International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

FRACTIONATION USING K MEANS CLUSTERING

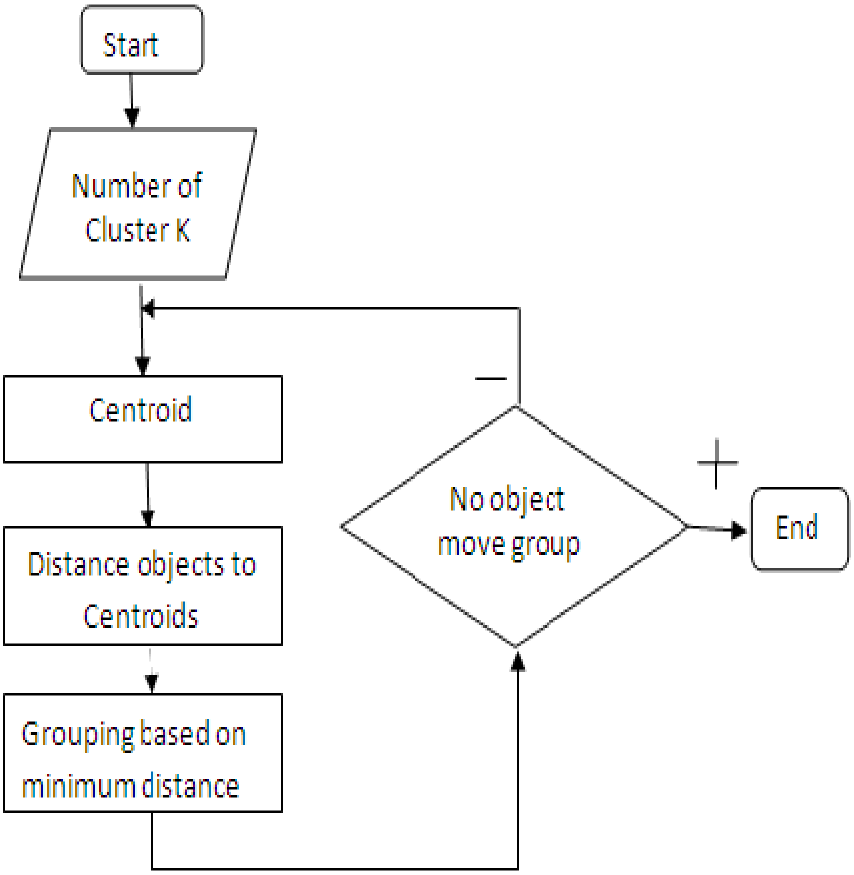
Kannammal A[1], Sindhu P[2], Santhiya R[3], Sujitha S[4], Yuvetha S[5]
[1] Assistant Professor - Computer Science Engineering, Paavai Engineering college, Namakkal, Tamil Nadu
[2] UG - Computer Science Engineering, Paavai Engineering College, Namakkal, Tamil Nadu
[3] UG - Computer Science Engineering, Paavai Engineering College, Namakkal, Tamil Nadu
[4] UG - Computer Science Engineering, Paavai Engineering College, Namakkal, Tamil Nadu
[5] UG - Computer Science Engineering, Paavai Engineering College, Namakkal, Tamil Nadu

ABSTRACT
The k-means algorithm is often used in clustering applications but its usage requires a complete data matrix. Missing data, however, is common in many applications. Mainstream approaches to clustering missing data reduce the missing data problem to a complete data formulation through either deletion or imputation but these solutions may incur significant costs. Our k-POD method presents a simple extension of k-means clustering for missing data that works even when the missingness mechanism is unknown, when external information is unavailable, and when there is significant missingness in the data.
Keywords—K-means clustering, k-POD

