis method that seeks to organize data into multiple clusters, grouping similar objects within the same cluster [1]. Maximizing homogeneity within a cluster and heterogeneity between distinct clusters is the primary goal of clustering [2]. Similarities and differences in clustering are usually based on distance calculations or the attribute values of objects [3]. In sports analytics, clustering helps to understand athlete performance and identify patterns in physical characteristics, so that it can optimize training strategies. This research applies clustering methods to analyze a comprehensive dataset of Olympic athletes, spanning over 120 years of history (1896-2016), focusing on their physical characteristics and performance in various disciplines.

A previous study employing a clustering method in sports data was conducted by [4]. This research applies the k-means clustering method for taekwondo selection in Porprov. This study's dataset was relatively small, consisting of only around 216 athletes.

Other research that has used clustering algorithms such as k-means, agglomerative clustering, and DBSCAN in different fields, for example, [5] examined the performance of DBSCAN and k-means for clustering Micro, Small, and Medium Enterprises (MMSEs) based on asset value and turnover, highlighting the strengths and weaknesses of each algorithm in handling different data structures. Similarly, [6] conducted a comparative analysis of k-means, DBSCAN, and hierarchical clustering for market segmentation, demonstrating variations in clustering outcomes depending on the method used. Additionally, [7] applied k-means and DBSCAN cluster techniques for credit and customer segmentation, focusing on expenditure levels and emphasizing the district characteristics captured by each algorithm.

However, these previous studies compared clustering algorithms across industries with relatively small samples. This paper seeks to contribute to the growing body of knowledge by applying these clustering methods to a large-scale dataset comprising over 200,000 rows of athlete data from the Olympic Games. Of course, this offers the challenge of a more robust analysis than previous studies.

https://ojs3.unpatti.ac.id/index.php/variance/ Submitted: 2025-01-14 Accepted: 2025-06-03

Statistically, this research evaluates clustering algorithm behavior on high-dimensional, historical, and large-scale data. This study uses three commonly applied algorithms (k-means, agglomerative, and DBSCAN) to assess their scalability and internal validity through Silhouette Score analysis. The comparison highlights how clustering performance can be affected by data volume, dimensionality, and cluster structure, which are critical concerns in statistical learning but rarely explored in sports datasets.

In addition, the study also applies clustering to identify meaningful patterns in the Olympic athlete dataset. Using these algorithms, we aim to group athletes based on shared characteristics such as age, height, weight, and medal achievements. Furthermore, we explore the potential of data visualization to interpret the clustering results, ensuring that the findings are understandable and actionable.

While clustering has proven valuable in various fields, using large, historical datasets in sports analytics remains underexplored. In particular, this paper focuses on evaluating the clustering techniques' ability to uncover trends that can be used to inform decision-making in sports management, coaching, and athlete development. We also assess the performance of each algorithm using standard evaluation metrics such as the silhouette score to ensure the robustness and interpretability of the clusters.

By leveraging this large-scale dataset and advanced clustering techniques, this study aims to contribute new insights into Olympic athlete performance and provide a foundation for future research in sports analytics. Furthermore, the findings may have practical implications for optimizing athlete selection, training regimens, and performance predictions, offering a real-world application of data science in sports.

2. METHODOLOGY

2.1. Data Collection

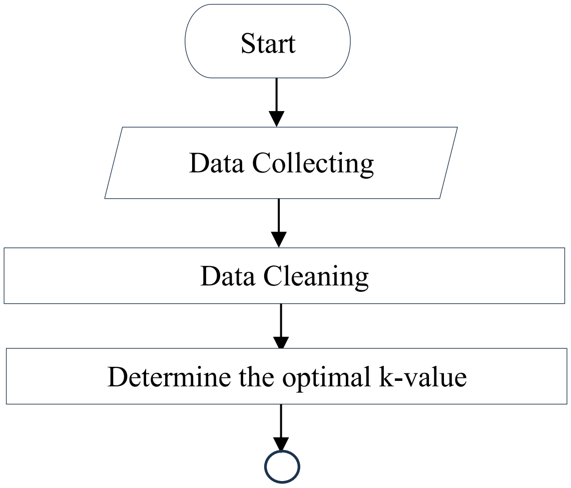
The data used in this research is a dataset of Olympic athletes for 120 years. This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. This data is secondary data, obtained from [8].

