International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT12173174

# Clustering Analysis using an Unsupervised Machine Learning Method

**Tashfin Ansari*[1], Dr. Almas Siddiqui[2], Awasthi G. K[2]**
*[1]Computer Science and Engineering, P.E.S. College of Engineering, Aurangabad, Maharashtra, India
[2]Assistant Professor, Vivekanand College, Aurangabad, Maharashtra, India

**Article Info**
Volume 7, Issue 3
Page Number: 602-609

**Publication Issue :**
May-June-2021

**Article History**
Accepted : 20 June 2021
Published : 30 June 2021

## ABSTRACT

Artificial Intelligence (AI) and Machine Learning (ML), which are becoming a part of interest rapidly for various researchers. ML is the field of Computer Science study, which gives capability to learn without being absolutely programmed. This work focuses on the standard k-means clustering algorithm and analysis the shortcomings of the standard k-means algorithm. The k-means clustering algorithm calculates the distance between each data object and not all cluster centres in every iteration, which makes the efficiency of clustering is high. In this work, we have to try to improve the k-means algorithm to solve simple data to store some information in every iteration, which is to be used in the next interaction. This method avoids computing distance of data object to the cluster centre repeatedly, saving the running time. An experimental result shows the enhanced speed of clustering, accuracy, reducing the computational complexity of the k-means. In this, we have work on iris dataset extracted from Kaggle.

Keywords : Machine Learning (ML), Artificial Intelligence (AI), K- Means Clustering, Classification, Unsupervised Learning.

## I. INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) are dynamic methods for predicting the future. In the recent era, it was observed fast advances in executing numerous modern and improved algorithms. Scholastic, researchers as well as high-tech businesses [1] proposed various innovative algorithms.

ML is the study of CS algorithms that improve automatically through experience, and it is the subset of Artificial Intelligence. ML algorithm builds a mathematical model based on sample data to create predictions or decisions without being explicitly programmed to do so. The advent of Data analytics leading to the use of Machine Learning and others can be attributed to the subsequent development of technologies such as Big Data, Business Intelligence, etc. Machine learning is the subset of Artificial Intelligence. The first thing machine learning works on a training dataset is choosing an algorithm to run on the training dataset. Therefore, this algorithm depends on the type of the data i.e. whether the data is labelled or unlabeled. The common types of algorithm that use labeled training data are Regression algorithm, decision trees, etc. On the contrary unlabeled training, datasets work on algorithms like Clustering, Association algorithms, and Neural Networks. Now,

there are different machine learning methods categorized as supervised learning, unsupervised learning, semi-supervised learning, etc. In this paper, we have used an unsupervised machine-learning algorithm like K- Means Clustering for the prediction of clusters in the IRIS dataset extracted from Kaggle.

### A. Supervised Learning:

Supervised Learning is the Machine Learning process that maps input/output processes based on examples and it infers a function from labelled training data consisting of a set of training examples. Supervised Learning is the process of its use of labelled datasets to train algorithms to classify data or predict outcomes accurately. It helps organizations to solve problems at scale, such as classifying spam in a separate folder from your inbox. Supervised learning uses a discipline to teach models to yield the desired output. This training of dataset includes inputs and approximately correct outputs, which allow the model to learn over time. Supervised learning can be separated into two types i.e. classification and regression. Various algorithms and computational techniques are used in supervised machine learning processes. Some learning methods are Neural networks, Naive Bayes, Linear regression, Logistic regression, Support vector machine (SVM), K-nearest neighbour, and Random Forest[2].

### B. Unsupervised Learning:

Un-supervised Learning is a type of ML that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.

It uses machine-learning algorithms to analyse unlabeled datasets. These algorithms explore hidden patterns or data groupings without the need for human intervention. In unsupervised learning, you only have input data(X) and no corresponding output variables. The ability of Un-supervised Learning is to discover similarities and a difference in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

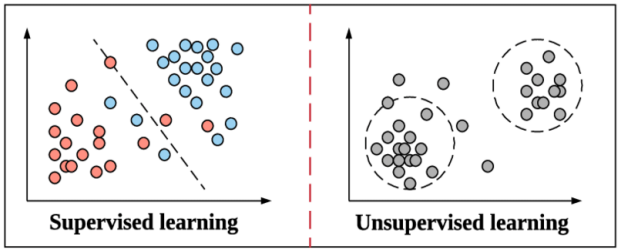



Fig.1. Comparison of supervised & unsupervised learning.