1.Introduction
The national education goal is to educate nations. It can't be separated from the role of teacher. In order to improve the quality of education, the development of curriculum, learning innovation, and education facilities contentmentare needed. To improve students' learning achievement, teacher should make innovative learning that support the students learn optimally, either in self-learning and in class learning. In an attempt to improve the quality of education, teachers are required to make learning more effective, innovative and fun. In learning process, the role of teachers is as motivator and facilitator. In consequence of learning paradigm change from instructional based learning into constructional based learning, the teachers should capable to design learning that activate students.To make it more effective, innovative, and fun, the teachers can utilize all school resources, either human resources, facilities and infrastructuresor other resources .
The survey result of this research showed that Sekolah Menengah Kejuruan (SMK) Negeri 2 Bengkulu Tengahhas had computer network and already utilize Information and communication Technology (ICT) by utilizee-learning in implementing learning into teaching and learning process. It was a step forwardthat capable to improve students' learning motivation and also to meet the shortage of meeting in classroom doe to the vocational curriculum that require field study, so that the students can conduct self learning without depend on teacher and books.. E-learning is learning media technology that utilize either electronic media or software.. E-learning can be developed by using Learning Management System (LMS)called Moodle.
Provided by Moodle include reading module, assignment module, chat module, forum module, option module, and quiz module. Once the pattern obtained, analyzing each pattern of subject taken by students was performed, so that it can describe the subject taken by students based on students' participation interest in following the lessons. Fuzzy K-Means algorithm in this research was used to cluster teaching and learning activities between teacher and students so that it resulted in information about activities group attended by students in e-learning.Besides that, k-means algorithm is also versatile which means that it easy to modify steps in the algorithm, such as in the initialization of function to calculate distance and also criterion on stopping iteration.
.

DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

2. Experimental Methods or Methodology



We propose an enhanced version of the k-means algorithm with simple partitioning to speed up the time in finding the final converged centroids. The pseudo-code of our algorithm. Table 1 shows the notation used in describing the algorithm. Our algorithm is composed of three parts.
We propose an enhanced version of the k-means algorithm with simple partitioning to speed up the time in finding the final converged centroids. The pseudo-code of our algorithm. The notation used in describing the algorithm.

3.Results and Discussion

Cluster analysis

Cluster analysis is based on various kinds of objects' differences and uses distance functions' regulations to make model classification . Whether the classification is really make a difference or not is rest with the distribution form of pattern character vectors. If the contributions of dots of vectors is clustered and sample dots in the same group are concentrated and sample dots in different groups are distant, it will be easy to use distance functions to classify the dots, which will as far as possible make statistics in the same group be similar and statistics in different group be different. The eigenvector of the whole sample pattern congregation can be treated as dots which distribute in feature space. The distance function between dots may act as the measure of similarity of patterns. According to the proximity of dots' distance, the measure can be used to classify patterns.

K-Means algorithm

K-Means algorithm based on dividing is a kind of cluster algorithm, and it is proposed. This algorithm which is unsupervised is usually used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion are foundations of this

DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

algorithm. To seek the optimal zing outcome, this algorithm tries to find K divisions to satisfy a certain criterion. Firstly, choose some dots to represent the initial cluster focal points(usually, we choose the first K sample dots of income to represent the initial cluster focal point); secondly, gather the remaining sample dots to their focal points in accordance with the criterion of minimum distance, then we will get the initial classification, and if the classification if unreasonable, we will modify it(calculate each cluster focal points again), iterate repetitively till we get a reasonable classification.

Data clustering techniques

Data clustering techniques are descriptive data analysis techniques that can be applied to multivariate data sets to uncover the structure present in the data. They are particularly useful when classical second order statistics (the sample mean and covariance) cannot be used. Namely, in exploratory data analysis, one of the assumptions that is made is that no prior knowledge about the dataset, and therefore the dataset's distribution, is available. In such a situation, data clustering can be a valuable tool. Data clustering is a form of unsupervised classification, as the clusters are formed by evaluating similarities and dissimilarities of intrinsic characteristics between different cases, and the grouping of cases is based on those emergent similarities and not on an external criterion. Also, these techniques can be useful for datasets of any dimensionality over three, as it is very difficult for humans to compare items of such complexity reliably without a support to aid the comparison

Very large datasets

For very large datasets that would make the computation of the previous algorithms too computationally expensive, it is possible to choose a random sample from the whole population of cases and apply the algorithm on the sample. If the sample is sufficiently large, the distribution of these initial reference points should reflect the distribution of cases in the entire set.

