

Article

Comparison of K-Means and Hierarchical Clustering Methods for Buffalo Milk Production Data

Lucia Trapanese ¹, Giovanna Bifulco ^{1,*}, Matteo Santinello ¹, Nicola Pasquino ² , Giuseppe Campanile ¹  and Angela Salzano ¹ 

¹ Department of Veterinary Medicine and Animal Production, University of Naples Federico II, 80137 Naples, Italy; lucia.trapanese@unina.it (L.T.); matteo.santinello@unina.it (M.S.); giucampa@unina.it (G.C.); angela.salzano@unina.it (A.S.)

² Department of Electrical Engineering and Information Technologies, University of Naples Federico II, 80125 Naples, Italy; nicola.pasquino@unina.it

* Correspondence: giovanna.bifulco@unina.it

Simple Summary

In Italy, buffalo farming has gained increasing importance in recent years, with a steady rise in the number of animals reared. To optimize herd management and extract meaningful insights from available datasets, the use of Machine Learning techniques has become increasingly valuable. Among these, unsupervised learning provides a relatively straightforward and efficient approach for analyzing unlabeled data. In this study, we applied two unsupervised algorithms to routine buffalo production data to assess their performance and to evaluate their potential with the aim of improving herd management practices. Our findings suggest that, despite certain limitations, clustering analysis can offer substantial advantages in understanding herd structure, consistency and management.

Abstract

This study investigated the use of K-means and hierarchical clustering, to group Italian Mediterranean buffalo using routinely collected test-day records. The analysis was first conducted on a combined dataset comprising three buffalo herds and subsequently on each herd individually. The main objective was to determine whether data-driven groupings could be implemented to support improvements in general herd management strategies. Results indicated that K-means consistently outperformed hierarchical clustering across all datasets, as reflected by average silhouette scores (0.17–0.18 vs. 0.10–0.12 for K-means and hierarchical, respectively), favorable Davies–Bouldin Index (DBI; 2.05–2.16 vs. 2.11–2.5 for K-means and hierarchical, respectively) and Calinski–Harabasz Index values (CHI; 1034–3877 vs. 729–2109 for K-means and hierarchical, respectively). K-means identified two clusters in the combined dataset and in two of the three herds, while three clusters were identified in the remaining herd. Cluster composition analysis revealed that days in milk and milk yield were the main discriminating factors when two clusters were formed. When three clusters emerged, K-means also identified a subgroup of animals that differed from the others in both age and lactation stage. These findings were supported by the analysis of variance (ANOVA), which showed statistically significant differences among most of the evaluated variables.



Academic Editors: Daniel Zaborski and Wilhelm Grzesiak

Received: 25 September 2025

Revised: 1 November 2025

Accepted: 4 November 2025

Published: 9 November 2025

Citation: Trapanese, L.; Bifulco, G.; Santinello, M.; Pasquino, N.; Campanile, G.; Salzano, A. Comparison of K-Means and Hierarchical Clustering Methods for Buffalo Milk Production Data. *Animals* **2025**, *15*, 3246. <https://doi.org/10.3390/ani15223246>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: machine learning; buffalo; clustering; K-means; unsupervised analysis

1. Introduction

The livestock sector is undergoing a profound transformation aimed at improving production efficiency while addressing environmental and economic sustainability challenges. As highlighted by Gong et al. [1], the growing global demand for animal-based products must be reconciled with the need to ensure animal welfare and the reduction in the environmental footprint, concerns increasingly prioritized by both researchers and consumers. In this regard, Peyraud et al. [2] observed that consumer preferences are progressively shaped by factors such as product quality, animal husbandry conditions and the nutritional value of processed animal products. Meeting these multifaceted expectations requires a shift from traditional livestock management toward the integration of advanced technologies and data-driven approaches. This transition is embodied in the concept of Precision Livestock Farming (PLF), which involves the adoption of real-time monitoring systems for animals, their products and their surrounding environment. The goal is to provide farmers with objective, timely and actionable data to enhance farm management practices [3].

Among emerging technologies, Machine Learning (ML) has gained prominence as a powerful tool for pattern recognition and decision support. In particular, ML techniques are well suited to handle large, complex, non-linear and noisy datasets, typical of livestock systems [4]. The ML algorithms are generally categorized into supervised and unsupervised, depending on whether the input data contains labeled outcomes [5]. Supervised algorithms are typically applied to classify events or predict values, such as nitrogen emissions [6] or reproductive efficiency [7]. Whereas unsupervised techniques are valuable for exploring data structures and detecting latent patterns. Indeed, Tremblay et al. [8] combined Principal Component Analysis (PCA) with the K-means algorithm to cluster cows affected by Poor Metabolic Adaptation Syndrome (PMAS). The five resulting clusters were classified into low, intermediate, or high PMAS categories based on their agreement with expected indicators and characteristics. Similarly, Franceschini et al. [9] assessed the overall health status of dairy cattle using large-scale milk-recording data, identifying four closely related clusters and one distinct group characterized by higher somatic cell score (SCS).

Although ML applications have become increasingly common in dairy cattle research, their use in other livestock species remains limited. Among these, dairy buffaloes represent a major component of the global milk market; however, the potential of ML in this field remains largely underexplored [10]. Buffalo farmers collect extensive routine data that, while already valuable for herd management and genetic improvement, still hold untapped potential for further optimization. In this context, unsupervised ML approaches can effectively exploit such information [11]. To address this gap, the present study has investigated the application of unsupervised ML techniques to routine test-day recording from three dairy buffalo farms. Specifically, two clustering algorithms, namely the K-means and hierarchical clustering, were applied to identify natural groupings of animals based on their production and physiological characteristics. The aim was to evaluate the potential of unsupervised learning to enhance herd management by grouping animals with similar physiological and productive profiles.

