Data Mining dengan Teknik Clustering Menggunakan Algoritma K-Means pada Data Transaksi Superstore

Priati*, Ahmad Fauzi
Sistem Informasi, FTIK
Universitas Buana Perjuangan Karawang
Jl. H.S Ronggowaluyo, Telukjambe Timur, Karawang
priati@ubpkarawang.ac.id*

***Abstrak*— Toko serba ada yang besar seperti halnya superstore yang memiliki data transaksi penjualan semakin lama semakin banyak sehingga akan terjadi penumpukan, penumpukan data tersebut dapat dimanfaatkan untuk mencari informasi yang bermanfaat dan tersembunyi sehingga hal tersebut dapat digunakan untuk melakukan strategi penjualan. Pengolahan penumpukan data yang demikian dapat dilakukan dengan penerapan data mining. Penelitian ini bertujuan untuk melakukan pengelompokan data superstore dengan menggunakan teknik clustering menggunakan algoritma K-Means. Sehingga akan diketahui empat kelompok order priority yaitu low, medium, high atau critical..**

Kata kunci*—— *data mining; data superstore; teknik clustering; algoritma K-means; kelompok order priority.

I. PENDAHULUAN

Data transaksi penjualan yang ada pada *superstore* semakin lama semakin banyak jumlahnya, hal tersebut sangat disayangkan jika tidak dimanfaatkan dengan baik karena hanya akan menjadi kuburan data alangkah lebih baik jika data tersebut dapat dimanfaatkan untuk mencari informasi yang bermanfaat untuk strategi penjualan sehingga dapat bersaing dengan *competitor*.

Pemahaman yang baik terhadap pelanggan dapat digunakan perusahaan untuk berinvestasi pelanggan yang potensial. Masalah yang sering dihadapi adalah kesulitan dalam menganalisa nilai pelanggan seperti penelitian yang telah diakukan[1]. Banyak pemasar mengalami kesulitan untuk mengidentifikasi pelanggan atau nasabah yang tepat, hal tersebut telah dilakukan penelitian oleh[2], hal tersebut dapat mengakibatkan perusahan kehilangan nasabah potensial dan tentunya akan sangat merugikan perusahaan.

Segmentasi pelanggan adalah metode yang populer yang digunakan untuk memilih pelanggan atau nasabah yang tepat untuk memulai promosi[2]. Dengan segmentasi nasabah berdasarkan prilakunya, kita dapat menargetkan tindakan mereka dengan lebih baik. Seperti peluncuran produk yang disesuaikan, target pemasaran dan untuk memenuhi harapan pelanggan[3].

Namun untuk menganalisa data pelanggan atau nasabah dalam jumlah besar memerlukan tenaga dan waktu yang banyak. *Clustering* data penjualan untuk menentukan *order priority* penting karena untuk meningkatkan penjualan dan terutama strategi pelayanan kepada *customer*. Penelitian terkait, *K-Means* merupakan suatu algoritma pengklasteran yang cukup sederhana yang mempartisi dataset kedalam beberapa klaster k. Algoritmanya cukup mudah untuk diimplementasi dan dijalankan, relatif cepat, mudah disesuaikan dan banyak digunakan[4].

Prinsip utama dari teknik ini adalah menyusun k buah partisi/pusat massa (*centroid*)/rata-rata (*mean*) dari sekumpulan data. Algoritma *K-Means* dimulai dengan pembentukan partisi klaster di awal kemudian secara iteratif partisi klaster ini diperbaiki hingga tidak terjadi perubahan yang signifikan pada partisi klaster[5].

Data Mining (DM) adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara *manual*. Beberapa teknik yang sering disebut-sebut dalam *literatur* DM antara lain: *clustering, classification, association rule mining, neural network,* dan *genetic algorithm*[6].

Clustering adalah salah satu sub kategori *data mining* dan merupakan proses di mana sampel yang sama dibagi menjadi kelompok-kelompok yang disebut *cluster*. Setiap *cluster* termasuk sampel dimana anggota yang mirip satu sama lain dan berbeda dengan sampel yang tersedia dari kelompok lain[7].

