ISSN: 0067-2904

Comparing K-Means, Nearest Neighbor, and Lloyd's Clustering Algorithms

Shaymaa Qasim Noor*, Tareef Kamil Mustafa
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 20/6/2023 Accepted: 6/10/2023 Published: 30/11/2024

Abstract

Clustering Organizing items into groups based on their properties such that the items in the same group are similar and those in other groups are distinct is known as clustering and is one method of unsupervised learning. The primary benefit of clustering is that, with little or no prior information, fascinating patterns and structures can be discovered directly from very large data sets. The most representative algorithms, the K-Means algorithm, the nearest neighbor algorithm, and Lloyd's algorithm, were explored and evaluated in this study based on their basic strategies. The proposed algorithms proved highly efficient in classifying data, as k-means results were high by creating data points and classifying that data into 10 groups, while the nearest neighbor algorithm proved highly effective in predicting new groups in light of pre-existing groups for new data points, and finally Lloyd's algorithm achieved high results through several iterations to reach the target groups. Using the random function, 100 random values for inputs as x and 100 for outputs as y are generated, and these values are grouped as points. Calculations of the mean and standard deviation for the data set indicate that 68% of the data points will fall within one standard deviation, 95% within two standard deviations, and 99.7% within three standard deviations. In the nearest neighbor algorithm, confidence in data points within a specified standard deviation number is determined by the coverage factor, or k value. For k = 10, 97% of the data points are expected to fall within one standard deviation. At Lloyd's, a normal distribution curve appears when generating calibration or measurement data.

Keywords: Data Mining; Cluster; K-means Clustering; Lloyd's Algorithm; Nearest Neighbor Algorithm.

مقارنة خوارزميات التصنيف *K-Means*، *Nearest Neighbor*، و*Lloyd's* لعنقدة البيانات

شيماء قاسم نور*، طريف كامل مصطفى
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

تنظيم العناصر في مجموعات بناءً على خصائصها، بحيث تكون العناصر الموجودة في نفس المجموعة متشابهة وتلك الموجودة في مجموعات أخرى مختلفة، يُعرف باسم التجميع، وهي إحدى طرق التعلم غير الخاضع للرقابة. تتمثل الفائدة الأساسية للتجميع في أنه، مع وجود معلومات مسبقة قليلة أو معدومة، يمكن اكتشاف الأنماط والهياكل الرائعة مباشرة من مجموعات بيانات كبيرة جدًا. تم استكشاف وتقييم الخوارزميات الأكثر تمثيلاً، خوارزمية K-Means، وخوارزمية NNA مجاور وخوارزمية Lloyd، في هذه الدراسة بناءً على

*Email: shaimaa.qasem1601b@sc.uobaghdad.edu.iq

استراتيجياتها الأساسية. أثبتت الخوارزميات المقترحة كفاءة عالية في تصنيف البيانات، حيث كانت نتائج k–means عالية من خلال إنشاء نقاط بيانات وتصنيف تلك البيانات إلى 10 مجموعات، بينما أثبتت خوارزمية أقرب جار فعالية عالية في التنبؤ بمجموعات جديدة في ضوء المجموعات الموجودة مسبقًا لمجموعات جديدة نقاط البيانات، وأخيراً حققت خوارزمية Lloyd نتائج عالية من خلال عدة تكرارات للوصول إلى الفئات المستهدفة. باستعمال الدالة العشوائية، يتم إنشاء 100 قيمة عشوائية للمدخلات x و100 للمخرجات y، ويتم تجميع هذه القيم كنقاط. تشير حسابات المتوسط والانحراف المعياري لمجموعة البيانات إلى أن 68% من نقاط البيانات ستقع ضمن انحراف معياري واحد، و95% ضمن انحرافين معياريين، و99.7% ضمن ثلاثة انحرافات معيارية. في NNA، يتم تحديد الثقة في نقاط البيانات ضمن رقم انحراف معياري محدد بواسطة عامل التغطية، بالنسبة لـ k=10، من المتوقع أن تقع 97% من نقاط البيانات ضمن انحراف معياري واحد. في Lloyd، يظهر منحنى التوزيع الطبيعي عند توليد بيانات المعايرة أو القياس.

1. Introduction

Data mining is a technique for finding patterns and important knowledge from a large data set, also referred to as knowledge discovery from data (KDD). Due to the emergence of big data and the development of data storage technologies, the use of data mining techniques has increased over the past two decades, helping those interested in this field transform their raw data into usable knowledge [1] and [2].

Data mining techniques are classified into two categories: the first classifies the target data into groups and has specific algorithms, and the other predicts the results and also has specific algorithms. All of these algorithms fall under the specialty of machine learning [3].
The most important unsupervised learning problem is clustering, which, like many other problems of this type, involves defining a data structure in a set of larger data. The process of grouping elements into groups whose members are related by certain properties might be a broad definition of agglomeration [4]. Therefore, a cluster is a group of elements that are similar to each other by a set of properties common to them and dissimilar to those of other groups [5]. Clustering lacks any preset classes, in contrast to classification, which assigns items to predetermined classes [6]. The fundamental benefit of clustering is that it allows for the direct discovery of intriguing patterns and structures from very huge data sets with little to no prior knowledge [7].

The cluster outcomes rely on the implementation and are arbitrary. A clustering technique's effectiveness depends on [8]:

- The method's similarity metric and how it was used.
- Its capacity to unearth all or some of the buried patterns.
- The cluster's definition and selected representation.

Numerous clustering techniques have been put forth, and this research presents a comparison of the K-Means, Nearest Neighbor Algorithm (NNA), and Lloyd's Algorithm.

2. Related Work

Many previous studies dealt with the issue of data mining through the use of clustering techniques and compared the results of each technique by evaluating those results.

In this part of the paper, we will discuss the most important previous studies similar to our current study.

Gregory A. Wilkin and Xiuzhen Huang [9] investigated two uses of the broad data clustering technique, k-means clustering. They conducted an experiment to examine the

speeds and distances covered by Progressive Greedy K-mean Clustering and Lloyd's K-mean Clustering. Lloyd's K-means clustering is more effective based on their implementation of the methods, not only in terms of processing time but also in terms of the mean squared difference (MSD). Both a gene expression level sample and randomly generated datasets in 3D space were used in this investigation.

The Qingying Yu project group in 2016 [10] offered an externally eliminated k-means of differential specificity (OEDP) method that protects privacy and boosts clustering effectiveness. To retain specificity, the suggested method introduced Laplacian noise to the original data and chose the raw center points based on the distribution density of the data points. Comparative experiments and theoretical analyses were both conducted. A theoretical examination revealed that the suggested algorithm satisfies differential specificity. Also, the results of the experiments showed that, compared to other methods, the suggested algorithm did a better job of keeping data private and improving clustering results in terms of accuracy, stability, and availability.

The Shahadat Uddin project group in 2022 [11] used eight common machine learning datasets from the Kaggle repository, UCI Machine Learning, and OpenML to evaluate clustering factors in depth. The datasets were connected to numerous settings for diseases. For the comparison study, they took into account accuracy and recall performance metrics. These variables' mean accuracy values varied from 64.22% to 83.62%. The KNN clustering approach has the second-highest average accuracy (82.34%), followed by the KNN Advances scale (83.62%). To assess each variable and compare outcomes, a relative performance index based on each performance measure is also suggested. Based on the accuracy-based version of this index, the study determined that the benefits of KNN were the top-performing variable, followed by the KNN group method. This study also provided a comparative analysis of KNN variables based on retrieval and accuracy metrics. The article concluded by summarizing which KNN variation offers the best chance of success when taking into account the three illness prediction performance metrics (accuracy and recall).