SOFTWARE CODE

```
import cv2
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from PIL import Image
import seaborn as sns
import numpy
import matplotlib.pyplot as plt
import pandas as pd
# Read Excel file
df = pd.read_excel('DATASET.xlsx')
print(df)
columns = ['Order Number','Order Status','Order Date']
df2 = pd.read_excel('DATASET.xlsx',
            header=None, names = columns)
plt.figure(figsize=(5.8, 4.2))
x = range(len(df['Order Number']))
plt.plot(x, df['Order Status'])
plt.xticks(x, df['Order Number'])
plt.xlabel('Order Number')
plt.ylabel('Order Status')
plt.show()

df = pd.read_csv('Mall_Customers.csv')

print(df.to_string())
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

```
ages = [10,15,70,40,30,45,50,45,43,40,44,
      60,7,13,57,18,90,77,32,21,20,30]

range = (0, 100)
bins = 10
plt.hist(ages, bins, range, color = 'BLUE',
      histtype = 'bar', rwidth = 0.8)

plt.xlabel('CustomerID')
plt.ylabel('Gender')
plt.title('Age')
plt.show()

x = [1,2,3,4,5,6,7,8,9,10]

y = [2,4,5,7,6,8,9,11,12,12]

plt.scatter(x, y, label= "stars", color= "green",
         marker= "*", s=30)


plt.xlabel('x - axis')

plt.ylabel('y - axis')

plt.title('CustomerID !')

plt.legend()

plt.show()
activities = ['CustomerID ', 'Gender', 'Age', 'Annual Income']


slices = [3, 7, 8, 6]


colors = ['r', 'y', 'g', 'b']


plt.pie(slices, labels = activities, colors=colors,
      startangle=90, shadow = True, explode = (0, 0, 0.1, 0),
      radius = 1.2, autopct = '%1.1f%%')


plt.legend()


plt.show()
from matplotlib import pyplot as plt
# Importing Numpy Library
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

```
import numpy as np
plt.style.use('fivethirtyeight')

mu = 50
sigma = 7
x = np.random.normal(mu, sigma, size=200)
fig, ax = plt.subplots()

ax.hist(x, 20)
ax.set_title('Order Product')
ax.set_xlabel('Order Number')
ax.set_ylabel('Order Status')

fig.tight_layout()
plt.show()
dataFrame = pd.DataFrame({
  "id": ['c_name', 'c_age', 'c_Gender', 'Order', 'Mercedes', 'Jaguar'],"Reg_Price": [7000,
1500, 5000, 8000, 9000, 6000]
})

plt.hist(dataFrame["Reg_Price"])
plt.show()
plt.rcParams["figure.figsize"] = [7.00, 3.50]
plt.rcParams["figure.autolayout"] = True
data = np.random.random(1000)
n, bins, patches = plt.hist(data, bins=25, density=True, color='red', rwidth=0.75)
col = (n-n.min())/(n.max()-n.min())
cm = plt.cm.get_cmap('RdYlBu')
for c, p in zip(col, patches):
  plt.setp(p, 'facecolor', cm(c))
plt.show()
```

