Data-driven business strategies with the power of the K-means Algorithm

Md Mizanur Rahman
School of Computing, Regent College London

Palto Datta
School of Business & Enterprise, Regent College London

Key words
Business Strategies, Competitive Advantages, Data-driven K-means Algorithm, K-means Algorithm, Pre-processing Techniques, Strategic Planning

Abstract

In today's dynamic business environment, Machine Learning (ML) or algorithm-based, data-driven models are essential for competitive advantages and strategic planning. This study aims to demonstrate the effectiveness of ML models - specifically the standard K-means clustering algorithm in identifying patterns that can inform strategic business decisions. A synthetic dataset was generated to simulate real-world business data scenarios, and the K-means algorithm was applied both with and without data pre-processing techniques such as scaling. The results indicate that although K-means remains a powerful and widely applicable clustering method, its performance is significantly improved by proper data scaling and identification of the optimal number of clusters. The findings of this study offer valuable insight how to develop business strategies over complex business scenarios.

Corresponding author: Md Mizanur Rahman
Email address for the corresponding author: MohammadM.Rahman@rcl.ac.uk
The first submission received: 18th of June 2025
Revised submission received: 20th of July 2025
Accepted: 25th of August 2025

Introduction

In the modern business landscape, data is often considered more valuable than anything else as it helps informed decision-making and strategic planning, reduces risks and increases the likelihood of success for businesses. A data-driven, algorithm-based model, such as, K-means clustering can be very effective in supporting the development of strategic business decisions.

Data is largely required by Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) models. AI is a branch of computer science that involves simulating human intelligence processed by a computer system to think, learn and perform specific tasks for problem-solving. For instance, speech recognition system is a simple AI model that converts human speech or spoken language as data into text using them to operate or control a machine or another system. Some other examples of AI models include spam filter, autonomous vehicles such as self-driving car, virtual assistance such as Siri, Alexa and Google Assistant, expert systems, Natural Language Processing (NLP), and chatbots.

ML is a subset of AI in which computer systems are trained to learn from data to create a model and improve their performances identifying patterns and making decisions based on these patterns. For instance, by learning past email selections for the spam box, newly receiving emails can be categorised as spam or not spam by a ML. Similarly, ML algorithms help diagnose diseases like cancer, diabetes and heart diseases; analyse customers behaviours and categorise them into different segments for personalised advertising by processing large volumes of relevant data.

DL is a further subset of ML that uses neural networks with multiple layers or deeps to model and identify complex patterns from large volume datasets. A common example is facial recognition system used in security systems and social media platforms is a deep learning model for detecting and verifying faces in images or video feeds. Other notable examples of DL models include ChatGPT, fraud detection systems, and predictive analytics tools.