In our current paper, Comparison between K-Means, NNA, and Lloyd's Data Clustering Algorithms, the results of the three algorithms will be compared to find the best way to predict new data and compare it with the data provided in the training phase.

3. Research Aim

Data points are grouped into k clusters using the unsupervised learning method K-Means. The system selects k cluster centroids at random, then assigns each data point to the nearest centroid. After that, the centroids are revised by averaging the data points given to each cluster, and the procedure is repeated until convergence [12].

NNA is a classifier that assigns a new data point a class based on the class of any nearby training data neighbors. The method first determines the distances between each point in the training set and the new data point, after which it chooses the class of the nearest neighbor(s) to categorize the new data point [13].

The k-means algorithm, commonly referred to as Lloyd's algorithm, is a K-means variation used to group data points into k clusters. The cluster centroids are chosen at random from a set of k locations, and each data point is then assigned to the nearest centroid. The process is repeated until convergence, at which time the centroids are updated by averaging the data points allocated to each cluster [14].

This paper aims to compare the results of the three algorithms and demonstrate the best of them in organizing data based on its distribution and grouping with the same number of nodes.

4. Methodology

In this study, the Python language and its libraries were used to calculate the distance between different input data points as a basis for analysis using three unsupervised clustering methods: K-Means, the Nearest Neighbor Algorithm (NNA), and Lloyd's Algorithm. Centroid clusters for each group were selected based on the separation between data points.

4.1 Dataset

In this paper, 100 random values of variable x and 100 other values of variable y were generated using the random function, where those values are grouped as points. Each value of x corresponds to a value of y; for example, a value of 4 is generated for variable x and a value of 20 for variable y, and then those two values are paired to form a point (4, 20).

Random data representing the x-value and the y-value of two dimensions were selected for the current study. These values will be paired using the proposed methods to produce points that reflect the two coordinates: the x-coordinate, represented by the point a, and the y-coordinate, represented by the value b, where the point appears as (a,b).

4.2 K-Means algorithm

K-Means is one of the simplest unsupervised learning methods to solve the well-known clustering problem. To categorize a given data set, the technique employs a specified number of clusters (let's say k clusters) that are defined a priori. The primary idea is to define k centroids, one for each cluster. These centroids need to be placed carefully since different positions have different consequences [14]. The first step in connecting each location in a given data set to the closest centroid is to situate them as far apart from one another as possible. The first step is finished when there are no open points and an early group is finished. Now that the clusters have centers, it is necessary to recalculate k new centroids and bound the identical data points to the centroids closest to them. This creates a loop [15]. This loop could lead to the observation that the k centroids gradually change places until no further changes are made. In other words, centroids are no longer moving. The procedure's main aim is to minimize an objective function, in this case, a squared error function. The main function is [16]:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \quad (1)$$

$\left\| x_i^j - c_j \right\|^2$ denotes the selected distance between a data point x_i^j and the cluster center c_j, and (n) denotes the number of data points. The distance between each data point and its corresponding cluster center. The following steps make up the algorithm [17]:

- K more points should be added to the region that the objects in the cluster represent. The first group's centroids are represented by these points.
- Each item should be distributed to the group whose centroid is closest to it.
- When all the objects have been distributed, update the positions of the k centroids.
- Until the centroids stop shifting, repeat steps 2 and 3.

In order to establish the measure that has to be lowered, the items are then separated into groups. The K-Means approach does not always pinpoint the configuration that relates to the minimum of the global objective function, despite the fact that it can be proved that the process will always come to an end [18]. The initial, randomly chosen cluster centers have a

considerable impact on how sensitive the algorithm is. To lessen this impact, the K-means method can be repeated several times. The straightforward algorithm K-Means has been applied to several problem fields. The program first chooses k random items. Because there is just one item in the cluster in this instance and each picked object represents a separate cluster, this object serves as the cluster's average or center.

4.3 Nearest Neighbor Algorithm NNA

NNA is one of the most basic but essential classification techniques in machine learning. Many data mining, intrusion detection, and pattern recognition applications are possible using it, which is categorized as unsupervised learning [19].

Due to its non-parametric character, which excludes any underlying assumptions about the distribution of data, it is usually useless in practical situations. We'll provide the training data that arranges coordinates according to an attribute.

Take the following two data points with two features as an illustration:

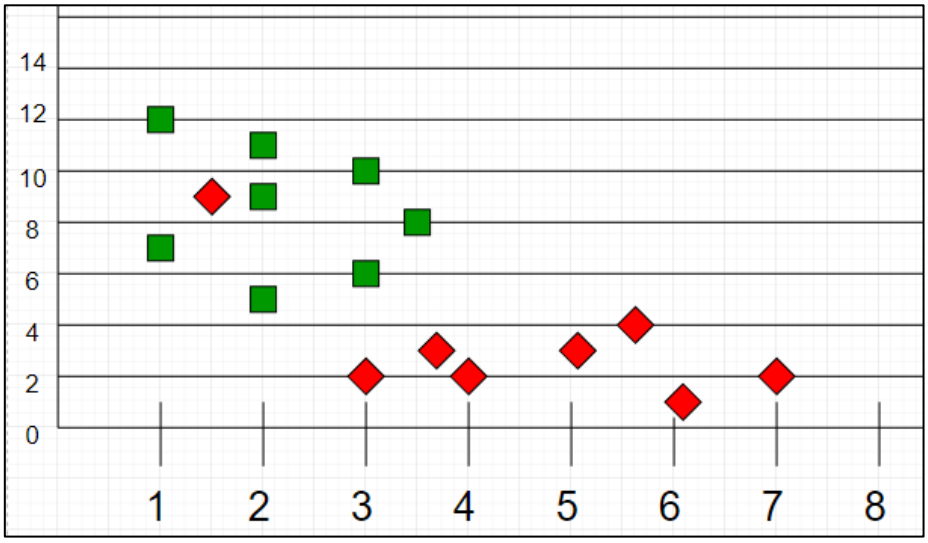

Figure 1: Two clusters of data points [20]

For another set of data points, assign these points to a group by looking at the training set (also known as testing data). White sites are those that are not categorized.

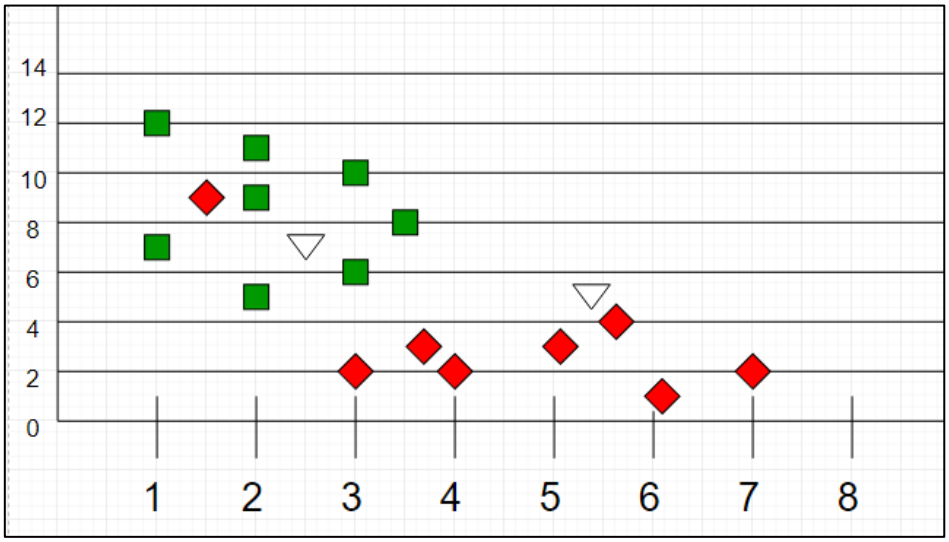

Figure 2: Two clusters of data points with two new data points [20]

These points can be put on a graph to reveal certain groups. It is now possible to assign a group to an unclassified site by identifying which group its closest neighbors belong to. This suggests that a point is more likely to be labeled red if it is adjacent to a collection of other points that have been selected [20].