2. Materials and Methods

2.1. Data Description and Preprocessing

Animal welfare approval was not required for this study because all data were obtained from pre-existing databases. The dataset consisted of routinely collected records from three commercial dairy buffalo farms located in southern Italy. Animal characteristics and production data have been collected from the national test-day recording system managed by the Italian Breeders' Association (AIA) and covered the period between 2013

and 2016. Only animals with fewer than six lactations were included. Before the analysis, a data-cleaning process was performed to ensure uniformity and reliability of the dataset. In this context, missing values, zeros and outliers were removed. Outlier detection thresholds were defined based on literature references and biological plausibility to ensure consistency [12]. The dataset comprised herds (h) containing a variable number of buffaloes m_h . For each animal (i) within herd h , multiple measurements $p_{h,i,j}$ were collected over time, with sampling frequencies varying according to the farm's recording schedule (one test-day per month). Overall, the dataset included 1,342 Italian Mediterranean buffaloes and a total of 16,702 test-day records. Data included information on:

- Animal and milk yield characteristics: parity order, daily milk yield, days in milk (DIM), age, lactation length and total lactation milk yield;
- Reproduction information: age at first calving (month);
- Milk quality: fat, protein, lactose, SCS and the fat/protein ratio.

The selection of the variables was aimed at capturing the multifaceted aspects of buffalo performance and management. Most of these variables were available from the routine recording system and were selected to represent three main areas of interest: animal and production traits, reproductive performance and milk quality. In addition, a correlation matrix including all variables in the dataset was computed to assess inter-variable relationships (Supplementary Materials Figure S1). The SCS was calculated from the Somatic Cell Count (SCC) using the formula proposed by Ali and Shook [13]:

$$SCS = 3 + \log_2 \left(\frac{SCC}{10^5} \right)$$

All variables were standardized using z-score normalization to ensure comparability across features and to improve the efficiency and accuracy of the clustering process [9]. The structure of the final dataset was illustrated in Figure 1.

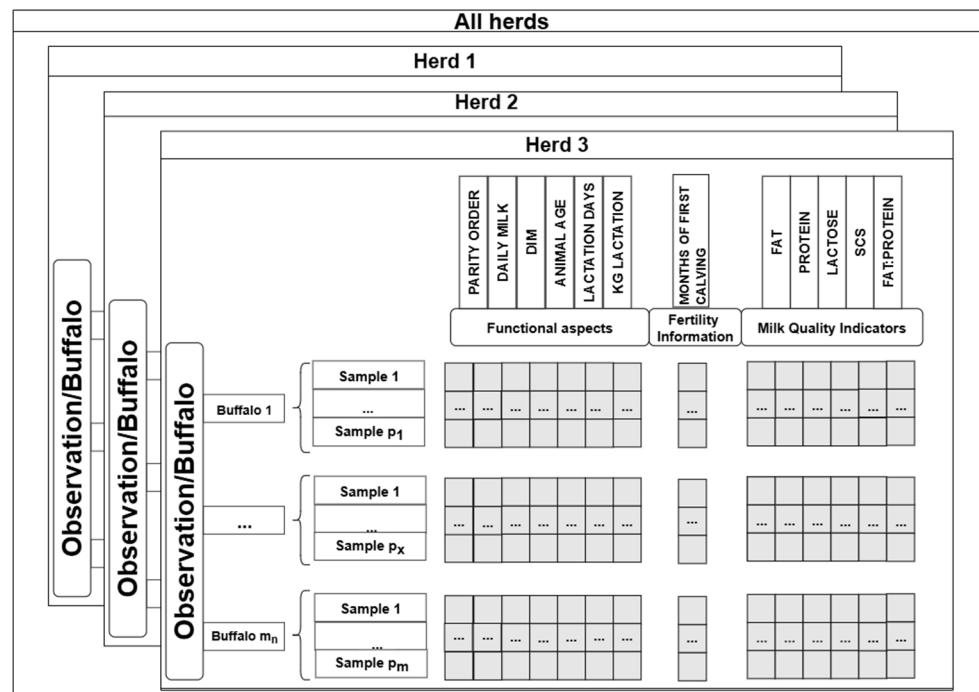


Figure 1. Schematic representation of the dataset structure used for clustering analysis.

2.2. Cluster Analysis

The analysis was performed using R software (version 4.5.1; <https://www.r-project.org/> accessed on 14 January 2025). A first analysis was initially carried out on the combined dataset (made of three different herds) to identify the most suitable clustering technique. Based on this preliminary approach, the best-performing algorithm was subsequently applied to each of the three herd-specific subsets to validate its effectiveness under different farm conditions. Cluster analysis was carried out using two unsupervised algorithms: K-means and hierarchical clustering. These methods were selected due to their previous application in cattle studies [14,15] and their compatibility with the structure of the dataset. Other unsupervised algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Self-Organizing Maps (SOM), were also evaluated. However, DBSCAN, being density-based, was excluded since it was more suitable for anomaly detection [16]. Moreover, even if the SOM method is useful for time-series analysis, it requires a large number of continuous observations and for this reason it was not applied in our case, as the dataset included few test-day records per animal within one lactation [17]. For hierarchical clustering a Ward's linkage method was employed, following recommendations from previous research [14]. A key challenge in such an unsupervised ML approach consists of determining the optimal number of clusters, which is not known a priori. To address this, the “NbClust” function from the NbClust R package (version 3.0.1) [18] was used. This function evaluates 30 different clustering indices to identify the most appropriate number of clusters for both algorithms. Clustering performance was assessed using four internal validation metrics: the average silhouette coefficient, the Davies–Bouldin Index (DBI), the Calinski–Harabasz Index (CHI), and the Dunn Index [19].