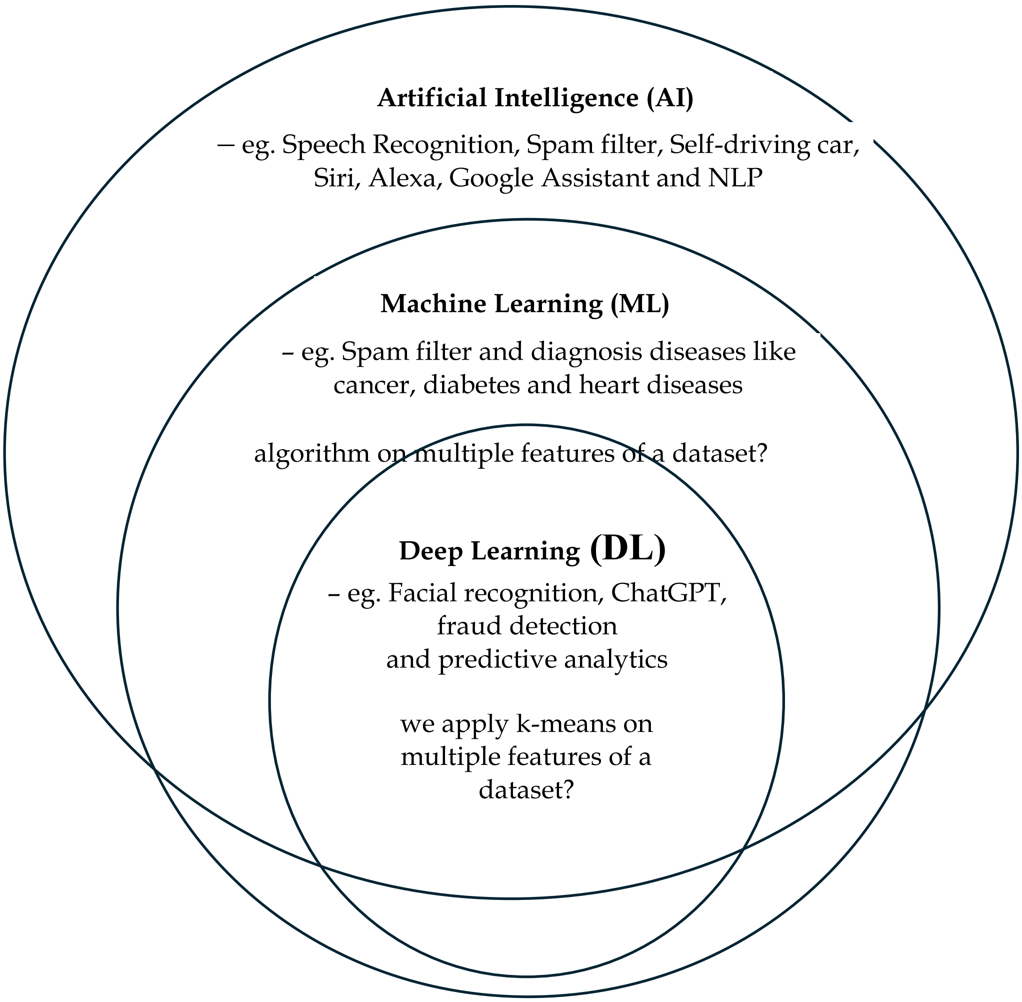

Figure 1: A relationship between AI, ML and DL models

AI, ML and DL models have clear relationships, and they are interconnected with each other, shown in Figure 1. The aim of AI model is to create an intelligent or smart machine where within the broader field of AI, ML model allows machines to learn from data and improve over time without being explicitly programmed. In turn, DL model is a specialised technique to handle complex and larger datasets using deep neural networks. For instance, all ML models fall under the umbrella of AI, but not all AI models involve ML. Similarly, all DL models are a subset of ML, but not all ML models use DL techniques. Such of these AI, ML and DL models relies on suitable and appropriate algorithms to their specific tasks and complexities.

There are four types of ML algorithms:

- **Supervised ML algorithms** – are trained labelled datasets mapping inputs to outputs in the training process to predict outcomes for new or unseen data (Hastie et al., 2009 and Bishop and Bishop, 2023). Some common ML algorithms are Linear Regression, Logistic Regression, Naïve Base, Support Vector Machine (SVM), Decision Trees, Random forests and K-Nearest Neighbours (KNN).
- **Semi-supervised ML algorithms** – are designed to work with a dataset that contains small amount of labelled data and a large amount of unlabelled data to create a predictive model and improve performance (Zhu and Goldberg, 2009). Self-training and Co-training are two common semi-supervised algorithms.
- **Unsupervised ML algorithms** – read data without labelling and find patterns or structures in the data, or group the similar instances together (Ikotun et al.., 2023; Bishop and Bishop, 2023 and Lam and Wunsch, 2014). Some common unsupervised ML algorithms are K-means clustering, Hierarchical clustering, Partitional clustering, Principal Component Analysis and Gaussian Mixture Models (GMM).
- **Reinforcement ML algorithms** – are based on agents that learn to make decisions and use feedback to find optimal solutions (Mnih et al. 2015 and Sutton and Barto, 2018). Common Reinforcement ML algorithms are Q-Learning, Deep Q-Networks, Actor-Critic and Policy Gradient Methods.