Analisa *cluster* merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis *cluster* mengklasifikasikan objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam *cluster* yang sama[8].

Algoritma K-Means merupakan salah satu algoritma dengan *partitional*, karena K-Means didasarkan pada penentuan jumlah awal kelompok dengan mendefinisikan nilai *centroid* awalnya[9]. Algoritma K–means akan mengelompokkan *item* data dalam suatu dataset ke suatu *cluster* berdasarkan jarak terdekat nilai *centroid* awal yang dipilih secara acak yang menjadi titik pusat awal[10], akan dihitung jarak dengan semua data menggunakan rumus *Euclidean Distance*. Data yang memiliki jarak pendek terhadap *centroid* akan membuat sebuah *cluster*. Proses ini berkelanjutan sampai tidak terjadi perubahan pada setiap kelompok[11].

Algoritma K-Means memiliki keuntungan yaitu :

1. Dalam implementasi menyelesaikan masalah, algoritma K-Means sangat *simple* serta *fleksibel*. Artinya perhitungan komputasinya tidak terlalu rumit dan algoritma ini dapat diimplementasikan pada segala bidang.
2. Algoritma K-Means sangat mudah untuk dipahami, terutama dalam implementasi data yang sangat besar serta dapat mengurangi kompleksitas data yang dimiliki[10].

Kelemahan yang dimiliki oleh algoritma K-Means yaitu :

1. Di Algoritma K-Means *user* memerlukan angka yang tepat dalam menentukan jumlah *cluster* sebanyak k karena terkadang pusat *cluster* awal dapat berubah sehingga kejadian ini dapat mengakibatkan pengelompokan data menjadi tidak stabil[12].
2. Algoritma K-Means tidak dapat maksimal dalam menentukan atau menginisialkan nilai *centroid* awalnya, karena pada pengelompokan data dengan algoritma K-Means sangat bergantung pada nilai *centroid*nya[13].
3. *Output* dari K-Means tergantung pada nilai – nilai pusat yang dipilih pada *clustering*. Sehingga pada algoritma ini nilai awal titik pusat *cluster* menjadi dasar dalam penentuan *cluster*. Pemilihan *centroid cluster* awal secara acak akan memberikan pengaruh terhadap kinerja *cluster* tersebut[14].

Beberapa penelitian dilakukan untuk mengatasi kelemahan yang ada pada Algoritma K-Means yaitu:

1. Perbaikan pada algoritma K-Means klasik untuk menghasilkan *cluster* yang lebih akurat. Algoritma yang diusulkan terdiri dari metode berdasarkan pemisahan data, untuk menemukan *centroid* awal sesuai dengan distribusi data[15]. Hasil penelitian ini menunjukkan bahwa algoritma yang diusulkan menghasilkan *cluster* yang lebih baik dalam waktu perhitungan yang singkat.
2. Ada beberapa cara untuk menentukan nilai k sebagai jumlah *cluster* yang dibentuk secara dinamis, salah satunya adalah dengan cara metode Elbow[16]. Penelitian ini menyatakan bahwa metode Elbow akan menentukan jumlah *cluster* yang sebenarnya pada satu dataset. Nilai k akan terus meningkat pada setiap langkahnya dan suatu saat nilai k akan mengalami penurunan dengan nilai yang besar, saat seperti itulah akan terbentuk siku dari semua nilai k yang didapat dan siku tersebut menjadi nilai k yang diinginkan.

II. METODE

Langkah-langkah melakukan *Clustering* dengan metode *K-Mean*[17] adalah sebagai berikut:

1. Pilih jumlah *cluster* K.
2. Inisialisasi K pusat *cluster* ini dapat dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara *random*. Pusat-pusat *cluster* diberi nilai awal dengan angka-angka *random*.
3. Alokasi semua data/objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagai berikut :

   $D\ ij = X1i - X1j2 + X2i - X2j\ 2 + \cdots + Xki - Xkj\ 2$ (1)

   Dimana :

   D (ij) = jarak data ke *(i)* ke pusat *cluster (j)*

   Xki = Data ke *(i)* pada atribut data ke *(k)*

   Xkj = Titik pusat *(j)* pada atribut *(k)*
4. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/ objek dalam *cluster* tertentu. Jika dikehendaki dapat juga menggunakan *median* dari *cluster* tersebut. Jadi rata-rata (*mean*) bukan satu-satunya ukuran yang dapat dipakai.