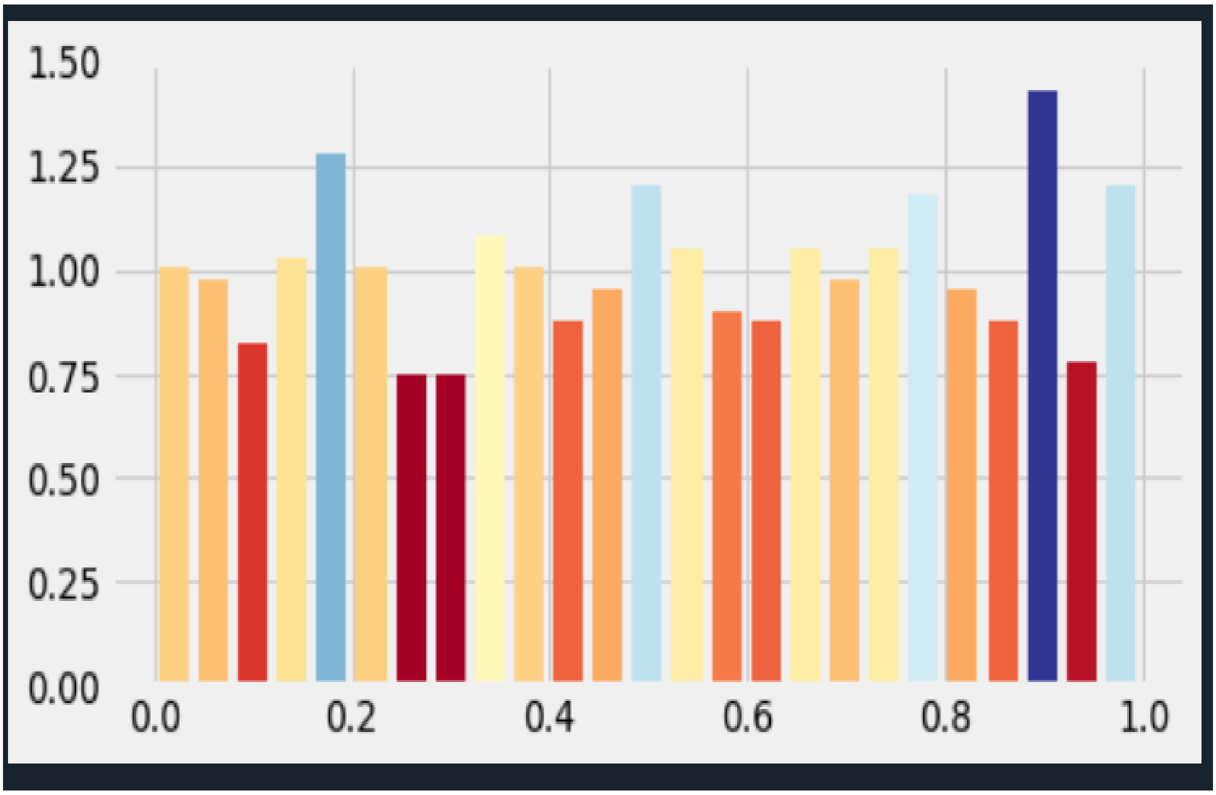
OUTPUT

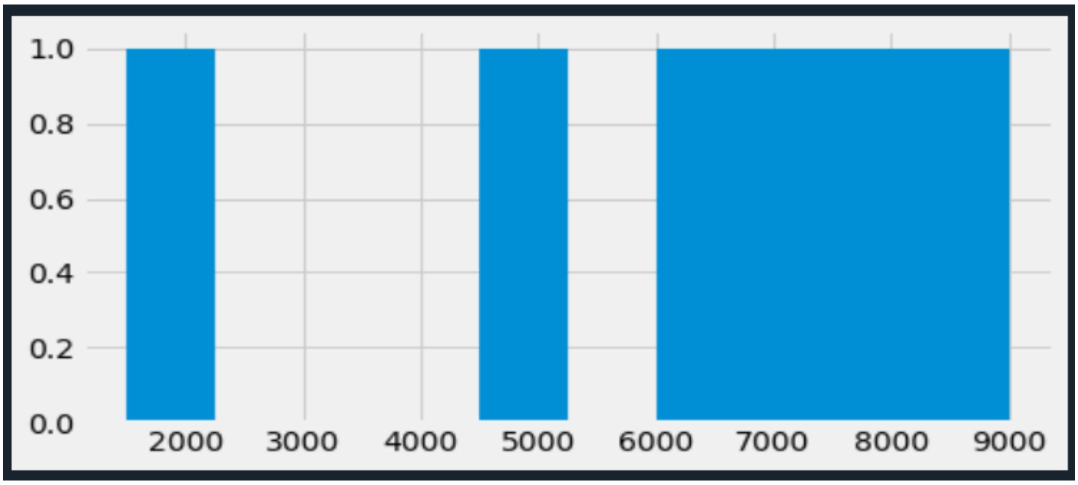
```
      Order Number     Order Status  ... Discount Amount Discount Amount Tax
0            30449  Pending payment  ...             NaN                 NaN
1            30449  Pending payment  ...             NaN                 NaN
2            30446       Processing  ...             NaN                 NaN
3            30445       Processing  ...             NaN                 NaN
4            30444  Pending payment  ...             NaN                 NaN
...            ...              ...  ...             ...                 ...
6109          5804        Cancelled  ...             NaN                 NaN
6110          5803           Failed  ...             NaN                 NaN
6111          5802           Failed  ...             NaN                 NaN
6112          5456        Cancelled  ...             NaN                 NaN
6113          5456        Cancelled  ...             NaN                 NaN

[6114 rows x 33 columns]
```