2.2. Variable Definition

The variables used in this research are determined by the research objectives, namely the physical characteristics and sports fields of athletes. Then, the achievements are based on the medals obtained. Thus, the variables used are Height, Weight, Age, Sport, Sex, and Medal.

2.3. Analytical Approach

The methods used in this research are k-means, agglomerative, and DBSCAN clustering. The steps for the analysis using Jupyter are as follows:



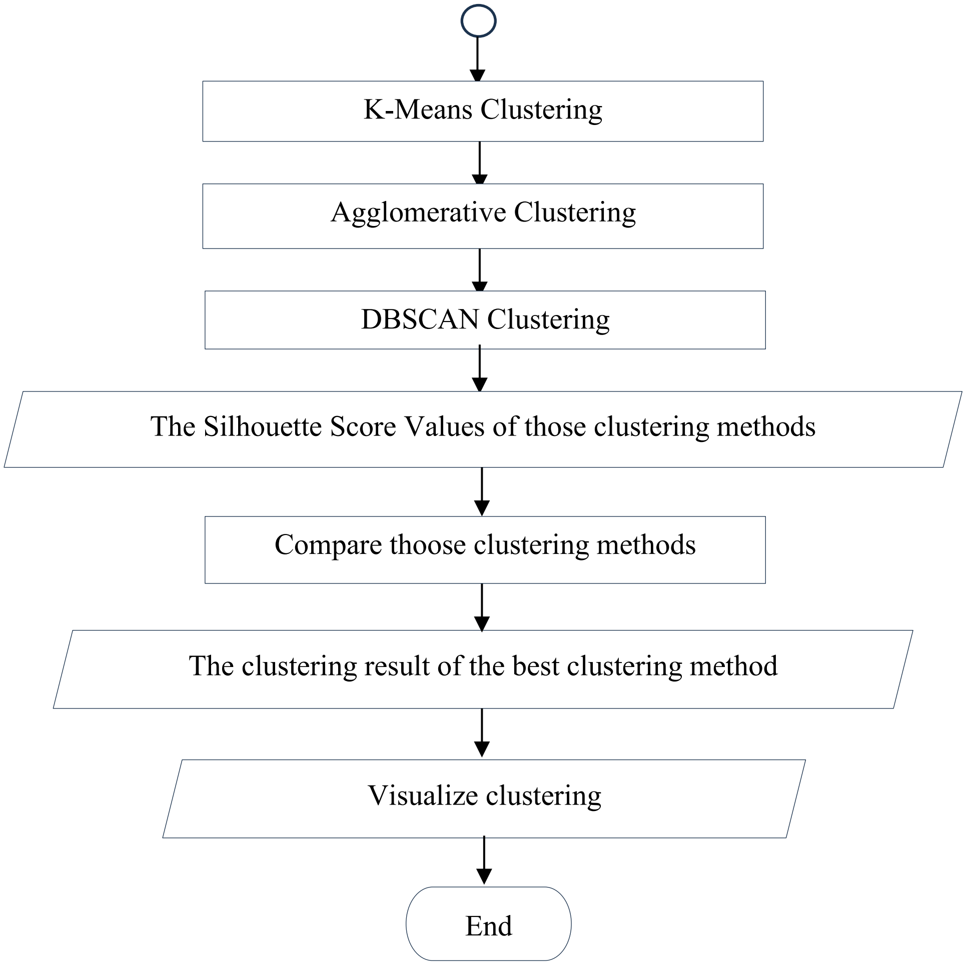

K-Means Clustering
Agglomerative Clustering
DBSCAN Clustering
The Silhouette Score Values of those clustering methods
Compare thoose clustering methods
The clustering result of the best clustering method
Visualize clustering
End

Figure 1. Research Flow Diagram
Source: Researcher

1) Clustering

Clustering or cluster analysis is an unsupervised classification mechanism in which a set of data is classified into clusters or groups, such that members of one cluster are similar to each other and different from other clusters [9].