An unsupervised learning model uses three main tasks clustering, association, and dimensionality reduction. Clustering is data mining technique that groups unlabeled data. Based on their similarities or differences. To process raw, un-classified data objects into groups represented by structures or patterns in the information Clustering algorithms are used. Clustering algorithms can be categorized as specifically exclusive, overlapping, hierarchical, and probabilistic. Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster. It is also known as “hard” clustering. K-Means Clustering is an example of exclusive clustering [3].

The reference[4], the authors compared their modern proposed Progressive method with five distinctive datasets to calculate the execution of different classification frameworks like Neural systems, KNN (K-Nearest Neighbors), and numerous more. Reliabilty was determined to utilize different classifiers on a distinctive dataset.

In the reference [5], the author gave a sincere thought of clustering strategy specifically Simple K-means, Density-Based spatial Clustering with Algorithm with noise (DBSCAN), Hierarchical Clustering Analysis (HCA), Make Density-Based The Clustering Algorithm (MDBCA) to calculate its

execution analysis and time complexity by testing it on diverse datasets specifically Abalone, Bankdata, Switch, SMS and Webtk dataset utilizing Weka tool.
In the reference[6], DR methods to be specific LDA and PCA recognition of Breast Cancer, Iris, Glass, Yeast, and Wine dataset is performed to improve the classified wrong information using the various classifiers. It make a difference in data optimization due to which the rate of correct classification increased.
The author referenced in [7], has displayed a web-based application on food photography for portable phone clients. This app is based on Thai foods and is of thirteen distinctive classes and an image recognition model is constructed utilizing Create Training and Test Data (CNN) and VGG19 model. This model is executed utilizing Tensorflow and Keras.

### C. Clustering

Clustering is one of the most common probing data analysis techniques used to get an instinct about the structure of the data. It is the function of identifying subgroups such that, data points in the same subgroup (cluster) similar and dis-similar data points. An unsupervised learning mechanism predicts the output data point and organizes it into a cluster format from the given input dataset. The output data points are grouped agreeing to the similarity and dissimilarity measures of the individuals to the dataset. Partitioning and Hierarchical strategies are essential sorts of clustering methods [8].

We try to find out homogeneous subgroups, such that data points in each cluster are similar according to a similarity measure, such as Euclidean-based distance or correlation-based distance. It is application-dependent that which similarity measure to use. Clustering analysis can be done on the basis of features, here we try to find subgroups of samples based on features, where we try to find subgroups of features based on samples. We'll cover here clustering based on features.

### D. K-means Algorithm:

K-means is the simplest and frequently utilized unsupervised learning method to solve clustering problems. K-means is additionally named Simple K-means since it is the easiest method to make clustering designs. The basic functionality of K-means is that it bundles clusters according to the dataset. This algorithm is an iterative algorithm to partition the dataset into K pre-defined distinct non-overlapping subgroups, where each data point belongs to only one group. It assigns data points to clusters such that the sum of the squared distance between the data points and the cluster's centroids is minimum. The less variation within the clusters, the more homogeneous data points are in the same cluster.

## II. DATABASE SPECIFICATION

In this research work, we have used the Iris flower species database from kaggal. The iris flower dataset is a multivariate dataset. There are variations of Iris flowers of three related species. The dataset contains 150 records under five attributes - Petal Length, Petal Width, Sepal Length, Sepal Width, and Class (Species). It is of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample i.e. sepal length, sepal width, petal length, petal width in centimetres [9].


Fig.2

## III.METHODOLOGY

**K- Means Clustering:**

K-Means Clustering the data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroids. The data points which are closest to a given centroids will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularities whereas a smaller K value will have larger groupings and fewer granularities. It is a centroids-based algorithm, in which each cluster is associated with centroids. [10], [11].

