Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

A Comparative Study of K-means Clustering Algorithms Using Euclidean and Manhattan Distance for Climate Data

Bakhshan Ahmed Hamad

Department of Mathematics, College of Education, Salahaddin University, Erbil, Iraq

Article information

Article history:
Received December 18,2024
Revised: April 20,2025
Accepted April 30,2025
Available June 1, 2025

Keywords:
K-means algorithms, Random, K-means++, Canopy, Farthest, Euclidean and Manhattan Distance

Correspondence:
Bakhshan Ahmed Hamad
Paxshan.ahmad1@su.edu.krd

Abstract

The K-means clustering algorithms (Random, K-means++, Canopy, and Farthest First) are unsupervised machine learning techniques designed to group data points based on their similarities. The study examined the effects of clustering algorithms and distance metrics on climate data analysis from meteorological stations in the Kurdistan Region of Iraq (2020–2022). 8-attribute dataset with 1,095 cases was clustered using Random, K-means++, Canopy, and Farthest First methods, evaluated with Euclidean and Manhattan distance metrics via the WEKA tools, which is a versatile and accessible open-source tool for machine learning and data mining. It features a user-friendly interface, a wide range of algorithms, robust pre-processing and visualization tools, and cross-platform compatibility.Focusing on efficiency and reducing variation within clusters, the results revealed that within Euclidean distance, all algorithms formed two clusters. Canopy required the most iterations, Farthest First the fewest. K-means++ was the fastest, Canopy the slowest. WCSS values were similar, with Random and Canopy scoring lowest, but within Manhattan Distance, all algorithms again formed two clusters. Canopy had the highest iterations, Farthest First the fewest and fastest, while Random was slowest. WCSS differences were negligible, with Random, Canopy, and Farthest First performing best.Graphs illustrate the highlighted differences in cluster distribution, iterations, execution time, and WCSS. Euclidean distance yielded lower WCSS, while interactive maps revealed clearer cluster distributions for most attributes compared to Manhattan distance. produced the lowest within-cluster sum of squared errors compared to the Manhattan distance.

DOI 10.33899/iqjoss.2025.187754 ,

1. Introduction

In today's world, driven by big data, businesses and individuals strive to make sense of the massive amounts of available data. Data mining employs analysis tools to unearth previously unknown relationships between data objects. Clustering, a prevalent data mining technique, entails grouping similar data points to reveal hidden patterns and insights. Clustering algorithms are instrumental in data organization and analysis, offering valuable insights for decision-making and comprehending complex datasets. (Gupta, et al., 2021)

Clustering algorithms are critical in today's data-driven society, enabling precise analysis and organization of massive datasets. Selecting clustering algorithm that can adapt to the quality and quantity is crucial. The k-means clustering algorithm is widely used in industries such as marketing, healthcare, and image processing.

The k-means clustering algorithm is one of many clustering algorithms used in machine learning. Each algorithm has advantages, disadvantages, and specific applications based on the similarities and differences between the K-means and other prominent clustering methods. Overall, K-means clustering is an effective and simple unsupervised learning approach with a wide range of applications. It is important to consider its limitations and select appropriate number of clusters for optimal outcomes (Karthikeyan, et al., 2020).

It partitions data points into clusters, enabling consumer segmentation, trend analysis in patient data, and image compression, and more. The k-means algorithm maximizes intra-cluster similarity while minimizing inter-cluster similarity. Understanding and utilizing the full potential of the k-means algorithm is essential for uncovering hidden insights within data. (Mohammed Elhassan & Ahamed, 2020).

This paper delves into various clustering algorithms, including Random, K-means++, Canopy, and Farthest First. It examines their methodologies and advantages in data analysis. Understanding these techniques is crucial for effectively extracting valuable insights from large datasets.

.K-means clustering algorithm:

K-means clustering is an unsupervised machine learning method that groups data points into clusters based on similarities. The "k" in K-means represents the number of clusters to identify. It aims to minimize variation within each cluster using the within-cluster sum of squares metric (Karthikeyan, et al., 2020). The algorithm works iteratively by:

1. Determining the number of clusters "k".
2. Assigning data points to their closest centroid can be done randomly or using approaches such as k-means++, canopy, and farthest first.
3. Recalculating cluster centers, the process repeats until cluster stability is achieved.

Using k-means clustering, a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$)is partitioned into with in cluster sum squares (WCSS) (i.e., variance) (Hochreiter, 2014). Formally, the objective is to find:

$$\text{wcss} = \underset{s}{\arg min} \sum_{i=1}^{k} \sum_{x \in S_i} \|\text{x} - \mu_i\|^2 \quad (1)$$

Where μ_i is the mean (also called centroid) of points in S_i, i.e.

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x_i \quad (2)$$

The size of S_i, denoted as $\|S_i\|$, represents the standard L2 norm. This standard minimizes the pairwise squared deviations of points within the same cluster.

$$\text{wcss} = \underset{s}{\arg min} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x,y \in S_i} \|x - y\|^2 \quad (3)$$

Because the overall variance remains constant, this is equivalent to maximizing the sum of squared deviations between points in distinct clusters.