Cluster analysis can be divided into two main groups based on their output structure, namely, hierarchical and non-hierarchical (partition) grouping methods. Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The clusters are combined (agglomeration method) or broken up (fragmentation method) step by step based on similarity measures. The result of the hierarchical clustering method requires agglomerative, and the division method can be displayed graphically using a tree diagram known as a dendrogram. The dendogram shows all the steps in a hierarchical procedure that includes the similarities or distances by which clusters are combined. As for the partition clustering method, it partitions a collection of data objects into clusters where each pair of object clusters is distinct or has several members in common. Partition clustering starts with an initial cluster partition that is iteratively scaled up, and locally optimal partitioning is achieved [9].

2) K-Means Clustering

K-means is a simple method for clustering analysis that aims to determine the best way to divide the n^{th} entity into groups called clusters [10]. This method starts by selecting the number of clusters. After the value is determined, the next step is to determine the cluster center, and continue by calculating the distance of each object to each cluster center. The distance between each existing data point and each centroid will be calculated using Euclidean analysis to determine the shortest distance between each data point and the centroid. Next, the objects are grouped based on the minimum distance to the cluster center. The cluster center is then temporarily used as the cluster center, or centroid, mean. If objects still need to be moved to another cluster, then the process is repeated, but if not, then the process is complete [11].

Each cluster formed will improve partition criteria, such as a distance-based dissimilarity function, so that objects within a cluster become similar, and objects in different clusters are found to be dissimilar in terms of dataset attributes. Euclidean distance is used as a distance measure in the k-means approach to highlight the similarities between each cluster with the smallest distance and highest similarity. The Euclidean distance between point $a = (a_1, a_2, \dots, a_k)$ and point $b = (b_1, b_2, \dots, b_n)$ can be calculated using formula [12]:

$$d(b_i, a_t) = \sqrt{\sum_{j=1}^{l} \left(b_{ij} - a_{tj}\right)^2} \tag{1}$$

where:

d : distance between the data value and the cluster center value

b_i : data value, $i = 1,2, \dots, n$, where $n =$ amount of data

a_t : cluster center value, $t = 1,2, \dots, k$, where $k =$ number of clusters

l : number of attributes or dimensions.

3) Agglomerative Clustering

The agglomerative hierarchical cluster method is often applied because the cluster process has a scientific nature. The following are the steps of the agglomerative algorithm [13]:

a) Determine the k clusters that will be formed, where k is the number of objects under study that have a distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - y_{jk}\right)^2} \tag{2}$$

where:

d_{ij} : distance between the object i and the object j

p : number of factors from the cluster

x_{ij} : data from subject i on variable k

y_{ik} : data from subject j on variable k.

b) Look for the distance matrix of the closest cluster pair, then determine the distance matrix.

c) Combining any clusters that are shown to be very close.

d) Then repeat stages 2 and 3 until all objects can form a cluster.

4) DBSCAN Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering technique that expands regions of significant density into clusters, allowing identification of clusters with arbitrary shapes in a spatial database that may include noise [14]. DBSCAN defines a cluster as a connected collection of maximum density points. Any object that does not belong to any cluster is considered noise [15].

It is a density-based clustering algorithm because it finds several clusters starting from an estimate of the density distribution of the associated nodes. This algorithm is based on connecting points within a certain distance threshold, similar to linkage-based clustering. However, it only connects points that meet the density criterion (minimum number of objects within the radius). A cluster of arbitrary shape is formed, consisting of all objects connected by density. DBSCAN separates data points into three classes:

a) Hub point: a point located in the inner part of the cluster (center).

b) Edge point: located in the neighborhood of a hub point that is not a hub point.

c) Noise point: any point that is neither a hub point nor an edge point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in the data set (D) and retrieves all instances of D with respect to epsilon (Eps) and minimum points (minPts). minPoints, defined as the minimum number of points required to exist neighborhood to be declared a cluster, and Eps defined as the neighborhood radius of a point based on a distance metric (Euclidean, Manhattan or Minkowski). This algorithm uses a spatial data structure to find points within Eps distance from the cluster core point [14].

3. RESULTS AND DISCUSSION

3.1. Data Cleaning

The data Used in this research is the Olympic athlete dataset for 120 years. This data consists of more than 200,000 athlete data with 15 variables, namely ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal. Following are these variables displayed using the Jupyter software.

