SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Exploring K-Means Clustering Efficiency: Accuracy and Computational Time across Multiple Datasets

Iliyas Karim Khan[1,*], Hanita Daud[1], Nooraini Zainuddin[1], Rajalingam Sokkalingam[1], Abdussamad[1], Abdus Samad Azad[1], Mudassar Iqbal[3], Mudasar Zafar[2], Atta Ullah[1], Musarat Elahi[4], Ahmad Abubakar Suleiman[1]

[1] Fundamental and Applied Sciences Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia
[2] School of Mathematics, Actuarial and Quantitative Studies (SOMAQS), Asia Pacific University of Technology & Innovation (APU), Bukit Jalil, 57000 Kuala Lumpur, Malaysia
[3] Department of Mathematical Sciences Faculty of Basic Sciences, Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Quetta 87300, Pakistan
[4] Shaheed Benazir Bhutto Women University Peshawar, Khyber Pakhtunkhwa 00384, Pakistan

ABSTRACT

In the realm of unsupervised machine learning, clustering stands as a pivotal method in data analysis. However, it grapples with challenges arising from diverse datasets, leading to certain algorithms displaying reduced effectiveness or prolonged execution times on specific data types. The performance of each clustering algorithms depends on both the dataset's sample size and its specific characteristics. Among these algorithms, K-means clustering stands out as a popular choice. It is essential to evaluate its accuracy levels and execution times across various datasets with different sample sizes and features. This paper assesses the precision and efficiency of the K-means clustering algorithm on three distinct datasets, namely seed data, iris data and well log data sourced from GitHub, each characterized by variations in both size and features. The Seed dataset represents three different varieties of wheat seeds, Iris dataset represents measurements of three different iris flowers species and Well log dataset represents Sonic log and Gamma ray data respectively. The aim is to analyse how accurate and efficient K-means algorithm performs across these data sets. The results show that K-means algorithm produces high accuracy and lower computational time to the Well log dataset.

Keywords:

Accuracy; efficiency; k-mean clustering; algorithm and dataset

1. Introduction

Clustering analysis belongs to the realm of unsupervised learning technique. The goal of clustering analysis is to categorize unlabelled data objects into groups where those within the same groups share high similarity, while those in different groups show low similarity [1,2]. Among the available clustering algorithms [3], k-means clustering stands out as one of the most utilized techniques, largely because of its simplicity, efficiency and effectiveness. Identifying inherent

* *Corresponding author.*
E-mail address: iliyas_22008363@utp.edu.my

https://doi.org/10.37934/araset.65.1.113

patterns within a given dataset through clustering is a widely favoured approach utilized across diverse fields such as psychology [4], biology, pattern recognition [5,6], image processing [7], computer security [8] and different types of drugs [9-12]. After a clustering algorithm has analysed a dataset and generated partitions, a significant question arises: Does this partition adequately address the dataset's underlying problem? Understanding the rationale behind this question holds considerable importance across numerous contexts. Additionally, there isn't a universally optimal clustering algorithm [13]. As a result, different algorithms or even variations of a single algorithm produces alternative divisions that may not universally be considered optimal in all scenarios [14]. Therefore, to create effective clusters, it's crucial to assess various partition segments and select the one that best fits the data. Additionally, many clustering algorithms cannot autonomously determine the number of inherent clusters in the dataset, requiring the specification of the k parameter. Typically, the approach involves running the algorithm multiple times with different k-values. Each resulting partition is then evaluated to determine the most suitable fit for the provided data. Clustering accuracy refers to how closely the generated clusters align with the true structure or ground truth present in the dataset [15]. Assessing the coherence of data points within clusters is crucial, often measured using metrics like the adjusted rand index and silhouette score. However, validating unsupervised results poses challenges due to the absence of labelled ground truth for gauging cluster accuracy. Ambiguity arises from differing cluster shapes, sizes and densities, making it challenging to determine the quality of clusters accurately. Furthermore, accurately evaluating clusters are complicated by sensitivity to algorithm parameters and the choice of evaluation metrics [16]. K-Means, is an extensively studied clustering method, focuses on reducing the overall variance within clusters [17]. Its popularity in diverse fields is attributed to its simplicity and efficiency in clustering tasks. However, a notable limitation of the widely employed K-means algorithm is its requirement for a predetermined number of clusters, K [18].

In prior studies, there is a deficiency in comparisons addressing both the accuracy and execution time of the K-means clustering algorithm across a range of datasets. Hence, this paper focuses on a detailed examination and comparison of the accuracy and execution time of the K-means clustering algorithm, utilizing three different datasets. In the examination of this paper, it is observed that K-means clustering lacks universal applicability across all datasets and its utilization has been restricted when applied to various datasets. However, this paper aims to utilize the K-means clustering algorithm across three distinct datasets that possess varying features and different numbers of observations.

This paper primarily focuses on enhancing accuracy while minimizing the time required for execution.

i. The accuracy of the k-means clustering algorithm tends to enhance with an increase in the sample size.
ii. The k-means clustering algorithm reveals a negative correlation between number of features and execution time.

This article is structured as follows:

i. Section 2 provides a literature review, focusing on clustering, especially K-means clustering.
ii. Section 3, the methodology explores the intricacies of K-means clustering, including a flowchart, equations and a step-by-step process.

iii. Section 4 presents the results and discussion, highlighting the accuracy and execution time performances obtained by applying diverse datasets to clustering methods.
iv. Section 5 encompasses concluding remarks of the paper.