The k-mean clustering algorithm has various advantages that make it a useful tool for data analysis. It offers a quick and effective method for analyzing huge datasets. By grouping comparable data points together, the method decreases data complexity and simplifies interpretation. This is especially valuable in exploratory data analysis, which requires a grasp of the underlying patterns and structure (Syakur, et al., 2018).

K-means clustering is an indispensable tool for data scientists exploring unsupervised machine learning and/or for data analysts seeking insights from unlabeled data, especially when acquiring labeled data is expensive and/or time-consuming.

The ability to effectively employ this potent strategy will depend on comprehension of the inner workings of the k-mean clustering algorithm, careful feature selection, data preparation, and clustering result analysis. The k-means clustering algorithm provides multiple initialization methods for cluster centroids, including random, k-means++, canopy, and farthest first.

These techniques are used to position the initial centroids in a way that increases the convergence and overall quality of the clustering. Here's a summary of each method:

1.1.1 Random clustering Algorithm:

The k-means clustering algorithm assigns data points to clusters based on their similarity. It randomly selects cluster centers and iteratively updates them until convergence is achieved. The steps involve choosing cluster centers, measuring distances

to data points, assigning points to clusters, and calculating new centroids. The steps in the k-means clustering algorithm are as follows:

1. Cluster centers 'c' are randomly chosen.
2. Calculate the separation between each data point and cluster center.
3. Allocate data points to cluster centers that are as close together as possible.
4. Calculate the new centroids using the assigned locations.
5. Repeat steps 2-4 until convergence is attained.

The algorithm tries to minimize the squared error function, which is given by

$$\text{sse} = \sum_{i=1}^{n} \min_{vi \epsilon V} \|x_i - v_i\|^2 \quad (4)$$

Where x_i is a data point, v_i is a cluster center, and S is the set of centers.This clustering algorithm is widely used in unsupervised machine learning to group observations into clusters. It is important to recognize that the algorithm's outcomes can be influenced by the initial cluster assignments. Hence, it is common practice to run the algorithm multiple times with different starting conditions (Edsa & Chandra, 2023).

1.1.2 k-means++ clustering Algorithm:

The K-means++ technique is an intelligent centroid that initializes centroids randomly in this algorithm. The goal is to position the centroids as far apart as possible, ensuring maximum distance between them. There are basic procedures to initialize centroids with K-means++:

1. Randomly select the initial centroid (C_1).
2. Calculate the separation between all data points and the first cluster center:

$$D_i = Max_{\ j:1 \geq k} \left\| x_i - c_j \right\|^2 \quad (5)$$

This formula represents the distance between the farthest centroid c_j and data point x_i.

3. Assign the data point x_i as a new centroid.
4. Re-initialize the centroids by determining the average of all the data points in that cluster.

$$\boldsymbol{C_j = \frac{1}{|N_j|} \sum x_j} \quad (6)$$

Repeat steps 3 and 4 until all of the defined K clusters have been found (Arthur & Vassilvitskii, 2007).

1.1.3 Canopy Clustering Algorithm:

Proposed by Andrew et al. (2000), the unsupervised pre-clustering method accelerates clustering operations on large datasets and is often used as a preliminary step for the K-means technique, also known as hierarchical clustering. The algorithm consists of two phases:

1. Phase 1: Use an "non-expansive" distance measurement to approximately divide the data into overlapping subsets known as canopies. This is done using two thresholds, T_1 (the loose distance) and T_2 (the tight distance), where $T_1 > T_2$.
2. Phase 2: Perform elaborate clustering within each canopy by using an "expensive" distance measurement.

The canopy clustering algorithm is a faster and simpler algorithm commonly used as an early stage in more stringent clustering algorithms. (Andrew, et al., 2000)

1.1.4 Farthest first clustering algorithm:

The farthest first clustering, also known as complete-linkage clustering, is a hierarchical clustering technique. It entails identifying the most distant pair of items, one from each group, and utilizing the distance between these two objects to calculate the distance between the groups. In complete-linkage clustering, the following formula can be used to get the distance between clusters X and Y:

$$\mathbf{D(X, Y) = Max_{x \in X, y \in Y} d(x, y)} \quad (7)$$

$d(x, y)$ It represents the distance between the elements $x \in X$ and $y \in Y$. X and Y represent two groups of items (clusters). The farthest neighbor strategy is part of the agglomerative hierarchical approach, which classifies each data point as a cluster before gradually joining the closest clusters. In which the distance between groups is presented as the distance between the most distant pair of clusters (Jayakameswaraia, et al., 2017).

1.2 Distance Metric:

Distance metrics are crucial in clustering algorithms such as K-means clustering because they determine the similarity between data points. The choice of distance metric has a significant impact on clustering outcomes. The Euclidean distance is typically employed in k-means clustering, but alternative measures may be preferable based on the data's characteristics. Study measures include:

1. Euclidean distance: Euclidean distance measures the straight-line distance between two points in Euclidean space and is commonly used as a distance metric. It is calculated using the square root of the sum of squared coordinate differences and determined using the formula (Edsa & Chandra, 2023):

$$\boldsymbol{d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \quad (8)$$

Where d is dimensional space for the similarity calculation distance, $x_i , y_i \; ith$ attribute of $x \; and \; y$ data points, and this distance measure is appropriate when the measurements are comparable and the data is continuous. (Amira , et al., 2023)

2. Manhattan Distance: Commonly referred to as the **city block distance**, it is appropriate for categorical or ordinal data. It **calculates** the total of the absolute coordinate differences, determined by summing the values of x and y. The formula is:

$$d_{man}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (9)$$

x_i and y_i are the variables of vectors X and Y in the two-dimensional vector space. (Gan, et al., 2021).