The first point (2.5, 7) clearly belongs in the green category, but the second point (5.5, 4.5) clearly belongs in the red category.

- Assume that p is an unknown location and that m is the number of training data samples.
- Save practice samples in a collection of data points called arr[]. This indicates that each component of this matrix is an array (x, y).
- The Euclidean separation will be calculated between the data points. The distance between two locations is known as the Euclidean distance, and it may be computed as follows:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

- Compile a list of the S shortest distances out of K. Each of these distances relates to an earlier-found data point.
- Choose the S label that appears most frequently.

As Red and Blue are the only two groups that may exist, we can calculate a clear majority by keeping K as an odd number. When K grows, we get more rounded, more clear boundaries between different groups. Also, the accuracy of the aforementioned classifier improves as we increase the number of data points in the training set [21].

4.4 Lloyd's Algorithm

Stuart P. Lloyd invented the method currently known as Lloyd's algorithm, also known as Voronoi iteration or relaxation, for finding regularly spaced groups of points in subsets of Euclidean spaces and separating these subsets into well-shaped and properly sized convex cells. It continually finds the centroid of each set in each partition, much like the closely related k-means clustering method, and then divides the input into new segments based on which of these centroids is closest. In this case, the nearest centroid operation yields Voronoi diagrams, but the mean operation generates an integral across a region of space [22].

The strategy is most immediately applicable in the Euclidean plane, but similar techniques may also be applied in higher-dimensional spaces or in settings with a variety of non-Euclidean metrics. Voronoi tessellations of the input can be roughly calculated using Lloyd's method and then used for quantization, dithering, and stippling. Another use for Lloyd's strategy is in the finite element method's triangle mesh smoothing [23].

In the first phase of Lloyd's approach, the input domain is initially populated with k-point locations. These would be the mesh's vertices in mesh-smoothing applications; in other cases, they may be randomly distributed or produced by intersecting a suitable-sized uniform triangular mesh with the input domain. The following stage of relaxation is then repeated [24]:

- The k-site Voronoi diagram is calculated.
- The centroid of the Voronoi diagram is calculated after the integration of each cell.
- After that, each site is shifted to the centroid of its own Voronoi cell.

Approximations can be used instead of the exact steps for making this diagram and finding the exact centers of its cells, since making a Voronoi diagram can be very hard, especially for inputs with more than two dimensions. But for our task, two-dimensional points will do, and this is what this study will focus on [25].

Although embedding in other spaces is also feasible, this explication uses the L2 norm and assumes Euclidean space while discussing the two most pertinent cases, which are two and three dimensions, respectively.
Since a Voronoi cell always surrounds its site and has a convex shape, there are minor faults that can easily merge [26] and [27]:

- The polygonal cell's site and edges are joined in two dimensions to form an umbrella-shaped collection of triangles.
- A weighted combination of the centroids of the simplest is now used to combine the cell and obtain the centroid (center of mass) (it will be called ci).
- Using Cartesian coordinates, for example, it is simple to calculate the centroid of a triangle.
- Simplex-to-cell area ratios are used to derive weighting.

For a 2D cell with overlapping area and triangular simplifiers $A_C = \sum_{i=0}^{n} a_i$ (where $a_i$is the area of a triangle simplex), the new cell centroid computes as:

$$C = \frac{1}{A_C} \sum_{i=0}^{n} c_i a_i \tag{3}$$

5. Results and Discussion

The results of the three suggested algorithms, which were put into practice using the Python programming language, will be given and discussed in this section of the study.

5.1 K-means Results

In order to cluster the data using K-means, we must define K, or how many clusters we wish to use. We may graph inertia—a measurement dependent on distance—using the elbow approach to see when it begins to linearly decline.

```
[(4, 21), (5, 19), (10, 24), (4, 17), (3, 16), (11, 25), (14, 24), (6, 22), (10,
21), (4, 21), (12, 21), (5, 19), (14, 24), (6, 17), (10, 16), (5, 25), (4, 24),
(6, 22), (3, 21), (10, 21), (11, 24), (6, 18), (3, 20), (4, 19), (6, 18), (5,
17), (10, 25), (4, 21), (4, 19), (11, 24), (10, 17), (3, 16), (10, 25), (6, 24),
(4, 22), (3, 21), (5, 21), (10, 24), (6, 18), (10, 20), (4, 19), (8, 18), (9,
17), (3, 25), (5, 29), (10, 29), (4, 21), (6, 19), (5, 24), (7, 17), (4, 16),
(10, 25), (9, 24), (4, 22), (5, 21), (10, 21), (4, 21), (3, 19), (11, 24), (14,
17), (6, 16), (10, 25), (4, 24), (12, 22), (5, 21), (14, 21), (6, 24), (10, 18),
(5, 20), (4, 19), (6, 18), (3, 17), (10, 25), (11, 21), (6, 19), (3, 24), (4,
17), (6, 16), (5, 21), (10, 19), (4, 24), (4, 17), (11, 16), (10, 25), (3, 24),
(10, 22), (6, 21), (4, 21), (3, 21), (5, 19), (10, 24), (6, 17), (10, 16), (4,
25), (8, 24), (9, 22), (3, 21), (5, 21), (10, 24), (6, 18)]
```

Figure 3: Dataset as data points

First, the system represented the data as points for the values of (x) and (y) and determined the distance between them formally by drawing a straight line between one point and another.

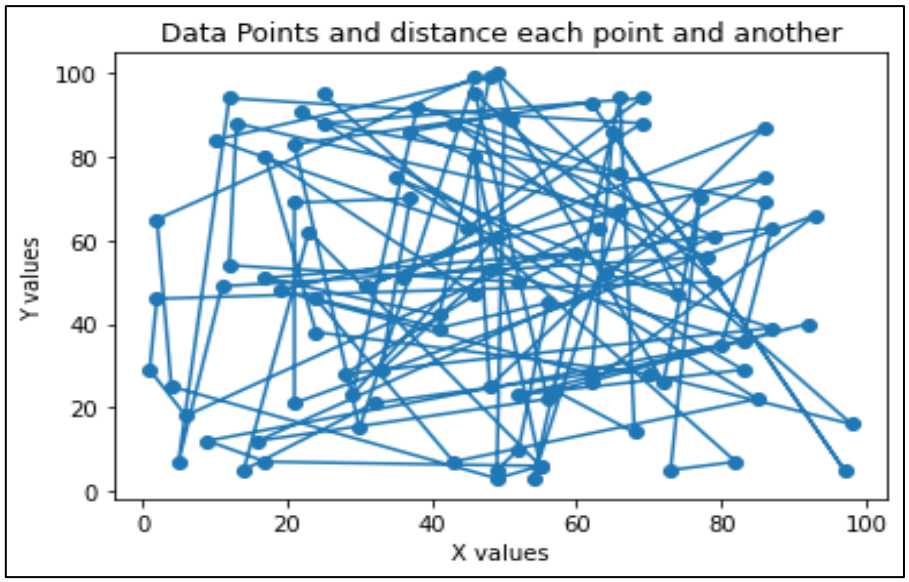

Figure 4: K-means data points and distances between each point and another

The inertia is now shown using the elbow approach for various values of K.

