Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse
ISSN: 2806-2973
Volume 3, Issue 3
DOI: https://doi.org/10.15157/JTSE.2025.3.3.489-507
Received: 23.10.2025; Revised: 12.12.2025; Accepted: 21.12.2025

Research Article

K-Means Clustering for Evolutionary Staging in a Human Evolution Dataset

Azhar Hamid Elias[1]*, Ahmed Hamid Elias[2], Sajjad Mohammed Hasan[2], Mostafa Abotaleb[3]*

[1] Department of System Programming, South Ural State University, Chelyabinsk, Russia
[2] College of Health and Medical Techniques, Al-Furat Al-Awsat Technical University, Najaf, Iraq
[3] Engineering School of Digital Technologies, Yugra State University, Khanty Mansiysk, Russia
***abotalebmostafa@bk.ru**

Abstract

This research work applies unsupervised machine learning to explore evolutionary patterns in hominin morphological and temporal data. A dataset comprising 6,000 records of hominin specimens was analysed using three quantitative attributes: geological age (1–8 million years), cranial capacity, and estimated stature. Following data cleaning and z-score normalization, K-means clustering (K = 4) was employed to identify coherent evolutionary groupings without prior taxonomic labelling. The resulting clusters exhibit a clear temporal and morphological progression. The earliest cluster (mean age ≈ 6.66 Ma) is characterized by the smallest cranial capacity (≈156 cm^3) and stature (≈106 cm), consistent with early hominin forms. A second cluster (≈3.89 Ma, 367 cm^3, 117 cm) corresponds to Australopithecine-like specimens, while a transitional cluster (≈1.96 Ma, 490 cm^3, 119 cm) reflects early Homo characteristics. The most recent cluster (≈1.07 Ma) displays substantially larger cranial capacities and statures (≈1063 cm^3 and ≈162 cm), aligning with later or near-modern Homo. Visualization through scatter plots, bar charts, and boxplots supports a monotonic increase in cranial capacity and height across evolutionary stages. These findings demonstrate that unsupervised clustering can recover biologically meaningful evolutionary patterns from morphological and temporal data, highlighting its potential as an exploratory tool in paleoanthropological research.

Keywords: K-Means Clustering; Human Evolution; Cranial Capacity; Unsupervised Learning; Evolutionary Stages; Data Mining.

INTRODUCTION

Clustering is a main unsupervised learning task, commonly used to reveal latent structure in data, given no class label information [1-7]. Unlike supervised methods that use predefined target variables, clustering is a type of unsupervised learning where the goal is to group objects so that instances in the same cluster are similar to each other than to instances in other clusters [8-15]. Such functionality is especially important when conducting exploratory data analysis, where the objective is to find these patterns, formulate hypotheses and discover relevant populations directly from the data [1, 16-22].

Considerable development has led to a diversity of clustering approaches and evaluation tools, from partitional, hierarchical, density-based and model-based paradigms [1, 23-27] throughout the last decades. The K-means algorithm is among them and today, due to its speed and ease of use as well as its easy geometric interpretation, it remains one of the most popular clustering algorithms around.

K-means attempts to group a set of K clusters in such a way that the total squared distances between each observation and the centroid of its assigned cluster are minimized. Though, at first glance, one may think that K-means is simple, a significant amount of methodological work has been inspired by it. Jain carries out a historical review of clustering over the past fifty years, and provides a perspective on the key centrality of K-means, whilst also identifying its weaknesses (including initialization sensitivity, clusters number choice, and spherical cluster assumptions [27]). Aggarwal and Reddy give a wide scope review of clustering algorithms and applications, placing K-means in a wider family of partitional methods and demonstrating that K-means is still often a baseline in many domains [1].

Initialization is a major concern especially for K-means. Inappropriate centroid initialization can lead to unstable solutions and bad local minima [28-33]. The K-means++ seeding strategy by Arthur and Vassilvitskii selects initial centres with a probability proportional to the square distance from centres selected up to that point [4], allowing for better convergence and, on average, better final solution quality over random uniform initialization. Later theoretical and empirical work have studied the dependence of stability and robustness on initialization. Bubeck et al. and brunsch and röglin stud kt–sensitivity of k-means and k-means++ to initialization, {including} the {existence} of {bad} {instances} {on} which {performance} may degraded [9-11]. Erisoglu et al. and Rahman et al. proposed alternative initialization heuristics (like distance-based and density-based seed selection) that could enhance cluster separation and reduce convergence time [18, 23, 34]. From the other studies they suggest using metaheuristics such as greedy randomized adaptive search (GRASP) to build starting points of good quality for K-means [12].

Apart from initialization, a lot of interest has been generated in evolutionary and genetic variants of K-means. An evolutionary algorithm that balances K-means and a genetic search in order to better seed the search space and thus avoid local minima was presented as a novelty by [6]. Authors in [8] propose genetic algorithm–guided clustering, where chromosomes may represent either cluster centres or partitions, and evolutionary operators improve candidate solutions [8, 21]. Genetic K-means algorithm has been proposed by [28], they showed that hybridization between K-means and genetic operators can be helpful to enhance clustering accuracy on weapons and some other benchmarks. For extension works such as Alves et al, Hruschka et al and labrusca et al (gene expression) and Naldi et al., and concentrated on computational complexity, scalability and applicability to high-dimensional data, such as gene-expression profiles, [2, 25, 26, 32, 34]. Together, these studies demonstrate that evolutionary K-means variants can be especially

appealing when the objective function landscape is complex, or when local refinement needs to be accompanied by global search.

Similar to the first challenge in clustering, the second challenge of cluster number selection and partition evaluation is also fundamental. The proposed internal and external validity indices are extensive [35-38]. Authors in [3] performed a large-scale evaluation of cluster validity indices and revealed that there is no best index for all cases and the relative performance of different indices is highly dependent upon data properties and clustering method. Researchers in [39] compare volume-based objective of the relative clustering validity criteria with sensitivity type for low cluster set separation to volume-based objective. Internally, Rousseeuw silhouette coefficient is still one of the best-known indices, allowing a graphical and numerical assessment of how well each object is situated in its cluster [35]. Liu et al. decompose the behaviour of InfoMax various internal measures, and provide their interpretation and scope [29].