The algorithm takes the un-labeled dataset as input and divides the dataset into a k-number of clusters, and repeats this process until it finds the best clusters. This algorithm should predetermine the value of k. The k-means clustering algorithm mainly performs two tasks i.e. 1. Determines the best value for k center points by an iterative process. 2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center creates a cluster. Hence, each cluster has data points with some resemblances, and it is away from other clusters. The following diagram explains the working of the K-means Clustering Algorithm.

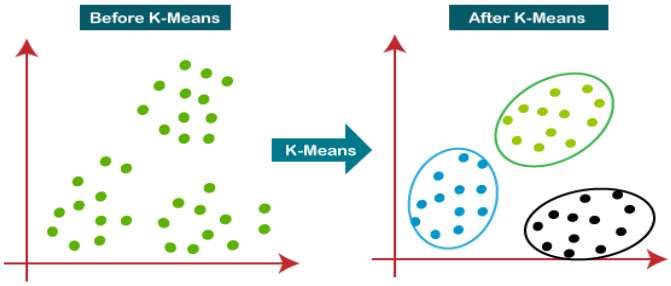



Fig.3

The K-means algorithm usually compares well to more refined and computationally expensive clustering algorithms concerning the quality of results. The range of possible values of the k parameter is sufficiently small so that you can examine this range by running the algorithm several times with different values of k [12].

1) Steps of k-means algorithm
2) Step 1: Choose the number of clusters k
3) Step 2: Select k random points as centroids
4) Step 3: Assign all the points to the centroids
5) Step 4: Recompute the centroids for newly formed clusters
6) Step 5: Repeat steps 3 and 4.

Stopping Criteria for K-Means Clustering is we can stop the algorithm if the centroids of newly formed clusters are not changing. Even after multiple iterations, if we found the same centroids for all the clusters, then the algorithm is not learning any new patterns, then stop the training. Another clear sign to stop the training process, if the points remain in the same cluster even after training. At last, we can stop the training if the maximum number of iterations is reached [13].

**The Elbow Method:**

For the k-means clustering method, the most common approach for this is the elbow method. It involves running the algorithm multiple times over a loop, with an increasing number of cluster choices, and then plotting a clustering score as a function of the number of clusters. The elbow method runs k-means clustering on the dataset for k range values (say from 1-10), and then for each value of k computes an average score for all clusters. By default, the distortion Score is computed, the sum of square distances from each point to its assigned center [14].

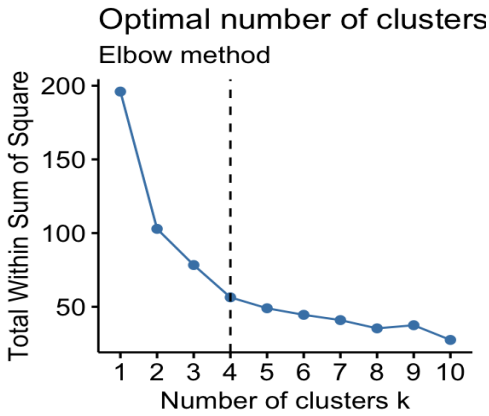



Fig.4

**The Silhouette Method:**

The silhouette Method is a method for finding the optimal number of clusters of consistency within clusters of data. The silhouette method computes coefficients for each point that measures a point is similar to its own cluster compared to other clusters. By providing a concise graphical representation of how well each object has been classified [15].