Table 1. Variables in The Dataset

Variable	Number of Missing Data
ID	0
Name	0
Sex	0
Age	9474
Height	60171
Weight	62875
Team	0
NOC	0
Games	0
Year	0
Season	0
City	0
Sport	0
Event	0
Medal	231333

Based on the image above, it can be seen that several variables have missing data. Apart from that, we will not use all the variables in this research. We only use a few required variables as mentioned in Section 2.2. Therefore, we need to carry out data cleaning with several processes carried out as follows:

1) Resolve missing data. At this stage, missing data is manipulated. Specifically for the Medal variable, data that does not have a value is considered to have no Medal.

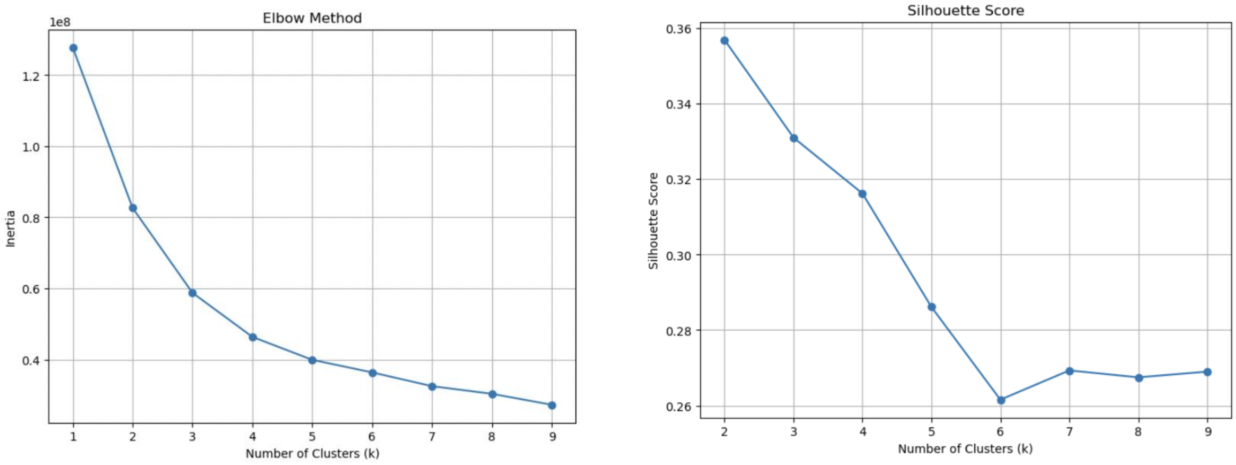
2) Select the required variables: Height, Weight, Age, Sport, Sex, and Medal.

3) Data transformation is changing the format, data type, or data structure to make it more consistent and ready for analysis.

4) Data normalization.

3.2. The Optimal k-Value

To determine the optimal k value here we use the Elbow method and the Silhouette score. Here are the results:



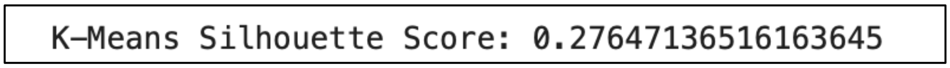
(a) **(b)**

Figure 2. K-Optimum Results, where (a) Uses Elbow Method and (b) Uses Silhouette Score

Based on the results of the image above, it can be seen that k −optimum is 2. Thus the number of cluster that will be used is 2.

3.3. K-Means Clustering

Using the Jupyter software, the Silhouette Score for K-Means is as follows:

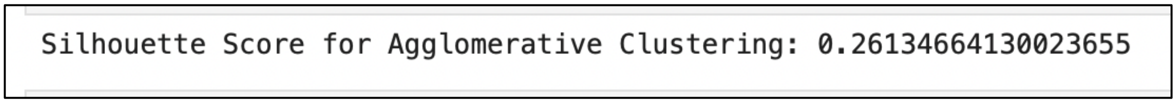


```
K-Means Silhouette Score: 0.27647136516163645
```

Figure 3. The Silhouette Score for K-Means

3.4. Agglomerative Clustering

Using the Jupyter software, the Silhouette Score for Agglomerative clustering is as follows:

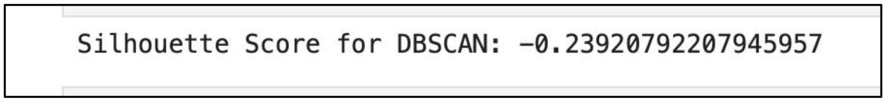


```
Silhouette Score for Agglomerative Clustering: 0.26134664130023655
```

Figure 4. The Silhouette Score for Agglomerative Clustering

3.5. DBSCAN Clustering

Using the Jupyter software, the Silhouette Score for DBSCAN is as follows:



```
Silhouette Score for DBSCAN: -0.23920792207945957
```

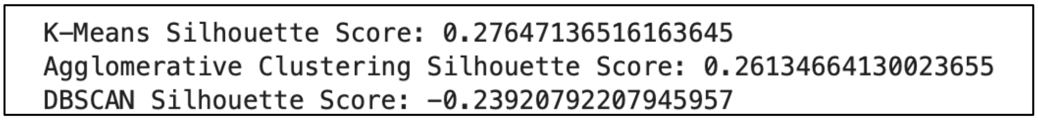
Figure 5. The Silhouette Score for DBSCAN Clustering

3.6. Comparison of Those Clustering Methods

Silhouette score is a metric used to measure the quality of clustering results. Silhouette score values range from -1 to 1, have the following interpretation:

1) Value close to 1: a very good cluster, because objects are closer to their cluster than to other clusters.
2) Value close to 0: clusters are unclear or overlap. Objects are on the border between two clusters.
3) Negative value (< 0): an object may be incorrectly clustered, being closer to other clusters than to its cluster.

The following are the Silhouette scores from the three clustering methods.



```
K-Means Silhouette Score: 0.27647136516163645
Agglomerative Clustering Silhouette Score: 0.26134664130023655
DBSCAN Silhouette Score: -0.23920792207945957
```

Figure 6. Comparison The Silhouette Score of The Clustering Methods

Based on Figure 7, k-means produced the highest Silhouette score among the three methods, although only slightly better than agglomerative. The Silhouette scores also show that the clustering results are suboptimal. There is an overlap between the clusters, and most data points are not very close to the center of their cluster. This score indicates that many data points overlap or are on the border between the two clusters.

It is necessary to re-evaluate the parameters or algorithm. It may be necessary to re-evaluate the number of clusters (k). The k value chosen may be too small. It could also be that k-means, agglomerative, and DBSCAN are not optimal algorithms for large datasets like this.

These findings suggest that traditional clustering algorithms, such as k-means, agglomerative, and DBSCAN, may face limitations when applied to high-dimensional and large-scale datasets like this. The relatively low Silhouette scores, particularly for DBSCAN (with a negative value), indicate that many data points do not fit well into the defines clusters, potentially due to high overlap, unbalanced cluster density, or insufficient feature differentiation.

For future improvements, the use of dimensionality reduction techniques such as Principal Component Analysis (PCA) prior to clustering may help improve separation. Additionally, advanced clustering models such as Gaussian Mixture Models (GMM), Spectral Clustering, or hybrid ensamble clustering could be explored to handle complex data structures more effectively.

3.7. Visualizing Clustering

The following is an example of visualization of the k-means cluster (as the best cluster among the three clustering methods).