2. Literature Review

This section examines pertinent studies related to clustering algorithms, with a specific emphasis on the K-means clustering algorithm. The objective of this article is to conduct a thorough analysis of studies, methodologies and advancements associated with K-means clustering, investigating its applications across diverse domains. This comprehensive review will shed light on the strengths, limitations and potential improvements associated with the K-means algorithm in the context of clustering analysis. Studied a new GPU based K−means; ASB-K−means is introduced which is faster than current GPU based k-means algorithms [19,20]. A Centroids-Guided Deep Multi-View K-means method linking DL with MVC is proposed by the paper. This centres on cluster centroids to help it with deep representation learning giving the K-means friendly representations. Its effectiveness in multi-view task is evaluated by showing that this approach leads to improved clustering and closer-to-cluster-semantic matching of representations across datasets [21]. This study evaluated the Kernelized Rank Order Distance (KROD) method for converting non-spherical data into spherical form using various datasets. By combining a rank order distance (ROD) equation with a Gaussian kernel, KROD calculated distances and assigned weights to data points for spherical transformation. Pairwise similarities were weighted using the Gaussian kernel, while actual distances were computed with ROD, capturing both global and local structures. Numerical results showed that increasing the sample size improved KROD's effectiveness in accurately transforming non-spherical data into spherical form [22]. The paper explored big data's importance in case clustering, proposing an improved Mahalanobis Distance-based K-Means scheme with enhanced precision for clustering similar data [23]. The ratio-cut polytope, crucial in K-means and spectral clustering, analysed algorithmic therefore a new linear programming relaxation for K-means consistently outperformed prior guarantees, demonstrating superior cluster recovery in experiments [24]. The examination of the paper revealed that Hierarchical++ exhibited better performance when compared to conventional hierarchical methods (such as single-link, complete-link, etc.), as well as K-means and K-means++ [25]. The document overviewed hierarchical clustering in astronomy, tracing its origins, discussing its applications across astronomical scales and explaining its role in revealing celestial hierarchies while classifying objects. It elaborated on algorithm functionalities, limitations and contributions to reliable astronomical discoveries [26]. The paper introduced HMC (Hierarchical Means Clustering), a novel technique using nested partitions and least squares to minimize within-cluster deviance across n partitions, resulting in a cascade of (n-1) divisions. Six case studies compared HMC to established hierarchical clustering algorithms like k-means, Ward's method and Bisecting k-means [27]. The article proposed a novel hierarchical clustering method to address small file issues in Hadoop Distributed File Systems (HDFS). Using Dendrogram analysis, it recommended efficient consolidation strategies, successfully identifying and merging seven specific files in a simulation of 100 CSV files. This demonstrated its effectiveness in enhancing file management efficiency [28]. The paper introduced HY-DBSCAN, a parallel DBSCAN algorithm incorporating a modified KD-tree, grid-based spatial indexing and a distributed merging scheme. It outperformed existing solutions up to 2048 cores, leveraging process and thread parallelization for implementing DBSCAN on scientific datasets [29]. The paper improved DBSCAN for efficient clustering in polygonal-shaped databases by reducing computation costs through targeted sampling. It demonstrated superior speed compared to recent approaches, with only a slight accuracy reduction from traditional DBSCAN [30]. The paper

introduced STRP-DBSCAN, a parallel method for clustering spatial-temporal trajectory data, reducing clustering time by up to 96.2%. It also presented PER-SAC, a deep reinforcement learning-based technique for tuning DBSCAN parameters, achieving 8.8% better accuracy than other strategies [31]. The paper introduced a modified DBSCAN algorithm for detecting anomalies in seasonally correlated time-series data, outperforming conventional DBSCAN by 2.16% in identifying abnormalities. The refined method efficiently detected both inter-annual and within-year anomalies, showcasing higher efficiency in detecting local anomalies in seasonal data [32]. Previous research saw a surge in high-dimensional datasets due to increased data volumes. Bisecting K-Means struggled as dimensions grew. To fix this, we combined stability-based measures and MSE in CHB-K-Means. Experiments, altering outlier detection, showed a 75% average clustering accuracy and decreased computation time with more outliers detected [33,34]. The paper proposed a modified DBSCAN for detecting anomalies in seasonally correlated time-series data, outperforming conventional DBSCAN by 2.16%. The refined method efficiently detected inter-annual and within-year anomalies, demonstrating higher efficiency in detecting local anomalies in seasonal data [35]. The study enhanced K-means++ by integrating variance from probability and statistics. Initial centres were selected based on minimum variance among high-density samples and subsequent centres used a weighted D2 method. Experimental results demonstrated increased accuracy and improved stability [36,37]. The paper analysed the issue of soil clustering and the spatial representation of results obtained from in-situ measurements of soil's physical and chemical traits. It adapted the K-means and fuzzy K-means algorithms for soil data clustering, utilizing a database of soil samples collected in Montenegro for comparative analysis. The classified soil data were displayed on a static Google map for visualization [38,39]. Evolutionary K-Means (EKM) merged K-Means with genetic algorithms, autonomously selecting parameters during partition evolution. While effective for distinct clusters, EKM struggled with noise. To improve, combined EKM with clustering stability-based analysis. The novel CSEKM method used matrices to capture clustering tendencies, enhancing robustness to noise across various datasets [40].