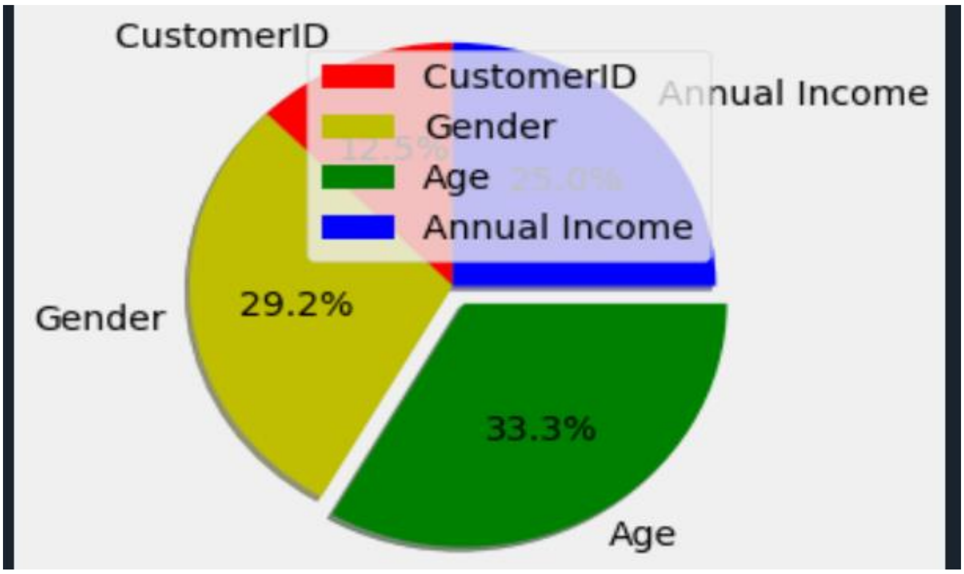
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14
11	12	Female	35	19	99
12	13	Female	58	20	15
13	14	Female	24	20	77
14	15	Male	37	20	13
15	16	Male	22	20	79
16	17	Female	35	21	35
17	18	Male	20	21	66
18	19	Male	52	23	29
19	20	Female	35	23	98
20	21	Male	35	24	35
21	22	Male	25	24	73
22	23	Female	46	25	5
23	24	Male	31	25	73
24	25	Female	54	28	14
25	26	Male	29	28	82