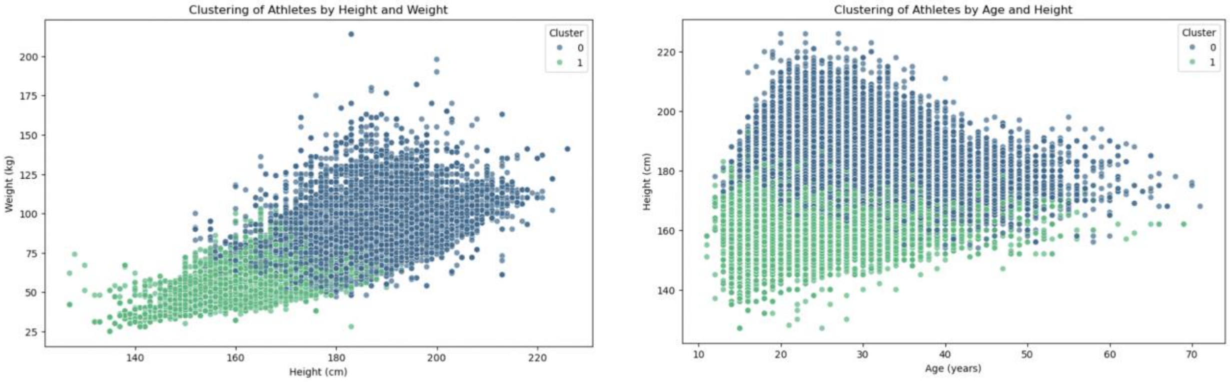
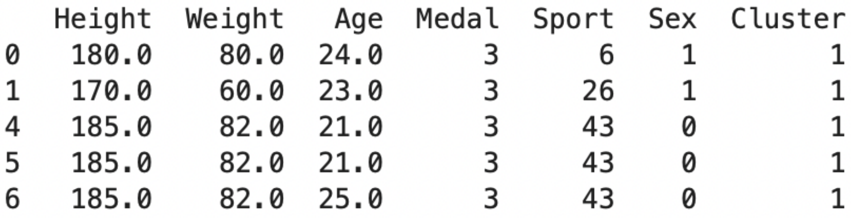


Figure 7. Clustering Visualization

Based on the figure, it can be clearly seen that there is indeed overlap between the clusters, and most of the data points are not very close to the center of their own cluster.

As for the cluster members, an illustration is as follows:



```
   Height  Weight   Age  Medal  Sport  Sex  Cluster
0   180.0    80.0  24.0      3      6    1        1
1   170.0    60.0  23.0      3     26    1        1
4   185.0    82.0  21.0      3     43    0        1
5   185.0    82.0  21.0      3     43    0        1
6   185.0    82.0  25.0      3     43    0        1
```

Figure 8. The Members of Clusters

Below is an interactive map to see the distribution of athlete data.

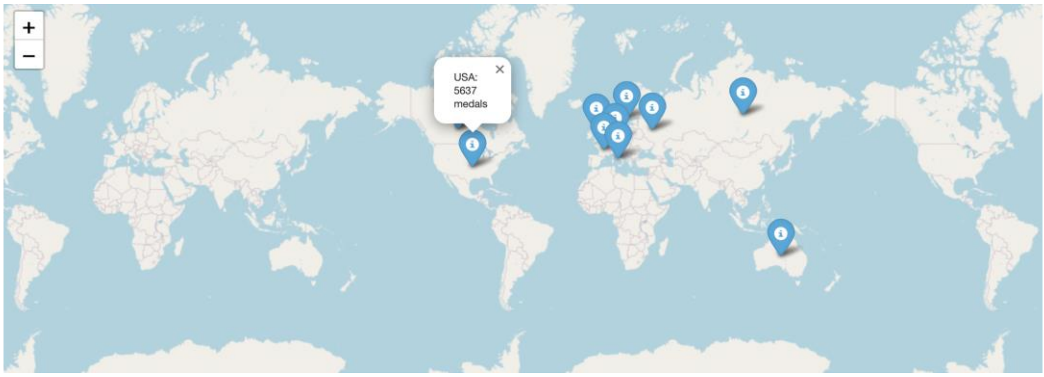


Figure 9. Interactive Map

Based on the map above, it can be seen that the USA won the most medals at 5637.

4. CONCLUSION

From this research, it can be concluded that the best clustering method for this large Olympic athlete dataset is k-means clustering. However, this method is also not optimal in forming clusters, because there is still significant overlap in the data, and the Silhouette score is still relatively low at 0.2764713651616345. For further research, clustering analysis on large datasets can be carried out using other methods to determine optimal clustering results.

In future research, exploring advanced clustering techniques such as Spectral Clustering, Gaussian Mixture Model (GMM), or ensemble-based approaches is recommended, which may provide better performance on large and complex datasets. In addition, incorporating dimensionality reduction methods such as Principal Component Analysis (PCA) before clustering may improve cluster separation. Researchers may also consider including more diverse variables or temporal dimensions to enhance the interpretability and relevance of clustering results in sport analysis.

ACKNOWLEDGMENT

The researcher expressed the gratitude for the help and support from many parties, especially to family and colleagues.