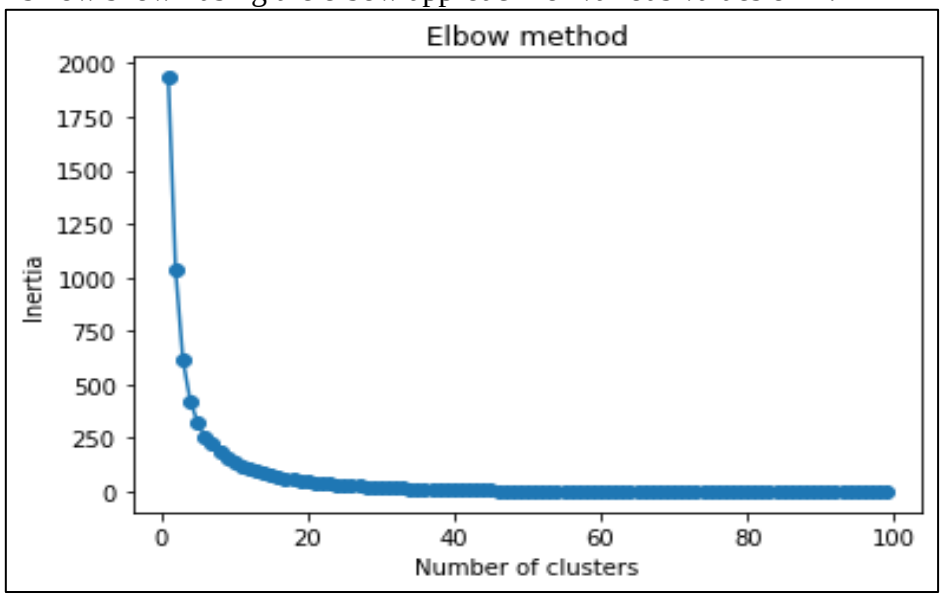

Figure 5: Data points after the elbow method

The x-axis in Figure 5 represents the number of data points, which is 100, and the y-axis represents inertia. The graph above shows an elbow at K = 10, when the interaction becomes more linear. Next, we can train our K-means algorithm again and visualize the various clusters that were given to the data:

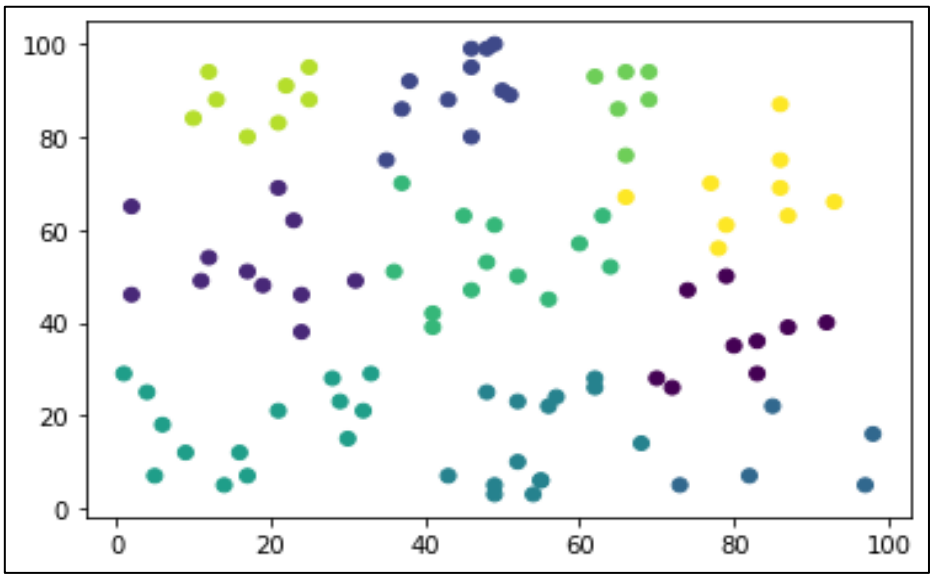

Figure 6: Data Point Clusters

When the k-means method is applied to the data points, as seen in Figure 6, the cluster is shown by ten distinct colors, with each color designating a set of data that belongs to the same cluster. By adjusting K's value, it is possible to manage the cluster count. By measuring the separation between two points, the data points were categorized and organized into groups based on the shortest distance.

5.2 Nearest Neighbor Algorithm NNA Results

In this case, the NN algorithm implementation will highlight how categorization works and how different K values influence the outcomes.

As in the K-means algorithm, in the ANN algorithm, the system represents the data as points for the values of (x) and (y) and determines the distance between them formally by drawing a straight line between one point and another.

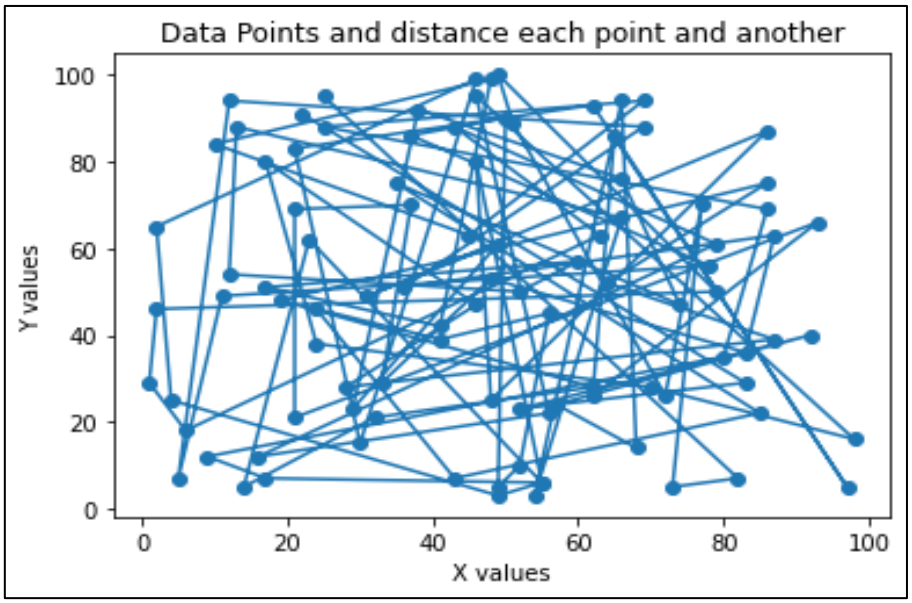

Figure 7: ANN data Points and distances between each point and another

K is the chosen value for the nearest neighbors. A majority vote determines which class a new observation should belong to for classification. Larger values of K often offer more reliable decision limits and are frequently more resistant to outliers than extremely small numbers (K = 1 would be better than K = 3).

The identical cluster from Figure 4's k-means algorithm was employed in this technique.

