A COMPARISON OF K-MEANS AND MEAN SHIFT ALGORITHMS

Mehak Nigar Shumaila
Dept. of Information Technology
The OWL "Technische Hochschule Ostwestfalen-Lippe"
Germany
mehak.shumaila@stud.th-owl.de

Abstract—Clustering, or otherwise known as cluster analysis, is a learning problem that takes place without any human supervision. This technique has often been utilized, much efficiently, in data analysis, and serves for observing and identifying interesting, useful, or desired patterns in the said data. The clustering technique functions by performing a structured division of the data involved, in similar objects based on their characteristics that it identifies. This process results in the formation of groups, and each group that is formed, is called a cluster. A single said cluster consists of objects from the data, that have similarities among other objects found in the same cluster, and resemble differences when compared to objects identified from the data that now exist in other clusters. The process of clustering is very significant in various aspects of data analysis, as it determines and presents the intrinsic grouping of objects present in the data, based on their attributes, in a batch of unlabeled raw data. A textbook or otherwise said, good criteria, does not exist in this method of cluster analysis. That is because this process is so different and so customizable for every user, that needs it in his/her various and different needs. There is no outright best clustering algorithm, as it massively depends on the user's scenario and needs. This paper is intended to compare and study two different clustering algorithms.The algorithms under investigation are k-mean and mean shift.These algorithms are compared according to the following factors: time complexity , training , prediction performance and accuracy of the clustering algorithms.

Index Terms—K-Mean, Mean-Shift, Performance, Accuracy

I. INTRODUCTION

The quickened progress of technology in recent time is encouraging a significant expansion in the measure of created and stored information in fields like education, engineering, training, medication, and trade, among others [1] [2]. Accordingly, there is an advocated interest in getting valuable information that can be extracted from those tremendous measures of information, to assist with settling on better choices and understanding the idea of information [3]. Clustering is one of the fundamental techniques for getting insight on the underlying nature and structure of data. Clustering, or otherwise known as cluster analysis, is a learning problem that takes place without any human supervision. This technique has often been utilized, much efficiently, in data analysis, and serves for observing and identifying interesting, useful, or desired patterns in the said data [4]. The clustering technique functions by performing a structured division of the data involved, in similar objects

based on their characteristics that it identifies. This process results in the formation of groups, and each group that is formed, is called a cluster. A single said cluster consists of objects from the data, that have similarities among other objects found in the same cluster, and resemble differences when compared to objects identified from the data that now exist in other clusters [5]. This process works to assemble the said data in well-structured classes or groups, that we call clusters, so as the objects from the data that are now in one cluster, resemble higher properties of having homogeneous attributes within themselves inside the cluster than the homogeneity representation when compared to objects existing in various other clusters in the process of the division [6]. The process of clustering is very significant in various aspects of data analysis, as it determines and presents the intrinsic grouping of objects present in the data, based on their attributes, in a batch of unlabelled raw data [7]. One of the clustering algorithms more widely used to date is K-means, due to its easiness for interpreting its results and implementation.

The term "k-means" was first used by James Mac queen in 1967 [8] as part of his paper on "some methods for classification and analysis of multivariate observations". The standard algorithm was also used in bell labs as part of a technique in pulse code modulation in 1957 [9]. it was also published by in 1965 by e. w. forgy and typically is also known as the lloyd-forgy method [10]. Clustering is the process of splitting a population or set of data points into several groups so that data points in the same group are more similar than data points in other groups. To put it another way, the goal is to separate groups with similar characteristics and assign them to clusters. The k-means algorithm's aim is to discover groups in data, with the variable k representing the number of groups. Based on the features presented, the algorithm assigns each data point to one of k groups iteratively. In the reference image below, k=2 and two clusters from the source data set have been highlighted.

II. MOTIVATION

K-means can typically be applied to data that has a smaller number of dimensions, is numeric, and is contin-

uous. Following is a list of some interesting use cases for k-means [11]:

- Document classification
- Delivery store optimization
- Identifying crime localities
- Customer segmentation
- Fantasy league stat analysis
- Insurance Fraud Detection

In order to detect frauds ,Telecom companies [12] use Call Detail Record which contains information of call, SMS, and Internet activity of a customer, in order to detect fraud detection by clustering the user profiles, reducing customer churn by usage activity.The customer activities for 24 hours were clustered using k means algorithm.