In the context of existing literature, there is a noticeable gap in studies that specifically examine the accuracy and execution times of the K-means clustering algorithm across diverse datasets featuring varying sample sizes and features.

3. Methodology

In this section, the paper explores the analysis of the K-means clustering algorithm, its associated flowchart and the process employed in clustering datasets with K-means. The primary goal is to assess the accuracy and quantify the execution time of K-means clustering algorithms.

3.1 K-Mean Clustering Algorithm

K-means clustering stands out as a top algorithm in the realm of unsupervised machine learning. The clustering process is visually represented in the flowchart below.

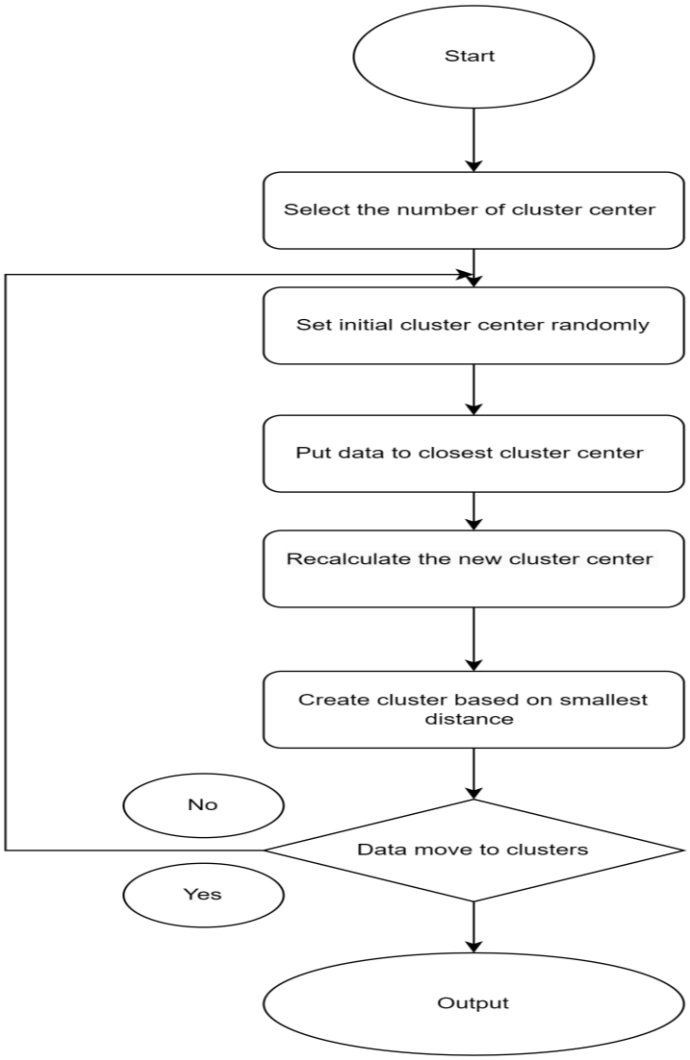


Fig. 1. Flow chart of k-mean clustering algorithm

The process begins with setting initial centroids for the clusters, serving as reference points. Subsequently, the data points are grouped into their respective clusters based on proximity to these predetermined centroids [18,41,42]. This assignment involves multiple sequential steps as outlined below.

Algorithm: Assigning data points to clusters

Given a collection of n data points represented by D and a set of k centroids denoted by C, where D contains elements d_1 to d_n and C includes elements c_1 to c_k, the provided input defines the dataset and the centroid references.

Input:

D = { $d_1, d_2, d_3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots d_n$} // set of n data points

C = { $c_1, c_2, c_3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots c_k$} // set of k centeriods

A set of k canter

Steps:

1. Calculate the distance between each data points d_i (1<=i<=n) to all the centroid c_j (1<=j<=k) as d(d_i , c_j);
2. Identify the closest centroid c_j for each data point d_i and allocate di to the corresponding cluster j.

3. Assign the cluster Id[i] =j, //j:Id of the closest cluster
4. The nearest Distance D(i)= d (d_i, c_j).
5. Each cluster j (I<=J<=k), recompute the centroids.
6. **Repeat**
7. For each datapoints d_i,

Determine the distance from the data point to the centroid of the currently nearest cluster.

Should this distance be less than or equal to the current closest distance, the data point remains assigned to the cluster.

Else

Calculate the distance, denoted as d(d_i, c_j), between each centroid c_j (I<=J<=k) and the data point d_i.

End for;

Allocate the data point d_i to the cluster possessing the closest centroid c_j.

Set clusterId[i]=j;

Set nearest distance[i]= d(d_i, c_j)

End for;

8. Iteratively recompute the centroids for each cluster, j(I<=j<=k), **until** the convergence condition is satisfied.

The iteration persists through steps 2 and 5 until stability is reached. Stability is achieved when the centroids exhibit minimal to no further change or after a specific number of iterations. Consequently, the outcome comprises clusters along with their individual centroids, signifying the arrangement of similar data points. This iterative method aims to reduce the total variance within clusters or the squared distances of data points to their respective centroids, ensuring the formation of coherent and distinct clusters [43]. In the above step 2 we assigned the data point by using Eq. (1).

$$C_i = arg.min_j \parallel x_i - \mu_j \parallel^2 \quad (1)$$

Where C_i: cluster to which data points. x_i μ_j: centroid of clusters. $\parallel x_i - \mu_j \parallel^2$: Euclidean distance

Updating cluster centroid by applying the formula as in Eq. (2).

$$\mu_j = \frac{1}{\|C_j\|} \sum_{x_i \subseteq c_j} X_i \quad (2)$$

Where, $\sum_{x_i \subseteq c_j} X_i$: summation of all data points in clusters j

The process will be iterative and convergence will be achieved when the assignments and centroid stop changing or if a stopping criterion is reached. K-means does not present a general equation which is also the case with linear regression and others [44,45].