2. Literature review:

First proposed by Andrew in 1967, the K-means clustering technique was introduced as a simple, unsupervised learning clustering system. Over time, researchers have proposed various enhancements and modifications to the K-means algorithm to address its limitations and improve its effectiveness (Andrew, et al., 2000).

Arthur and Vassilvitskii (2007) studied the selecting initial centroids of k-means algorithm that are far apart, the algorithm delivers better clustering outcomes than random initialization.

Furthermore, Jain (2008) discussed the k-mean algorithm's history, modifications, and applications across numerous areas.

Rodriguez and Luciano da F.'s (2019) study, "Clustering algorithms: A comparative approach" evaluates and contrasts various performance metrics for clustering algorithms, including the K-means approach. The study highlights the merits and weaknesses of the K-means algorithm and its performance across different dataset variants.

Sinaga & Yang (2020) suggested that the unsupervised K-means algorithm does not require any initializations or parameters. The U-k-means algorithm can determine the ideal number of clusters automatically. The approach was adaptable to various cluster sizes and forms. Experimental results show that the U-k-means algorithm outperforms other known algorithms.

Ghazal et al. (2021) assessed the K-means clustering method using both Euclidean and Manhattan distances. The performance of the K-means algorithm was measured in terms of execution time and the number of clusters formed. The results indicated that using Manhattan distance measurement metrics yielded the best outcomes.

Tabianan et al. (2022) conducted a study on clustering analysis using the K-means algorithm and the Elbow method to determine the optimal number of clusters. The study employed pre-processing techniques for data cleaning on repository questionnaires for a quantitative study, utilizing Malaysia's e-commerce dataset.

Hamad (2023) developed a predictive model for variables impacting evaporation, using climatic data from meteorological stations in Iraq's Kurdistan Region from January 2020 to December 2022. The study utilized cluster analysis and regression to highlight the strengths of each technique. It also explored the application of hierarchical cluster analysis using Euclidean and Manhattan distances. Two clusters were created, and the results were used to compare K-means clustering algorithms in the current study.

The body of work on the K-means clustering algorithm demonstrates the research community's dedication to enhancing and refining its capabilities. These efforts aim to improve the algorithm's adaptability and effectiveness across various disciplines. By expanding its capabilities and addressing its limitations, researchers are making K-means a more powerful tool for data analysis.

3. Experiments and Results:

The variables impacting evaporation rate were discovered in this study utilizing climate data from meteorological stations in the Kurdistan Region of Iraq from January 2020 to December 2022.

A series of experiments were conducted on an 8-attribute dataset consisting of 1,095 cases. The dataset was divided into a predetermined number of clusters using the Random, K-means++, Canopy, and Farthest First methods.

The initial centroids for the K-means approach were randomly determined using Euclidean and Manhattan distance metrics. Clustering accuracy and algorithmic proficiency were evaluated. The testing examined the accuracy and speed of simple K-means algorithms using the Weka tool, version 3.8.6, with predefined datasets. This includes a data mining document that outlines and thoroughly explains all the techniques covered. The attributes used in the K-means algorithm analysis are displayed in Table 1.

Table 1: Attributes of the Climate Data Set

	Variables		Variables
	AE-Amount of Evaporation		SP-Pressure at Sea Level
	T-Temperature		WD-Wind Direction
	WT-Wet Temperature		RT-Relative Humidity
	SS-Sunshine		WS-Wind Speed

3.1 Result Analysis:

The following table represents the analysis process of all algorithms, displays the findings for the dataset, which has 8 attributes and 1,095 instances, and compares the clustering algorithms using Euclidean distance metrics. The parameters chosen were the number of clusters produced, the number of iterations, the time (seconds) required to construct the model, and the within cluster sum of squared errors (WCSS).

Table 2 shows that the number of clusters for each algorithm was two (2) when using the Euclidean distance metric. The K-means++ and Canopy techniques required the most iterations, while the Farthest First algorithm required the fewest.
The k-means++ approach takes the least amount of time, while the canopy algorithm takes the most. The within-cluster sum square (WCSS) of errors showed no significant variations, there is a slight difference between the lowest value and the highest value of WCSS. The random and canopy algorithms had the lowest WCSS scores, while the k-means++ and farthest first algorithms had the highest.

Table 2: Clustering algorithms result for Climate Dataset with Euclidean distance

Algorithm	No. of Clusters	Cluster instance	No. of Iteration	Time taken (seconds)	WCSS
Random	2	513 (47%) 582 (53%)	8	0.25	154.867
K-means++	2	515 (47%) 580 (53%)	9	0.10	154.869
Canopy	2	582 (53%) 513 (47%)	9	0.31	154.867
Farthest Firs	2	515 (47%) 580 (53%)	6	0.14	154.869

Table 3 presents the results for the same attributes using Manhattan distance metric to compare the clustering algorithms. The chosen parameters include the number of clusters formed, the number of iterations, the time (in seconds) taken to build the model, and the within-cluster sum of squared errors.