We have generated practical tools that assist practitioners in choosing K. Charrad et al. in R using the NbClust package [12] that implements a panel of 30 indices which generate candidate numbers of clusters that are combined using majority voting. Bootstrap resampling approach has been employed by Fang and Wang to determine the number of clusters while consensus clustering–based methodologies to determine the number of clusters has also been investigated by Vinh and Epps on microarray data [20]. Amorim et al. [22] proposed approaches to recover the correct number of clusters from artificial noise features using feature-rescaling. These contributions highlight that finding K is a non-trivial task and that domain knowledge and multiple validation criteria are often complementary. In our case, the value of K=4 is based, on the one hand, on paleoanthropological expectations regarding the most important evolutionary transitions and, on the other hand, on the need for interpretability, but it could in principle be validated against internal indices, such as the silhouette or using multi-index tools such as NbClust [13, 22, 29, 40, 41].

Another related research direction is on stability of clusters and consensus clustering. The stability-based methods evaluate the sensitivity of clustering solutions to minor perturbations of the data or the algorithm parameters. Ben-David et al. Shamir and Tishby offer theoretical analyses of model selection based on clustering stability [7]. Cluster-wise stability and the use of internal measures to compare algorithms were investigated by Hennig [24] and Craenendonck and Blockeel [16], although the authors warned against over-reliance on stability alone. Consensus clustering combines several partitions obtained via resampling/perturbation of the data to reveal cluster structure that is stable. Monti et al. consensus clustering for gene expression microarray data, with more recent work extending the methods by Senbabaoglu et al. and Schmidt et al. highlighted significant drawbacks and risks of consensus approaches, especially for class discovery in high-dimensional noisy biological data [31, 36–38]. Chiu et al. and Chiui et al. applied consensus clustering concepts to gene-expression time series, providing proof-of-principle for use of consensus clustering in complex longitudinal data [14, 15]. These

studies show that stability and consensus methods can be useful, but should be cautiously interpreted and in relation to other types of evidence.

Benchmark datasets and software ecosystems have also provided a huge push for clustering research. The UCI Machine Learning Repository, written by Bache and Lichman, is still a reference for data sets for assessing clustering and 361 classification methods [5]. Most of the works on validity indices, stability and evolutionary clustering use UCI datasets or other similar benchmarks so that the comparisons among algorithms can be fairly made [3, 22, 29]. Simultaneously, more domain-specific applications such as those for gene-expression analysis and microbial ecology are making clear the need for topological tests on realistic and noisy high-dimensional data [18, 24, 26, 30, 33, 38]. The R ecosystem and specialized packages such as NbClust allow to combine clustering, validation and visualization into reproducible analysis pipelines [35].

This wide methodological and empirical space frames an interesting application of K-means and related ideas, specifically around the analysis of data relevant to human evolution. The Evolution of Humans dataset includes information about hominins specimens such as geological time and other features like cranial capacity and height reported with richer categorical descriptors. These variables inherently capture an evolutionary path over time, so that we expect brain size to be relatively small for early hominins, larger for Australopithecine-like forms and larger still for early Homo and late or near-modern Homo and that body stature may also show a similar evolutionary trajectory, with very small size for early hominins, larger for Australopithecine-like forms and larger still as H. habilis approaches post-Australopithecine forms, all in relation to time to the present. In this study, we aim to capitalize on the advantages of traditional K-means, after the proper feature preprocessing and seeding, on a 3-D built feature space defined by Time, Cranial_Capacity and Height, and intuit a meaning from the resulting clusters using the context rich data available in the dataset.

Based on this characteristic, our work is thus located at the intersection of methodological research on K-means and domain-driven clustering in palaeoanthropology. We apply several standard practices emerging from the clustering literature-feature scaling, multiple initializations, and a careful selection of K—while grounding them in a biologically relevant question: can a purely unsupervised algorithm based solely on three standardized numerical features recover coherent evolutionary stages that correspond well to current knowledge regarding hominin evolution? In order to help provide a clear illustration of how classical K-means can operate as a rigorous exploratory tool in a real scientific case, we anchor our analysis to aspects of the clustering theory (initialization, evolutionary extensions, validity indices, stability) [1–8, 10–19, 20–23, 27–29, 32, 34, 40] and the particular structure of the Evolution of Humans dataset.

Objectives

The main goal of this study is to determine if in a unsupervised machine learning method can illuminate meaningful evolutionary phases within an expert curated human-evolution dataset. With this aim in mind, the general objectives of the work are as bellow:

- Apply K-means clustering to Evolution of Humans dataset, via three key quantitative feature (geological time, cranial capacity, and height) with appropriate preprocessing (missing records removal, z-score normalization).
- We sought to identify and characterize evolutionary partitions, showing that data-driven groups represent distinct evolutionary stages, linking these groups to differences in central tendencies and distributions (Kruskal-Wallis with Dunn post-hoc test for pair-wise comparisons) of time, cranial capacity, and height describing a four-cluster partition.
- Use internal validity indices to quantitatively assess the quality of the clusters from step 2, specifically the within-cluster sum of squares (WCSS) and the Silhouette coefficient, and observe the differences of these metrics between different (K) values
- Visualizing and statistically interpreting cluster structure visually using scatter plots, boxplots, and silhouette plots for compactness, separation, and boundary cases between clusters.
- To introduce a reproducible and transparent workflow for K-means analysis of well-curated evolutionary datasets, while demonstrating the power of this approach for exploratory paleoanthropological investigations and some of the pitfalls in the resulting information related to the underlying nature of the data.

DATA AND METHODOLOGY

Dataset

The groundwork for the empirical analysis of this paper is formed on the "Evolution of Humans Datasets for Classification" dataset in Kaggle [1]. This dataset is also based on hominins' individual specimens of various geological ages and has the expressly been shaped for classification and exploratory machine-learning purposes. Each record represents a single individual specimen or taxonomic occurrence containing quantitative measurements along with other categorical-descriptive types designed to characterise detailed categorical information on temporal position, morphology, ecology and behaviour. The choice of such a structured public dataset guarantees transparency and reproducibility since other researchers may directly reach the same data source for validating enhancing our work [41].