The silhouette coefficient evaluated how well each observation fits within its assigned cluster compared to other clusters. For each observation i , the silhouette coefficient is defined as:

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

where $a(i)$ represents the average distance between a sample and all other points within the same cluster (a measure of intra-cluster cohesion), and $b(i)$ represents the minimum average distance between a sample and all points in other clusters (a measure of inter-cluster separation). Specifically, $a(i)$ is defined as:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\substack{x_j \in C_i \\ j \neq i}} d(x_i, x_j)$$

$b(i)$ is defined as:

$$b(i) = \min_{j \neq i} \left(\frac{1}{|C_j|} \sum_{x_j \in C_j} d(x_i, x_j) \right)$$

where $d(x_i, x_j)$ is the Euclidean distance between points x_i and x_j and $|C_i|$ is the total number of points in cluster i . The average silhouette coefficient (AVG_SIL) was then computed for each cluster as:

$$\text{AVG_SIL} = \frac{1}{N} \sum_{i=1}^N \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where N is the total number of observations (within a cluster or across all clusters); $a(i)$ is the average distance between observation i and all other points in the same cluster, while $b(i)$ is the minimum average distance between i and points in any other cluster (i.e., the nearest neighboring cluster). The silhouette score ranges from -1 to 1 . Values close to 1 indicate

that the observation is well clustered, while negative values imply that the observation may be misclassified. The Davies–Bouldin Index (DBI) quantifies the compactness and separation of clusters, with lower values reflecting superior clustering performance. The DBI was computed following Davies et al. [20]:

$$\text{DBI} = \frac{1}{M} \sum_{j=1}^M \max_{j \neq k} R_{j,k},$$

where M is the total number of clusters, and $R_{j,k}$ is the similarity between clusters j and k , typically defined as the ratio of within-cluster dispersion to between-cluster distance.

The CHI is determined by the ratio of inter-cluster dispersion to intra-cluster dispersion across all clusters:

$$\text{CHI} = \frac{\text{SSB}}{\text{SSW}} \cdot \frac{(N - M)}{(M - 1)}$$

where N is the total number of observations of the dataset, SSB is the between-cluster sum of squares and SSW the within-cluster sum of squares, computed as

$$\text{SSB} = \sum_{i=1}^M |C_i| \cdot d(c_i, c)$$

where c_i is the centroid of cluster i and c the overall dataset centroid.

The within-cluster sum of squares (SSW) is calculated as:

$$\text{SSW}_K = \sum_{i=1}^{|C_i|} d(x_i, c_k)$$

which measures the intra-cluster dispersion for each cluster, leading to:

$$\text{SSW} = \sum_{i=1}^M \text{SSW}_i$$

Finally, the Dunn Index measures the relationship between the minimum separation among clusters and the maximum dispersion within clusters. Higher Dunn Index values correspond to better clustering performance. It is defined as:

$$\text{Dunn} = \frac{\min_{1 \leq i < j \leq M} \delta(C_i, C_j)}{\max_{1 \leq l \leq M} \Delta(C_l)}$$

where $\delta(C_i, C_j)$ is the distance between clusters i and j and $\Delta(C_l)$ is the diameter of cluster l , defined as the maximum distance between any two points within the same cluster.

Figure 2 summarizes the workflow of the study. Starting from the dataset, the first step involved applying z-score standardization to all variables. Then, the optimal number of clusters was determined based on the value most frequently suggested by thirty different validation indices. Subsequently, both K-means and hierarchical clustering were performed using the selected number of clusters, and their performance was compared using the Silhouette coefficient and the DBI, CHI and Dunn indices. The final step consisted of exploring the cluster composition and interpreting the resulting groups.

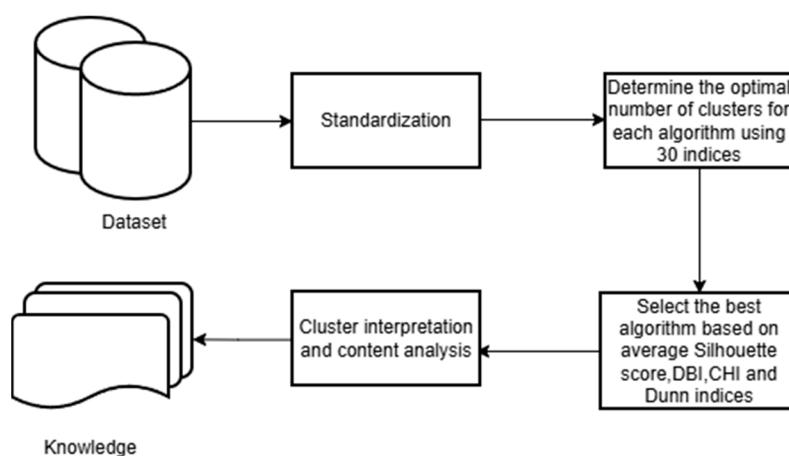


Figure 2. Overview of the analytical workflow adopted in the study. DBI = Davies–Bouldin Index, CHI = Calinski–Harabasz Index.

2.3. Statistical Analysis and Cluster Exploration

Once the algorithm showing the best performance indices for each trial was identified, cluster validation was performed. Specifically, the validation aimed to assess whether the mean values of each feature differed significantly among the identified clusters. To account for repeated measurements within animals, a repeated-measures analysis of variance was conducted using mixed-effects models, with the animal included as a random effect. When more than two clusters were identified, pairwise comparisons between clusters were performed using the Sidak adjustment for multiple testing. Differences were considered statistically significant at $p < 0.05$.

3. Results

Table 1 summarizes the clustering performance of the K-means and hierarchical algorithms across the different datasets. K-means outperformed hierarchical clustering in terms of the average silhouette, DBI and CHI, yielding higher average silhouette scores (0.17–0.18 vs. 0.10–0.12 for K-means and hierarchical, respectively), higher CHI values (1034–3877 vs. 729–2109 for K-means and hierarchical, respectively), and lower DBI values (2.05–2.16 vs. 2.11–2.5 for K-means and hierarchical, respectively). Conversely, the Dunn Index showed better results for the hierarchical algorithm, with values ranging from 0.03 to 0.05.

Table 1. Comparison of clustering performance ¹ between hierarchical and K-means clustering algorithms applied to the complete dataset and three individual herds.