The K-means clustering or simply K-means is a popular unsupervised ML algorithm that partitions a dataset into a pre-defined number of clusters or groups. This is one of the widely used unsupervised learning algorithms due to its simplicity, effectiveness and computational efficiency in partitioning data. For the implementation of K-means algorithm, first k number of data points as initial centroids are selected randomly or manually for k clusters. Then Euclidean distance is calculated from each data point to the nearest centroid or the centre of the cluster. The centroids are recalculated by computing the mean of the data points assigned to each cluster. The process is repeated until the centroids of the clusters remain unchanged.

K-means clustering is valuable in business intelligence and strategic decision making as it can discover hidden patterns from the datasets over customer behaviours, market trends, churn prediction, customer retention and business operations. For instance, the K-means algorithm can segment customers based on similar characteristics such as demographics (e.g., age and gender) and purchasing behaviours (e.g., frequency and value). The algorithm can also identify customer segments with a high risk of churn based on usage patterns and satisfaction levels. To get these benefits, K-means based models requires large and comprehensive datasets.

In this paper, we will apply the K-means algorithm and evaluate its performance on a case study by a few tools both with and without pre-processing or scaling focusing how it helps businesses to make their effective data-driven strategies. The rest of this paper is structured as follows. Section 2 presents the literature reviews over the K-means algorithm including different researchers' proposals and modifications over this algorithm. We provide the details of the methodology in Section 3, implementation of data-driven K-means models in Section 4, and findings of the data-driven K-means model in Section 5. Section 6 discusses the practical implications of the results, and Section 7 concludes the paper.

Literature review

K-means clustering algorithm is one of the oldest algorithms in computing history, proposed by Stuart Lloyd of Bell Labs in 1957, based on an idea by Hugo Steinhaus in 1956, and first used by James MacQueen in 1967. Since 1990s, when ML algorithms gained significant momentum in the market, the K-means became popular grouping objects by their properties such as size, colour, weight, shape, length and so on. Over the years many researchers evaluated the algorithm including Suyal and Sharma (2024), Dalmaijer (2022), Oti et al. (2021), Sinaga and Yang (2020), Li and Wu (2012) and Napolean and Pavalakodi (2011), and some of them including Zubair et al. (2024), Annas and Wahab (2023), Hossain et al. (2019) and Xiao et al. (2018) also proposed to modify the algorithm for different purposes and/or categorise the algorithm based on the originality.

There are many challenges with the number of clusters in the K-means algorithm - if the number of clusters is not optimal, it affects the performance or results of the algorithm. One of the major drawbacks of this algorithm is the initially random selection of cluster centres, which is greedy in nature and leads to poor clustering results. Due to this drawback, the original K-Means algorithm may not be useful or cannot identify patterns correctly in various real-world problems such as data mining, image segmentation, medical diagnostic and economics. Hossain et al. (2019) proposed a new method to cluster data dynamically setting threshold value or Euclidean distance between two datapoints when number of clusters is not set correctly. Frost et al. (2020) proposed X-means method to determine number of clusters efficiently by making local decisions with splitting themselves for cluster centres in each iteration. Sinaga and Yang (2020) proposed a modified version of K-means algorithm, called U-K-means to determine correct number of clusters from a noisy and large dataset.

Ikotun et al. (2023) identified that the overlapping clustering behaviour of some data points or ambiguous nature limits the performance and robustness of the algorithm. Pasin and Gonec (2023) applied the Fuzzy K-means algorithm, originally developed by Bezdek in 1981, to complex datasets like COVID-19 where cluster boundaries are not clearly defined to understand the data structures better. Arthur and Vassilvitskii (2007) proposed K-Means++, an improved version of K-means algorithm, choosing initial cluster centres smatter or better way instead of choosing them randomly for accuracy of the results, higher speed and stability of the algorithm.