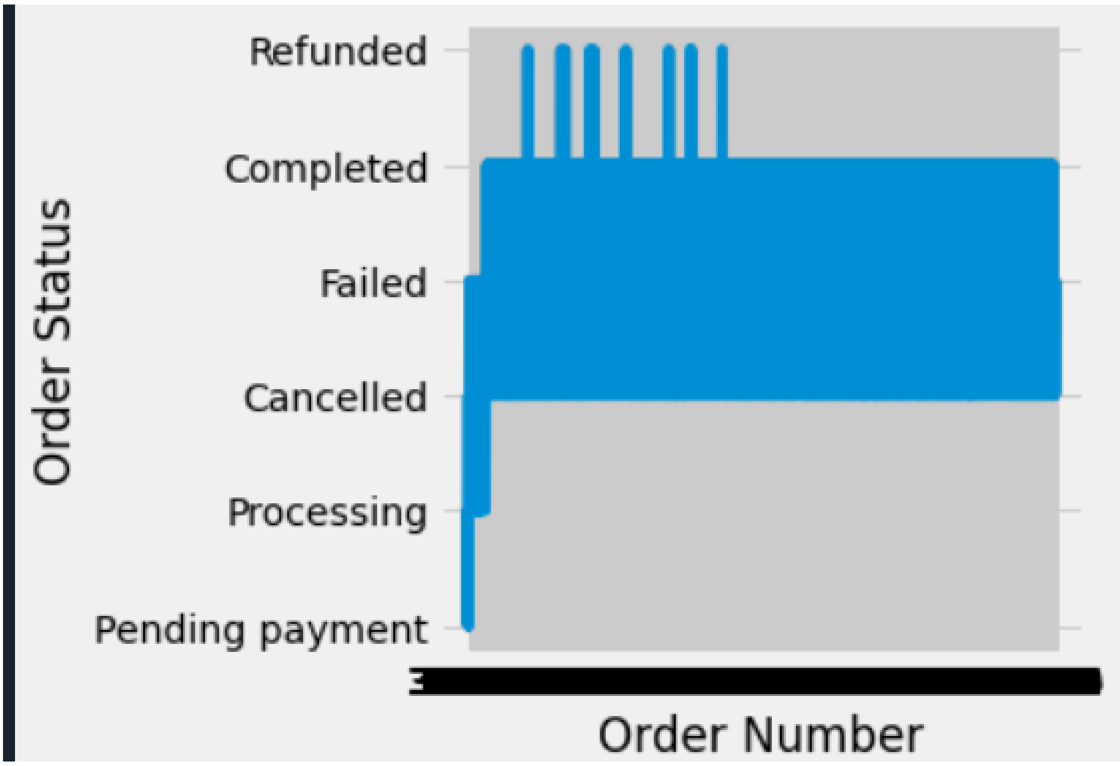



1.0
0.8
0.6
0.4
0.2
0.0
2000
3000
4000
5000
6000
7000
8000
9000


Order Product
Order Status
Order Number
0
10
20
30
30
40
50
60
70


CustomerID
Gender
Age
Annual Income
12.5%
25.0%
29.2%
33.3%

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621



CONCLUSION

Our algorithm is very well suited for handling outliers - in fact, it becomes simpler. Using the notion of balanced clusters in conjunction with Lemma 2.2, by eliminating at most $(1+\mu)\gamma|P|$ outliers, we can approximate the cost of the optimal k-means clustering with at most $\gamma|P|$ outliers. An interesting direction for further research is to extend our methods for other clustering problems. Also, it is an open problem to get a polynomial time $(1 + \varepsilon)$- approximation algorithm for the k-means clustering problem when n, k and d are not constants

The numerical results demonstrate that k-POD is not only accurate, but also fast particularly at higher levels of overall missingness. There are two reasons for this. First, the majorization step consists of simply copying the relevant entries of the centroid estimates from the k-means step into the missing entries in the data matrix. Second, the minimization step consists of running k-means, which is fast in itself; indeed, each iteration requires O(knp) computations, which is linear in the data. Particularly when moving to larger data, the setup and computational costs required to obtain reasonable imputations may become prohibitively expensive, as exhibited in the experiments with 500 observations on 100 variables.

References

[1] P.K. Agarwal and C.M. Procopiuc, ªExact and Approximation Algorithms for Clustering,º Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 658-667, Jan. 1998.
[2] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998.
[3] S. Arora, P. Raghavan, and S. Rao, ªApproximation Schemes for Euclidean k-median and Related Problems,º Proc. 30th Ann. ACM Symp. Theory of Computing, pp. 106-113, May 1998.
[4] S. Arya and D. M. Mount, ªApproximate Range Searching,º Computational Geometry: Theory and Applications, vol. 17, pp. 135- 163, 2000.
[5] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu, ªAn Optimal Algorithm for Approximate Nearest Neighbor Searching,º J. ACM, vol. 45, pp. 891-923, 1998.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.079 ISSN: 2581-4621

[6] G.H. Ball and D.J. Hall, ªSome Fundamental Concepts and Synthesis Procedures for Pattern Recognition Preprocessors,º Proc. Int'l Conf. Microwaves, Circuit Theory, and Information Theory, Sept. 1964.
[7] J.L. Bentley, ªMultidimensional Binary Search Trees Used for Associative Searching,º Comm. ACM, vol. 18, pp. 509-517, 1975.
[8] L. Bottou and Y. Bengio, ªConvergence Properties of the k-means Algorithms,º Advances in Neural Information Processing Systems 7, G. Tesauro and D. Touretzky, eds., pp. 585-592. MIT Press, 1995.
[9] P.S. Bradley and U. Fayyad, ªRefining Initial Points for K-means Clustering,º Proc. 15th Int'l Conf. Machine Learning, pp. 91-99, 1998.
[10] P.S. Bradley, U. Fayyad, and C. Reina, ªScaling Clustering Algorithms to Large Databases,º Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining, pp. 9-15, 1998.
[11] V. Capoyleas, G. Rote, and G. Woeginger, ªGeometric Clusterings,º J. Algorithms, vol. 12, pp. 341-356, 1991.
[12] J.M. Coggins and A.K. Jain, ªA Spatial Filtering Approach to Texture Analysis,º Pattern Recognition Letters, vol. 3, pp. 195-203, 1985.
[13] S. Dasgupta, ªLearning Mixtures of Gaussians,º Proc. 40th IEEE Symp. Foundations of Computer Science, pp. 634-644, Oct. 1999.
[14] S. Dasgupta and L.J. Shulman, ªA Two-Round Variant of EM for Gaussian Mixtures,º Proc. 16th Conf. Uncertainty in Artificial Intelligence (UAI-2000), pp. 152-159, June 2000.
[15] Q. Du, V. Faber, and M. Gunzburger, ªCentroidal Voronoi Tesselations: Applications and Algorithms,º SIAM Rev., vol. 41, pp. 637-676, 1999