Dataset	K-Means				Hierarchical					
	N of Clusters	Average Silhouette Score	DBI	CHI	Dunn	N of Clusters	Average Silhouette Score	DBI	CHI	Dunn
Total dataset	2	0.18	2.07	3877	0.0213	3	0.11	2.30	2109	0.0300
Herd 1	2	0.18	2.08	1307	0.0272	3	0.11	2.50	992	0.0343
Herd 2	2	0.17	2.16	1191	0.0298	3	0.10	2.29	729	0.0534
Herd 3	3	0.18	2.05	1034	0.0241	3	0.12	2.11	951	0.0451

¹ Clustering performance: Silhouette score: indicates how well each observation fits within its cluster; DBI: measures cluster compactness and separation. CHI: evaluates the ratio of between-cluster dispersion to within-cluster dispersion, Dunn: assesses the minimum inter-cluster distance relative to the maximum intra-cluster distance.

Based on these results, K-means was selected as the most appropriate unsupervised learning algorithm for this study, as three out of four indices indicated superior performance.

Figure 3 provided a graphical representation of the silhouette analysis for the clusters generated by the K-means algorithm, both for the combined dataset and for each individual herd. The silhouette analysis revealed an overall coherent cluster structure, with only a small proportion of observations exhibiting negative silhouette values, indicative of potential misclassifications.

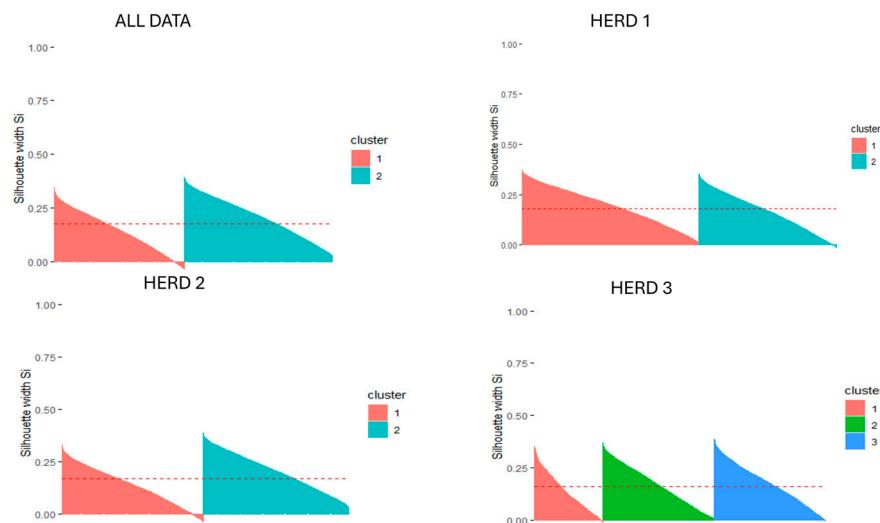


Figure 3. Silhouette analysis obtained after performing the K-means algorithm for the combined dataset and for herds 1, 2 and 3. The red dashed line is the average silhouette.

In the combined dataset, 3.63% of the observations were identified as misclassified, mainly concentrated in Cluster 1, suggesting that this cluster was less defined than the others. This interpretation was supported by the lower average silhouette value for Cluster 1 (0.14) compared with Cluster 2 (0.21). At the herd level, for Herd 1 Cluster 1 had no misclassified observation, while Cluster 2 contained 65 misclassified observations (1.13%). In Herd 2, only Cluster 1 showed a misclassification rate of 4%, corresponding to 223 observations. Finally, in Herd 3, Cluster 1 included 35 misclassified cases (2.75%), whereas Cluster 2 included only 13 observations (0.63%).

Table 2 reports the mean \pm standard deviation of each variable for the clusters identified by the K-means algorithm applied to the combined dataset. The latter comprised 16,702 test-day records from 1341 buffaloes. Tables 3–5 present the corresponding results for Herds 1, 2, and 3, respectively, based on separate K-means analyses carried out within each herd. In Table 2, referring to the combined dataset analysis, Cluster 1 included 7825 observations (46%), whereas Cluster 2 contained 8877 observations (54%). Most of the features showed statistically significant differences between clusters ($p < 0.001$).

Table 2. Animal production traits, reproductive performance and milk quality for each cluster identified by the K-means clustering applied on the combined dataset. Data were shown as means \pm standard deviation. NS = not significant ($p > 0.05$).

Traits	Cluster 1	Cluster 2	p-Value
N. of samples	7825	8877	
Animal characteristics:			
Parity order	1.58 \pm 0.77	1.68 \pm 0.85	<0.0001
Primiparous (%)	52.65	55.98	
Days in milk (d)	213 \pm 70	89 \pm 55	<0.0001
Lactation length (d)	324 \pm 75	293 \pm 60	<0.0001
Age at first calving (mo)	33 \pm 5	34 \pm 4	NS
Age (y)	4.18 \pm 1.13	3.95 \pm 1.17	<0.0001

Table 2. Cont.

Traits	Cluster 1	Cluster 2	p-Value
Milk production and composition:			
Lactation milk yield (kg)	2476 ± 758	2638 ± 736	<0.001
Daily milk yield (kg/d)	6.18 ± 2.40	10.89 ± 3.00	<0.0001
Fat (%)	9.74 ± 1.18	7.47 ± 1.05	<0.0001
Protein (%)	4.99 ± 0.36	4.53 ± 0.33	<0.0001
Fat: Protein	1.96 ± 0.22	1.65 ± 0.23	<0.0001
Lactose (%)	4.58 ± 0.37	4.86 ± 0.27	<0.0001
Somatic cell score	4.16 ± 1.57	2.99 ± 1.54	<0.0001

Table 3. Animal production traits, reproductive performance and milk quality for each cluster identified by K-means clustering applied on Herd 1. Data were shown as means ± standard deviation. NS = not significant ($p > 0.05$).