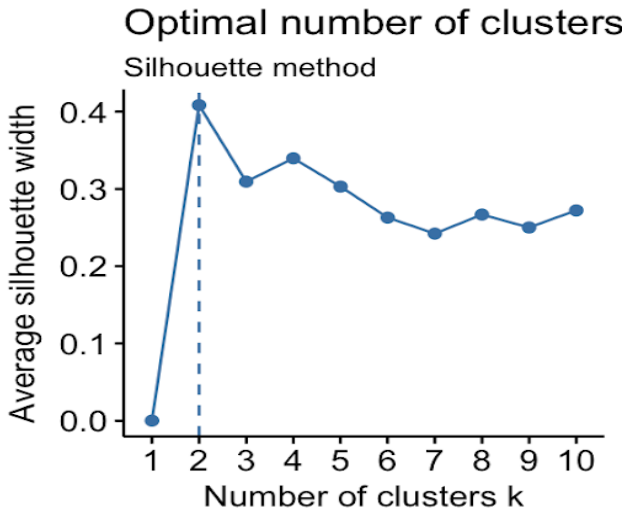
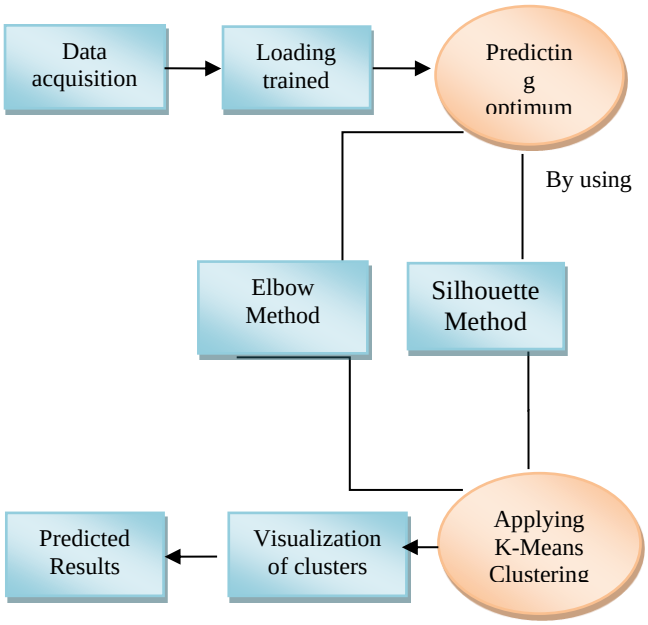



Fig.5

## IV.PROPOSED METHODOLOGY




## V. IMPLEMENTATION

**Python:**

Python is an object-oriented, dynamic data type of high-level programming language. Its style is simple, clear and it contains powerful different kinds of classes. It is easy to understand. In this research paper, we have used python programming language for the implementation of machine learning algorithms on the iris database.

1. Importing the required libraries:

The first and foremost step is to import all the required libraries for training the data. We have imported Numpy, which is mainly used for the matrix calculation, matplotlib creates a plotting area in a figure, pandas are used for data manipulation and analysis, and sklearn which contains tools like classification, regression, clustering, and dimensionality reduction.

```
In [1]: #Importing  the required libraries

        import numpy as np
        import matplotlib.pyplot as plt
        import pandas as pd
        from sklearn import datasets
```

Fig.6

2. Loading the Iris dataset:

```
In [2]: # Loading the iris dataset

        iris = datasets.load_iris()
        iris_df = pd.DataFrame(iris.data, columns = iris.feature_names)
```

Fig.7

3. Displaying the first five rows of dataset:

```
In [3]: # displaying the first 5 rows

        iris_df.head()
```

Out[3]:

| | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) |
|---|---|---|---|---|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 |

Fig.8

4. To find the optimum no. of clusters by using Elbow Method:

Using Elbow Method :

It is the variations in between the no. of clusters within the certain range

```
In [4]: # Calculating the within cluster sum of squares of k- means classification

        x = iris_df.iloc[:, [0, 1, 2, 3]].values

        from sklearn.cluster import KMeans
        wcss = []

        for i in range(1, 11):
            kmeans = KMeans(n_clusters = i, init = 'k-means++',
                            max_iter = 300, n_init = 10, random_state = 0)
            kmeans.fit(x)
            wcss.append(kmeans.inertia_)
```

```
In [5]: # Plotting the "Within cluster sum of squares" on line graph
        # 'The elbow Method'
        plt.plot(range(1, 11), wcss)
        plt.title('The elbow method')
        plt.xlabel('Number of clusters')
        plt.ylabel('WCSS')
```