There can be improvements that can be done to IT operations using data science techniques. The major goal of a data science-driven IT process is to be able to perform automated root cause finding and failure prediction. To achieve this, Derek et al., [13] used K-Means clustering algorithm to develop and IT alert system that notifies the users about different incidents based on their category.

As internet usage is increasing rapidly , researcher studied [14] about the log activity alternate to user behaviour.To find a solution for the analysis of user behavior clustering technique was carried using a k-mean algorithm for the cyber profiling process. The study revealed that Internet users in higher educational institutions are more accessible to websites for searching information and also social media has a high-level visit after website search engine. Research related to profiling, among others, performed by [15]. In these studies, used machine learning to help the process of profiling to assist the experts in analyzing the crime.

Mean-Shift has been widely used in many applications such as:

- Image Segmentation
- Outlier Detection Methods
- Tracking objects

There many improvements that can be done to rectify the outlier detection methods.Traditional modals were created which had many outliers which later actually polluted the models making them inefficient and unreliable.Panel et al., [16] worked on improving the modal and proposed mean-shift outlier detector which modify data and cancel the bias caused by the outliers.

Object tracking is a critical mechanism that allows a variety of applications, including retail space instrumentation, which analyzes customer shopping behavior and improves building and environment design. One of the primary goals of object tracking is to estimate the object's position in pictures or video. For this researchers [17] worked on object detection using mean shift algorithm and implemented the mean shift on each window. In this method, a rectangular target window for a moving target in a video is identified in the first frame, and the tracked object is then separated

from the background by processing the data inside that window.

III. METHODOLOGY

This research process has the objective of formulating a comparative study between two algorithms, that of K-means and Mean Shift. This process will focus significantly on obtaining research data on both algorithms in various ways and the goal is to construct a reasonable and accurate comparison between the two algorithms that are utilized widely. For the task of running these algorithms to obtain data on them, I need a data-set to run the said algorithm. For this particular research, we have decided on running K-means and Mean Shift algorithm on the data-set known as IRIS (Iris Flower Data set), also known as the Fisher's Iris data set and wine data set. I feel that this particular data-sets serves us perfectly for the purpose of this research.
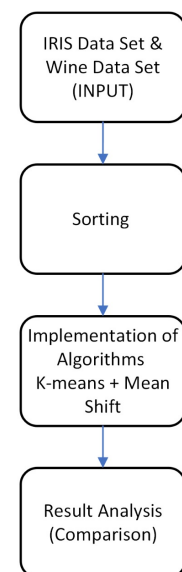

Fig. 1: Flow Diagram of Proposed Work Flow

The methodology process proposed is as follows:

1) I will be testing K-means and Mean Shift algorithms on IRIS and WINE , pre-selected data set for this task. I will be obtaining data from this test which will be used to construct a comparative study between the two said algorithms.

2) The data obtained from the testing will be divided into two particular phases, for the sake of this research. The testing phase and the training phase. Running the data through the testing and training phases will display a better understanding and analysis of our data. One of the integral and most vital type of data that assists the machines to study, learn and formulate the predictions, is Training data. In this process, arguably the most crucial section to the assessment of the data mining models is to dissect and separate the data in two sets, the training set and the testing

set. Generally, in this process, when you dissect and separate the data in the two aforementioned steps, the majority of the data is normally used for the training phase, and a smaller proportion is reserved for testing.
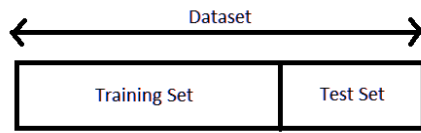

Fig. 2: Train and Test Set

I have created a class for this purpose which when initialized takes in a iris data set and wine data set and divides it into train and test data set.

3) The results obtained after following the aforementioned steps will be accumulated. This data will be the spine of this research process as we will be using this to form a meaningful, fair, and accurate comparison of the strengths, uses, and advantages of the said algorithms. The focus during this entire methodology has been to make the comparison as fair, useful, and accurate as possible, and I will be making a specific effort to ensure that this objective will be achieved in this step. The comparative study will be devised, and the results, points, and conclusions obtained will be discussed.