The NNA technique is needed to identify the cluster of the two new data points because they were introduced without a cluster (i.e., of an unknown class). As seen in Figure 8, the new points are identified by their placement inside a black ring.

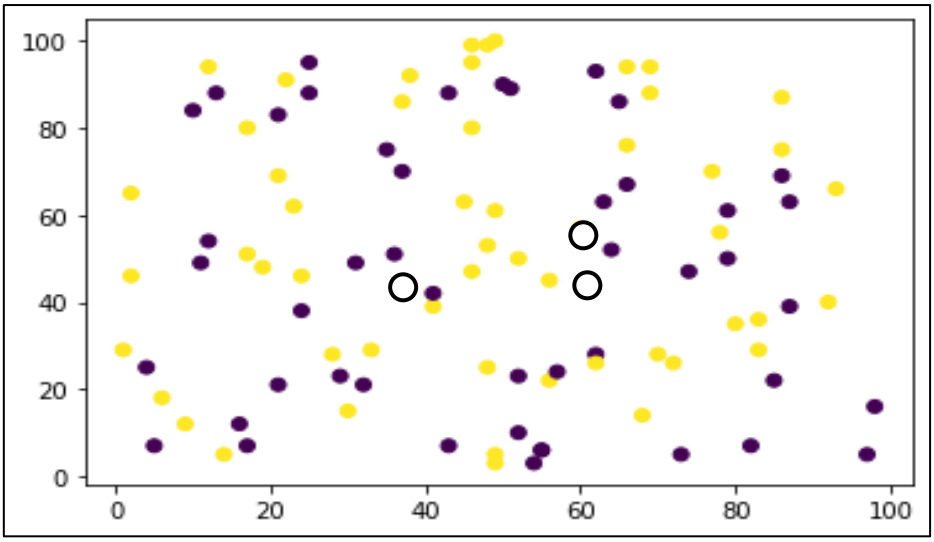

Figure 8: Data points cluster with three new points

The class of unexpected new data points may then be predicted using the same KNN object. A class of 0 or 1 is obtained by using knn.predict() on the new data point after creating new x and y features.

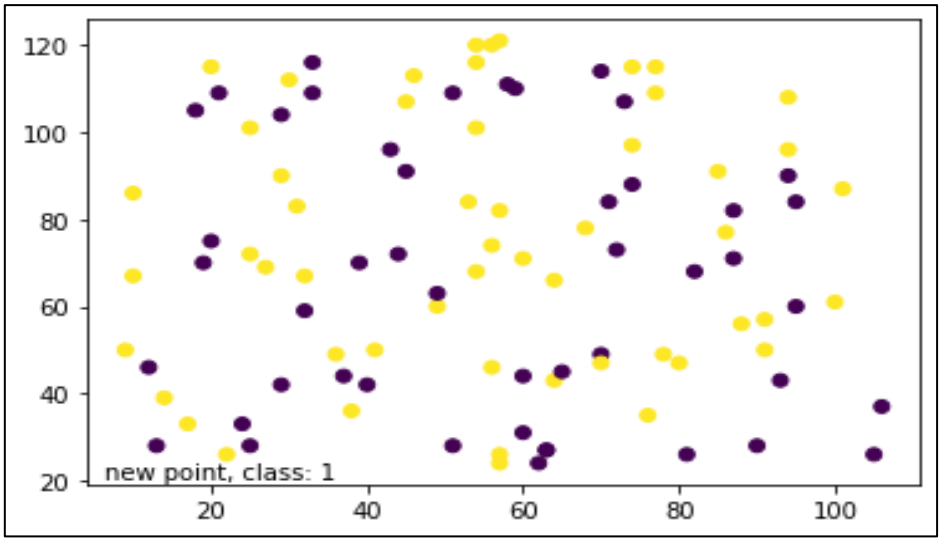

Figure 9: Data points cluster with two new points

The number of points utilized to categorize our new points changes as the neighbors increase to 5. The new point is therefore also categorized as follows:

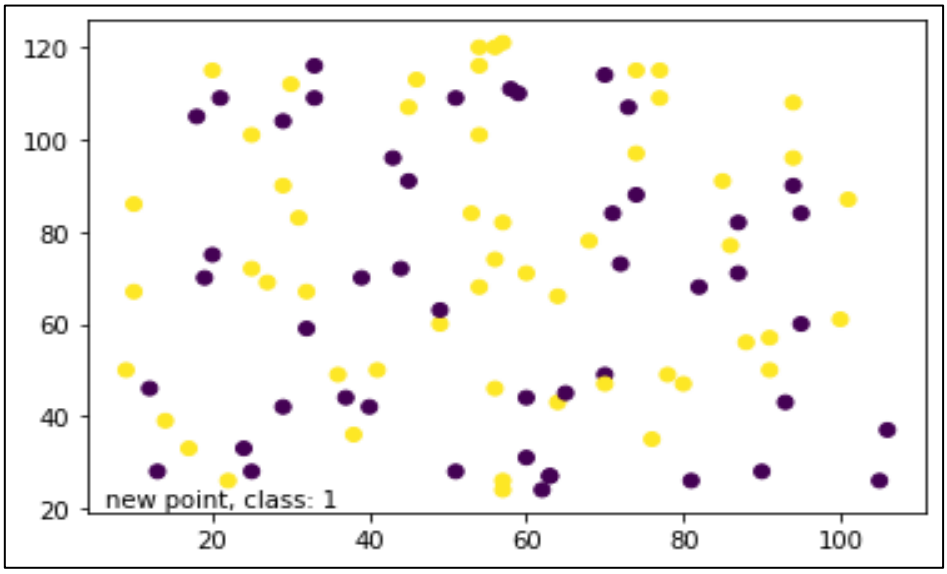

Figure 10: Data points clusters with two new points and class = 1

5.3 Lloyd's Algorithm Results

K-Means Clustering uses the Expectation-Maximization Algorithm to tackle the issue. Iteratively approaching the greatest likelihood function is how it is done.

These procedures enable K-means clustering to locate cluster centroids iteratively. The k-means technique may be used to cluster data points into groups that are each represented by a centroid using Lloyd's two-step implementation. This method is used in many different areas of machine learning, including dimensionality reduction issues and unsupervised learning methods. Although being NP-hard, the iterative process always converges, albeit to a local minimum, on the generic clustering problem. It's crucial to initialize the centroids properly. The ideal K for the k-means cannot be determined using this algorithm; it must be determined using other techniques. As seen in Figure 11, data are sorted into unclassified categories when Lloyd's technique is used:

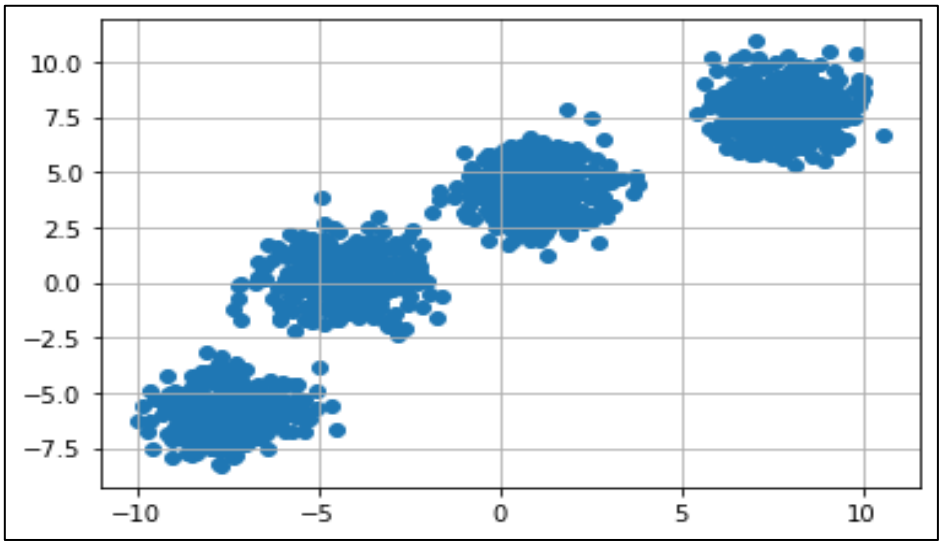

Figure 11: Dataset with initial uncategorized clusters

The approach is demonstrated by setting the k-means function to find four groups for a collection of 100 randomly chosen points on a plane. 10 iterations after startup with random centers, the algorithm converges. The stars in the following plots stand in for the two points μ_k from the clusters, while the dots in the plots correspond to the target data points X. A distinct color represents each group.