   $Rk = 1NkX1k + X2k + \cdots + Xnk$ (2)

   Dimana :

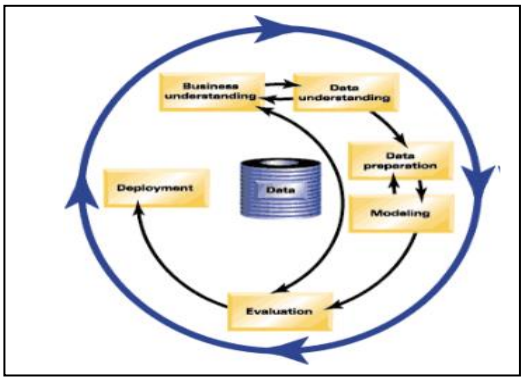
   Rk = Rata-rata baru.

   Nk = Jumlah *training pattern* pada *cluster (k)*.

   Xnk = Pola ke *(n)* yang menjadi bagian dari *cluster (k)*.

   Tugaskan lagi tiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai atau kembali ke langkah nomor 3 sampai pusat *cluster* tidak berubah lagi.

Penelitian ini menggunakan metode CRISP-DM. Dalam CRISP-DM sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase seperti pada Gambar 1. Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan bergantung kepada keluaran dari fase sebelumnya. Hubungan penting antar fase digambarkan dengan panah. Sebagai contoh, jika proses berada pada fase *modeling*. Berdasar pada perilaku dan karakteristik model, proses mungkin kembali kepada fase data *preparation* untuk perbaikan lebih lanjut terhadap data atau berpindah maju kepada fase *evaluation*[18].



Gambar 1. Proses CRISP-DM

III. HASIL DAN DISKUSI

A. *Fase Pemahaman Bisnis (Business Understanding Phase)*

1. Menentukan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan. Tujuan proyek adalah menganalisis data transaksi di sebuah *Superstore*, akan dilakukan 4 cluster *Order Priority* yaitu *Low, High, Medium, Critical*, berdasarkan variabel *Order Quantity* dan *Sales*. Data diperoleh dari http://community.tableau.com. Jumlah *record* dari data tersebut adalah 8399 dengan 21 *variable*. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining.* Pada tahap ini dilakukan pemahaman terhadap tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau *unit* penelitian secara keseluruhan dan menerjemahkannya kedalam tujuan *data mining.*
2. Menyiapkan strategi awal untuk mencapai tujuan.
   Strategi awal untuk mencapai tujuan adalah melakukan pencarian data di internet.

B. *Fase Pemahaman Data (Data Understanding Phase)*

Dataset *superstore sales* yang didapat dari internet berupa dokumen *spreadsheet*.

1. Mengumpulkan data.
   Sumber data utama yang digunakan dalam penelitian ini adalah dataset *Superstore Sales.*
2. Menggunakan analisis penyelidikan data untuk mengenali data lebih lanjut dan pencarian pengetahuan awal.
   Dataset *superstore sales* terdiri dari beberapa atribut antara lain *RowID, OrderID, Order Date, Order Priority, Order Quantity, Sales, Discount, Ship Mode, Profit, Unit Price, Shipping Cost, Customer Name, Province, Region, Customer Segment, Product Category, Product Sub-Category, Product Name, Product Container, Product Base Margin, Ship Date.*
3. Mengevaluasi kualitas data.
   Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau dobel dan ditemukan banyak nilai kosong atau *null* yang disebut sebagai *missing value*. Memilih dan menciptakan satu dataset untuk mendukung proses penemuan pengetahuan yang akan dilakukan. Melakukan *preprocessing* dan *cleansing* seperti menangani data yang tidak lengkap dan menghilangkan gangguan atau *outlier*.
4. Jika diinginkan pilih sebagian kecil kelompok data yang mungkin mengandung pola dari permasalahan.Variabel yang digunakan adalah *Order Priority*, *Order Quantity*, dan *Sales*.