A. Selection of Data sets

In order to test the efficiency of both algorithms Iris data set and winde is taken and been used in this experiments.The properties of Iris data set is mentioned in Table.

TABLE I: IRIS DATA SET

Data Set	No of Classes	No of Attributes	Instances
Iris Data	classes 1.Iris Setosa, 2. Iris Versicolour 3.Iris Virginica	1. sepal length in cm 2. sepal width in cm 3. petal length in cm 4. petal width in cm 5. class	150 for each attribute

This data set is often used in data mining , clustering , classification mostly to test the algorithms.Due to its popularity I used this data set.The data set contains fours measures for 150 flowers.Four features like sepal length , sepal width , petal length and petal width [18]. It contains 50 samples of three species of iris:

1) Iris Setosa
2) Iris Versicolour
3) Iris Virginica

Overall , this is a good data set to test time series clustering algorithms.

TABLE II: WINE DATA SET

Data Set	Types	No of Attributes	Instances
Wine Data	3 types of wine cultivars	1)Alcohol 2) Malic acid 3) Ash 4) Alcalinity of ash 5) Magnesium 6) Total phenols 7) Flavanoids 8) Nonflavanoid phenols 9) Proanthocyanins 10)Color intensity 11)Hue 12)OD280/OD315 of diluted wines 13)Proline	178 for each attribute

The data set is most of the time used for clustering. The analysis determined the quantities of 13 constituents found in each of the three types of wines [19].

B. How Algorithms are implemented?

I decided to use Python language and implemented the algorithms in Google Colab.

1) Python: For this particular research project, I will be using the very powerful, dynamic, and multipurpose coding language, "Python". While there are many programming languages out there with their own advantages, perks, and features, we feel that python suits this project the best and is certainly powerful enough to be the spine of this work [20]. There are many specific reasons why I have decided upon using python for this specific project, which are given below in some detail:

- Simplicity: One of the main reasons why python is so sought-after in the world right now is that, it Skip the complex codes and save the time , to focus instead, on bettering the task at hand, and so did I, saving them time and extra effort from coding and putting the time to good use on making the quality of this research and data obtention better. Python is written in uncomplicated code, making it easier, convenient, and more efficient for us to create models for machine learning [21].
- Flexibility: The sheer resilience and flexibility that python provides is largely unmatched and most importantly, highly valuable, especially for projects such as ours that involve machine learning. From getting the choice between scripting or OOP's to saving the effort of recompiling the source code as we could alter and redesign whenever we want, and lastly, the significant feature allowing us to integrate python with other programming languages if we ever feel the need to do so in our given project [22].
- Platform Adaptability: A very underrated and useful aspect of python is that I needed to make only minor

3

adjustments and alterations to run the code on mac from a windows pc.
- Extensive Library of Libraries: The last of the reasons given here, and arguably the most important, is that there is a wide range of module libraries available on python, and that of high quality and usability too. This is very significant as I was able to obtain a sizeable portion of our basic code from the libraries and this saved us a lot of time [23]. Sckikit-learn, Pandas, Keras, and TensorFlow are just some of the more popular and widely used ones for those doing projects that involve machine learning, but the shelf is huge and there are libraries for every kind of task there is [24].In this experiment used scikit.

2) Google Colab: I searched the internet extensively to find data clustering algorithm implementation for testing and then came across Google Colab. Google Colab is public domain software.

Reason behind choosing this software:
- Popular and powerful for implementing different data clustering algorithms.
- Allows to write and execute python code.
- Well suited for data analysis and machine learning.
- Allows access to Python libraries for analyzing and visualising the data.

C. Which Algorithms are compared?

For this research task, I shortlisted two contrasting clustering algorithms, for the objective of studying, investigating, researching, exploring, and finally, comparing them. The algorithms in question are K-Means and Mean Shift. The main grounds on which I choose these particular algorithms are:
- Flexibility
- Popularity
- Applicability

1) K-Means Algorithm: K-Means is a widely used data mining algorithm that can be used to classify data by analyzing and comparing numerical values from the data itself. It is a well know partitioning method and this algorithm is an iterative process.In this objects are classified as belonging to one of the k groups and than are allocated to the nearest cluster. A cluster means joining the items that present similarities by simply diving the data into groups.