Here, Euclidian distance is employed. Yet, there are several techniques we may employ to determine distance.

The data is categorized in the initial iteration, but with very little precision because the clusters overlap, as seen in Figure 12.

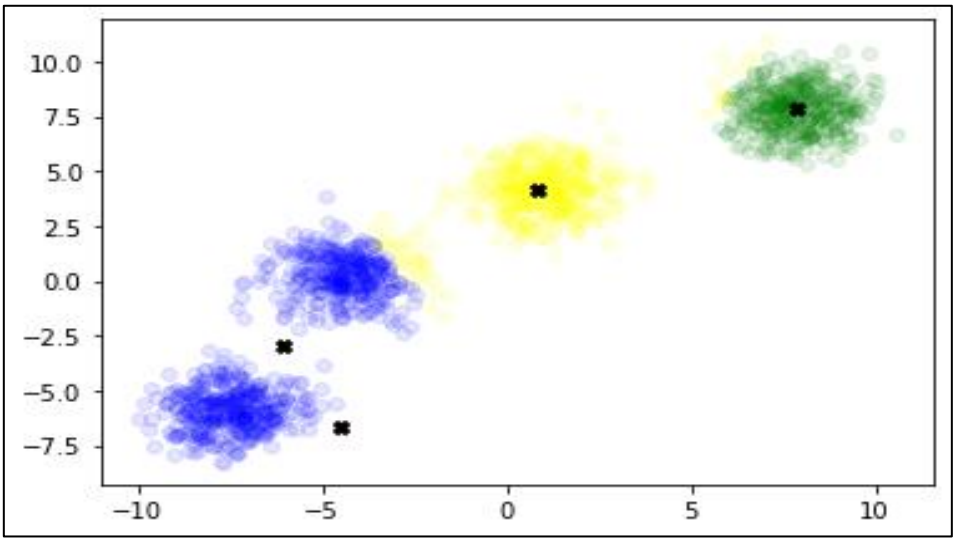

Figure 12: Dataset clusters with initial iteration

The risk arises that the local minimum achieved is not the best option if the target distribution is disjointedly clustered and just one instance of Lloyd's procedure is used. This is seen in the example below, where starting data is created using extremely peaked Gaussians:

The technique frequently requires more iterations to converge for a setup with twice as many points and a target of four clusters.

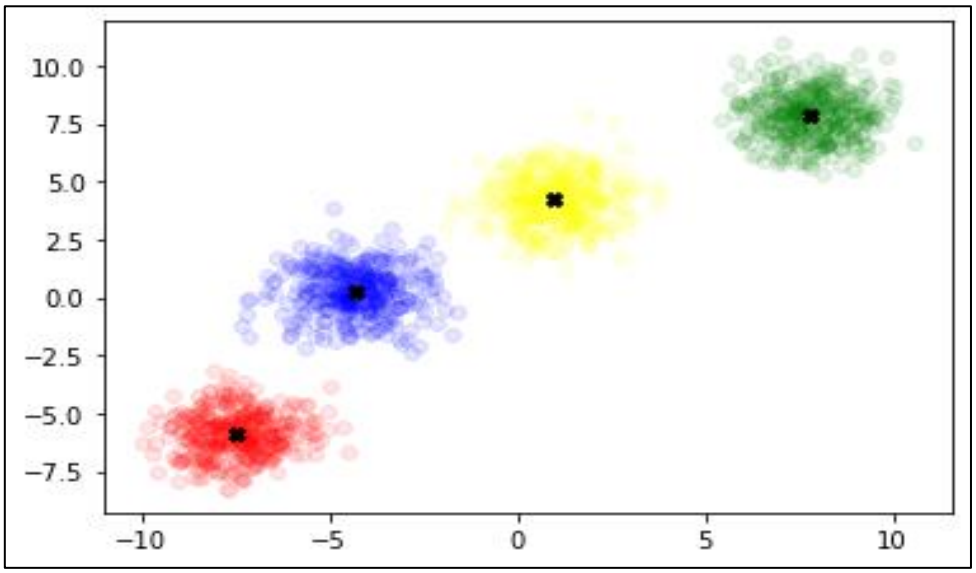

Figure 13: Dataset Clusters with Final Iteration

In Figure 13, it can be seen that after the final iteration, the data clusters were separated by four distinct colors, with each hue denoting a distinct cluster from the others.

Plotting the heights of several randomly chosen individuals is a nice example of a statistically valid data set with a normal distribution, also known as a Gaussian distribution, which will result in a bell curve. The data set's mean and standard deviation calculations reveal that roughly 68% of the data points will be within one standard deviation of the mean, 95% of them will be within two standard deviations, and 99.7% of them will be within three standard deviations.

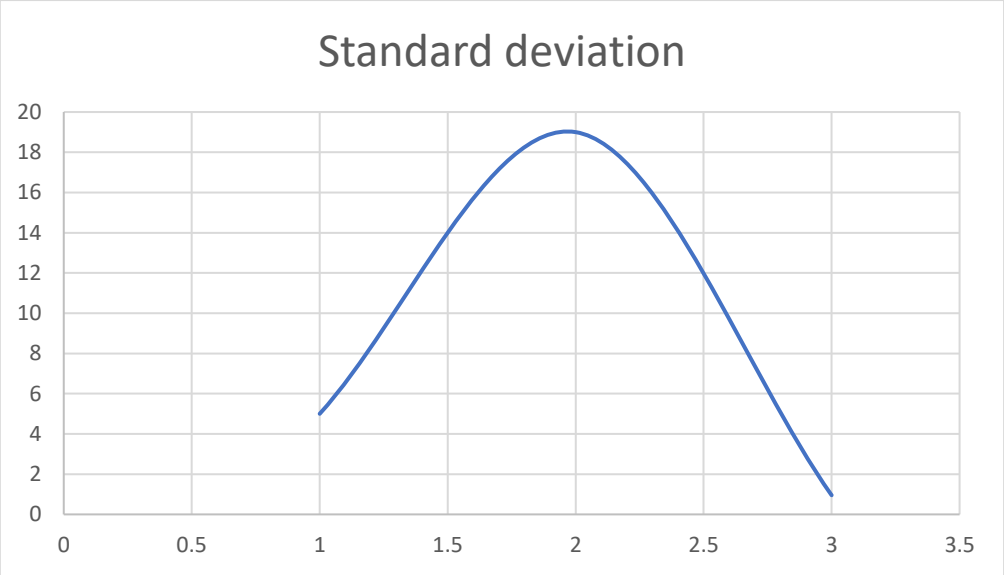

Figure 13: Standard deviation for k-means with 10 clusters

A normal distribution curve looks exactly like this when calibration or measurement data are produced. The confidence in the data points within a specific standard deviation number is determined by the coverage factor, or k-value. 95% of the data points for k = 10 are predicted to be within one standard deviation.