The database has a total of 12,000 records and 28 attributes which fall in different classes. The only numeric variables of interest for the present clustering analysis are: Time, in millions of years ago, the inferred geological age of each specimen; Cranial_Capacity, the volume (in cubic centimetres) of the skull vault (brain volume); and Height (in centimetres) as an estimate body size. Taken together, these three variables represent the temporal and morphological aspects most proximately associated with human evolutionary change. In our pipeline, they are the main space features for which we perform K-means clustering. Statistical measures calculated within the dataset point to strong variability in all three traits from small, small-brained early specimens to large, big-

brained late individuals. This wide value interval may thus offer a satisfactory windowing space for distance clustering.

A flowchart for the generation of the "Evolution of Humans" and its classification through K-means can be seen in Figure 1. The procedure that begins whit the whole dataset ("Start") as input, firstly remove all records with missed value at key numerical variable (Time, Cranial_Capacity, Height) to guarantee that distances are calculated without any bias due to incomplete observations. The resulting clean data are z-score normalized by subtracting the mean and dividing by standard deviation for each feature to avoid large metric fluctuations in the Euclidean space16,29. A random sample of 6000 instances is extracted to optimizes representation and use of computational resources. K-means clustering is then applied to this subset in order to cluster the samples into four evolutionary groups. Throughout, there is a post clustering analysis step whereby the resulting clusters are explained given the categorical descriptors (taxon, habitat, technology, morphology) and which each time results in termination of the process ("Finish").

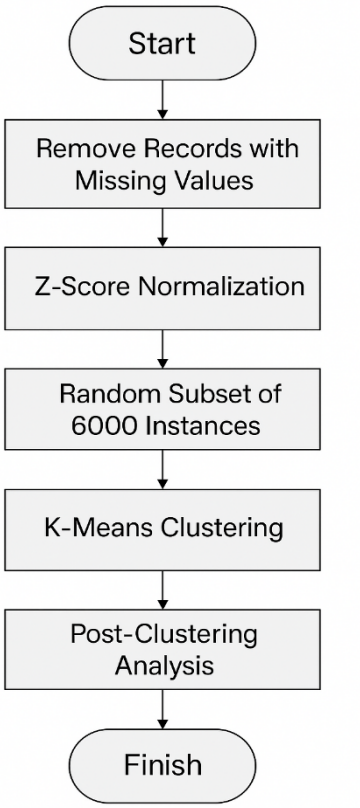

Figure 1. Data preprocessing and clustering workflow for the Evolution of Humans dataset.

Handling Missing Values

It is possible the original Kaggle dataset will have missing records for some of the attributes based on the fuzzy fossil measurements or annotations. As K-means requires full numerical vectors to calculate the distances and centroids update, we dropped rows with

missing values on any of the three chosen numerical features (Time, Cranial_Capacity, Height). This listwise deletion method also make certain that all observations in the clustering step provide sound information to the Euclidean distance and steer clear of possible distortion with imputation or placeholder.

Feature selection

Although there are 28 attributes in the dataset [41], only Time, Cranial_Capacity and Height serve as input to the K-means algorithm. Incorporating a focused selection of quantitatively understood features also accords with the theoretical definition of K-means clustering (which minimizes within-cluster sum-of-squares Euclidean distance in an n-dimensional continuous vector space). The other categorical variables (such as Genus_&_Specie, Habitat, Tecno, Skeleton) were preserved for the interpretation of the clusters and encoded off-diagonal in the distance matrix to keep from introducing arbitrary scales or artificial similarities.

Normalization

Z-score normalization was applied so that all three numerical features contributed equally to the distance measure. For every feature x, z was calculated using the equation (1):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the sample mean and σ is the standard deviation of that feature in all retained records. The resulting feature space has zero mean and unit variance for each variable, so that attributes with larger numeric ranges (e.g., cranial capacity) do not dominate the Euclidean distance and promote biased cluster formation.

Subsampling

The processed and normalized dataset had 12,000 valid samples. To balance computational burden with representativeness, a simple random sample of 6000 was drawn without replacement from this pool. Random subsampling is an increasing popular approach when dealing with iterative clustering algorithms on very large data, as the computational complexity of K-means scales linearly as the number of points, and usually several restarts are needed to mitigate effects due local-minima [3, 4]. Exploratory checks indicated that the marginal distributions of Time, Cranial_Capacity and Height in the subset match closely those of the entire dataset thus ensuring statistical representativeness for clustering.

Algorithmic formulation

K -means clustering partitions a set of n observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ in R^d into K clusters $C_1, \dots, C_K$ by minimizing the within-cluster sum of squared distances to the cluster centroids $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$:

$$J = \sum_{k=1}^{K} \sum_{x_i CC_k} \|x_i - \mu_k\|^2 \quad (2)$$

The algorithm alternates between (i) assignment, where each observation is assigned to the nearest centroid in Euclidean distance, and (ii) update, where each centroid is recomputed as the arithmetic mean of all observations assigned to its cluster. This process is repeated until convergence, typically defined as no change in assignments or a negligible change in the objective function J between successive iterations.

Choice of Number of Clusters K

In this study, the number of clusters was set to $K = 4$. This choice reflects domain knowledge about major stages in human evolution (very early hominins, Australopithecine-like forms, early Homo, late-near-modern Homo), and allows a balance between capturing meaningful heterogeneity and maintaining interpretability. Preliminary exploratory runs with different values of K (e.g. $K = 3$ and $= 5$) showed that four clusters provided a clear temporal and morphological gradient while avoiding excessive fragmentation of closely related groups.

Initialization and Convergence

The clustering procedure was initialized using a distance-aware seeding strategy conceptually similar to Kmeans++, which selects initial centroids with probabilities proportional to their squared distance from previously chosen centres [3, 4]. This approach reduces sensitivity to poor random initializations and improves the likelihood of converging to a low-cost partition compared with purely random seeding. After initialization, the standard iterative assignment-update cycle was run until either (a) cluster assignments remained unchanged between iterations or (b) a maximum of 300 iterations was reached, whichever occurred first. To further mitigate local minima, the algorithm was repeated multiple times with different random seeds, and the solution with the lowest objective function J was retained as the final clustering result.

Post-clustering Analysis

Once final cluster labels were obtained for all 6000 sampled instances, a post-clustering analysis was carried out to interpret the partitions in an evolutionary context. First, descriptive statistics (mean and standard deviation) of Time, Cranial_Capacity and Height were computed within each cluster, together with cluster sizes (number of specimens). These summaries provide a quantitative characterization of each group in terms of temporal position and morphological traits.