4. Results and Discussion

In this section, the research provides an overview of the datasets, evaluation criteria, clustering algorithms targeted for optimization, initial methodologies used as a benchmark and the configurations of parameters applied in the study.

4.1 Datasets

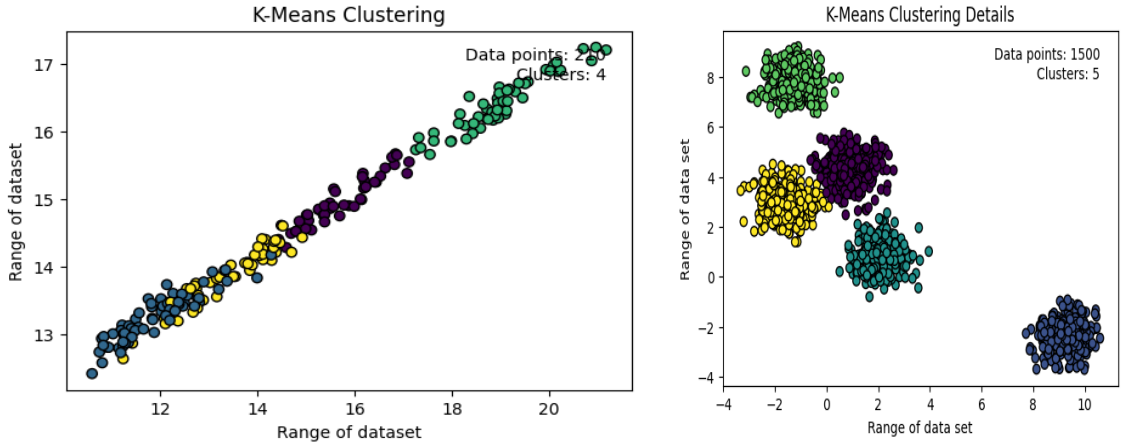
Table 1 summarizes three distinct datasets: seed data, iris data and well log data sourced from GitHub, each varying in size and features. The table includes details such as mean, standard deviation, kurtosis, symmetry and the total number of observations for each dataset. The seed dataset includes three types of wheat seeds: Kama, Rosa and Canadian each with 210 observations. The standard deviation of the Kama seeds is higher compared to the other two varieties. All three types of data exhibit a platykurtic distribution, indicated by a kurtosis value below 3. Kama and Rosa seeds display positive skewness, while Canadian seeds show negative skewness. The Iris dataset comprises three species: Iris setosa, Iris versicolor and Iris virginica, each with 1500 observations. The standard deviations of both datasets are moderate. Both setosa and versicolor datasets exhibit a slightly platykurtic pattern, indicated by a kurtosis below 3 and all skewness values greater than 0 indicate positive skewness. The well log dataset includes gamma ray data and Sonic data, comprising 2435 observations each. There's a disparity in the means of the two logs: the gamma ray log has a mean of 26.60, while the sonic log has a mean of 0.06. Both datasets display a platykurtic pattern and demonstrate positive skewness in their distributions.

Table 1
The description of dataset

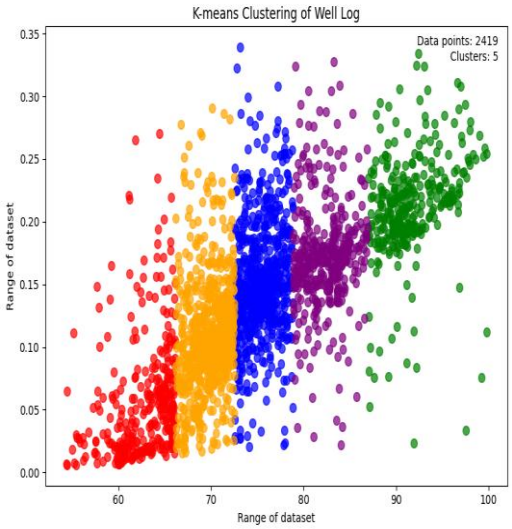
Measure	Seed Data			Iris Data			Well Log Data	
	Kama	Rosa	Canadian	Setosa	versicolor	virginica	Gamma Ray	Sonic log
Mean	14.84	14.55	0.87	5.76	3.07	3.46	44.72	0.06
S. D	2.90	1.30	0.02	0.80	0.53	1.71	26.60	0.064
Kurtosis	-1.08	-1.10	-0.14	-0.19	-0.16	-1.43	0.72	1.131
Skewness	0.39	0.38	-0.53	0.46	0.31	0.13	1.08	1.21
Total observations	210			1500			2435	

4.2 K-Means Clustering Algorithm Across Different Datasets

The K-means unsupervised machine learning algorithm's performance is assessed across diverse datasets, revealing a correlation between the number of clusters and the data diversity and size. Larger and more diverse datasets often require a greater number of clusters. K-means clustering leverages similarity and prevalence within the data to form clusters. Figure 2 visually represents K-means clustering applied to diverse datasets, showcasing how the algorithm organizes data based on their similarities and distributions.