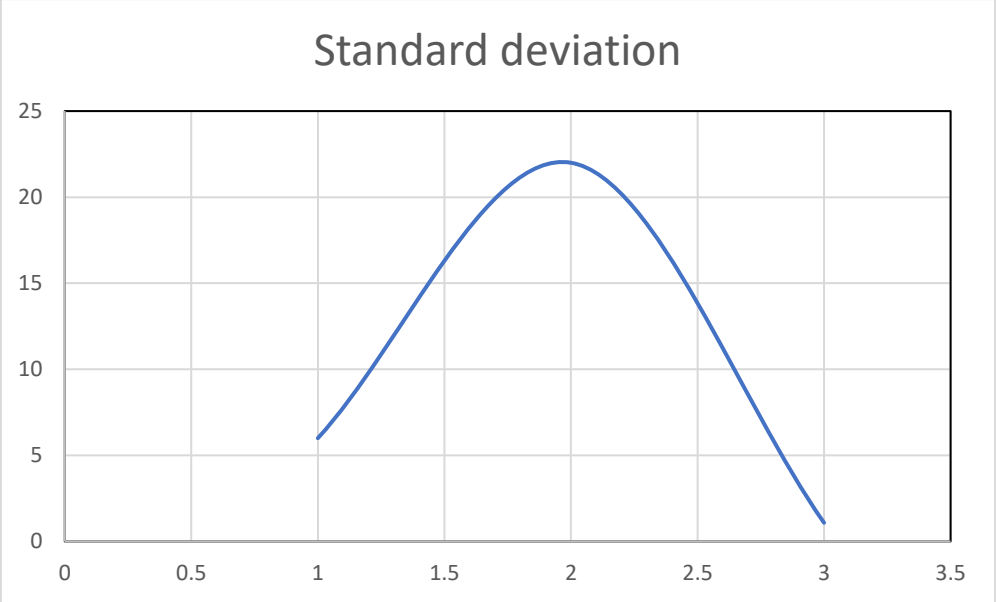

Figure 14: Standard deviation for NNA with 10 clusters

For NNA, this is a representation of a normal distribution curve for calibration or measurement data. The coverage factor, often known as the k-value, establishes the degree of confidence in the data points falling inside a given standard deviation. For k = 10, it is expected that 95.8% of the data points will be within one standard deviation.

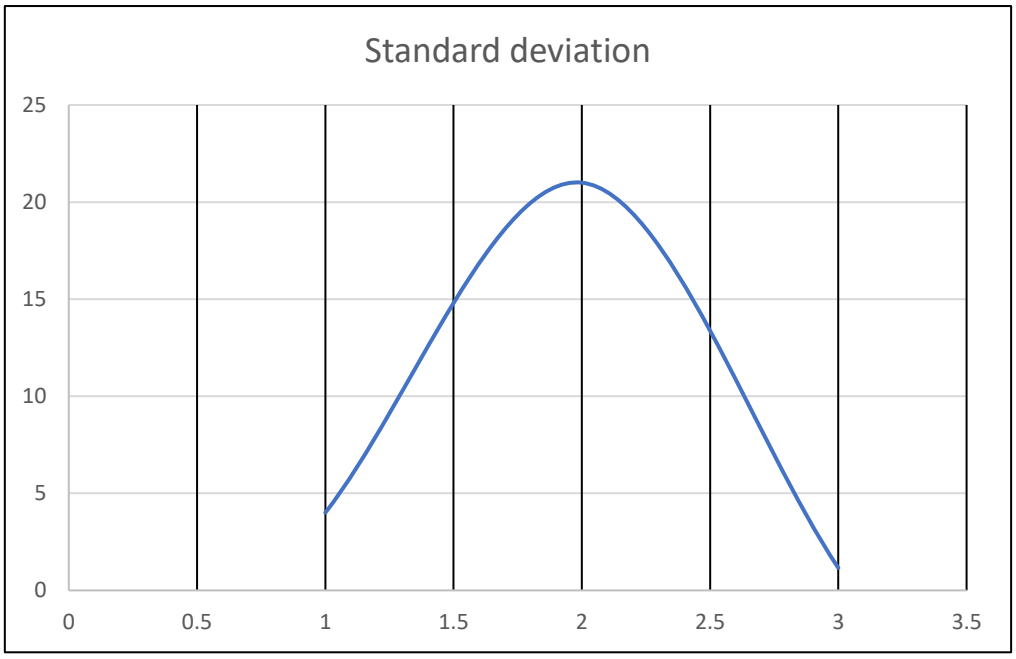

Figure 15: Standard deviation for Lloyd's with 10 clusters

For Lloyd's, when calibration or measurement data are generated, a normal distribution curve appears just like this. The coverage factor, or k-value, determines how confident one may be in the data points falling within a given standard deviation. For k = 10, 97% of the data points are expected to fall within one standard deviation.

6. Conclusion

The paper provides a comprehensive analysis of these three prominent clustering algorithms. Throughout the paper, several problems and limitations in each of these algorithms are identified. To enhance the effectiveness of these algorithms, it is necessary to address these issues.

In the K-means algorithm, it is very sensitive to the choice of the initial centroid. To improve its performance, researchers can explore more robust initialization methods, such as K-means, to reduce the probability of convergence to suboptimal solutions. Incorporating outlier detection mechanisms or using modified distance metrics that are less sensitive to outliers can make K-means more robust in scenarios where outliers are prevalent.

In the nearest neighbor algorithm, the nearest neighbor can become computationally intensive as the size of the data set grows. Implementing approximate neighbor techniques or using tree-based data structures such as KD trees can greatly enhance scalability. Addressing the curse of dimensionality by using dimensionality reduction techniques such as PCA or considering more advanced distance metrics that consider high-dimensional spaces can improve the accuracy of nearest neighbor clustering.

As for the Lloyd's clustering algorithm, it is known that it converges slowly or gets stuck at local minima. Improvements could include exploring alternative optimization techniques such as mini-batch K-means or using convergence acceleration methods.

To improve their ability to deal with non-spherical clusters, researchers can investigate the integration of advanced distance metrics, such as the Mahalanobis distance, or adopt model-based clustering techniques, such as Gaussian mixture models.
Furthermore, it should be noted that the choice of clustering algorithm should be guided by the specific characteristics of the dataset and the objectives of the clustering task. There is no single algorithm that is universally superior, and choosing the most appropriate algorithm requires careful consideration of the nature of the data and the problem at hand.

In future research, exploring hybrid or combined clustering methods that combine the strengths of multiple algorithms could be a promising direction to overcome the inherent limitations of individual methods. In addition, leveraging the power of deep learning and neural network-based clustering techniques may offer innovative solutions to address some of the persistent challenges in clustering.

In summary, while K-Means, NNA, and Lloyd's Clustering algorithms have their own limitations, continued research and innovation hold the potential to improve their performance, making them more versatile and effective tools for a wide range of clustering applications.