Traits	Cluster 1	Cluster 2	p-Value
N. of samples	2505	3223	
Animal characteristics:			
Parity order	1.61 ± 0.77	1.59 ± 0.78	NS
Primiparous (%)	55.96	55.83	
Days in milk (d)	211.31 ± 61.57	86.94 ± 50.77	<0.0001
Lactation length (d)	302.33 ± 59.42	283.57 ± 52.48	<0.0001
Age at first calving (mo)	34.72 ± 2.91	34.67 ± 2.68	NS
Age (y)	4.20 ± 1.05	3.75 ± 0.96	<0.0001
Milk production and composition:			
Lactation Milk yield (kg)	3034 ± 692	3024 ± 655	<0.001
Daily milk yield (kg/d)	8.19 ± 2.40	12.67 ± 2.75	<0.0001
Fat (%)	9.44 ± 1.10	7.23 ± 1.03	<0.0001
Protein (%)	4.90 ± 0.37	4.43 ± 0.32	<0.0001
Fat: Protein	1.93 ± 0.20	1.64 ± 0.23	<0.0001
Lactose (%)	4.64 ± 0.27	4.85 ± 0.27	<0.0001
Somatic cell score	3.96 ± 1.33	3.08 ± 1.51	<0.0001

Table 4. Animal production traits, reproductive performance and milk quality for each cluster identified by K-means clustering applied on Herd 2. Data were shown as means ± standard deviation. NS = not significant ($p > 0.05$).

Traits	Cluster 1	Cluster 2	p-Value
Number of samples	2741	2833	
Animal characteristics:			
Parity order	1.55 ± 0.78	1.57 ± 0.80	NS
Primiparous (%)	57.03	56.49	
Days in milk (d)	229 ± 72	86.00 ± 54.63	<0.0001
Lactation length (d)	352.90 ± 90.94	314.76 ± 83.22	<0.0001
Age at first calving (mo)	32.10 ± 6.48	31.26 ± 4.83	NS
Age (y)	4.05 ± 1.22	3.60 ± 1.19	<0.0001
Milk production and composition:			
Lactation Milk yield (kg)	2214 ± 718	2107 ± 733	<0.0001
Daily milk yield (kg/d)	4.78 ± 2.06	8.64 ± 2.80	<0.0001
Fat (%)	10.08 ± 1.25	8.02 ± 1.10	<0.0001
Protein (%)	5.02 ± 0.36	4.71 ± 0.31	<0.0001
Fat: Protein	2.01 ± 0.23	1.71 ± 0.24	<0.0001
Lactose (%)	4.49 ± 0.48	4.87 ± 0.27	<0.0001
Somatic cell score	4.77 ± 1.62	3.51 ± 1.59	<0.0001

Table 5. Animal production traits, reproductive performance and milk quality for each cluster identified by K-means clustering applied on Herd 3. Data were shown as means \pm standard deviation. Means not sharing a common letter are significantly different at $p < 0.05$. NS = not significant ($p > 0.05$).

Traits	Cluster 1	Cluster 2	Cluster 3	<i>p</i> -Value
Number of samples	1270	2067	2063	
Animal characteristics:				
Parity order	2.89 \pm 0.65 A	1.31 \pm 0.48 C	1.45 \pm 0.60 B	<0.0001
Primiparous (%)	0	60.56	56.74	
Days in milk (d)	122.37 \pm 64.22 B	84.04 \pm 53.62 C	219.95 \pm 59.30 A	<0.001
Lactation length (d)	270.83 \pm 37.84 C	303.91 \pm 48.30 B	311.77 \pm 49.80 A	<0.0001
Age at first calving (mo)	35.78 \pm 3.39	37.17 \pm 3.72	36.64 \pm 3.57	NS
Age (y)	5.82 \pm 0.82 A	3.69 \pm 0.73 C	3.69 \pm 0.73 B	<0.001
Milk production and composition:				
Lactation Milk yield (kg)	2556 \pm 610 A	2480 \pm 539 B	2441 \pm 548 B	<0.0001
Daily milk yield (kg/d)	10.07 \pm 3.06 A	9.95 \pm 2.44 A	6.18 \pm 1.97 B	<0.0001
Fat (%)	8.06 \pm 1.15 B	7.25 \pm 1.01 C	9.70 \pm 1.16 A	<0.0001
Protein (%)	4.59 \pm 0.35 B	4.54 \pm 0.33 C	5.04 \pm 0.34 A	<0.0001
Fat/Protein	1.76 \pm 0.23 B	1.60 \pm 0.22 C	1.93 \pm 0.21 A	<0.0001
Lactose (%)	4.76 \pm 0.31 B	4.87 \pm 0.26 A	4.58 \pm 0.28 C	<0.0001
Somatic cell score	3.18 \pm 1.42 B	2.36 \pm 1.46 C	3.55 \pm 1.47 A	<0.0001

Herd 1 included 493 animals and 5728 test-day records. The optimal number of clusters was two. Cluster 1 comprised 2505 observations (43%), whereas Cluster 2 included 3223 observations (57%). The corresponding results are reported in Table 3. All features showed statistically significant differences between clusters ($p < 0.001$), except parity order and age at first calving.

Herd 2 included 421 animals, and the clustering analysis was performed on 5574 observations. Cluster 1 comprised 2741 observations (49%), whereas Cluster 2 contained 2833 observations (51%). The corresponding results are reported in Table 4. All features showed statistically significant differences between clusters, except parity order and age at first calving.

Herd 3 consisted of 427 animals and 5400 test-day records. Cluster 1 comprised 1270 observations (23%), Cluster 2 included 2067 (38.6%), and Cluster 3 contained 2063 (38.4%). The corresponding results are reported in Table 5. Many features showed statistically significant differences ($p < 0.001$) among the three clusters, whereas daily milk yield did not differ significantly between Clusters 1 and 2 while lactation milk yield did not differ significantly between Clusters 2 and 3. Supplementary Materials (Figure S2) supported the evidence presented in Tables 2–5.