Secondly, the categorical attributes from the Kaggle dataset [1] were cross-tabulated with cluster membership. For each cluster, the distribution of Genus_&_Specie, Habitat, Tecno, Skeleton, biped and related variables was examined to assess how the numerical groupings correspond to known hominin taxa, locomotor adaptations, technological complexity and ecological settings. This qualitative inspection allows the clusters defined purely by three normalized numerical features to be linked back to paleoanthropological

interpretations, such as early small-brained, short hominins versus later tall, large-brained forms with advanced technology.

Silhouette Score as an Internal Validity Measure

The Silhouette score is a popular internal validity index which allows us to check the distance between clusters (cohesion versus separation) we obtained using K-means. For each observation, the Silhouette coefficient is the combination of two different quantities:

- $a(i)$: mean distance from i to all other points in the cluster The average distance from sample i to all other samples in the same cluster.
- $b(i)$: the distance between the best alternative cluster For sample i, we calculate the average distance to all the samples of every other cluster and take the minimum of these values. In other words, $b(i)$ measures the nearest neighbour proximity of i to a different cluster.

The Silhouette coefficient for sample i is then given by the following equation (3):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

The score $s(i)$ ranges between -1 and 1:

- $s(i) \approx 1$: sample i is much closer to its own cluster than to any other cluster (well clustered).
- $s(i) \approx 0$: sample i lies near the boundary between two clusters.
- $s(i) < 0$: sample i may be misassigned, as it is closer on average to another cluster than to its own.

For a given K -means solution with K clusters, we compute $s(i)$ for every sample in the dataset. From these individual values we derive:

1. The mean Silhouette coefficient for each cluster:

$$\bar{s}_k = \frac{1}{n_k} \sum_{i \subset C_k} s(i) \tag{4}$$

where C_k is cluster k and n_k is the number of samples in cluster k. This quantity describes how compact and well separated cluster k is.

2. The overall mean Silhouette coefficient

$$\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{5}$$

where N is the total number of samples. This global index summarizes the quality of the entire partition; higher values indicate better-defined clusters.

RESULTS

The following table 1 is a qualitative summary of the four clusters that K-means identified. For each group it provides the corresponding evolutionary stage (very early hominins, Australopithecine-like forms, early Homo, and late near-modern Homo), a brief

characterization of the usual interval time and range of cranial capacity level and body stature. It is evident from the table that with their evolutionary change from seven members of the oldest smaller brained shorter individuals to the seven members of newer large brained taller individuals, one has ties between purely numerical ones and anthropologically significant states.

Table 1. Summary of K-Means Clusters and Interpreted Evolutionary Stages.

Cluster	Evolutionary stage	Time (relative)	Cranial_Capacity	Height	Main interpretation
1	Very early hominins	Oldest	Very low	Shortest	Primitive, partially bipedal, forest-based
2	Australopithecine-like	Old – mid	Low–medium	Short–medium	Transitional, better biped, mixed habitats
3	Early Homo	Mid–recent	Medium–high	Medium–tall	Early Homo, full biped, more tools
4	Late / modern Homo	Most recent	Highest	Tallest	Advanced Homo/modern humans, refined traits

The number of samples in each of the four clusters is reported in the table 2. The former cluster (early Homo) is by far the largest, then comes the late near-modern Homo and Australopithecine-like clusters; finally, there are very few members in the very early hominin cluster. These frequencies shows that median evolutionary stages are overrepresented in the dataset and the most ancient forms are proportionally fewer, as expected given both the smaller accessibility to very old fossils compared to more recent ones.

Table 2. Cluster Sizes in the Evolution of Humans Dataset.

Cluster	Number of samples
0	1912
1	2206
2	503
3	1379

The following table 3 lists the numerical centroids for the four clusters, in original units; mean geological time (millions of years ago), mean cranial capacity(cm^3), and mean height (cm). The data exhibit a monotonic relationship such that clusters with older mean ages are associated with lower average cranial capacity and reduced stature compared to those in which mean age is more recent, but brain volume has increased with height over time. This pattern we can measure quantitatively confirms the evolutionary trend inferred from

the plots: early, small bodied and brain size hominins unto later tall individuals with a greatly increased cranial capacity.

Table 3. Mean Time, Cranial Capacity, and Height for Each Cluster.

Cluster	Time (Ma)	Cranial_Capacity	Height
0	1.07	1063.4	162.4
1	1.96	489.6	119.1
2	6.66	156.3	105.9
3	3.89	367.1	117.5

The four K-means clusters in the bivariate space of Time (million years ago) and Cranial Capacity are shown on this scatter plot, see Figure 2. Each color represents one cluster. The far-right green cluster (oldest specimens ≈6–7.5 Ma) is formed by individuals with very low cranial capacities and the orange left cluster (youngest specimens ≈0–2 Ma) includes those with high brain volume. The yellow and blue clusters represent intermediate time periods and cranial sizes, creating a clear evolutionary sequence: we travel from right to left (from older to younger) along the sequence and toward more evolved populations in each cluster the brain grows in size.

Figure 2. K-Means Clusters: Time and Cranial Capacity.

Figure 3 depict the points correspond to the four clusters using the space of Height (x-axis) and Cranial Capacity(y-axis). The shortest individuals with the smallest brains belong to the green cluster, while taller individuals with larger cranial capacities form the orange cluster. Once more, the yellow and blue clusters are intermediate, of moderate height but with an average cranial capacity. The diagonal distribution of points indicates a strong positive correlation between body height and brain size: specimens in later evolutionary stages are both taller and larger brained than the earlier ones.