References

[1] Dunham M., Data Mining: Introductory and Advanced Topic, *1st Edn.* USA : Prentice Hall, ISBN: 10: 0130888923, 2020, pp. 315.

[2] Tareef K. Mustafa, Ammar A, Abdul Razzaq, and Ehsan A. Al-Zubaidi, "Authorship Arabic Text Detection According to Style of Writing by using (SABA) Method," *Asian Journal of Applied Sciences,* vol. 5, no. 2, April 2017. DOI: https://doi.org/10.24203/ajas.v5i2.4750.

[3] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, *2nd Edn.* New Delhi : Morgan Kaufmann Publishers, ISBN: 978-81-312-0535-8., 2006.

[4] Enan A. Khalil, and Bara'a A. Attea, "Energy-aware evolutionary routing protocol for dynamic," *Swarm and Evolutionary Computation, Elsevier, ed. 1,* 2011, pp. 195-203. DOI: https://doi.org/10.1016/j.swevo.2011.06.004.

[5] Park, H.S., J.S. Lee and C.H. Jun., "A K-means like algorithm for K-medoids clustering and its performance," *POSTECH,* South Korea, 2006. https://api.semanticscholar.org/CorpusID:9220990.

[6] Bara'a A. Attea, "Improving the Performance of Evolutionary Multi-objective Co-clustering Models for Community Detection in Complex Social Networks," *Swarm and Evolutionary Computation, Elsevier,* vol. 26, pp. 137–156, 2016. DOI: 10.1016/j.swevo.2015.09.003.

[7] Rakhlin, A. and A. Caponnetto, "Stability of k-Means clustering," *Adv. Neural Inform. Process. Syst.*, vol. 12, pp. 216-222, 2007. https://doi.org/10.1007/978-3-540-72927-3_4.

[8] Xiong, H., J. Wu and J. Chen., "K-Means clustering versus validation measures: A data distribution perspective," *IEEE Trans. Syst., Man, Cybernet. Part B,* vol. 39, pp. 318-331, 2009. https://doi.org/10.1145/1150402.1150503.

[9] Gregory A Wilkin and Xiuzhen Huangcorresponding, "A practical comparison of two K-Means clustering algorithms," *BMC Bioinformatics,* vol. 9, no. 6, p. S19, 2008. https://doi.org/10.1186/1471-2105-9-S6-S19.

[10] Qingying Yu, Yonglong Luo, Chuanming Chen, and Xintao Ding, "Outlier-eliminated k-means clustering algorithm based on differential privacy preservation," *Applied Intelligence* vol. 45, no. 4, pp. 1179–1191, 2016. https://doi.org/10.1007/s10489-016-0813-z.

[11] Shahadat Uddin, Ibtisham Haque, Haohui Lu, Mohammad Ali Moni, and Ergun Gide., "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scince Report,* vol. 12, p. 6256, 2022. DOI: https://doi.org/10.1038/s41598-022-10358-x.

[12] Saadoon Abdoon R., "Watershed Transform Based on Clustering Techniques to Extract Brain Tumors in MRI," *Iraqi Journal of Science*, vol. 57, no. 1B, pp. 540–555, 2023. https://ijs.uobaghdad.edu.iq/index.php/eijs/article/view/9257.

[13] Salman, N. H., and Mohammed, S. N., "Image Segmentation Using PSO-Enhanced K-Means Clustering and Region Growing Algorithms," *Iraqi Journal of Science*, vol. 62, no. 12, pp. 4988–4998, 2022. https://doi.org/10.24996/ijs.2021.62.12.35.

[14] D. S. Al-Azzawy and Faiez M. L. Al-Rufaye, "Arabic words clustering by using K-means algorithm," *IEEE, Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, Iraq*. pp. 263-267, 2017. Doi: 10.1109/NTICT.2017.7976098.

[15] S. Mahmood and Faiez M. L. Al-Rufaye, "Arabic text mining based on clustering and coreference resolution," *International Conference on Current Research in Computer Science and Information Technology (ICCIT),* Sulaymaniyah, Iraq, pp. 140-144, 2017. Doi: 10.1109/CRCSIT.2017.7965549.

[16] Jain, A.K., M.N. Murty and P.J. Flynn, "Data clustering: A review," *ACM Comput. Surveys* vol. 31, pp. 264-323, 1999. DOI: https://doi.org/10.1145/331499.331504.

[17] Khan, S.S. and A. Ahmad., "Cluster center initialization algorithm for K-Means clustering," *Patt. Recog. Lett.*, vol. 25, pp. 1293-1302, 2004. https://doi.org/10.1016/j.patrec.2004.04.007.

[18] Sarab M. Hameed, Rasha A.M. Altea, and Bara'a A. Attea, "Fuzzy Based Clustering For Grayscale Image Steganalysis," *Iraqi Journal of Science*, vol. 56, no.2A, pp. 1161-1175, 2015. https://www.ijs.uobaghdad.edu.iq/index.php/eijs/article/view/10266

[19] F. M. Lahmood Al-Rufaye and H. I. Mhaibes, "Decision Tree Technique for Arabic Sentences Classification with Preprocessing of NLP by Using of Words Features," *Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT),* Bhilai, India. pp. 1-6, 2022. Doi: 10.1109/ICAECT54875.2022.9808024.

[20] Ben-David, S., Pál, D., Simon, H.U, "Stability of k-Means Clustering," In: *Bshouty, N.H., Gentile, C. (eds) Learning Theory. COLT 2007. Lecture Notes in Computer Science(), vol 4539. Springer, Berlin, Heidelberg*, 2007. https://doi.org/10.1007/978-3-540-72927-3_4

[21] Mohammed. M. I. Al-Sagheer and F. M. Lahmood. Alrufaye, "Data Mining and RBF Neural Networks to Analyze Data from COVID-19 Patients and Predict New Cases Based on Symptoms," *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA),* Ankara, Turkey, pp. 1-6, 2022. Doi: 10.1109/HORA55278.2022.9799979.

[22] H. Xiong, J. Wu and J. Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 318-331, April 2009, Doi: 10.1109/TSMCB.2008.2004559.

[23] F. Soleymani, S. Miah and D. Spinello, "Temporal Difference Learning of Area Coverage Control with Multi-Agent Systems," *IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, Abu Dhabi, United Arab Emirates, 2022, pp. 1-8, Doi: 10.1109/ROSE56499.2022.9977412.

[24] C. Taylor and M. Gowanlock, "Accelerating the Yinyang K-Means Algorithm Using the GPU," *IEEE 37th International Conference on Data Engineering (ICDE)*, Chania, Greece, pp. 1835-1840, 2021. DOI: 10.1109/ICDE51399.2021.00163.

[25] X. Wu, "On convergence of Lloyd's method I," in *IEEE Transactions on Information Theory*, vol. 38, no. 1, pp. 171-174, Jan. 1992, Doi: 10.1109/18.108266.

[26] Z. Jin, T. Tillo, J. Xiao and F. Cheng, "3-D video depth map quantization based on Lloyd's algorithm," *IVMSP 2013*, Seoul, Korea (South), 2013, pp. 1-4, Doi: 10.1109/IVMSPW.2013.6611903.

[27] Ghathwan, K. I., & Mohammed A. J., "Intelligent Bat Algorithm for Finding Eps Parameter of DbScan Clustering Algorithm," *Iraqi Journal of Science*, vol. 63, no. 12, pp. 5572–5580, 2022. DOI: https://doi.org/10.24996/ijs.2022.63.12.41.