Many other researchers also proposed better and memory efficient K-means algorithm. Lang and Schubert (2024) proposed K-means clustering with Cover Tree index employing upper and lower bounds on point-to-cluster distances and the triangle inequality to accelerate the computation process of the algorithm. Mohammadi et al. (2021) proposed an improved version of K-means algorithm to use the most significant data distribution axis to split the clusters incrementally into better fits or detect automatically number of clusters for the accuracy and speed. Lattanzi and Sohler (2019) proposed a better K-means algorithm using local search strategy and Beretta et al. (2023) improved the algorithm again considering local search neighbourhoods allowing to swap multiple centres at the same time. Xie et al. (2020) also proposed an improved version of K-means algorithm based on density selection, and Lee and Lin (2012) proposed the same using selection and Erasure rules.

The authors conducted an extensive review over more than 50 published articles on K-Means clustering algorithm to justify its accuracy and outcomes over large datasets for business applications, particularly data-driven business strategies. Many studies highlight the weaknesses of the algorithm and propose a wide range of versions or improvements to overcome its shortcomings. However, very rare of them points to the importance of pre-processing before implementing the algorithm on large datasets and most of them are interested in proposing counter approaches to modify this powerful algorithm. Although the K-means algorithm is valuable for many business applications, but it is sensitive to the larger magnitude of the features that disproportionately influence the outcomes of the algorithm. Following a pre-processing step, the algorithm can significantly improve the performance and accuracy of the outcomes and can handle a large dataset efficiently to support more effective and data-driven business strategies.

Methodology

This research uses a quantitative research design using synthetically generated primary dataset. The K-means algorithm identifies patterns considering the most important two features from a dataset that may have multiple features. The methodology consists of data collection, pre-processing, implementing and reviews stages.

- **Data collection** – For this research, we generated a dataset of 1,000 random numbers using uniform manner. The dataset has two numerical features – "No of Visits" and "Total Sales" where random numbers were used between 1 and 50, and 20 and 450 respectively. The synthetic data was generated to reflect realistic business scenarios while minimising potential biases. Though synthetic, the dataset exhibits characteristics of big data -specially Volume, Variety and Velocity (3V) and makes it suitable for analytical purposes.
- **Pre-processing** – Pre-processing can be done for many reasons including missing values, categorical variables, outliers and feature scaling (Wongoutong, 2024). In this study, min-max scaling was applied for feature scaling to convert all required feature values between 0 and 1 to ensure all features contribute equally during clustering. The pre-processing technique such as min-max scaling can ensure that all features contribute equally for quality outcomes or data-driven models. Due to the lack of feature scaling, the K-means algorithm, particularly when handle a large dataset shows poor results or cannot identify any patterns.
- **Implementing** – Python IDLE with Scikit-learn, Pandas and Matplotlib libraries and Orange data mining tool were used to implement the K-means algorithm over the pre-processed dataset. Then Elbow method (Herdiana et al. 2025) was employed to determine the optimal number of clusters, and visualisation was used to show clustering outcomes both with and without pre-processed dataset in relation to making business strategies.
- **Reviews** – Based on the clustering outcomes, reviews were conducted to support data-driven business strategies. The review process focused on evaluating the effectiveness of the clustering in segmenting data in a way that could inform strategic business decisions and strategic planning.

Data-driven K-means models

In this research, we will create a data-driven standard K-means algorithm-based model to evaluate how effective strategies can be made for the benefits of businesses. Programming tools such as Python with different libraries and Orange data mining tool were used to create a data-driven K-means model. We first applied the K-means algorithm using "No of Visits" and "Total Sales" columns or features to the randomly generated dataset comprising 1,000 data points. The Figure 2 represents Scatter graph of "No of Visits" vs "Total Sales".