C. *Fase Pengolahan Data (Data Preparation Phase)*

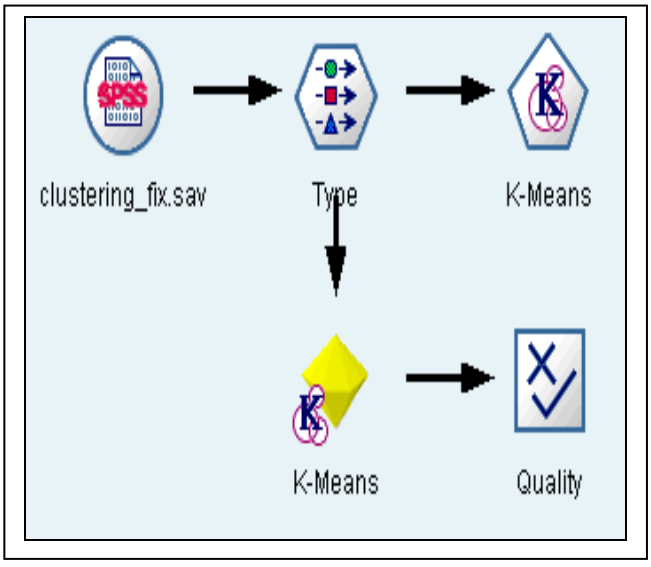
1. Siapkan data awal yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilakukan secara intensif. Persiapan data mencakup semua kegiatan untuk membangun dataset *superstore sales* yang akan diterapkan kedalam alat pemodelan dari data mentah awal berupa dataset *superstore sales* dan selanjutnya akan dilakukan proses *data mining*.
2. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai dengan analisis yang akan dilakukan.
   Variabel yang digunakan adalah *Order Priority, Order Quantity*, dan *Sales*.
3. Siapkan data awal sehingga siap untuk perangkat pemodelan.
   Tahap ini merupakan tahap untuk memastikan data *superstore sales* yang dipilih telah layak untuk dilakukan pengolahan seperti terlihat pada Tabel 1.

TABEL 1. DATASET SIAP PEMODELAN

ORDER PRIORITY	ORDER QUANTITY	SALES
LOW	6	261,54
HIGH	49	10123,02
HIGH	27	244,57
HIGH	30	4965,7595
HIGH	12	93,54
HIGH	22	905,08
HIGH	21	2781,82
LOW	44	228,41
MEDIUM	45	196,85
MEDIUM	32	124,56
MEDIUM	46	1815,49
CRITICAL	44	4462,23
CRITICAL	11	663,784

D. *Fase Pemodelan (Modeling Phase)*

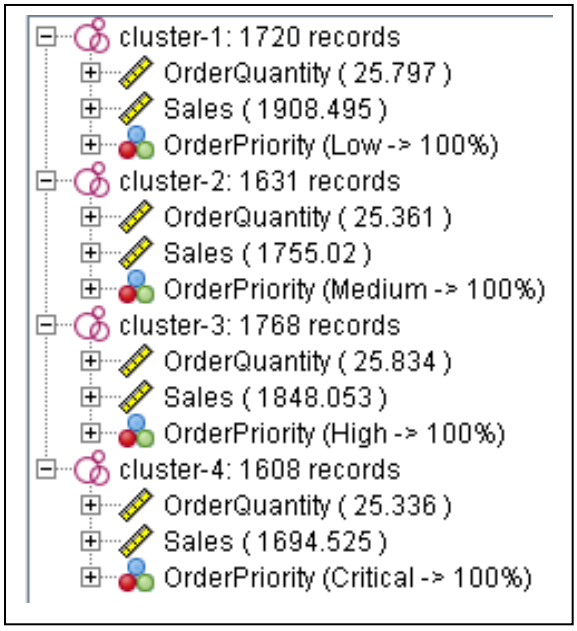
Fase pemodelan adalah fase yang secara langsung melibatkan teknik data mining dan menentukan algoritma yang akan dilakukan. Dalam hal ini teknik data mining yang digunakan adalah teknik *Clustering* dengan algoritma K-Means. Seperti terlihat pada Gambar 2.