(a) (b)



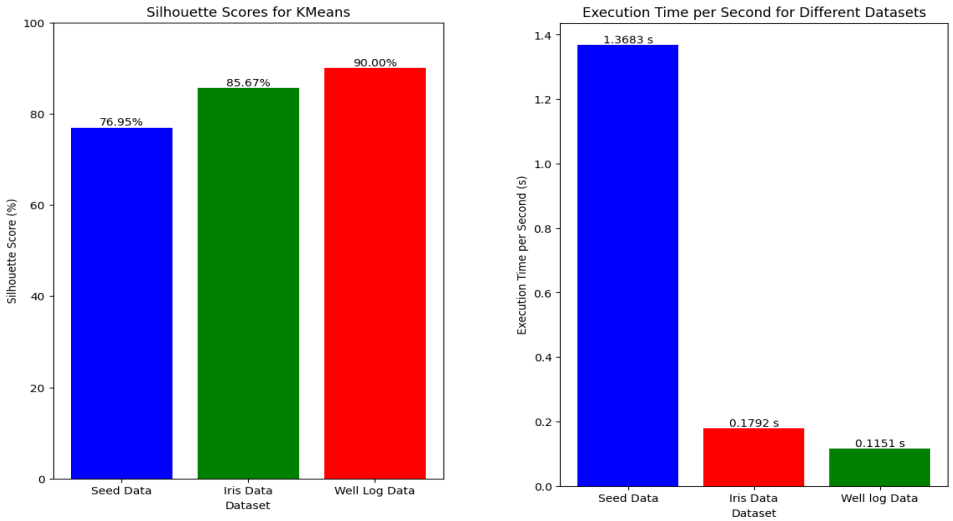
(c)

Fig. 2. The K-means clustering on the seed, iris and well log datasets (a) K-means clustering using seed datasets (b) K-means clustering using Iris dataset (c) K-means clustering using well log data

Figure 2(a) illustrates the application of the K-means clustering algorithm to select clusters based on seed data. In this scenario, the seed dataset is structured in a way that distributes the data in a linear formation within the clusters. The algorithm operates by iteratively assigning data points to the nearest cluster centroid and refining these assignments until convergence. Moving on to Figure 2(b), it showcases K-means clustering applied to the Iris dataset. Here, the K-means algorithm organizes observations into circular clusters, each represented by distinct colours. The Iris dataset, known for its floral species classification, demonstrates how K-means can delineate different species into cohesive clusters, aiding in their visual differentiation. Figure 2(c) depicts K-means clustering applied to well log data. These well log data, known for their quality and significance in various analyses, showcase a linear formation of clusters. The abundance of observations in the Well log dataset allows K-means to effectively delineate the data points into well-defined clusters, offering valuable insights into the underlying structure of the data. Each representation highlights how the K-means clustering algorithm adapts to different datasets, shaping clusters based on the inherent distribution and characteristics of the data, whether linear, circular or otherwise, aiding in pattern identification and analysis.

4.3 The Accuracy and Computational Efficiency of K-Means Clustering Algorithm

Evaluating the accuracy and efficiency of K-means clustering across varied datasets is crucial. The objective is to assess its performance on different datasets and observe variations in both accuracy and execution time. The figures below showcase this evaluation, aiming to quantify the accuracy levels achieved and the time taken for execution across various datasets. Each dataset introduces distinctive results in terms of accuracy and execution time, providing insights into how well K-means adapts to different data structures. These assessments serve to highlight the variability in K-mean performance across diverse datasets, shedding light on its efficacy and computational efficiency under varying circumstances.



(a) (b)

Fig. 3. Accuracy and Execution Time of the algorithm (a) Accuracy in the K-means clustering algorithm (b) Execution time in the K-means clustering algorithm

In Figure 3(a), the accuracy assessment of the k-means clustering algorithm demonstrates a 76.95% accuracy level for the seed dataset, 86.67% for the Iris dataset and a remarkable 90.00% accuracy for the well log data. Now, focusing on Figure 3(b), which illustrates the execution times of the k-means clustering algorithm across three diverse datasets with varying sample sizes. For the seed dataset, the clustering process takes 1.3683 seconds. The clustering of the Iris dataset using k-means requires only 0.1792 seconds, while for the well log dataset, it takes a mere 0.1151 seconds. These results underscore that the k-means algorithm showcases superior performance in both accuracy and execution times when applied to the well log data.

4.4 The Precision and Effectiveness in Carrying Out the K-Means Clustering Algorithm

In Table 2, we can find accuracy and execution time metrics for the K-means clustering algorithm applied to three distinct datasets: Seed, Iris and Well log. Within this table, the Seed dataset, configured with 4 clusters, attained an accuracy level of 76.95%, accompanied by an execution time of 1.3683 seconds per iteration. The Iris dataset, employing 5 clusters, attained an accuracy of 86.67% with an execution time of 0.1792 seconds. Notably, the Well Log dataset, utilizing 6 clusters, demonstrated an exceptional 90.00% accuracy within a mere 0.1151 seconds per iteration.