Steps of K Mean Algorithm:
- Initialization : First select number of classes or groups. Choose the number of number of clusters.
- Initialize respective center points from the data as centroids.
- Centroid assignment : Each point is classified by computing the distance. Those points are than assign to the closest cluster centroid.
- Centroid correction : The next step is to compute the centroids of newly formed cluster.
- Difference comparison : Recompute the group center by taking the mean of all the vectors in the group.

- Repeat these steps for a set number of iterations or until the group centers don't change much between iterations.

The choice of calculating the distance is very critical steps.The two main distance technique used for calculating distances is:
- Euclidean distance
- Manhattan distance

Euclidean Distance : This research utilizes Euclidean distance technique, due to its simplicity , to calculate the distance of resemblance. It concern distance between two data points or distance between a data point and a cluster centre.The Euclidean distance between two points, a and b, with k dimensions [25] is calculated as : The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. It is the straight line distance between two points.

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m}(x_{ik} - y_{ik})^2}$$

Fig. 3: Formula of Euclidean distance

K mean is extremely sensitive to initialization and poor initialization can even lead to poor convergence speed and poor overall clustering. So it is best to initialize the data randomly and try to find out the "k-spread out" points. Here is a flow diagram for better understanding:
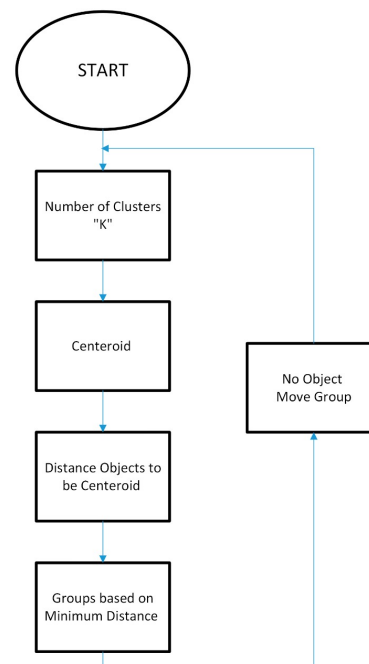

Fig. 4: Flow Diagram of K-Mean Algorithm

2) Mean Shift Algorithm: Mean Shift is a non-parametric clustering algorithm that is used to analyze discrete data

and sets to find the density points. It an advance and versatile technique for clustering based segmentation. The mean shift is a centriod-based algorithm that locates the centers of each group / cluster. The main difference between k mean and mean shift is that one does not need to define clusters in advance. This is because the number of clusters will be defined by algorithm with respect to the data.At every individual point, Mean-Shift calculates its affiliated peak.The algorithm, then, displaces the window to the mean point, and executes the repetition of this process until convergence occurs. On every iteration, the window will be displaced or shifted to a section of the data set that is more densely populated than the previous section, till the point of peak is achieved, on which point the data in question is evenly distributed. Steps of Mean Shift Algorithm:

- Initialization: Initialize with the data points assigned to a cluster of their.
- Pick any random point, and place the window on that data point.
- Calculate the mean of the points that are inside the window.
- Move/Shift the window so that it is at the point of the mean.
- Move the sliding window further according to the mean value until there is no more direction in which a shift can accommodate more points in the kernel.
- This process of steps 1 to 3 is carried out with many sliding windows until all points are within one window.
- The data points are then grouped according to the sliding window in which they are located.