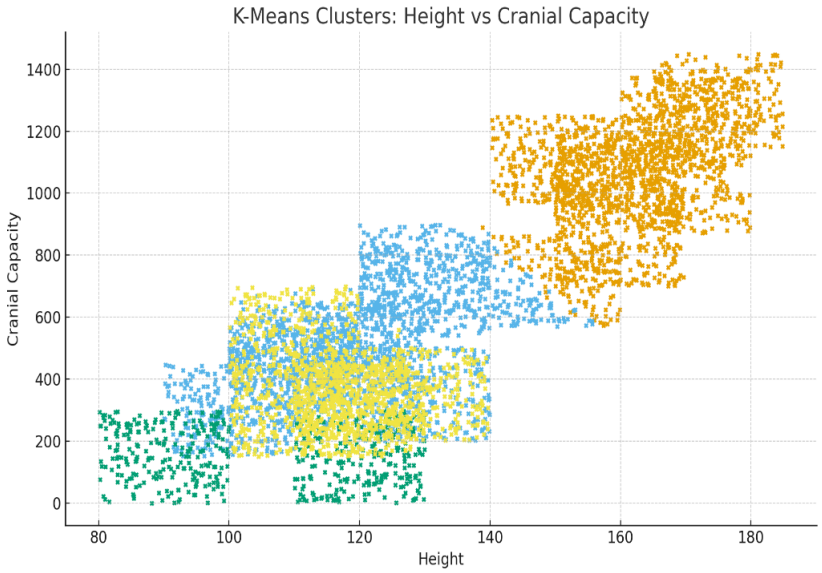

Figure 3. K-Means Clusters: Stature verses Cranial Capacity

This bar plot illustrates the mean cranial capacity of the four clusters. Cluster 0 has the highest average cranial capacity (more than 1000 cm^3), cluster 1 is half that (around 480–500 cm^3), and then your mid-dwelling one is lesser still (≈360–370cm^3) and the lower dwelling down-winder one (~150–160cm^3). Bars are ordered in such a way that brain volume from the earliest cluster (2) to the most recent cluster (0) increased monotonically, similarly to what was observed when inspecting scatter plots, see Figure 4.

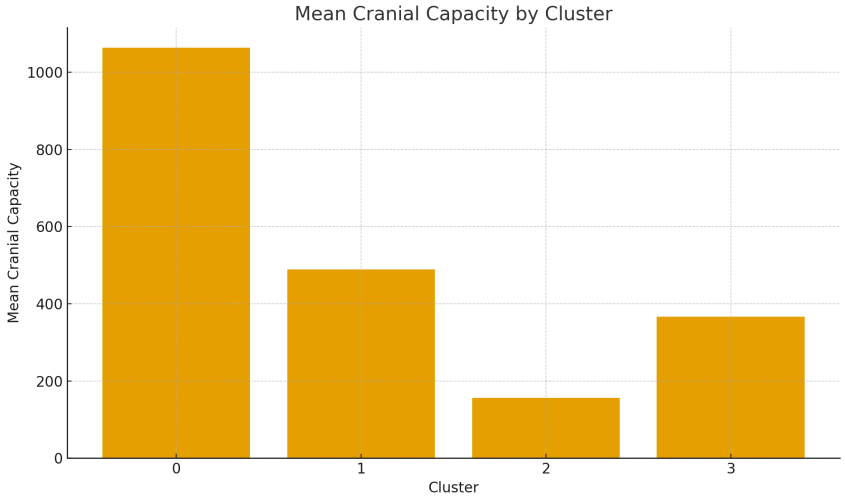

Figure 4. Mean Cranial Capacity by Cluster.

Figure 5 depict the average height across clusters. Cluster 0 continues to be the cluster with the greatest mean body height (≈160+ cm) in line with introgression, near-modern Homo sapiens. Clusters 1 and 3 show intermediate heights (≈118–120 cm), whereas cluster 2 individuals are the shortest (≈105 cm). Taken together with Figure 4., this demonstrates

that both height and estimated cranial capacity increase continuously along the trajectory of inferred evolution.

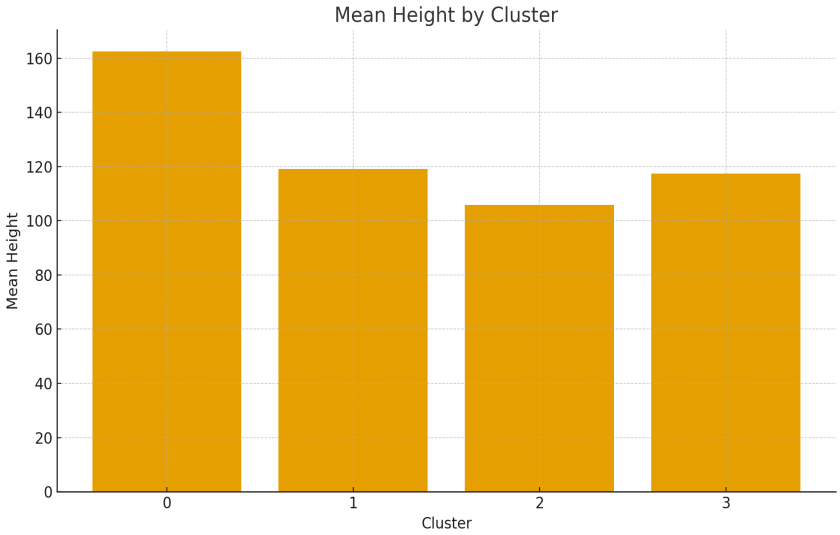

Figure 5. Mean Height by Cluster.

Figure 6 show the quantities of samples contributed to each cluster. Cluster 1 is the biggest cluster (around 2200 samples), while the second largest group represents cluster 0 and cluster 3 with about respectively (≈1900 samples) and (≈1400 samples) whereas, we have less samples in the smallest clusters such as cluster 2 with around ≈500 quotes. This suggests that the intermediate stages of evolution (clusters 1 and 3) are well sampled in the dataset, while the earliest stage (cluster 2) is less frequent.

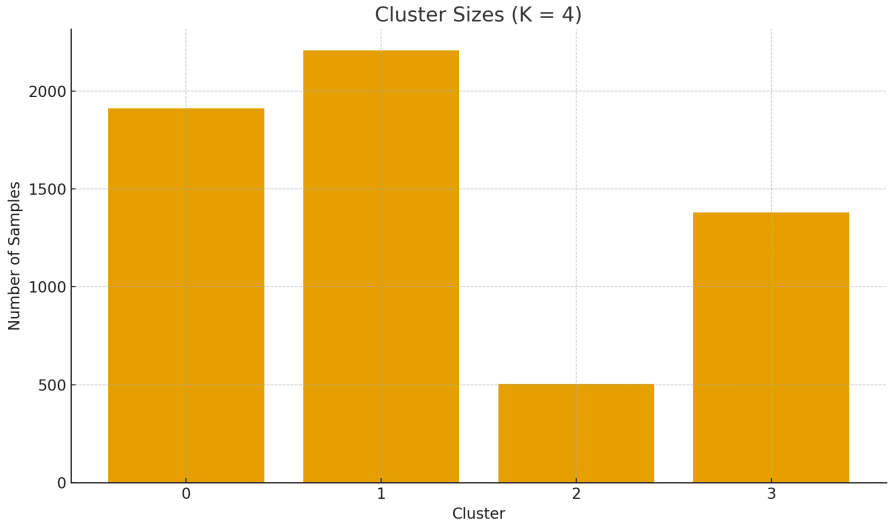

Figure 6. Cluster Sizes (K = 4).