Fig.9

The plot looks like following:

```
Out[5]: Text(0, 0.5, 'WCSS')
```

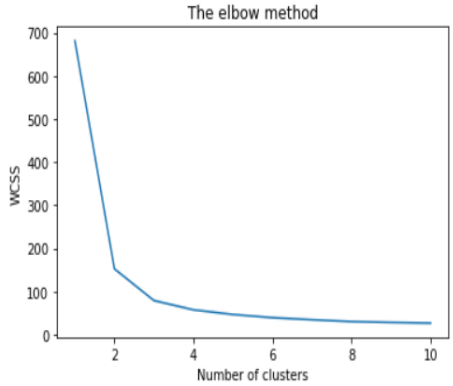



Fig.10

From above it is clear that 3 will be better option than 2 as the optimum no. of clusters.

Using Silhouette Method for comparison:

```
In [16]: #Predicting the optimum no. of clusters using silhouette Analysis
         from sklearn.metrics import silhouette_score
```

```
In [18]: sse_ = []
         for k in range(2, 11):
             kmeans = KMeans(n_clusters=k).fit(x)
             sse_.append([k, silhouette_score(x, kmeans.labels_)])
```

```
In [19]: #Plotting the results
         plt.plot(pd.DataFrame(sse_)[0],pd.DataFrame(sse_)[1]);
         plt.title('The silhouette method')
         plt.xlabel('Number of clusters')
         plt.ylabel('silhouette score')
         plt.show()
```

Fig.11

The plot looks like following,

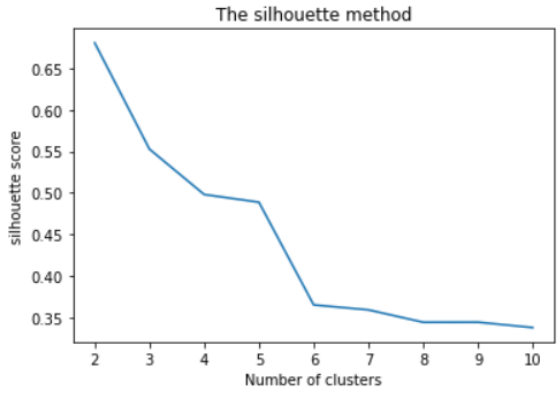



Fig.12

So from above fig.1 and fig.2, we have analyzed to choose optimum no. of clusters as 3.

Finally applying K-Means clustering on data and visualizing the clusters:

```
In [6]: #Creating the k means classifier:

kmeans = KMeans(n_clusters = 3, init = 'k-means++',
                max_iter = 300, n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(x)
```

```
In [12]:
# visualisation of clusters on first two coloumns

plt.figure(figsize=(8,6))
plt.scatter(x[y_kmeans == 0, 0], x[y_kmeans == 0, 1],
            s = 100, marker = '*', c = 'brown', label = 'Iris-setosa')
plt.scatter(x[y_kmeans == 1, 0], x[y_kmeans == 1, 1],
            s = 100, marker = '*', c = 'indigo', label = 'Iris-versicolour')
plt.scatter(x[y_kmeans == 2, 0], x[y_kmeans == 2, 1],
            s = 100, marker = '*', c = 'darkgreen', label = 'Iris-virginica')

# Plotting the centeroids of clusters

plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1],
            s = 100,marker =',', c = 'yellow', label = 'Centroids')

plt.legend()
```

Fig.13

The plot looks like the following,

```
Out[12]: <matplotlib.legend.Legend at 0x23d65f8d2e0>
```