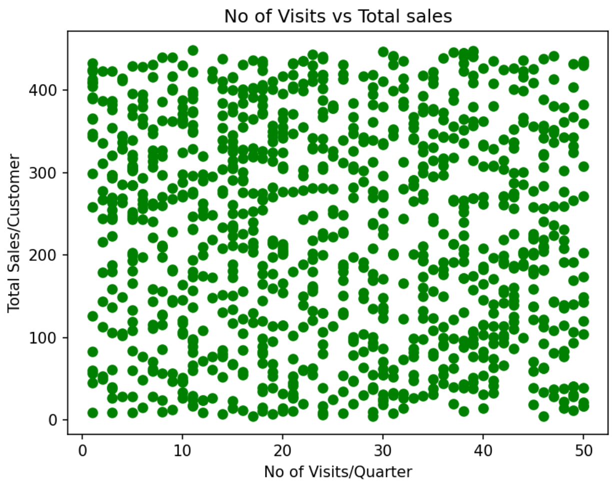

Figure 2: Scatter graph of No of Visits vs Total Sales

4 clusters were identified as optimal by the Elbow method. The algorithm was applied without any pre-scaling or pre-processing of data in Python. Figure 3 represents the output of the algorithm.

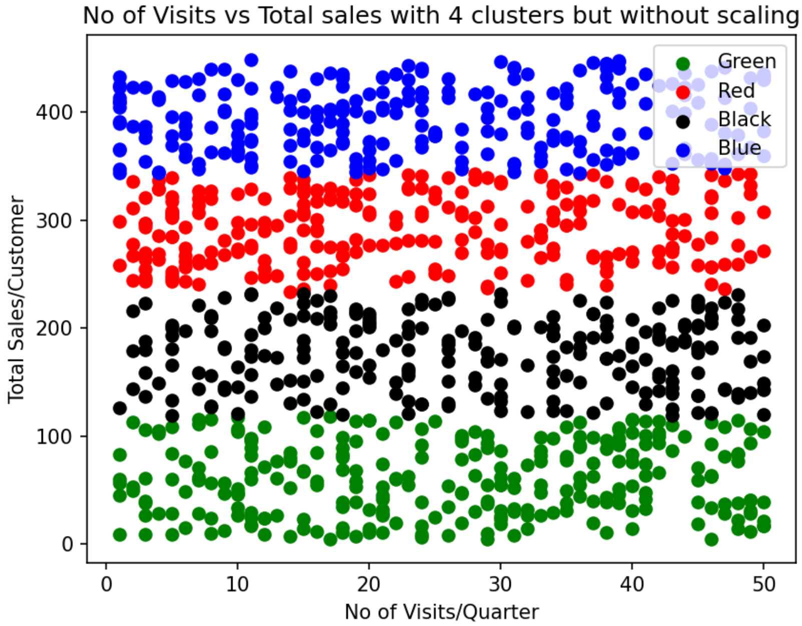

Figure 3: Output of K-means algorithm without scaling

Figure 4 represents the output of the algorithm while it was applied with min-max scaling in Python.

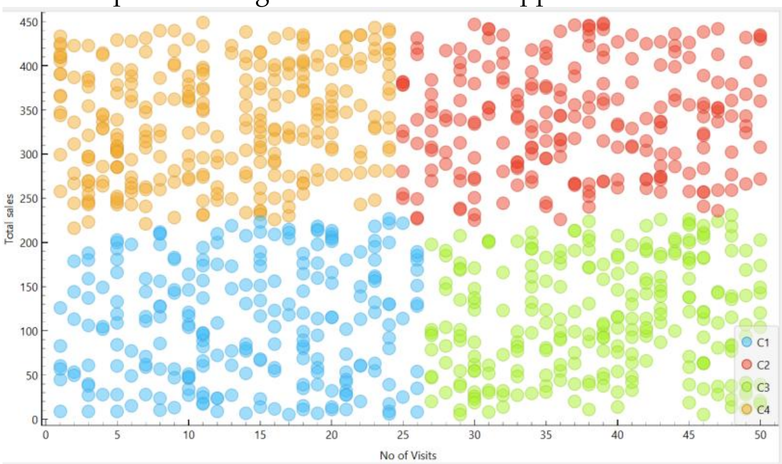

Figure 4: Output of K-means algorithm with scaling

Findings

Without applying scaling or pre-processing, the output or the 4 clusters provided less meaningful differentiation. The algorithm divides the total sales per quarter into 4 clusters or categories – roughly up to £100, between £100 and £200, £200 and £300, and £300 and £450. These categories/groups might suggest preliminary business strategies or ideas on possible business strategies. For example, customers up to total sales of £100 (Green cluster) can be given more discounts or higher priority to increase their expenses per visit. Customers total sales between £100 and £200 (Black cluster) can be given a moderate discount or a bit lower priority (than the Green cluster) to increase their expenses and so on. This strategy solely depends on customers' total sales but does not consider other column or feature, "No of Visits" per quarter, which is a critical factor for understanding customers' behaviours. As a result, this strategy may not provide effective solutions for the business.