When using the Manhattan distance measure, the number of clusters for each algorithm was two (2). The Canopy algorithm exhibited the highest number of iterations, while the Farthest First algorithm required the fewest iterations. The Farthest First algorithm also had the shortest time to build the model, whereas the Random algorithm took the longest. There were no significant differences in the sum of squared errors within the clusters. The lowest within-cluster sum of squares (WCSS) values were observed for the Random, Canopy, and Farthest First algorithms, with the highest value for the K-means++ algorithm.

Table:3 Clustering algorithms result for Climate Dataset with Manhattan distance

Algorithm	No. of Clusters	Cluster instance	No. of Iteration	Time take (seconds)	WCSS
Random	2	547 (50%) 548 (50%)	7	0.14	835.429
K-means++	2	551 (50%) 544 (50%)	6	0.13	835.448
Canopy	2	547 (50%) 548 (50%)	9	0.10	835.429
Farthest Firs	2	548 (50%) 547 (50%)	5	0.05	835.429

Figure 1 shows a comparison graph for Euclidean distance between the first and second cluster instances for all K-means methods. The graph indicates that the Canopy algorithm produces different outcomes compared to the Random, K-means++, and Farthest First algorithms. The Canopy algorithm has 582 instances in the first cluster and 513 instances in the second cluster, as shown in Table 2, whereas the other algorithms have fewer instances in the first cluster and more instances in the second cluster.

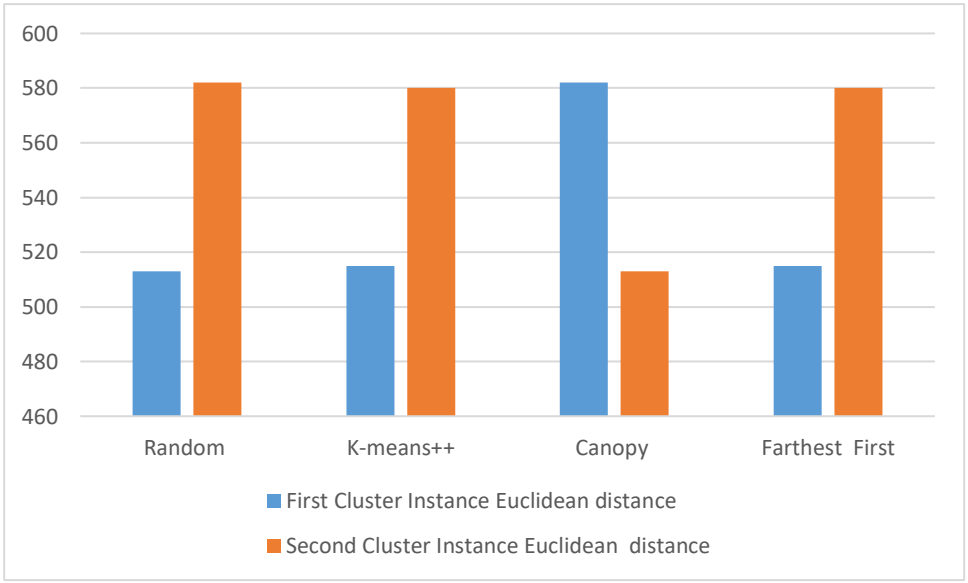

Figure 1: Comparison of First and Second Cluster Instance for Euclidean Distance

Figure 2 shows a comparison graph of the first and second cluster instances for all K-means algorithms using the Manhattan distance. The graph reveals that the K-means++ algorithm produces different outcomes compared to the Random, Canopy, and Farthest First algorithms. This algorithm has 551 instances in the first cluster and 544 instances in the second cluster, as shown in Table 3, whereas the other algorithms have fewer instances in the first cluster and more instances in the second cluster.