Gambar 2. Teknik clustering algoritma K-Means pada SPSS clementine

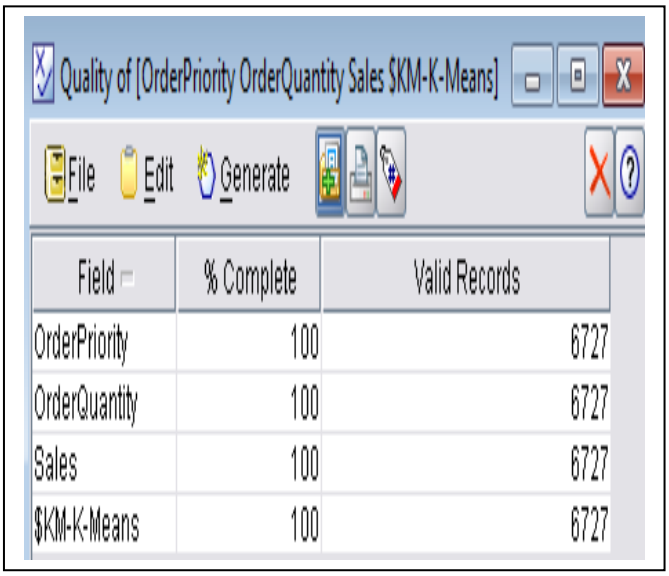
E. Fase Evaluasi (Evaluation Phase)

1. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
   Semua model sesuai dengan tujuan.
2. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
   Tidak terdapat permasalahan penting yang tidak tertangani.
3. Mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining*. Terdapat pada Gambar 3.
   Pengetahuan yang didapatkan dalam interpretasi sebagai berikut :
   - Cluster pertama *Order Priority Low* → cluster ini memiliki 1720 *record* merupakan kelompok pelanggan yang melakukan *Order Quantity* (25,797) dan *Sales* (1908,495)
   - Cluster kedua *Order Priority Medium* → cluster ini memiliki 1631 *record* merupakan kelompok pelanggan yang melakukan *Order Quantity* (25,361) dan *Sales* (1755,02)
   - Cluster ketiga *Order Priority High* → cluster ini memliki 1768 *record* merupakan kelompok pelanggan yang melakukan *Order Quantity* (25,834) dan *Sales* (1848,053)
   - Cluster keempat *Order Priority Critical* → cluster ini memiliki 1608 *record* merupakan kelompok pelanggan yang melakukan *Order Quantity* (25,336) dan *Sales* (1694,525)



Gambar 3. Hasil pengolahan data teknik clustering

4. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektifitas. Hasil evaluasi mendapatkan kualitas seperti Gambar 4.



Field	% Complete	Valid Records
OrderPriority	100	6727
OrderQuantity	100	6727
Sales	100	6727
$KM-K-Means	100	6727

Gambar 4. Kualitas hasil pengolahan

F. Fase Penyebaran (Deployment Phase)

Dalam hal ini penyebaran dilakukan dengan pembuatan laporan dan artikel jurnal.