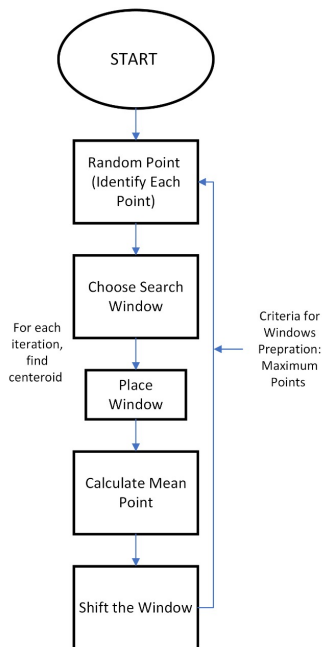Here is a flow diagram for better understanding:

Fig. 5: Formula of Euclidean distance

IV. EXPERIMENTAL RESULTS

The two algorithms are compared according to the following factors:

- Time Complexity
- Training
- Prediction
- Accuracy

A. Time Complexity

For the k-means algorithm, if the number of clusters(k) and the dimension (d) are fixed, considering n as the total number of entities, the problem can be exactly solved in time [26] [27]:

$$O(n^{dk+1})$$

For the mean shift algorithm, since the implementation uses a flat kernel and a Ball Tree to look up members of each kernel, the complexity will tend towards:

$$O(T * n * log(n))$$

in lower dimensions, with n the number of samples and T the number of points. In higher dimensions the complexity will tend towards [28]:

$$O(T * n^2)$$

.

B. Training

The graph below 6 displays the data that was accumulated after the plotting of IRIS and WINE data-sets on both K-means and Mean-shift algorithms. The time that the algorithms proceeded to take for the training phase of the data, or otherwise known as the "training time", was recorded and serves as the y-axis of the graph. The following data can be obtained from the graph:

- K-means completes the training phase of the IRIS data-set in 0.11 seconds while the same process is completed by Mean-shift in 0.71 seconds, 6.5x more time, which is significantly longer.
- K-means completes the training phase of the WINE Data-set in 0.09 seconds, while the same process is completed by Mean-Shift on WINE Data-set in 2.054 seconds, 22x more time, which deems the Mean-shift algorithm much slower than K-means.

The WINE data-set is somewhat larger than the IRIS data-set, and we can see that both algorithms take longer to complete the training phase for IRIS data-set than WINE data-set. This shows that the larger the size of the data, the more time required for the training phase. We can safely deduce that if we increase the number of objects in the respective data-sets, or in other words "increase their size", the time each algorithm takes to complete the training phases of the data will increase accordingly. We can state that the size of the data-sets is directly proportional to the time needed to complete the training phase for the data.
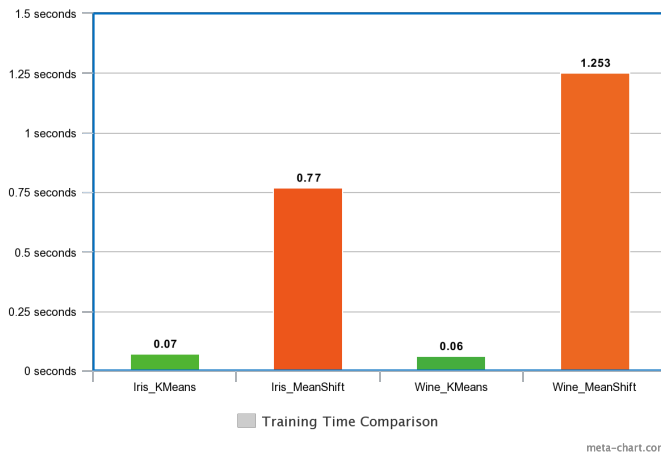
Fig. 6: Time Comparison of Training with K-Means and Mean Shift Algorithm i.e Iris and Wine data sets

C. Prediction

The prediction time that we obtained for these data-sets, using these particular algorithms, is displayed in the graph. The following data can be obtained from the graph 8.

- The prediction time for the IRIS data-set, using K-means, is 0.07 seconds, while 0.77 seconds using Mean-shift algorithm. This is 11x longer, which is a significant increase.
- The prediction time for WINE data-set using K-means is 0.06 seconds, while 1.253 seconds using Mean-shift algorithm, and this is again a significant increase in the prediction time, almost 21 times longer.