Table 2
The accuracy and execution time of k mean clustering algorithms

Dataset	Number of clusters	Accuracy (%)	Execution Times/sec
Seed	4	76.95	1.3683
Iris	5	86.67	0.1792
Well Log	5	90.00	0.1151

4.5 Discussion

Clustering stands as a pivotal component in unsupervised machine learning, pivotal for handling varied and diverse datasets. Its primary aim is to group similar data points together, laying the groundwork for deeper analysis. With a multitude of clustering algorithms available, selecting the

most appropriate method for different datasets presents a significant challenge. The primary focus of this research is the investigation of the K-means clustering algorithm, utilizing three distinct datasets: the seed dataset, Iris dataset and a well log dataset acquired from GitHub. The table 1 presents descriptive statistics of the three datasets, outlining the count of observations, mean, standard deviation, kurtosis and skewness for each dataset. The K-means clustering algorithm is applied to all the datasets, as illustrated in Figure 1. The results of the clustering process are depicted in Figure 2, providing a detailed representation of the obtained clustering outcomes. Figure 2 visually represents our evaluation of accuracy and execution time across a range of datasets through the application of the K-means clustering algorithm. A comprehensive summary of all the findings is presented in detail within Table 2. Research findings reveal that K-means performs notably well in clustering the well log dataset, achieving a remarkable accuracy of 90.00% with an execution time of 0.1151 seconds. This study highlights two significant observations: K-means clustering is better suited for managing larger datasets. Moreover, we noted a positive correlation between sample size and accuracy, along with a negative correlation between number of features and execution time.

5. Conclusions

Clustering plays a vital role in unsupervised machine learning, especially when dealing with extensive datasets and intricate feature spaces. The performance of clustering algorithm depends on the nature of dataset. Despite the availability of various clustering algorithms in unsupervised learning, each comes with its own limitations. Among these algorithms, K-means clustering stands as a widely used and efficient method. However, assessing its accuracy and efficiency across diverse datasets remains a significant challenge for researchers. To fill this research gap, the primary objective of this study is to assess and compare the performance of the K-means clustering algorithm across diverse datasets. This study investigates varied outcomes observed when applying the K-means clustering algorithm to datasets, including Seed, Iris and well log, each characterized by unique sample sizes and attributes. The goal is to evaluate how accurately and efficiently the K-mean clustering algorithm performs across various datasets. The findings emphasize the outstanding performance of the K-means algorithm, achieving an accuracy of 90.00% and execution times of 0.1151 seconds when applied to the well log dataset. This highlights a positive correlation between sample size and enhanced accuracy. Moreover, we noted a positive correlation between sample size and accuracy, along with a negative correlation between number of features and execution time.

Acknowledgments

This project received funding from YUTP grant with cost centre 015LC0-432 and Centre of Graduate Studies (CGS) at Universiti Teknologi PETRONAS, Malaysia. The authors extend their appreciation to the unnamed reviewers and the editor for their thorough assessment of this paper, as well as for their invaluable recommendations and insights provided in their comments.

References

[1] Li, Wenjun, Zikang Wang, Wei Sun and Sara Bahrami. "An ensemble clustering framework based on hierarchical clustering ensemble selection and clusters clustering." *Cybernetics and Systems* 54, no. 5 (2023): 741-766. https://doi.org/10.1080/01969722.2022.2073704

[2] Li, Hongmin, Xiucai Ye, Akira Imakura and Tetsuya Sakurai. "LSEC: Large-scale spectral ensemble clustering." *Intelligent Data Analysis* 27, no. 1 (2023): 59-77. https://doi.org/10.3233/IDA-216240