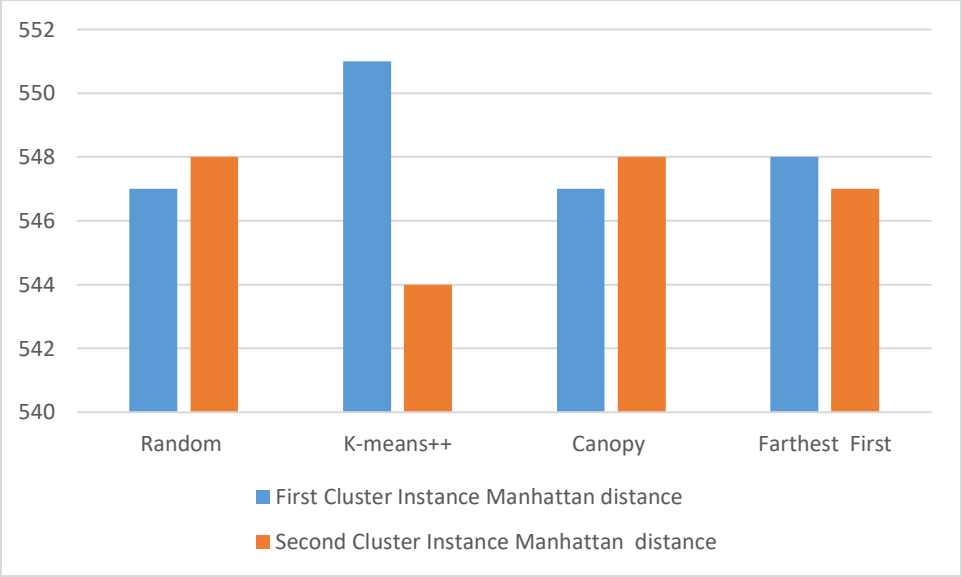

Figure2 : Comparison of First and Second Cluster Instance for Manhattan Distance

Figure 3 shows a comparison graph of the number of iterations for Euclidean and Manhattan distances across all K-means algorithms. The Farthest First and Random algorithms have the fewest iterations for both distances, while the Canopy algorithm has the most iterations for both distances. The K-means++ algorithm exhibits clear differences in the number of iterations between the two distances metrics.

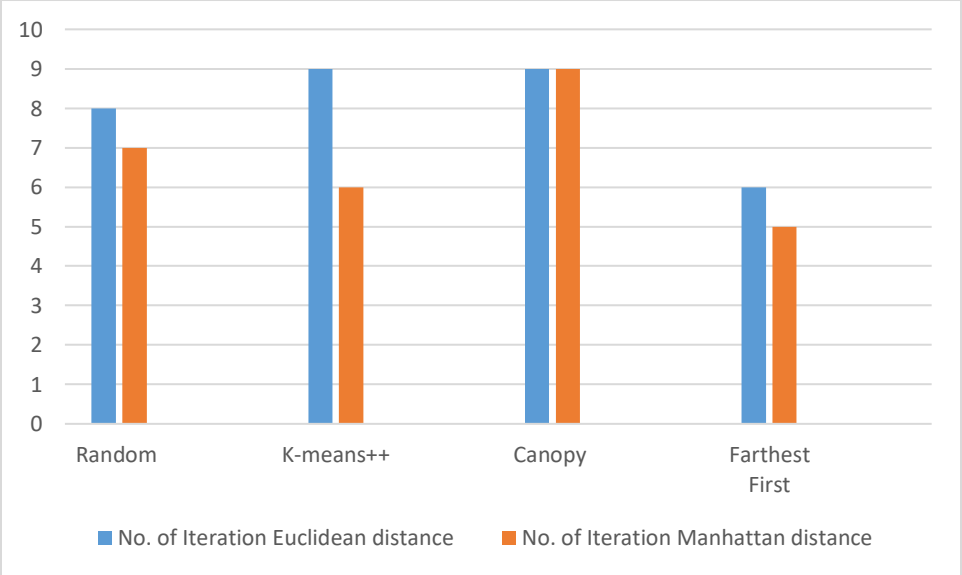

Figure3 : Comparison of Number of Iteration for Euclidean and Manhattan Distance

Figure 4 presents a comparison graph of the time (seconds) taken for Euclidean and Manhattan distances using all K-means algorithms. The K-means++ algorithm using Euclidean distance takes the shortest time, while the Canopy algorithm takes the longest. For Manhattan distance, the Farthest First algorithm takes the shortest time, while the Random algorithm takes the longest.

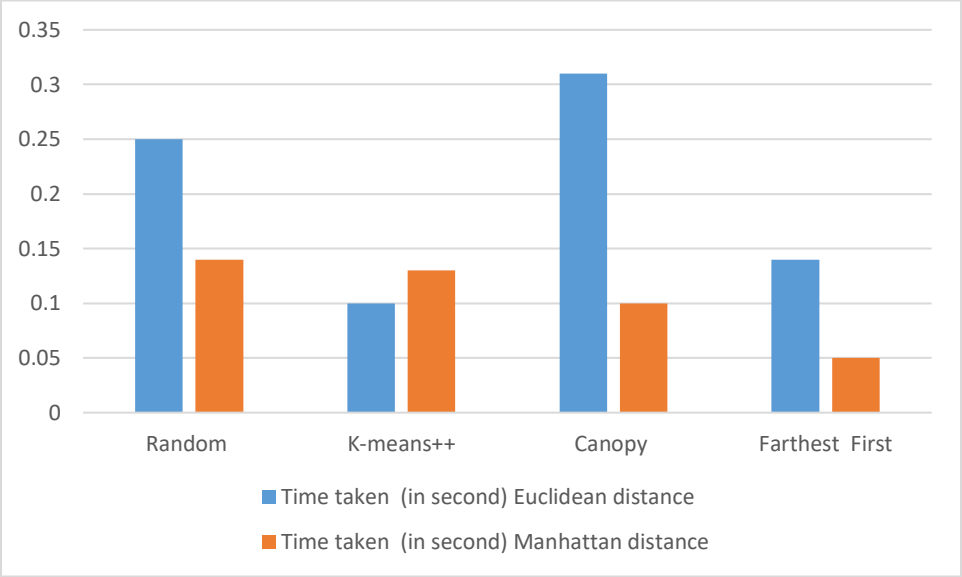

Figure 4: Comparison for Time Taken by k-means algorithms Using Euclidean and Manhattan Distance

Figure 5 shows a comparison graph of the sum of squared errors within each cluster for both Euclidean and Manhattan distances using all K-means algorithms. The graph indicates that Euclidean distance results in the lowest sum of squared errors within the clusters compared to Manhattan distance.