Both data-sets have a much lesser prediction time when plotted on K-means, while the time increases significantly when K-means is replaced by Mean-shift. This shows that K-means is indeed significantly faster than Mean-shift.
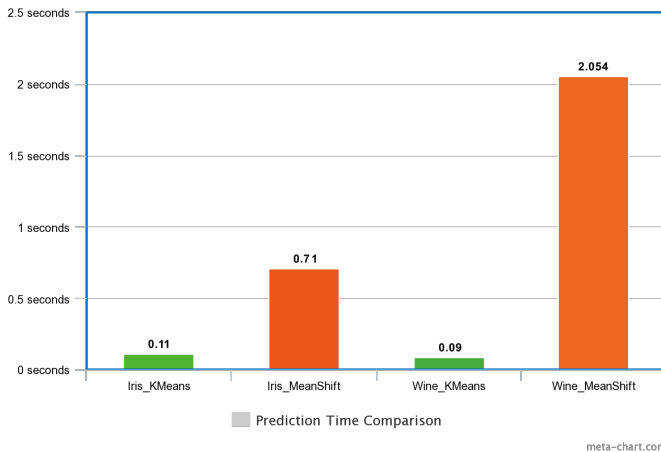

Fig. 7: Time Comparison for Prediction with K-Means and Mean Shift Algorithm i.e Iris and Wine data sets

D. Accuracy

Accuracy is one metric for evaluating. Accuracy is the fraction of predictions. Formally, accuracy has the following
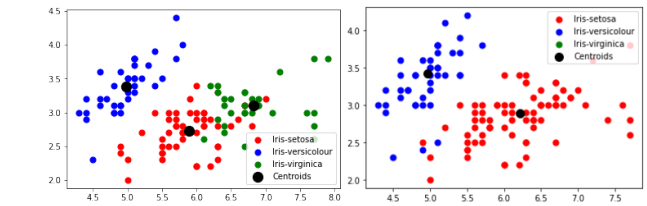

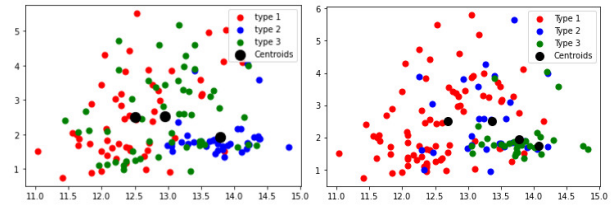Fig. 9: K Mean and Mean Shift Clustering on Iris

Fig. 10: K Mean and Mean Shift Clustering on Wine

definition:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Fig. 8: Formula of Accuracy

The table explains about the accuracy scored obtained from both of the algorithms: As can be seen , the k-Mean

TABLE III: Accuracy Score

Accuracy Score	IRIS DATASET	WINE DATASET
K MEAN	0.36	0.527
MEAN SHIFT	0.533	0.13

has lower score on iris compared to the Wine data set, while Mean shift, on the other hand, has a higher score for Iris which is a small data set but has a low rating for Wine which is comparatively larger than the iris sample is.

V. CONCLUSION

In conclusion, after carrying out the research process described in the paper, in its entirety, I have concluded that, although both data analysis techniques have their perks and advantages, in the majority of the user-cases and situations, the technique "K-means" is better compared to mean shift due to some reasons. Firstly, the training time that is needed to carry out K-means is significantly less than it is needed in the mean shift, and especially where large sets of data and clusters are concerned, this advantage of K-means is absolutely integral. This perk saves much-needed time and effort and makes the whole process much quicker and more efficient. Second, the time complexity of the mean shift is much higher than that of K means. The mean accuracy rating was high for small data sets but very low for a

larger sample, while the k mean accuracy rating became more efficient for larger samples. Lastly, but perhaps the most important advantage that gives K-means the edge over Mean shift, is the ability it gives us to define, specify, or input the number, or set of clusters that we want to involve in the process. This feature gives us control over the whole data analysis process and increases efficiency.Apart from this if in k Mean the number of clusters are unknown than additional algorithms such elbow ,silhouette algorithm are required. Mean shift, however, fails to provide us with any such feature that allows us to set the number of clusters. Due to these reasons, I have concluded, after going through research and proper study of the techniques, that K-means has an edge over Mean shift, and is comparatively better than Mean shift.

REFERENCES

[1] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *Journal of parallel and distributed computing*, vol. 74, no. 7, pp. 2561–2573, 2014.