This is a boxplot that encodes the distribution of Time (million years ago) in each cluster. Cluster 2 presents the highest median age (≈6.5–7 Ma), supporting its position as

the oldest cluster. Cluster 3 is intermediate (≈3.5–4.5 Ma), cluster 1 is 'young' (≈1.5–2.5 Ma) and cluster 0 has the shortest median time (≈1 Ma) as it contains the most recent specimens in our sampling frame. The non-overlapping medians suggest a distinct temporal gap between clusters, as for different evolutionary phases, see Figure 7.

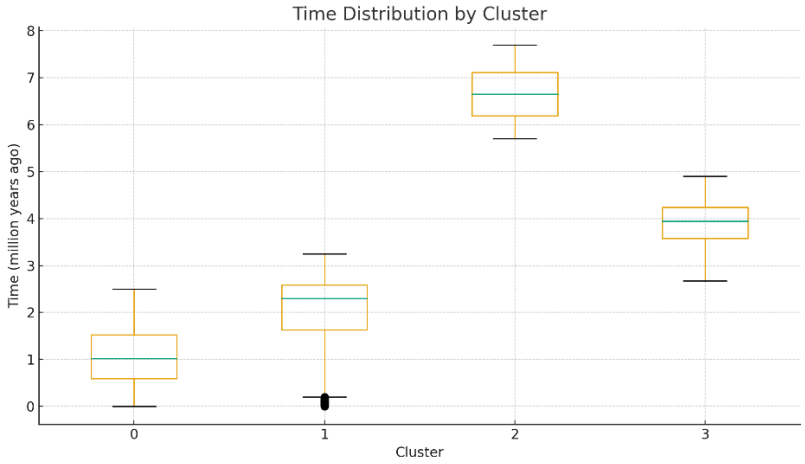

Figure 7. Time Distribution by Cluster.

Figure 8 shows the distribution of cranial capacity for each cluster. Custer 2 shows very low values with a small median, Custer 3 and cluster 1 are medium-high while cluster 0 has the highest median and the largest variation in cranium capacities. The incremental increase in medians from cluster 2 to 0 is evident and represents a sharp evolutionary tendency for brain size enlargement in the four groups.

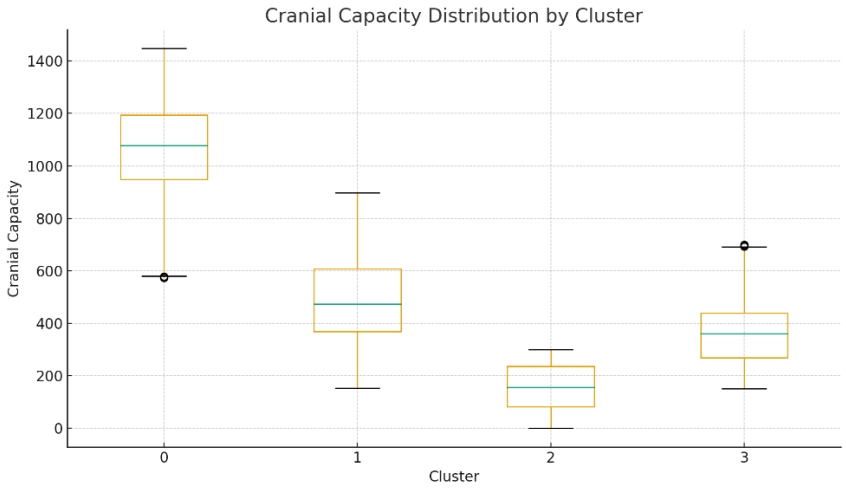

Figure 8. Cranial Capacity Distribution by Cluster.

Figure 9 depict the height distribution within clusters. The median height of cluster 2 is the shortest and its upper range has only a few positive cases, whereas clusters 1 and 3 have intermediate medians across which a wide variability in all cases is observed; cluster 0 has the highest median fold height group and also appears to present the largest dispersion towards higher values. These trends correspond to the distributions of cranial

capacity and reinforce that subsequent clusters are taller as well as larger in encephalic volume, following expectations from human evolutionary biology.

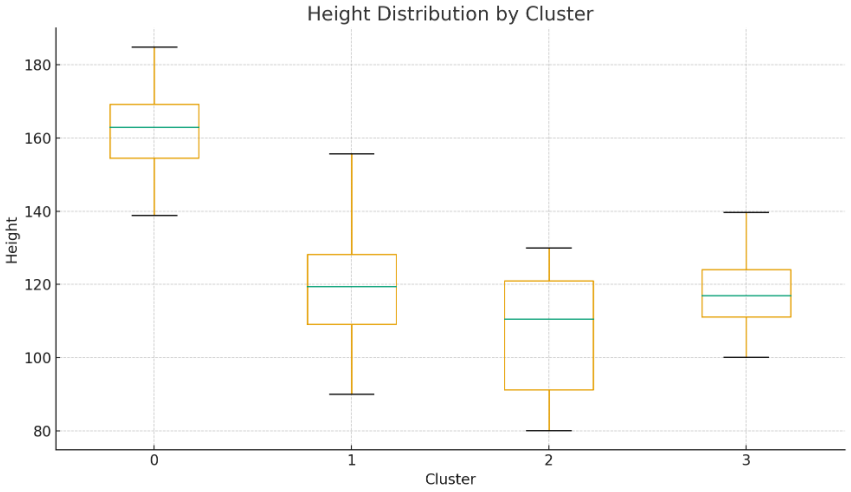

Figure 9. Height Distribution by Cluster

Table 4 provides an overview of the extent to which the four K-means clusters are shaped according to Silhouette optimum. For each cluster, the table indicates the number of samples, the mean and standard deviation of the Silhouette value as well as its minimum and maximum scores. Clusters 0, 1 and 2 have a relatively high mean Silhouette value (from 0.47–0.62) with largely positive values suggesting that their members are close to the centroid of their own cluster as opposed to other clusters. In contrast, cluster 3 has a lower average Silhouette value (≈0.21) and negative scores as well, indicating that some specimens of this group equally lie close to the decision boundary between clusters or are not so clearly assigned. The last row indicates a mean Silhouette coefficient of 0.419, truth that the four-cluster solution yields clusters that are on average quite homogeneous (although one cluster (Cluster 3) is less compact and more mixed than the others).