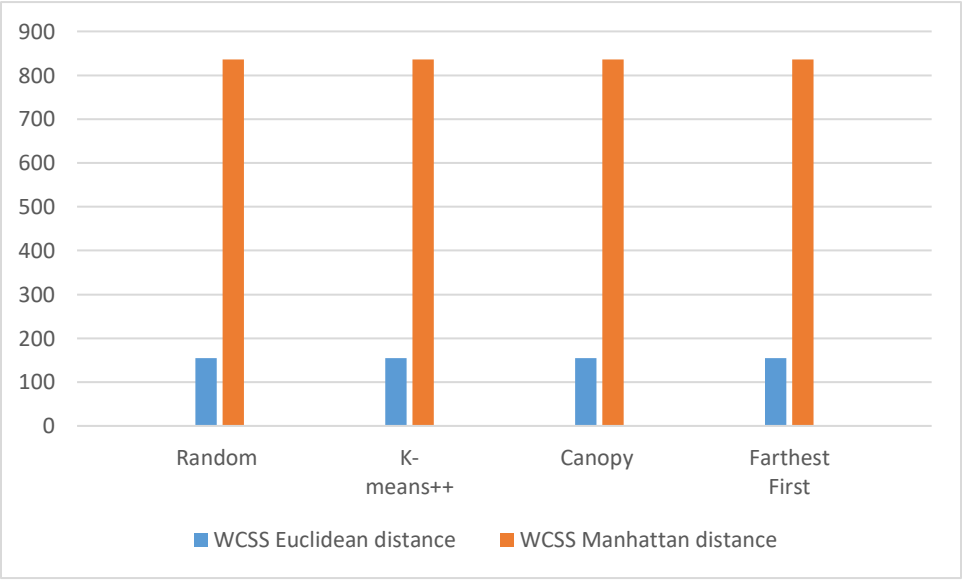

Figure5 : Comparison of Clustering Algorithms With in Sum Square Error (WCSSE)

Figure 6 illustrates insights into the clusters within the climate dataset using the K-means clustering technique with Euclidean distance. It visualizes the cluster data as an interactive map, offering an overview and detailed information on attributes and conceptually related clusters. The distribution of cluster instances for attributes (WT, RH, SP, SS, T) between the first and second clusters is clearer, while the distribution of attribute (RH) instances shows overlap compared to Manhattan distance, as depicted in Figure 7.