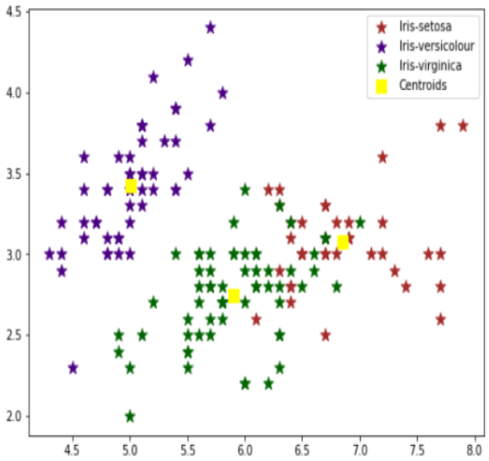



Fig.14

## VI. CONCLUSION

In this research work, we have presented an algorithm for performing K-means clustering. Our experimental result shows that our scheme can improve the direct K-means algorithm. The above scatter Plot is the visual representation of Iris Dataset where it observed as optimum no. of clusters is 3. We can clearly see that there are 3 groups of species of Iris setosa, Versicolor, and virginica. The yellow mark represents the centroids of each cluster. According to the academic analysis and result of the experiment, the improved K-Means not only keep the high efficiency of standard K-Means but also raises the speed of convergence effectively by improving the way of selecting the initial cluster focal point. The improved K-Means are obviously better than standard K-Means in both cluster precision and stability.

## VII. REFERENCES

[1]. Bhattacharya, Sambit & Czejdo, Bogdan & Agrawal, Rajeev & Erdemir, Erdem & Gokaraju, Balakrishna. (2018). 1-4. 10.1109/SECON.2018.8479098. Sambit Bhattacharya,

[2]. Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science” January 2020 DOI: 10.1007/978-3-030-22475-2_1 In book: Supervised and Unsupervised Learning for Data Science (pp.3-21)

[3]. https://www.ibm.com/cloud/learn/unsupervised-learning

[4]. L. B. Goncalves, M. M. B. R. Vellasco, M. A. C. Pacheco and Flavio Joaquim de Souza, "Inverted hierarchical neuro-fuzzy BSP system: a novel neuro-fuzzy model for pattern classification and rule extraction in databases," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 36, no. 2, pp. 236-248, March 2006.

[5]. P. H. Ahmad and S. Dang, "Performance evaluation of clustering algorithm using different datasets", Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 3, no. 1, pp. 167-173, 2015.

[6]. Panahi N, Shayesteh MG, Mihandoost S, Zali Varghahan B, "Recognition of different datasets using PCA, LDA, and various classifiers", In 5th International Conference on Application of

Information and Communication Technologies (AICT), Baku, Azerbaijan, 2011; 1– 5.

[7]. U. Tiankaew, P. Chunpongthong and V. Mettanant, "A Food Photography App with Image Recognition for Thai Food," 2018 Seventh ICT International Student Project Conference (ICT-ISPC), Nakhonpathom, 2018, pp. 1-6.

[8]. Dang, Shilpa. (2015). Performance Evaluation of Clustering Algorithm Using Different Datasets. IJARCSMS. 3. 167-173. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[9]. https://www.kaggle.com/arshid/iris-flower-dataset

[10]. JAIN A K, DUBES R C. Algorithms for clustering data[M].New Jersey:Prentice-Hall,1988.

[11]. ZhangYufang etc. A kind of improved K-means algorithm [J]. Computer Application,p3133, 2003, (8).

[12]. Yu Yang “A study of pattern recognition of Iris flower based on Machine Learning” Degree Program: Information Technology | Specialization: Internet Technology 2013

[13]. Dataset Tanvi Gupta, Supriya P. Panda2 “A Comparison of K-Means Clustering Algorithm and CLARA Clustering Algorithm on Iris” International Journal of Engineering & Technology, 7 (4) (2018) 4766-4768 International Journal of Engineering & Technology Website: www.sciencepubco.com/index.php/IJET doi: 10.14419/ijet. v7i4.21472

[14]. K. Maheswari, “Finding Best Possible Number of Clusters using K-Means Algorithm”, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9, Issue-1S4, December-2019.

**Cite this article as :**

Tashfin Ansari, Dr. Almas Siddiqui, Awasthi G. K, "Clustering Analysis using an Unsupervised Machine Learning Method", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 602-609, May-June 2021.
Available at
doi : https://doi.org/10.32628/CSEIT12173174
Journal URL : https://ijsrcseit.com/CSEIT12173174