Table 4. Silhouette statistics for the K-means clustering solution (K = 4).

Cluster	Number of specimens (N)	Mean Silhouette coefficient	Standard deviation	Minimum Silhouette	Maximum Silhouette
0	1251	0.621	0.137	0.014	0.746
1	908	0.468	0.125	0.013	0.638
2	345	0.472	0.136	0.032	0.631
3	1496	0.210	0.133	–0.165	0.423

Figure 10 show the distribution of Silhouette coefficients for all samples when the dataset is splatted into four clusters with K-means. Every colored band represents one cluster (0–3) with its samples' Silhouette values sorted from low too high in the horizontal direction. Considering clusters 0, 1 and 2 most of the points show obviously positive Silhouette values suggesting compact and well separated clusters. Cluster 3 is composed by some samples that have low or nearly negative values, which indicates that not all

observations are completely in one cluster. The dashed vertical line is the composite mean Silhouette score (~0.42) which shows that our K = 4 solution has overall a good (but not perfect) separation of data points.

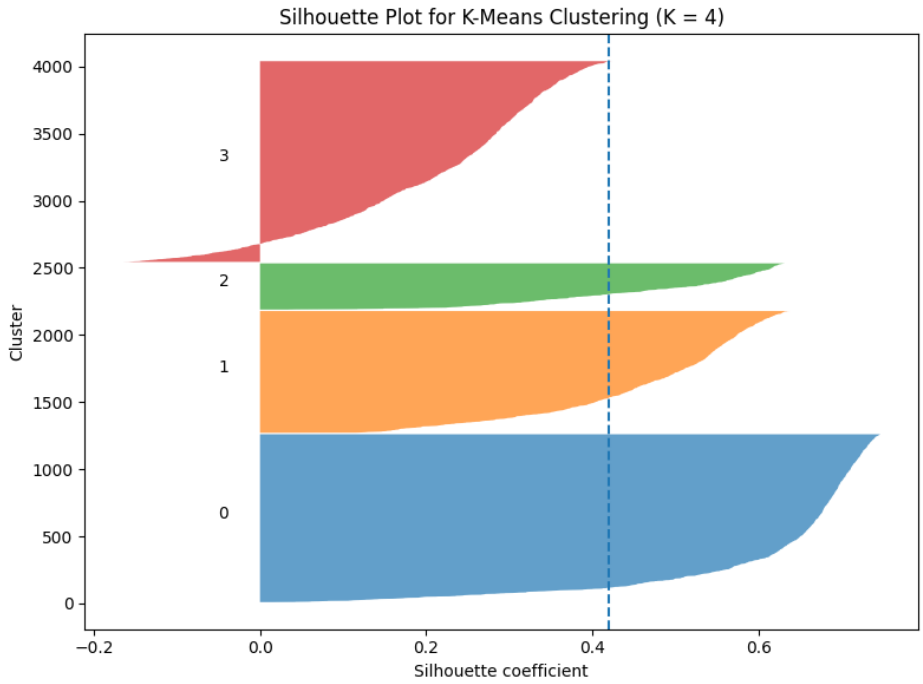

Figure 10. Silhouette for K-Means Clustering (K = 4).

CONCLUSION

This work showed that it is still possible to successfully retrieve meaningful evolution structure from the Evolution of Humans dataset after applying K-means clustering using only 3 numerical features: Time, Cranial_Capacity and Height. The algorithm, by dividing the data into four clusters, showed a clear gradient from very early small-bodied and small-brained hominins to later tall individuals with greatly enlarged cranial capacities. The temporal distribution of cluster centroids combined with monotonic ascent in both cranial capacity and height give evidence that unsupervised clustering corresponds to expectations based on known human evolutionary trends.

The study also revealed that the biggest clusters in the dataset are intermediate stages in evolution, while the most primitive remnants are among those with smallest membership. This concentration is consistent with the rarity of very old fossils, compared to more recent material. Most importantly, it is to note that the clusters were achieved without any a priori taxonomic label and hence the derived evolutionary consideration is based on numerical data only. This underscores the utility of K-means as an exploratory tool which can be used to identify coherent stages or regimes within paleoanthropological datasets before engaging more focused, hypothesis-driven analyses.

At the same time, it's an approach that has its limitations we can't ignore. It is a low-dimensional-space-based clustering with only three standardised inputs and Euclidean distance, relying on the assumption of spherical cluster. Contemporaneous features of hominin biology that were not included in the distance metric – such as specific cranio-dental traits, postcranial morphology or ecological context and technology use –were employed solely for qualitative interpretation following clustering. Furthermore, K-means offers a unique hard assignment of each sample to one cluster and does not account for potential overlaps or transitional forms between stages.

This study could be extended in various directions in future work. To further improve the performance of the model, additional dataset features could be included using mixed numeric categoric variable treatment approaches or thoughtful feature engineering. Secondly, since K-means results may be sensitive to the type of algorithm choice, other clustering methods such as Gaussian mixture models, hierarchical and density based can be compared to that of the K-means in order to validate the robustness of collected evolutionary stages. Ultimately, a more direct integration with expert paleoanthropological knowledge (to include taxonomic and anatomical annotations thereof) would facilitate fine-grained evaluation of the degree to which unsupervised clusters map onto formally recognized species and lineages. Our results are however limited by these issues; nevertheless, they suggest that K-means is promising as a straightforward initial step in analysing structure and temporal patterns within real human evolution data.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Aggarwal, C.C.; Reddy, C.K. Data Clustering: Algorithms and Applications; CRC Press: Boca Raton, FL, USA, **2014**.
2. Alves, V.; Campello, R.J.G.B.; Hruschka, E.R. Towards a fast evolutionary algorithm for clustering. *In Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2006)*; **2006**, pp. 1776–1783.
3. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit*. **2013**, *46*, 243–256.
4. Arthur, D.; Vassilvitskii, S. K-Means++: The advantages of careful seeding. *In Proceedings of the 18th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA)*; **2007**; p. 1027–1035.
5. Bache, K.; Lichman, M. UCI Machine Learning Repository; University of California, School of Information and Computer Science: Irvine, CA, USA, 2013; Available online: http://archive.ics.uci.edu/ml (accessed on 29 July 2025).
6. Bandyopadhyay, S.; Maulik, U. An evolutionary technique based on K-Means algorithm for optimal clustering. *Inf. Sci.* **2002**, 146, 221–237.