REFERENCES

[1] W. Wahyuri, U. Athiyah, I. Puspitasari and Y. Nita, "Clustering of Drug Sampling Data to Determine Drug Distribution Pattern with K-Means Method: Study on Central Kalimantan Province, Indonesia," *J. Inf. Syst. Eng. Bus. Intell.,* vol. 5, no. 2, pp. 208-218, Oct. 2019.

[2] H. Cui, W. Wu, Z. Zhang, F. Han and Z. Liu, "Clustering and Application of Grain Temperature Statistical Parameters Based on The DBSCAN Algorithm," *J. Stored Prod. Res.,* vol. 93, pp. 1-9, Sep. 2021.

[3] S. Nurdiani, S. Linawati, R. A. Safitri and P. Sari, "Pengelompokan Perilaku Mahasiswa pada Perkuliahan E-Learning dengan K-Means Clustering," *J. Kajian Ilmiah,* vol. 19, no. 2, pp. 126-133, May 2019.

[4] M. T. A. Gultom and R. A. Putri, "Penerapan Metode K-Means Clustering untuk Seleksi Taekwondo Porprov," *Kajian Ilmiah Inform. dan Komput.,* vol. 4, no. 3, pp. 1539-1550, Dec. 2023.

[5] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi and I. W. A. S. Darma, "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping Based on Aset Value and Turnover," *J. Inf. Syst. Eng. Bus. Intell.,* vol. 10, no. 1, pp. 13-24, Feb. 2024.

[6] S. D. K. Wardani, A. S. Ariyanto, M. Umroh and D. Rolliawati, "Perbandingan Hasil Metode Clustering K-Means, DBSCANNER & Hierarchical untuk Segmentasi Pasar," *J. Inform. dan Komput.,* vol. 7, no. 2, pp. 191-201, Sep. 2023.

[7] H. Ramadhan, M. R. A. Kamaludin, M. A. Nasrullah and D. Rolliawati, "Comparison of Hierarchical, K-Means and DBSCAN Clustering Methods for CreditCard Customer Segmentation Analysis Based on Expenditure Level," *J. Appl. Inform. Comput.,* vol. 7, no. 2, pp. 246-251, Dec. 2023.

[8] R. Griffin, "120 years of Olympic history: Athletes and results," *Kaggle*, May 2018. [Online]. Available: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results. [Accessed: 20 Sep. 2024].

[9] E. U. Oti, M. O. Olusola, F. C. Eze and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithm," *Int. J. Adv. Sci. Res. Eng.,* vol. 7, no. 8, pp. 64-68, Aug. 2021.

[10] E. A. Saputra and Y. Nataliani, "Analisis Pengelompokan Data Nilai SIswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means," *J. Inf. Sys. Inform.,* vol. 3, no. 3, pp. 424-439, Sep. 2021.

[11] Y. Darmi and A. Setiawan, "Penerapan Metode Clustering K-Means dalam Pengelompokan Penjualan Produk," *J. Media Infotama,* vol. 12, no. 2, pp. 148-157, Sep. 2016.

[12] F. Akhmatshin, P. Egarmin, M. Gerasimova, I. Petrova and S. Mikitchak, "Clustering of K-Means Based on Euclidean Distance Metric and Mahalanobis Metric," *EDP Sci.,* vol. 531, no. 03002, pp. 1-6, 2024.

[13] E. U. Oti and M. O. Olusola, "Overview of Agglomerative Hierarchical Clustering Methods," *Br. J. Comput., Netw. Inf. Technol.,* vol. 7, no. 2, pp. 14-23, 2024.

[14] S. Dang and P. H. Ahmad, "Performance Evaluation of Clustering Algorithm using Different Datasets Computer Science and Management Studies Performance Evaluation of Clustering Algorithm using Different Datasets," *J. Inf. Eng. Appl.,* vol. 5, no. 1, pp. 39-46, Jan. 2015.

[15] L. Mochurad, A. Sydor and O. Ratinsky, "A FAst Parallalized DBSCAN Algorithm Based on OpenMp for Detection of Criminals on Streaming Services," *Frontiers,* vol. 6, pp. 1-12, 2023.