On the same dataset, when the algorithm was applied with min-max scaling and 4 clusters, we got the following 4 clear distinct divisions or segments of customers, shown in Figure 4:

- C1 – Low number of visits (less than 27 visits per quarter) with lower total sales (roughly up to £220)
- C4 - Low number of visits (less than 26 visit per quarter) with higher total sales (roughly between £220 and £450)
- C3 – High number of visits (more than 26 visits per quarter) with low total sales (roughly up to £220)
- C2 – High number of visits (more than 24 visits per quarter) with higher total sales (roughly between £220 and £450)

For the following 4 clusters/segments, the business can make different targeted strategies to increase their revenues. The types of customers suggested strategies and two targets are outlined in Table 1, where organisational decisions are guided by data driven segmentation by K-means algorithm (John et al., 2024; Husein et al., 2021).

	C1	**C4**	**C3**	**C2**
Type of customers	Low valued/Less frequent	High valued/Less frequent	Low valued/High frequent	Highest valued/High frequent
Strategies	Discount on each visit by the customers and/or discount after reaching a set total sale	Discount on each visit by the customers	Discount or free item after reaching a set total sale	Special offer
Target 1	Increase number of visits and total sales per quarter	Increase number of visits per quarter	Increase total sales per quarter	Retain customers longer
Target 2	Turn it to C4 or at least C3	Turn it to C2	Turn it to C2	Turn it to a special customer group

Table 1: Formation of suggested strategies based on 4 clusters

Customers of C1 and C4 clusters have a fewer number of visits per quarter but their spending behaviours differ significantly. They need different approaches on the discount strategies as C4 customers spend much more than C1 customers and the same discount strategy would be ineffective. The objectives of C1 customers are twofold where the targets are to encourage customers more frequent visits and increase their overall expenses per quarter with a long-term goal of turning them into high-valued customers C4 or at least to turn them into C3. The targets of C4 customers are simply to increase the number of visits per quarter, with the potential goal of turning them into highest-value C2 customers.

Customers of C3 and C2 clusters have a higher number of visits per quarter but their spending patterns differ significantly. They need different approaches on the discount strategies as C2 customers spend much more than C3 customers. The targets of C3 customers are to increase their expenses for reaching a predefined spending threshold, with the potential goal of turning them into highest-value, C2 customers. The business should not offer C3 customers any discounts per visit like C1 and C4 customers unless they meet a predefined spending threshold as they spend much less money per frequent visit. The business may need special or tailored retention strategies for C2 customers because they already have high spending and frequent visits per quarter, such as exclusive offers or royalty incentives to engage them and reinforce their values to the business more.

Implications

- K-means is one of the useful unsupervised ML algorithms for helping businesses to make their effective strategies and business planning for competitive advantages, but the algorithm needs some kind of pre-processing or scaling of some features and optimal number of clusters where the algorithm will be applied on.
- The standard form of the algorithm can be applied almost any fields or domains including business, healthcare, finance, education and governance to make effective strategies. However, the business will benefit well if the strategies are aligned with competitive advantages and innovation.
- The standard K-means algorithm still performs well despite numerous enhancements or alternative approaches by many researchers, mentioned in the literature review. This is the enduring power of this algorithm, even after more than 60 years for making effective data-driven business development strategies.
- The algorithm was applied to a randomly generated dataset comprising 1,000 data points over many times using a few different tools and the outputs were obtained consistently similar. However, in the real-world scenarios, the algorithm may present meaningless categories or groups due to lacking scaling or pre-processing, and as a result, different improved K-means algorithm may be required to achieve for meaningful and optimal solutions.