[3] Hamid, Hamzah Abdul, Yap Bee Wah, Khatijahhusna Abdul Rani and Xian Jin Xie. "The Effect Of Divisive Analysis Clustering Technique on Goodness-Of-Fit Test for Multinomial Logistic Regression." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 48, no. 2 (2024): 39-48. https://doi.org/10.37934/araset.48.2.3948
[4] Shanmugam, Gowri, Tamilvizhi Thanarajan, Surendran Rajendran and Sadish Sendil Murugaraj. "Student Psychology based optimized routing algorithm for big data clustering in IoT with MapReduce framework." *Journal of Intelligent & Fuzzy Systems* 44, no. 2 (2023): 2051-2063. https://doi.org/10.3233/JIFS-221391
[5] Li, Yang, Mingcong Wu, Shuangge Ma and Mengyun Wu. "ZINBMM: a general mixture model for simultaneous clustering and gene selection using single-cell transcriptomic data." *Genome Biology* 24, no. 1 (2023): 208. https://doi.org/10.1186/s13059-023-03046-0
[6] Singh, Surender and Koushal Singh. "Novel fuzzy similarity measures and their applications in pattern recognition and clustering analysis." *Granular Computing* 8, no. 6 (2023): 1715-1737. https://doi.org/10.1007/s41066-023-00393-y
[7] Flores, Marco A., Fernando E. Serrano, Carlos Cadena and Jose C. Alvarez. "Thermographic image processing analysis in a solar concentrator with hard C-means clustering." *Energy Reports* 9 (2023): 312-321. https://doi.org/10.1016/j.egyr.2023.05.261
[8] Kiran, Ajmeera, Prasad Mathivanan, Miroslav Mahdal, Kanduri Sairam, Deepak Chauhan and Vamsidhar Talasila. "Enhancing data security in IoT networks with blockchain-based management and adaptive clustering techniques." *Mathematics* 11, no. 9 (2023): 2073. https://doi.org/10.3390/math11092073
[9] Ullah, Atta, Hamzah Sakidin, Kamal Shah, Yaman Hamed and Thabet Abdeljawad. "A mathematical model with control strategies for marijuana smoking prevention." *Electronic Research Archive* 32, no. 4 (2024): 2342-2362. https://doi.org/10.3934/era.2024107
[10] Ullah, Atta, Afnan Ahmad, Umair Khan and Abdussamad Abdussamad. "A Time-Fractional Model for Brinkman-Type Nanofluid with Variable Heat and Mass Transfer." *City University International Journal of Computational Analysis* 5, no. 1 (2022): 11-30. https://doi.org/10.33959/cuijca.v5i1.56
[11] Ullah, Atta, Hamzah Sakidin, Shehza Gul, Kamal Shah, Mohana Sundaram Muthuvalu, Thabet Abdeljawad and Mudassar Iqbal. "Sensitivity analysis-based validation of the modified NERA model for improved performance." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 3 (2023): 1-11. https://doi.org/10.37934/araset.32.3.111
[12] Ullah, Atta, Hamzah Sakidin, Shehza Gul, Kamal Shah, Yaman Hamed and Thabet Abdeljawad. "Mathematical model with sensitivity analysis and control strategies for marijuana consumption." *Partial Differential Equations in Applied Mathematics* 10 (2024): 100657. https://doi.org/10.1016/j.padiff.2024.100657
[13] Atienza, Casey H., Sean David R. Aggabao, Thrisha Mae T. Banguis, Larie Joseph R. Lacsina, Vianca D. Manalo, Rhiz John P. Susi, Emmanuel T. Trinidad and Lawrence Materum. "MIMO Multipath Component Clustering using k-Deep Autoencoder." *Journal of Advanced Research in Applied Mechanics* 119, no. 1 (2024): 27-36. https://doi.org/10.37934/aram.119.1.2736
[14] Wiroonsri, Nathakhun. "Clustering performance analysis using a new correlation-based cluster validity index." *Pattern Recognition* 145 (2024): 109910. https://doi.org/10.1016/j.patcog.2023.109910
[15] Ahmadinejad, Navid, Yunro Chung and Li Liu. "J-score: A robust measure of clustering accuracy." *PeerJ Computer Science* 9 (2023): e1545. https://doi.org/10.7717/peerj-cs.1545
[16] Li, Qi, Shuliang Wang, Xianjun Zeng, Boxiang Zhao and Yingxu Dang. "How to improve the accuracy of clustering algorithms." *Information Sciences* 627 (2023): 52-70. https://doi.org/10.1016/j.ins.2023.01.094
[17] Alam, Afroj and Muhammad Kalamuddin Ahamad. "K-Means Hybridization with Enhanced Firefly Algorithm for High-Dimension Automatic Clustering." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 33, no. 3 (2023): 137-153. https://doi.org/10.37934/araset.33.3.137153
[18] Pham, Duc Truong, Stefan S. Dimov and Chi D. Nguyen. "Selection of K in K-means clustering." *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219, no. 1 (2005): 103-119. https://doi.org/10.1243/095440605X8298
[19] Li, Mi, Eibe Frank and Bernhard Pfahringer. "Large scale K-means clustering using GPUs." *Data Mining and Knowledge Discovery* 37, no. 1 (2023): 67-109. https://doi.org/10.1007/s10618-022-00869-6
[20] Sobran, Nur Maisarah Mohd and Zool Hilmi Ismail. "A Systematic Literature Review of Unsupervised Fault Detection Approach for Complex Engineering System." *Journal of Advanced Research in Applied Mechanics* 103, no. 1 (2023): 43-60. https://doi.org/10.37934/aram.103.1.4360
[21] Liu, Jing, Fuyuan Cao and Jiye Liang. "Centroids-guided deep multi-view K-means clustering." *Information Sciences* 609 (2022): 876-896. https://doi.org/10.1016/j.ins.2022.07.093
[22] Khan, Iliyas Karim, Hanita Binti Daud, Rajalingam Sokkalingam, Nooraini Binti Zainuddin, Abdussamad Abdussamad, Noor Naheed and Mudassar Iqbal. "Numerical solution by kernelized rank order distance (KROD) for non-spherical

data conversion to spherical data." In *AIP Conference Proceedings*, vol. 3123, no. 1. AIP Publishing, 2024. https://doi.org/10.1063/5.0223847

[23] Brown, Paul O., Meng Ching Chiang, Shiqing Guo, Yingzi Jin, Carson K. Leung, Evan L. Murray, Adam GM Pazdor and Alfredo Cuzzocrea. "Mahalanobis distance based k-means clustering." In *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 256-262. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-12670-3_23

[24] De Rosa, Antonio and Aida Khajavirad. "The ratio-cut polytope and K-means clustering." *SIAM Journal on Optimization* 32, no. 1 (2022): 173-203. https://doi.org/10.1137/20M1348601

[25] Pinheiro, Wallace Anacleto and Ana Bárbara Sapienza Pinheiro. "Hierarchical++: improving the hierarchical clustering algorithm." *International Journal of Data Mining, Modelling and Management* 15, no. 3 (2023): 223-239. https://doi.org/10.1504/IJDMMM.2023.10058462