IV. KESIMPULAN

Teknik *clustering* dengan algoritma K-Means dapat membantu pengelompokkan data *superstore* dalam pengambilan keputusan untuk *order priority* menjadi empat kelompok yaitu, *low, medium, high,* dan *critical.*

Pengetahuan yang didapat dari penelitian adalah Kelompok *Order Priority Low*. Kelompok ini memiliki anggota 1720 *record*, dan merupakan kelompok yang memiliki *Order Quantity* 25.797 dan *Sales* 1908.495.

a. Kelompok *Order Priority Medium*
   Kelompok ini memiliki anggota 1631 *record*, dan merupakan kelompok yang memiliki *Order Quantity* 25.361 dan *Sales* 1755.02.
b. Kelompok *Order Priority High*
   Kelompok ini memiliki anggota 1768 *record*, dan merupakan kelompok yang memiliki *Order Quantity* 25.834 dan *Sales* 1848.053.
c. Kelompok Order Priority Critical
   Kelompok ini memiliki anggota 1608 *record* dan merupakan kelompok yang memiliki *Order Quantity* 25.336 dan *Sales* 1694.525.

Dengan kualitas hasil pengolahan data dari masing-masing kelompok adalah 100%, seperti terlihat pada Gambar 4. Demikian pengetahuan yang didapat dari penelitian ini akan bermanfaat untuk pengambilan keputusan dalam pemberian pelayanan terhadap pelanggan *(customers).* Kualitas pelayanan yang baik akan meningkatkan penjualan.

DAFTAR PUSTAKA

[1] Xing, B. I. (2010). The evaluation of customer potential value based on prediction and cluster analysis, international conference on management science & engineering 17th, Melbourne, Australia 613–618.

[2] Chai, C., & Chan, H. (2008). Intelligent value-based customer segmentation method for campaign management : a case study of automobile retailer, expert system with application, 34, 2754–2762.

[3] Balaji, S., & Srivatsa, S. K. (2012). Customer segmentation for decision support using clustering and association rule based approaches, international journal of computer science & engineering technology, 3(11), 525–529.

[4] Wu, Xindong & Kumar, Vipin. (2009). The top ten algorithms in data mining. London: crc press.

[5] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: practical machine learning and tool. Burlington: morgan kaufmann publisher.

[6] Lindawati (2008), data mining dengan teknik clustering dalam pengklasifikasian data mahasiswa studi kasus prediksi lama studi mahasiswa universitas bina nusantara, seminar nasional informatika (semnasif 2008), issn :1979-2328.

[7] Farhad Soleimanian Gharechopogh, Yasin Rahimpur, Seyyed Reza Khaze (2014), combining clustering algorithhms for provide marketing policy in electronic stores, international journal of programmning languages and applications (ijpla), volume : 4, nomor: 1.

[8] Ediyanto Et Al, (2013). "pengklasifikasian karakteristik dengan metode k means cluster analysis", buletin ilmiah mat. Stat. Terapannya (bimaster), volume 02, no.2, (2013).

[9] Madhulatha, T.S., 2012. An overview on clustering methods. Iosr journal of engineering, ii(1), pp.719-25.

[10] Bangoria, B., Mankad, N. & Pambhar, V., 2013. A survey on efficient enhanced K-Means clustering algorithm. International journal for scientific research & development, i(9), pp.1698-700.

[11] Agrawal, A. & Gupta, H., 2013. Global K-Means (gkm) clustering algorithm: a survey. International journal of computer applications, lix(2), pp.20-24.

[12] Joshi, K.D. & Nalwade, P.S., 2013. Modified K-Means for better initial cluster centres. International journal of computer science and mobile computing, ii(7), pp.219-23.

[13] Ahmed, A.H. & Ashour, W., 2011. An initialization method for the K-Means algorithm using rnn and coupling degree. International journal of computer applications, xxv(1), pp.1-6.

[14] Singh, H. & Kaur, K., 2013. New method for finding initial cluster centroids in K-Means algorithm. International journal of computer applications, lxxiv(6), pp.27-30.

[15] Kaur, K., Dhaliwal, D.S. & Vohra, K.R., 2013. Statistically refining the initial points for kmeans clustering algorithm. International journal of advanced research in computer engineering & technology, ii(11), pp.2972-77.

[16] Kodinariya, T.M. & Makwana, P.R., 2013. Review on determining number of cluster in kmeans clustering. International journal of advance research in computer science and management studies, i(6), pp.90-95.

[17] Larose, Daniel. T. 2005. Discovering knowledge in data: an introduction to data mining. John willey & sons. Inc.