4. Discussion

The main novelty of this study relied on the application of unsupervised ML algorithms to buffalo breeding, using routine test-day records. Unsupervised learning approach was chosen because the objective of the work was to uncover natural patterns and groupings within the data, without relying on predefined categories or labels. This methodology made it possible to identify latent structures that improve the understanding of production dynamics and support data-driven decision-making in herd management [21]. Unsupervised algorithms have previously been applied to a range of livestock datasets. For example, Lee et al. [22] performed k-medoids clustering on Holstein cows based on lactation curve characteristics, identifying six groups that included four typical and two atypical lactation curve shapes. Their model highlighted physiological differences among clusters, suggesting possible implications for individual cow management. Moreover, Brotzman et al. [14]

successfully applied PCA and hierarchical algorithm to differentiate dairy farms in the US according to productivity and breeding systems. Finally, PCA followed by a hierarchical clustering algorithm has been applied to assess Australian dairy farmers' preferences for cow trait improvements [23]. Farmers' preferences were collected through a survey in which each participant was asked to indicate the traits they considered most important. These findings were used to guide trait and index selection.

4.1. Clustering Performance

To identify patterns in buffalo characteristics and test-day records, both K-means and hierarchical clustering algorithms were applied to the combined dataset, as well as to the three individual herds. This double approach enabled the evaluation of algorithm's robustness under different herd conditions. Overall, K-means consistently outperformed hierarchical clustering, as evidenced by the average silhouette and CHI values, and DBI values. For K-means, silhouette scores ranged from 0.17 to 0.18, indicating limited cohesion within clusters. These results suggested that intra-cluster similarity was only slightly greater than inter-cluster similarity [24], implying that many data points were positioned near cluster boundaries. Approximately 4% of the observations displayed negative silhouette values, indicating potential misclassifications. Although most points had positive scores, many were close to zero, further highlighting the blurred boundaries, particularly among clusters representing animals in mid-lactation. Similarly, DBI values ranged from 2.05 to 2.16, and CHI values confirmed modest clustering performance. These results indicated the presence of overlapping or weakly defined clusters [19]. It is worth noting that both metrics can be influenced by factors such as data noise, skewed distributions, internal substructures and differences in cluster density [20], which may also have affected clustering performance in this study. The structure and distribution of the data further contribute to these results. Specifically, the data points formed a dense, nearly circular distribution, leading to poor separation between clusters and therefore lower values for separation-based indices. Conversely, the clusters exhibited relatively high internal cohesion. Despite these limitations, the clusters generated by K-means were meaningful from a domain-specific standpoint and provided valuable insights into the variability of animal characteristics and production traits.

4.2. Clustering Ability to Create Homogenous Groups

Overall, DIM was useful for clustering purposes, effectively grouping animals into two or three distinct lactation phases. Moreover, in the combined dataset, the K-means algorithm separated observations according to milk yield, confirming this trait as one of the most influential in distinguishing clusters. Cluster 1 comprised buffaloes in late lactation, with an average of 213 DIM, having the lowest daily milk yield (6.18 kg) but the highest fat (9.74%) and protein (4.99%) contents. The same cluster also showed the highest SCS (4.16), a relatively long lactation length (324 days), and a total lactation milk yield of 2576 kg. In contrast, Cluster 2 included animals in early lactation (average 89 DIM), characterized by the highest daily milk yield (10.89 kg) and the lowest fat (7.47%) and protein (4.53%) contents, consistent with the expected dilution effect near the peak lactation [25]. This cluster also exhibited the lowest SCS (2.99), suggesting better udder health. A comparable clustering pattern was observed in Herds 1 and 2, where two distinct clusters emerged. In Herd 1, Cluster 1 consisted of late-lactation animals (average 211 DIM), showing the highest fat (9.94%) and protein (4.90%) contents, the lowest daily milk yield (9.44 kg), and the highest SCS (3.96), reflecting declining productivity and potential deterioration of udder health. The higher SCS in late-lactation animals may indicate cumulative stress or subclinical mastitis, as previously reported in previous research [26]. Cluster 2 grouped

animals in early lactation (86 DIM) that showed higher daily milk yield (12.67 kg) and lower fat (7.23%) and protein (4.43%) contents.

Concerning Herd 2, the pattern was similar: K-means successfully clustered animals into two main lactation stages, grouping buffaloes that exhibited high daily milk yield and lactose content with lower fat and protein content. SCS also followed this trend, increasing in the late lactation cluster.

Consistent with clustering analysis, the ANOVA results showed significant differences among clusters for most of the evaluated features. In Herds 1 and 2, similar trends were observed, except for parity order and age at first calving, which did not differ significantly. A possible explanation is that these variables contributed less to cluster discrimination.

In Herd 3, the K-means analysis identified three distinct clusters, consistently separating animals according to lactation stage. In this case, lactation was divided into early, mid-, and late stages based on DIM. As in the other herds, animals in each lactation stage exhibited different milk production patterns. This result clearly reflected the typical lactation curve [27], where milk yield increased progressively during early lactation (86 DIM; 9.95 kg), remained stable during mid-lactation (122 DIM; 10.07 kg), and gradually declined in late lactation (219 DIM; 6.18 kg). The early-lactation cluster also showed higher lactose concentrations, consistent with the osmotic role of lactose in milk secretion [28]. Furthermore, an inverse relationship between lactose and SCS was observed, confirming previous findings [29]. Interestingly, Cluster 1 grouped older animals compared with Clusters 2 and 3 (5.82, 3.69, and 3.69 years, respectively), and these animals also exhibited the highest total lactation milk yield. This association aligned with physiological expectations, as the genetic potential for milk production is typically fully expressed from the second lactation onward [30]. The ANOVA for Herd 3 revealed statistically significant differences among clusters for almost all traits.