Conclusions

The K-means clustering algorithm is still highly valuable in many fields for developing effective data-driven business strategies, even it was introduced more than 60 years ago. The standard K-means algorithm required pre-processing or scaling and a predefined optimal number of clusters in advance. The algorithm can handle a huge dataset for optimal solutions without the need for any modifications or enhancements. In this research, a randomly generated dataset comprising 1,000 data points is used for avoiding bias many times and they show consistent results to make effective business strategies. For evaluating how effective the data-driven K-means strategy is for business, researchers need a strong understanding of the relevant business domains rather than just finding better solutions using an enhanced or modified version of K-means algorithm.

References

Annas, M. and Wahab, S. N., 2023. Data mining methods: K-means clustering algorithms. *International Journal of Cyber and IT service management,* 3(1). ISSN: 2797-1325.

Arthur, D. and Vassilvitskii, S., 2007. K-Means++: The advantage of careful seeding in Proc. *18th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA.

Bereta, L., Cohen-Addad, V., Lattanzi, S. and Parotsidis, N., 2023. Multi-swap K-means++ in Proc. of the *37th International Conference on Neural information processing Systems,* Article no: 1135. pp. 26069-26091. New Orleans LA USA.

Bishop, C. M. and Bishop, H., 2023. *Deep Learning: Foundations and Concepts*. Springer. Cham, Switzerland.

Dalmaijer, E. S., Nord, C. L and Astle, D. E., 2022. Statistical power for cluster analysis. *BMC Bioinformatics,* vol. 23. Article number 205.

Frost, N., Moshkovitz, M and Rashtchian, C., 2020. ExKMC: Expanding Explainable k-Means clustering. https://arxiv.org/abs/2006.02399

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer. New York, NY.

Herdiana, I., Kamal, M. A., Triyani, Estri, M. N. and Renny, 2025. A more precise elbow method for optimum K-means clustering. arXiv. 22 pages. https://arxiv.org/abs/2502.00851

Hossain, M. Z., Akhtar, M. N., Ahmad, R.B., and Rahman, M., 2019. A dynamic K-Means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science,* 13(2). pp. 521 – 526.

Husein, A.M., Waruwu, F.K., Batu Bara, Y.M.T., Donpril, M. and others, 2021. Clustering Algorithm For Determining Marketing Targets Based on Customer Purchase Patterns And Behaviors. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 6(1), pp.137–143. Available at: https://www.researchgate.net/publication/355591874_Clustering_Algorithm_For_Determining_Marketing_Targets_Based_Customer_Purchase_Patterns_And_Behaviors

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J., 2023. K-means clustering algorithms: A comprehensive review, variants, and advances in the era of big data, Scientific Reports, vol. 622, April 2023, pp. 178-210. https://www.nature.com/articles/s41598-023-33214-y

John, J.M., Shobayo, O. and Ogunleye, B., 2024**.** An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *arXiv preprint arXiv:2402.04103*. Available at: https://arxiv.org/abs/2402.04103

Lam, D. and Wunsch, D., 2014. Clustering. Signal Processing: Signal processing Theory and Machine Learning, *Signal Processing Theory and Machine Learning*, vol. 1, pp. 1115-1149. https://www.researchgate.net/publication/285180280_Clustering, DOI: DOI:10.1016/B978-0-12-396502-8.00020-6

Lang, A. and Schubert, E., 2024. Accelerating K-Means clustering with Cover Trees* in Proc. of *International Conference on Similarity and Search Applications (SISAP 2023)*. Lecture Notes in Computer Science (LNCS, Volume 14289).

Lattanzi, S. and Soler, C., 2019. A better K-means++ algorithm via local search in Proc. of the *36th International Conference on Machine Learning, PMLR97*. Volume 97, pp. 3662—3671. Found at https://proceedings.mlr.press/v97/lattanzi19a.html. San Diego CA, USA.

Lee, S. S. and Lin, J. C., 2012. An accelerated K-means clustering algorithm using selection and erasure rules. *Journal of Zhejiang university SCIENCE* C, vol. 13, pp. 761-768. Springer nature link found at https://doi.org/10.1631/jzus.C1200078.