7. Ben-David, S.; von Luxburg, U.; Pál, D. A sober look at clustering stability. *In Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*; **2006**; p. 5–19.
8. Bezdek, J.C.; Boggavarapu, S.; Hall, L.O.; Bensaid, A. Genetic algorithm guided clustering. *In Proceedings of the First IEEE Conference on Evolutionary Computation*; **1994**; p. 34–39.
9. Brunsch, T.; Röglin, H. A bad instance for k-means++. *Theor. Comput. Sci.* **2013**, *505*, 19–26.
10. Bubeck, S.; Meilă, M.; von Luxburg, U. How the initialization affects the stability of the K-Means algorithm. *ESAIM Probab. Stat.* **2012**, *16*, 436–452.
11. Cano, J.R.; Cordón, O.; Herrera, F.; Sánchez, F. A greedy randomized adaptive search procedure applied to the clustering problem as an initialization process using K-means as a local search procedure. *J. Intell. Fuzzy Syst.* **2002**, 12, 235–242.
12. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **2014**, *61*(6), 1–36.
13. Chen, S.; Chao, Y.; Wang, H.; Fu, H. A prototypes-embedded genetic K-Means algorithm. *In Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*; **2006**; p. 724–727.
14. Chiu, T.Y.; Hsu, T.C.; Wang, J.S. AP-based consensus clustering for gene expression time series. *In Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*; **2010**; p. 2512–2515.
15. Chiui, T.Y.; Hsu, T.C.; Yen, C.C.; Wang, J.S. Interpolation based consensus clustering for gene expression time series. *BMC Bioinform.* **2015**, *16*, 117.
16. Craenendonck, T.V.; Blockeel, H. Using internal validity measures to compare clustering algorithms. In ICML 2015 AutoML Workshop; 2015; Available online: https://lirias.kuleuven.be/bitstream/123456789/504712/1/automl_camera.pdf (accessed on 28 July 2025).
17. de Amorim, R.C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci.* **2015**, *324*, 126–145.
18. Erisoglu, M.; Calis, N.; Sakallioglu, S. A new algorithm for initial cluster centers in K-Means algorithm. *Pattern Recognit. Lett.* **2011**, *32*, 1701–1705.
19. Famili, A.F.; Liu, G.; Liu, Z. Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics* **2004**, *20*(10), 1535–1545.
20. Fang, Y.; Wang, J. Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.* **2012**, *56*(3), 468–477.
21. Hall, L.O.; Özyurt, I.B.; Bezdek, J.C. Clustering with a genetically optimized approach. *IEEE Trans. Evol. Comput.* **1999**, *3*(2), 103–112.
22. Handl, J.; Knowles, J. An evolutionary approach to multiobjective clustering. *IEEE Trans. Evol. Comput.* **2007**, *11*(1), 56–76.
23. He, Z. Evolutionary K-Means with pair-wise constraints. *Soft Comput.* **2016**, *20*(1), 287–301.
24. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **2007**, *52*(1), 258–271.
25. Hruschka, E.R.; Campello, R.J.G.B.; de Castro, L.N. Evolving clusters in gene-expression data. *Inf. Sci.* **2006**, *176*, 1898–1927.
26. Hruschka, E.R.; Campello, R.J.G.B.; Freitas, A.A.; Carvalho, A.C.P.L.F. A survey of evolutionary algorithms for clustering. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2009**, *39*(2), 133–155.

27. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*(8), 651–666.
28. Krishna, K.; Murty, M.N. Genetic K-Means algorithm. *IEEE Trans. Syst. Man Cybern. B Cybern.* **1999**, *29*(3), 433–439.
29. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. *In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)*; **2010**; p. 911–916.
30. Möller, U. Resampling methods for unsupervised learning from sample data. In Mellouk, A.; Chebira, A., Eds.; Machine Learning; InTech: Cape Town, South Africa, 2009; pp. 289–304. Available online: http://cdn.intechweb.org/pdfs/6069.pdf (accessed on 28 July 2025).
31. Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **2003**, *52*, 91–118.
32. Naldi, M.C.; Campello, R.J.G.B.; Hruschka, E.R.; Carvalho, A.C.P.L.F. Efficiency issues of evolutionary K-Means. *Appl. Soft Comput.* **2011**, *11*, 1938–1952.
33. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2015; Available online: https://www.R-project.org/ (accessed on 28 July 2025).
34. Rahman, M.A.; Islam, M.Z.; Bossomaier, T. DenClust: A density based seed selection approach for K-Means. *In Proceedings of the 13th International Conference on Artificial Intelligence and Soft Computing (ICAISC); Lecture Notes in Computer Science*, Vol. 8468; **2014**; p. 784–795.
35. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
36. Schmidt, T.S.B.; Matias Rodrigues, J.F.; von Mering, C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ. Microbiol.* **2015**, *17*(5), 1689–1706.
37. Senbabaoglu, Y.; Michailidis, G.; Li, J.Z. Critical limitations of consensus clustering in class discovery. *Sci. Rep.* **2014**, 4, 6207.
38. Shamir, O.; Tishby, N. Stability and model selection in K-Means clustering. *Mach. Learn.* **2010**, 80(2–3), 213–243.
39. Vendramin, L.; Campello, R.J.G.B.; Hruschka, E.R. Relative clustering validity criteria: A comparative overview. *Stat. Anal. Data Min.* **2010**, *3*(4), 243–256.
40. Vinh, N.X.; Epps, J. A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. *In Proceedings of the 9th International Conference on Bioinformatics and Bioengineering (BIBE)*; **2009**; p. 84–91.
41. Evolution of Humans Datasets for Classification. Kaggle. Available online: https://www.kaggle.com/datasets/santiago123678/evolution-of-humans-datasets-for-clasification (accessed on 10 July 2025).