4.3. Practical Application

The results obtained with the K-means algorithm were consistent with a recent study [31], which demonstrated the effectiveness of clustering techniques in capturing meaningful variation in dairy production traits. K-means successfully differentiated physiological stages based on milk yield and composition in dairy cows, with cluster separation typically occurring after the lactation peak, between 70 and 90 DIM [32]. Despite this, it is important to highlight that the results obtained are valid for the Italian Mediterranean buffaloes, which are the only breed included in the dataset. Other buffalo breeds may exhibit different productive and reproductive performances, and therefore the results could differ [33]. However, since the data used in the present analysis is readily available and commonly recorded in the livestock sector, the proposed methodology could also be applied to other breeding systems and species. In a study by Trapanese et al. [34] a similar approach was applied to Saanen and Camosciata goat breeds, where, as in the present study, DIM strongly influenced cluster formation. These findings suggested that the proposed approach can provide valuable insights not only for Italian buffalo breeding but also for other livestock species. From a practical perspective, understanding the physiology of lactation is essential for evaluating herd genetic potential, managing reproductive and nutritional strategies and optimizing overall milk yield. Clustering analysis offers a data-driven framework for grouping animals according to physiological and production parameters rather than relying solely on farmer experience or intuition. This approach is particularly relevant to nutritional management [35]. It is well established that animals at different DIM have distinct energy and nutrient requirements, which should be addressed using precision feeding strategies. Our findings indicate that K-means clustering can support more precise and objective animal grouping based on routinely collected data. Moreover, according to

Kalantari et al. [36], data-driven grouping strategies can improve income over feed costs by simultaneously increasing milk output and reducing unnecessary feed expenditures, resulting in a significant positive economic impact.

5. Conclusions

This study proposed an automated approach to exploring routine buffalo milk data, offering a potential tool to support and enhance herd management practices. Although the evaluation indices were suboptimal, the K-means clustering produced coherent groupings that reflected the physiological patterns of lactation. One of the main advantages of this method lies in the accessibility of the data, which are routinely collected by farmers, technicians and automated systems, thereby eliminating the need for additional data collection efforts. The findings suggested that clustering techniques, particularly K-means, can serve as a foundation for modernizing buffalo breeding strategies. By grouping buffaloes according to lactation stage and performance profiles, it may be possible to optimize nutritional plans, improve resource allocation and enhance overall productivity. Looking ahead, this methodology could be further refined by integrating detailed information on milk protein and lipid fractions in relation to feeding regimes. Such integration would enable a more precise characterization of milk composition dynamics across lactation stages and nutritional contexts, ultimately contributing to the design of diet plans tailored to maximize both milk quality and animal health. Finally, the recent adoption of automatic milking systems could serve as a continuous data source, providing daily milk records. Such detailed and frequent data could improve clustering performance and allow the use of more advanced or alternative clustering algorithms.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ani15223246/s1>, Figure S1: Correlation matrix of the combined dataset. Figure S2: Density plot of Days in Milk among clusters.

Author Contributions: Conceptualization, L.T., N.P. and A.S.; methodology, N.P.; software, L.T.; validation, M.S., G.B. and G.C.; data curation, L.T. and M.S.; writing—original draft preparation, L.T. and A.S.; writing—review and editing, A.S., N.P. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1032 17/06/2022, CN00000022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gong, L.; Luo, L.; Gao, J.; Xiong, Y.; Chen, C.; Gan, H.; Song, H.; Morrone, S.; Dimauro, C.; Gambella, F.; et al. Industry 4.0 and Precision Livestock Farming (PLF): An up to Date Overview across Animal Productions. *Sensors* **2022**, *22*, 4319. [[CrossRef](#)]
2. Peyraud, J.-L.; MacLeod, M. *Future of EU Livestock: How to Contribute to a Sustainable Agricultural Sector?* Final Report; European Commission: Brussels, Belgium, 2020.
3. Trapanese, L.; Bifulco, G.; Calanni Macchio, A.; Aragona, F.; Purrone, S.; Campanile, G.; Salzano, A. Precision livestock farming applied to the dairy sector: 50 years of history with a text mining and topic analysis approach. *Smart Agric. Technol.* **2025**, *10*, 100827. [[CrossRef](#)]
4. Pugliese, R.; Regondi, S.; Marini, R. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Sci. Manag.* **2021**, *4*, 19–29. [[CrossRef](#)]