Li, Y. and Wu, H., 2012. A clustering method based on K-Means algorithm. *2012 International Conference on Solid State Devices and Materials Science. Physics Procedia,* vol. 25 (2012), pp.1104 – 1109.

Manish, S. and Sanjay, S. (2024) A review on analysis of K-Means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems,* vol. 6, pp. 85-95. ISSN: 2642-2859.

Mnih, V., Kavukcuoglu, K., Silver, D, and Rusu, A. A. (2015) Human-level control through deep reinforcement learning. *Nature,* 518(7540), pp. 529-533. DOI: 10.1038/nature14236 https://pubmed.ncbi.nlm.nih.gov/25719670/

Mohammadi, S. O., Kalhor, A., and Bodaghi, H. (2021) H-Splits: Improved K-Means clustering algorithm to automatically detect the number of clusters. *Computer Vision and Pattern Recognition.* University of Tehran, Iran.

Napolean, D. and Pavalakodi, S. (2011) A new method for dimensionality reduction using K-Means clustering algorithm for high dimensional data set. *International Journal of Computer Applications,* 13(7). DOI: 10.5120/1789-2471 https://www.ijcaonline.org/archives/volume13/number7/1789-2471/

Oti, E. U., Olusola, M. O., Eze, F. C. and Enogwe, S. U. (2021) Comprehensive Review of K-Means Clustering Algorithms, *International Journal of Advances in scientific Research and Engineering,* 7(8). pp. 64—69. E-ISSN: 2454-8006. DOI: 10.31695/IJASRE.2021.34050

Pasin, O. and Gonenc, S. (2023) An investigation into epidemiological situations of COVID with fuzzy K-means and K-prototype clustering methods. *Scientific Reports,* vol. 13. Article number 6255. https://www.nature.com/articles/s41598-023-33214-y.

Sinaga, K. P. and Yang, M. S. (2020) Unsupervised K-Means Clustering Algorithm. *IEEE Access*, vol. 10, pp. 80716 – 80727, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988796 https://www.scirp.org/reference/referencespapers?referenceid=3464953

Singh, S. and Gill, N. S., (2013) Analysis and study of K-means clustering algorithm. *International Journal of Engineering Research and Technology,* 2(7). ISSN: 2278-0181.

Sutton, R. S. and Barto, A. G., (2018) Reinforcement Learning: An introduction. *MIT Press*. Second Edition. Cambridge, MA. http://incompleteideas.net/book/the-book-2nd.html

Suyal, M. and Sharma, S. (2024) A review on analysis of K-Means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems*, vol. 6, pp. 85-95. ISSN: 2642-2859.

Wongoutong, C. (2024) The impact of neglecting features scaling in K-means clustering. *PLoS One,* 19(12). Mae Fah Luang University, Thailand. https://pmc.ncbi.nlm.nih.gov/articles/PMC11623793/.

Xiao, B., Wang, Z., Liu, Q. and Liu, X. (2018) SMK-means: an improved mini batch K-means algorithm based on mapreduce with big data. *Tech Science Press,* 1(1), pp. 1-5.

Xie, W., Wang, X. and Xu, B. (2020) An improved K-means clustering algorithm based on density selection in *Proc. of the 2020 international Conference on Machine Learning and big data Analytics for IoT Security and Privacy,* 2. Shanghai, China November 2020.

Yu, l. (2024) The application of K-means clustering algorithm in the evaluation of e-commerce websites. *J. Electrical Systems,* 20(6).

Zhu, X. and Goldberg, A. B. (2009) Introduction to Semi-Supervised Learning. *Morgan and Claypool Publishers*. Series: Synthesis Lectures on Artificial Intelligence and Machine Learning , ISBN: 978-1-59829-547-4. San Rafael, CA.

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A. and Sarkar, I. H. (2024) An improved K-means clustering algorithm towards an efficient data-driven modelling. *Annals of Data Science (2024)*, 11(5), pp. 1525 – 1544.