[26] Yu, Heng and Xiaolan Hou. "Hierarchical clustering in astronomy." *Astronomy and Computing* 41 (2022): 100662. https://doi.org/10.1016/j.ascom.2022.100662

[27] Vichi, Maurizio, Carlo Cavicchia and Patrick JF Groenen. "Hierarchical means clustering." *Journal of Classification* 39, no. 3 (2022): 553-577. https://doi.org/10.1007/s00357-022-09419-7

[28] Koren, Oded, Aviel Shamalov and Nir Perel. "Small Files Problem Resolution via Hierarchical Clustering Algorithm." *Big Data* 12, no. 3 (2024): 229-242. https://doi.org/10.1089/big.2022.0181

[29] Wu, Guoqing, Liqiang Cao, Hongyun Tian and Wei Wang. "HY-DBSCAN: A hybrid parallel DBSCAN clustering algorithm scalable on distributed-memory computers." *Journal of Parallel and Distributed Computing* 168 (2022): 57-69. https://doi.org/10.1016/j.jpdc.2022.06.005

[30] Hanafi, Nooshin and Hamid Saadatfar. "A fast DBSCAN algorithm for big data based on efficient density calculation." *Expert Systems with Applications* 203 (2022): 117501. https://doi.org/10.1016/j.eswa.2022.117501

[31] An, Xiaoya, Ziming Wang, Ding Wang, Song Liu, Cheng Jin, Xinpeng Xu and Jianjun Cao. "Strp-dbscan: A parallel dbscan algorithm based on spatial-temporal random partitioning for clustering trajectory data." *Applied Sciences* 13, no. 20 (2023): 11122. https://doi.org/10.3390/app132011122

[32] Jain, Praphula Kumar, Mani Shankar Bajpai and Rajendra Pamula. "A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality." *Int. Arab J. Inf. Technol.* 19, no. 1 (2022): 23-28. https://doi.org/10.34028/iajit/19/1/3

[33] Aparna, K. and Mydhili K. Nair. "Effect of outlier detection on clustering accuracy and computation time of CHB K-means algorithm." In *Computational Intelligence in Data Mining—Volume 2: Proceedings of the International Conference on CIDM, 5-6 December 2015*, pp. 25-35. Springer India, 2016. https://doi.org/10.1007/978-81-322-2731-1_3

[34] Abdussamad, Abdul Museeb and Agha Inayat. "Addressing limitations of the K-means clustering algorithm: Outliers, non-spherical data and optimal cluster selection." *AIMS Math* 9 (2024): 25070-25097. https://doi.org/10.3934/math.20241222

[35] Liu, Zhe, Jianmin Bao and Fei Ding. "An improved k-means clustering algorithm based on semantic model." In *Proceedings of the International Conference on Information Technology and Electrical Engineering 2018*, pp. 1-5. 2018. https://doi.org/10.1145/3148453.3306269

[36] Min, Zhang and Duan Kai-fei. "Improved research to K-means initial cluster centers." In *2015 Ninth international conference on frontier of computer science and technology*, pp. 349-353. IEEE, 2015. https://doi.org/10.1109/FCST.2015.61

[37] Khan, Iliyas Karim, Hanita Binti Daud, Nooraini Binti Zainuddin, Rajalingam Sokkalingam, Muhammad Farooq, Muzammil Elahi Baig, Gohar Ayub and Mudasar Zafar. "Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm." *Egyptian Informatics Journal* 27 (2024): 100504. https://doi.org/10.1016/j.eij.2024.100504

[38] Hot, Elma and Vesna Popović-Bugarin. "Soil data clustering by using K-means and fuzzy K-means algorithm." In *2015 23rd Telecommunications Forum Telfor (TELFOR)*, pp. 890-893. IEEE, 2015. https://doi.org/10.1109/TELFOR.2015.7377608

[39] Sivasankari, K. and Uma Maheswari KM. "Privacy Preserving Using K Member Gaussian Kernel Fuzzy C Means and Self Adaptive Honey Badger for Online Social Networks." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 45, no. 2 (2025): 25-37. https://doi.org/10.37934/araset.45.2.2537

[40] He, Zhenfeng and Chunyan Yu. "Clustering stability-based evolutionary k-means." *Soft Computing* 23, no. 1 (2019): 305-321. https://doi.org/10.1007/s00500-018-3280-0

[41] Bansal, Arpit, Mayur Sharma and Shalini Goel. "Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining." *International Journal of Computer Applications* 157, no. 6 (2017): 0975-8887. https://doi.org/10.5120/ijca2017912719

[42] Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine Piatko, Ruth Silverman and Angela Y. Wu. "The analysis of a simple k-means clustering algorithm." In *Proceedings of the sixteenth annual symposium on Computational geometry*, pp. 100-109. 2000. https://doi.org/10.1145/336154.336189

[43] Yuan, Chunhui and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." *J* 2, no. 2 (2019): 226-235. https://doi.org/10.3390/j2020016

[44] Brusco, Michael J., Emilie Shireman and Douglas Steinley. "A comparison of latent class, K-means and K-median methods for clustering dichotomous data." *Psychological methods* 22, no. 3 (2017): 563. https://doi.org/10.1037/met0000095

[45] Sinaga, Kristina P. and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." *IEEE access* 8 (2020): 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796