5. Le Duc, T.; Leiva, R.G.; Casari, P.; Östberg, P.O. Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey. *ACM Comput. Surv.* **2019**, *52*, 94. [[CrossRef](#)]
6. Can Machine Learning Algorithms Perform Better than Multiple Linear Regression in Predicting Nitrogen Excretion from Lactating Dairy Cows | Scientific Reports. Available online: <https://www.nature.com/articles/s41598-022-16490-y> (accessed on 1 August 2025).
7. Wang, J.; Bell, M.; Liu, X.; Liu, G. Machine-Learning Techniques Can Enhance Dairy Cow Estrus Detection Using Location and Acceleration Data. *Animals* **2020**, *10*, 1160. [[CrossRef](#)]
8. Tremblay, M.; Kammer, M.; Lange, H.; Plattner, S.; Baumgartner, C.; Stegeman, J.A.; Duda, J.; Mansfeld, R.; Döpfer, D. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. *J. Dairy Sci.* **2018**, *101*, 7311–7321. [[CrossRef](#)]
9. Franceschini, S.; Grelet, C.; Leblois, J.; Gengler, N.; Soyeurt, H. Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? *J. Dairy Sci.* **2022**, *105*, 6760–6772. [[CrossRef](#)]
10. Matera, R.; Pierro, F.; Santinello, M.; Iraci Fuintino, A.; Pacelli, G.; Norton, T.; Neglia, G. Precision livestock farming in buffalo species: A sustainable approach for the future. *Smart Agric. Technol.* **2025**, *11*, 101060. [[CrossRef](#)]
11. Warner, D.; Vasseur, E.; Lefebvre, D.M.; Lacroix, R. A machine learning based decision aid for lameness in dairy herds using farm-based records. *Comput. Electron. Agric.* **2020**, *169*, 105193. [[CrossRef](#)]
12. Costa, A.; Negrini, R.; De Marchi, M.; Campanile, G.; Neglia, G. Phenotypic Characterization of Milk Yield and Quality Traits in a Large Population of Water Buffaloes. *Animals* **2020**, *10*, 327. [[CrossRef](#)]
13. Ali, A.K.A.; Shook, G.E. An Optimum Transformation for Somatic Cell Concentration in Milk1. *J. Dairy Sci.* **1980**, *63*, 487–490. [[CrossRef](#)]
14. Brotzman, R.L.; Cook, N.B.; Nordlund, K.; Bennett, T.B.; Gomez Rivas, A.; Döpfer, D. Cluster analysis of Dairy Herd Improvement data to discover trends in performance characteristics in large Upper Midwest dairy herds. *J. Dairy Sci.* **2015**, *98*, 3059–3070. [[CrossRef](#)]
15. Abreu, B.D.S.; Barbosa, S.B.P.; Silva, E.C.D.; Santoro, K.R.; Batista, Â.M.V.; Martinez, R.L.V. Principal component and cluster analyses to evaluate production and milk quality traits. *Rev. Ciéncia Agronómica* **2020**, *51*, e20196977. [[CrossRef](#)]
16. Çelik, M.; Dadaşer-Çelik, F.; Dokuz, A. Anomaly detection in temperature data using DBSCAN algorithm. In Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15–18 June 2011; pp. 91–95.
17. Atif, M.; Farooq, M.; Abiad, M.; Shafiq, M. The least sample size essential for detecting changes in clustering solutions of streaming datasets. *PLoS ONE* **2024**, *19*, e0297355. [[CrossRef](#)]
18. 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*, 1–36. [[CrossRef](#)]
19. 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. [[CrossRef](#)]
20. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [[CrossRef](#)]
21. Wong, P.C. Unsupervised Machine Learning. In *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python*; von Davier, A.A., Mislevy, R.J., Hao, J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 173–193. ISBN 978-3-030-74394-9.
22. Lee, M.; Lee, S.; Park, J.; Seo, S. Clustering and Characterization of the Lactation Curves of Dairy Cows Using K-Medoids Clustering Algorithm. *Animals* **2020**, *10*, 1348. [[CrossRef](#)]
23. Martin-Collado, D.; Byrne, T.J.; Amer, P.R.; Santos, B.F.S.; Axford, M.; Pryce, J.E. Analyzing the heterogeneity of farmers' preferences for improvements in dairy cow traits using farmer typologies. *J. Dairy Sci.* **2015**, *98*, 4148–4161. [[CrossRef](#)]
24. Khan, I.K.; Daud, H.B.; Zainuddin, N.B.; Sokkalingam, R.; Farooq, M.; Baig, M.E.; Ayub, G.; Zafar, M. Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm. *Egypt. Inform. J.* **2024**, *27*, 100504. [[CrossRef](#)]
25. Silvestre, A.M.; Martins, A.M.; Santos, V.A.; Ginja, M.M.; Colaço, J.A. Lactation curves for milk, fat and protein in dairy cows: A full approach. *Livest. Sci.* **2009**, *122*, 308–313. [[CrossRef](#)]
26. Kirsanova, E.; Heringstad, B.; Lewandowska-Sabat, A.; Olsaker, I. Alternative subclinical mastitis traits for genetic evaluation in dairy cattle. *J. Dairy Sci.* **2019**, *102*, 5323–5329. [[CrossRef](#)]
27. Luna-Palomera, C.; Domínguez-Viveros, J.; Aguilar-Palma, G.N.; Castillo-Rangel, F.; Sánchez-Dávila, F.; Macías-Cruz, U.; Luna-Palomera, C.; Domínguez-Viveros, J.; Aguilar-Palma, G.N.; Castillo-Rangel, F.; et al. Analysis of the Lactation Curve of Murrah Buffaloes with Mixed Non-Linear Models. *Chil. J. Agric. Anim. Sci.* **2021**, *37*, 200–208. [[CrossRef](#)]
28. Gargiulo, J.I.; Garcia, S.C.; Hovey, R.C. Sources of variation underlying the production of lactose by dairy cows. *J. Dairy Sci.* **2025**, *108*, 4403–4421. [[CrossRef](#)]
29. Alessio, D.; Velho, J.; Mcmanus, C.; Knob, D.; Vancin, F.; Antunes, G.; Busanello, M.; Carli, F.; Thaler-Neto, A. Lactose and its relationship with other milk constituents, somatic cell count, and total bacterial count. *Livest. Sci.* **2021**, *252*, 104678. [[CrossRef](#)]

30. Cattaneo, L.; Piccioli-Cappelli, F.; Minuti, A.; Trevisi, E. Metabolic and physiological adaptations to first and second lactation in Holstein dairy cows. *J. Dairy Sci.* **2023**, *106*, 3559–3575. [[CrossRef](#)]
31. Rebuli, K.B.; Ozella, L.; Vanneschi, L.; Giacobini, M. Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation periods. *Comput. Electron. Agric.* **2023**, *211*, 108002. [[CrossRef](#)]
32. Ghavi Hossein-Zadeh, N. Comparison of non-linear models to describe the lactation curves for milk yield and composition in buffaloes (*Bubalus bubalis*). *Animal* **2016**, *10*, 248–261. [[CrossRef](#)] [[PubMed](#)]
33. Minervino, A.H.H.; Zava, M.; Vecchio, D.; Borgheese, A. *Bubalus bubalis*: A Short Story. *Front. Vet. Sci.* **2020**, *7*, 570413. [[CrossRef](#)]
34. Trapanese, L.; Bifulco, G.; Aragona, F.; GianMaria, P.; Pedota, G.; Pasquino, N.; Salzano, A. Explorative analysis of Saanen and Camosciata goats data through an unsupervised machine learning approach. In Proceedings of the 2024 International Workshop Measurements and Applications Inveterinary and Animal Sciences, Torino, Italy, 22–24 April 2024.
35. Barrientos-Blanco, J.A.; White, H.; Shaver, R.D.; Cabrera, V.E. Graduate Student Literature Review: Considerations for nutritional grouping in dairy farms. *J. Dairy Sci.* **2022**, *105*, 2708–2717. [[CrossRef](#)]
36. Kalantari, A.S.; Armentano, L.E.; Shaver, R.D.; Cabrera, V.E. Economic impact of nutritional grouping in dairy herds. *J. Dairy Sci.* **2016**, *99*, 1672–1692. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.