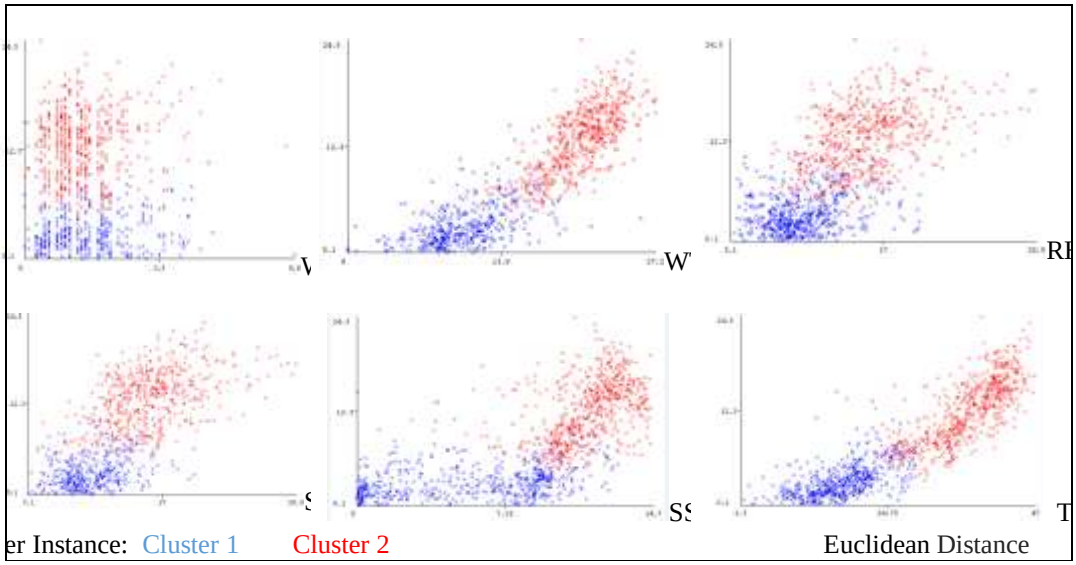

Figure 6: Distribution of Cluster Instances using Euclidean Distance

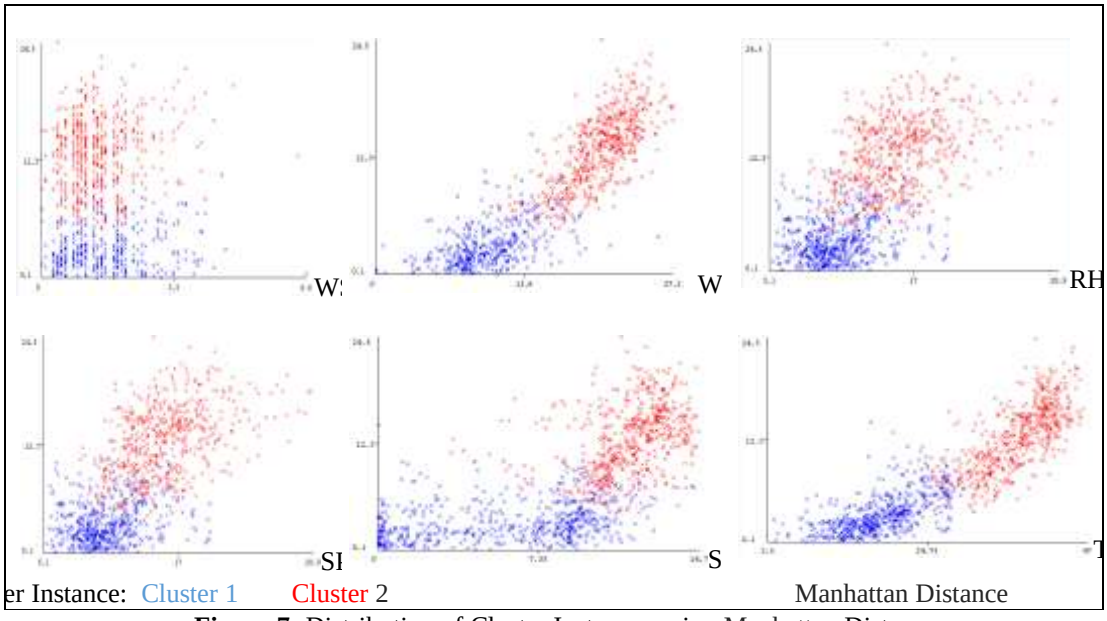

Figure 7: Distribution of Cluster Instances using Manhattan Distance

4. Conclusions:

In this study, a comparative analysis was conducted on the K-means clustering algorithms (Random, K-means++, Canopy, and Farthest First) using Euclidean and Manhattan distances. The climate dataset was analyzed using the WEKA tool, and the results are provided.

The comparative analysis focused on efficiency factors, including the number of iterations, time taken, and the within-cluster sum of squared errors (WCSS). The K-means clustering algorithms demonstrated good performance in clustering datasets, with improved results using the Euclidean distance. This distance metric showed the lowest WCSS and the shortest time taken compared to the Manhattan distance.

For achieving high precision and speed with continuous data, the Farthest First algorithm paired with Euclidean distance is the most suitable option. It is particularly effective when efficiency and rapid processing are essential, such as in real-time analysis or handling large datasets.

Additionally, the Farthest First algorithm outperformed the Random, K-means++, and Canopy algorithms when the primary objective was to minimize the within-cluster sum of squared errors (WCSS), achieve the fastest time (in seconds), and reduce the number of iterations, ensuring superior clustering quality for both Euclidean and Manhattan distances.

Graphs illustrate and emphasize significant differences in clustering outcomes, efficiency, and performance across algorithms and distance metrics. These variations are demonstrated through cluster distribution insights, the number of iterations, execution time, and within-cluster sum of squared errors for each algorithm.

5. Recommendations:

The insights from this study on K-means clustering algorithms, specifically the effectiveness of the Farthest First approach and the application of Euclidean distance measures, have potential applications for enhancing weather forecasting or modeling in various ways:

1-Integrating clustering results into hybrid models that combine machine learning with physical weather models to enhance forecasting precision and efficiency for sectors like agriculture, disaster planning, and water resource management.

2-Selecting distance measures based on data characteristics, with Euclidean distance suitable for continuous, scaled attributes and Manhattan distance better for datasets with categorical or non-linear relationships.

3-Using localized climate data clustering to improve region-specific forecasts and support cross-regional advancements in agriculture, water management, and environmental monitoring.

4-

Acknowledgment

The author is very grateful to the University of Mosul and the College of Computer Science and Mathematics, which helped improve this work.

Conflict of interest

The author has no conflict of interest.