[2] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm," *PloS one*, vol. 11, no. 9, p. e0162259, 2016.

[3] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big data*, vol. 2, no. 1, pp. 1–32, 2015.

[4] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[5] R. Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.

[6] A. Kapoor and A. Singhal, "A comparative study of k-means, k-means++ and fuzzy c-means clustering algorithms," in *2017 3rd international conference on computational intelligence & communication technology (CICT)*. IEEE, 2017, pp. 1–6.

[7] O. A. Abbas, "Comparisons between data clustering algorithms." *International Arab Journal of Information Technology (IAJIT)*, vol. 5, no. 3, 2008.

[8] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[9] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[10] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *biometrics*, vol. 21, pp. 768–769, 1965.

[11] K. Raghupathi. 10 interesting use cases for the k-means algorithm. [Online]. Available: https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm

[12] S. C. Nair, M. S. Elayidom, and S. Gopalan, "Call detail record-based traffic density analysis using global k-means clustering," *International Journal of Intelligent Enterprise*, vol. 7, no. 1-3, pp. 176–187, 2020.

[13] D. LIN. Using data science techniques for the automatic clustering of it alerts. [Online]. Available: https://tanzu.vmware.com/content/blog/using-data-science-techniques-for-the-automatic-clustering-of-it-alerts

[14] M. Zulfadhilah, Y. Prayudi, and I. Riadi, "Cyber profiling using log analysis and k-means clustering," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 430–435, 2016.

[15] A. Chakravarthy, "Analysis of cyber-criminal profiling and cyber-attacks: A comprehensive study," in *3rd World Conference on Applied Sciences, Engineering and Technology, Kathmandu, Nepal*, 2014.

[16] J. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognition*, p. 107874, 2021.

[17] A. Shivhare and V. Choudhary, "Object tracking in video using mean shift algorithm: A review," *International Journal of Computer Science and Information Technologies*, 2015.

[18] Fisher. Iris data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/iris

[19] W. D. Set. Wine data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/wine

[20] starweaver. Python is a powerful programming language of choice. [Online]. Available: https://starweaver.com/why-python-is-a-powerful-programming-language-of-choice/

[21] M. Learning. Why is python used for ai(artificial intelligence) machine learning? [Online]. Available: esparkinfo.com/why-python-is-used-for-ai-and-machine-learning.html

[22] Blogarama. Why is python used for ai(artificial intelligence) machine learning? [Online]. Available: https://www.blogarama.com/software-blogs/1070228-learnprogramingbyluckysir-blog/22404363-why-python-powerful-for-data-science

[23] P. Piotrowski, "Build a rapid web development environment for python server pages and oracle," *Oracle Technology Network*, 2012.

[24] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, 2020.

[25] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," *International Journal of Computer Applications*, vol. 67, no. 10, 2013.

[26] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted voronoi diagrams and randomization to variance-based k-clustering," in *Proceedings of the tenth annual symposium on Computational geometry*, 1994, pp. 332–339.

[27] M. K. Pakhira, "A linear time-complexity k-means algorithm using cluster shifting," in *2014 International Conference on Computational Intelligence and Communication Networks*, 2014, pp. 1047–1051.

[28] Scikit-Learn. Meanshift algorithm sci-kit learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

APPENDIX

Here are some of the libraries used in the experiment:

- NumPy : It is universal array processing package. It provides a powerful multidimensional array object and tools for working with those arrays. It is the basic package for scientific computing with Python.
- Pyplot: It is a collection of functions in the popular visualization package Matplotlib.Its functions process elements of a figure, e.g. B. Creating a figure, creating a plot area, plotting lines, adding plot labels, and more.
- Time : The time function returns the number of seconds passed.
- Pandas: It is the most popular Python library used for data analysis. It offers highly optimized performance as the back-end source code is written entirely in C or Python.
- load_iris: This is a function used from sklearn ,which will import the iris data.
- load_wine :This is a function used from sklearn ,which will import the wine data.
- train_test_split: It is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data.
- accuracy_score : It is used to calculate the accuracy score.
- math: This is used so that math functions can be implemented.