References

1. Amira , W. O., Shahen , M. .. & Mohamad, S. . H., 2023. An Application Of Two Classification Methods: Hierarchical Clustering And Factor Analysis To The Plays PUBG. *Iraqi Journal of Statistical Sciences,* 20(1), pp. 25-42.
2. Andrew , M., Kamal , N. & Lyle , H. U., 2000. Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,*, pp. 169-178.
3. Anil , K. J., 2009. Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters,* Volume 31, pp. 651-666.
4. Arthur, D. & Vassilvitskii, S., 2007. K-Means++: The Advantages of Careful Seeding. pp. 1027-1035.
5. Edsa, S. A. A. & Chandra, G., 2023. Modified K-Means Clustering with Semi Grouping Perspective. *Indonesian Journal of Computer Science,* 12(2), pp. 493-500.
6. Gan, G., Jianhong , W. & Chaoqun , M., 2021. Data Clustering. In: *Theory ,Algorthm and Applications Second Edition.* Philadephia: Math Works ,Inc., pp. 68-69.
7. Ghazal, T. M., Hussain, M. Z. & Said, R. A., 2021. Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation & Soft Computing,* 30(2).
8. Gupta, A., Sharma, H. & Akhtar, A., 2021. A comparative Analysis of K-means and Hierarchical Clustering. *EPRA International Journal of Multidisciplinary Research (IJMR),* 7(8), pp. 412-418.
9. Hamad, B. A., 2023. Combining Cluster Analysis with Multiple Linear Regression Analysis to Create the Most Accurate Prediction Model for Evaporation in the Kurdistan Region of Iraq. *Iraqi Journal of Statistical Sciences,* 20(2), pp. 188-199.
10. Hochreiter, S., 2014. Summarizing Multivariate Data. In: *Basic Methods of Data Analysis.* Linz: Institute of Bioinformatics, Johannes Kepler University, Austria, pp. 76-77.
11. Jayakameswaraia, M., Reddy, K. K. & Ramakrishna, S., 2017. Performance Assessment of Improved Farthest Firset Cluster Algorithm on Smartphone Senso Data. *International Journal of Creative Research Thoughts (IJCR),* 5(4), pp. 3302-3305.
12. Karthikeyan, B., George, D. J. & Manikandan, G., 2020. A Comparative Study on K-Means Clustering and Agglomerative Hierarchical Clustering. *International Journal of Emerging Trends in Engineering Research,* 8(5), pp. 1600-1604.
13. Mohammed Elhassan, F. D. & Ahamed, Y. M., 2020. Compare Clustering Algorithms of Weka Tool. *International Journal of Innovative Science, Engineering & Technology,* 7(10), pp. 276-290.
14. Rodriguez, M. Z., Comin, C. H. & Luciano da, . F., 2019. Clustering algorithms: A comparative approach. *PLOS ONE,* 14(1).
15. Sinaga, K. P. & Yang, . M.-S., 2020. Unsupervised K-Means Clustering Algorithm. *IEEE,* pp. 80716 - 80727.
16. Syakur, M. A., Khotimah, B. K. & Rochman, E. M. S., 2018. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *Materials Science and Engineering* , 336(9).
17. Tabianan , K., Velu, S. & Ravi , V., 2022. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability,* 14(12).

دراسة مقارنة لخوارزميات التجميع طريقة K-means باستخدام المسافة الإقليدية ومانهاتن لبيانات المناخ

بخشان احمد حمد

قسم الرياضيات، كلية التربية ،جامعة صلاح الدين

الخلاصة :خوارزميات التجميع K–means (العشوائية، ++K–means، المظلة، والأبعد أولاً) هي تقنيات تعلم آلي غير خاضعة للإشراف، مصممة لتجميع نقاط البيانات بناءً على أوجه التشابه بينها. وقد تناولت الدراسة تأثير خوارزميات التجميع ومقاييس المسافة على تحليل بيانات المناخ من محطات الأرصاد الجوية في إقليم كردستان العراق (2020–2022). تم تجميع مجموعة بيانات مكونة من 8 سمات تحتوي على 1095 حالة باستخدام طرق (العشوائية، ++K–means، المظلة، والأبعد أولاً)، وتم تقييمها باستخدام مقاييس المسافة الإقليدية ومانهاتن عبر أدواتWEKA ، وهي أداة مفتوحة المصدر ومتعددة الاستخدامات للتعلم الآلي واستخراج البيانات. تتميز بواجهة سهلة الاستخدام، ومجموعة واسعة من الخوارزميات، وأدوات معالجة مسبقة قوية، بالإضافة إلى توافقها بين الأنظمة الأساسية.مع التركيز على الكفاءة وتقليل التباين داخل العناقيد، كشفت النتائج أنه ضمن المسافة الإقليدية، شكلت جميع الخوارزميات عنقودين. تطلبت خوارزمية المظلة أكبر عدد من التكرارات، بينما تطلبت الأبعد أولاً أقل عدد من التكرارات. كانت ++K–means الأسرع، في حين كانت خوارزمية المظلة الأبطأ. كانت قيم مجموع مربعات الأخطاء داخل العنقود متشابهة، حيث سجلت خوارزميتا العشوائية والمظلة أدنى القيم. أما ضمن مسافة مانهاتن، فقد شكلت جميع الخوارزميات مرة أخرى عنقودين. سجلت خوارزمية المظلة أعلى عدد من التكرارات، بينما كانت الأبعد أولاً الأقل والأسرع، في حين كانت العشوائية الأبطأ. وكانت الاختلافات في مجموع مربعات الأخطاء داخل العنقود ضئيلة، مع أداء مميز للعشوائية والمظلة والأبعد أولاً.توضح الرسوم البيانية الفروقات المميزة في توزيع العناقيد، وعدد التكرارات، ووقت التنفيذ، وقيم مجموع مربعات الأخطاء داخل العنقود. أظهرت المسافة الإقليدية القيم الأقل لمجموع مربعات الأخطاء، بينما كشفت الخرائط التفاعلية عن توزيعات عناقيد أوضح لمعظم السمات مقارنة بمسافة مانهاتن. وأسفرت المسافة الإقليدية عن أقل مجموع مربعات أخطاء داخل العنقود مقارنة بمسافة مانهاتن.

الكلمات المفتاحية: خوارزميات K– means، العشوائية، ++K–means، المظلة، الأبعد أولاً، المسافة الإقليدية، مسافة مانهاتن.