Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Implementasi K-Means *Clustering* pada *Online Retail* berdasarkan *Recency*, *Frequency*, dan *Monetary*

(*Implementation of K-Means Clustering in Online Retail based on Recency, Frequency, and Monetary*)

Karima Marwazia Shaliha[1], Angelyna Angelyna[2], Arham Aulia Nugraha[3], Muhammad Humam Wahisyam[4], Tri Kurnia Sandi[5]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050057@student.uinsgd.ac.id
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050016@student.uinsgd.ac.id
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050017@student.uinsgd.ac.id
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050069@student.uinsgd.ac.id
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050115@student.uinsgd.ac.id

Abstrak

Di masa pandemi seperti saat ini banyak sekali perubahan yang terjadi, salah satunya yaitu semakin maraknya situs jual beli secara online. Setiap Online Shop menawarkan berbagai produk dan jasa dengan berbagai penawaran menarik, bersaing dengan ketat untuk menarik para peminatnya. Dengan terjadinya pola perubahan pada masyarakat tersebut perlu dilakukan sebuah pengelompokan untuk mendapatkan informasi guna menentukan strategi penjualan yang lebih baik. Proses pengelompokan tersebut menggunakan teknik dari data mining yakni Clustering dengan algoritma K-Means berdasarkan *Recency Frequency Monetary* (RFM), maka diharapkan dengan menganalisa tiga atribut tersebut dan pengimplementasian algoritma K-Means ini dapat memberikan sebuah keluaran yang akurat dan sesuai dengan tujuan penelitian ini.

Kata kunci: K-Means, Klasterisasi, Online Shop, Recency Frequency Monetary (RFM)

Abstract

During a pandemic like today, many changes have occurred, one of which is the increasing number of online buying and selling sites. Each Online Store offers a variety of products and services with a variety of attractive offers, competing fiercely to attract enthusiasts. With the occurrence of a pattern of change in society, it is necessary to carry out a grouping to obtain information in order to determine a better sales strategy. The grouping process uses techniques from data mining, namely Clustering with the K-Means algorithm based on the Recency Frequency Monetary (RFM) algorithm, it is hoped that by analyzing the three attributes and implementing the K-Means algorithm, it can provide an accurate output and in accordance with the objectives of this study.

Keywords: *Clustering, K-Means, Online Shop, Recency Frequency Monetary (RFM)*

1 Pendahuluan

Dunia saat ini tengah waspada dengan adanya virus COVID-19. Penyebaran yang sangat cepat membuat peningkatan jumlah kasus yang melesat, bahkan hampir tidak ada negara yang bisa

memastikan terhindar dari virus ini [1]. Tentunya dengan adanya pandemi ini banyak sektor yang terkena imbas, salah satunya yaitu sektor ekonomi. Di Indonesia tidak sedikit perusahaan yang sudah menutup usahanya serta menggelar PHK [2]. Akan tetapi dengan keadaan seperti ini membuat banyak usaha yang bertahan dengan mengalih fungsikan usaha nya menjadi kegiatan jual beli online. Hal ini membuat momentum yang lebih siap bagi pelaku usaha yang sudah eksis lebih awal dalam jual beli online (*online shop)* dan momentum baru bagi pebisnis yang baru memulai. Di sisi konsumen pun belanja *online* (*online shopping*) semakin meningkat [3]. Semakin maraknya jual beli online, para *Online Shop* menawarkan berbagai produk dan jasa dengan berbagai penawaran-penawaran menarik, bersaing dengan ketat untuk menarik para peminatnya. Dengan terjadinya persaingan tersebut perlu merumuskan strategi penjualan yang lebih baik agar bisa tepat sasaran kepada pelanggannya. Algoritma k-means dan k-medoids dari teknik clustering dapat membantu dalam mengklasifikasi pelanggan berdasarkan RFM sehingga perusahaan dapat menargetkan pelanggannya secara efisien.

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (unsupervised) serta tidak memerlukan target output. Dalam data mining ada dua jenis metode clustering yang digunakan dalam pengelompokan data, yaitu hierarchical clustering dan non-hierarchical clustering) [4].

Pada penelitian sebelumnya terkait dengan *clustering* ataupun penerapan algoritma *K-means* yang telah banyak dilakukan, seperti yang telah dilakukan oleh nurul rohmawati, dkk membuat cluster dengan algoritma *k-means* untuk penerima beasiswa berdasarkan ukt serta sks yang diambil [5]. Fauziah Nur, dkk dengan algoritma *k-means* juga membuat cluster untuk pengelompokan jurusan berdasarkan beberapa kriteria dari calon siswa [6]. kemudian Anindya Khrisna membuat cluster dengan algoritma *k-means* untuk pengelompokan penyakit pasien puskesmas Kajen Pekalongan berdasarkan data umur, kode penyakit dan lama mengidap penyakit [7], Asroni dan Ronald Adrian juga melakukan penelitian untuk membuat *cluster* mahasiswa berdasarkan nilai akademik dengan Weka Interface dengan studi kasus jurusan Teknik Informatika UMM Magelang [8], dan selanjutnya Ade Bastian, dkk melakukan *clustering* menggunakan algoritma *k-means* untuk pengelompokan penyakit menular manusia berdasarkan data penyakit menular di kabupaten Majalengka [9].

Pada penelitian ini kami menggunakan algoritma *K-means* untuk menentukan cluster pelanggan untuk strategi pemasaran berdasarkan data *recency*, *frequency* dan *monetary*.

2 Metodologi

Metode penelitian yang digunakan dalam penelitian ini diantaranya merumuskan masalah, tujuan penelitian, pengumpulan data, perancangan sistem serta implementasi sistem menggunakan algoritma *K-means*.

1. Merumuskan Masalah

Banyaknya online shop yang ada pada masa pandemi serta dimudahkan nya pembeli dalam membeli atau menggunakan jasa maka persaingan semakin ketat dan penjual diharuskan memiliki strategi penjualan yang lebih baik.

2. Merumuskan Tujuan Penelitian

Membangun beberapa cluster berdasarkan data RFM (Recency, Frekuensi dan Moneter), yang selanjutnya hasil cluster tersebut akan digunakan untuk pemilihan target pelanggan pada suatu perusahaan.

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

3. Pengumpulan data

Data yang digunakan merupakan dataset Online Ritel yang telah di publish di kaggle, ritel online adalah kumpulan data transnasional yang berisi semua transaksi yang terjadi antara 01/12/2010 dan 09/12/2011 untuk ritel daring yang berbasis non-toko yang berada di inggris dan terdaftar. Inti dari perusahaan ini yaitu menjual hadiah unik untuk semua acara dan banyaknya pelanggan perusahaan tersebut adalah para pedagang grosir.

4. Perancangan Sistem

Tahap ini merupakan tahap perancangan sistem untuk mengimplementasikan k-means hierarchical clustering pada online retail berdasarkan *recency,frequency* dan *monetary* yang dibangun dengan menggunakan bahasa pemrograman phyton. pada proses ini terdiri dari beberapa tahap, yaitu:

5. Preprocessing

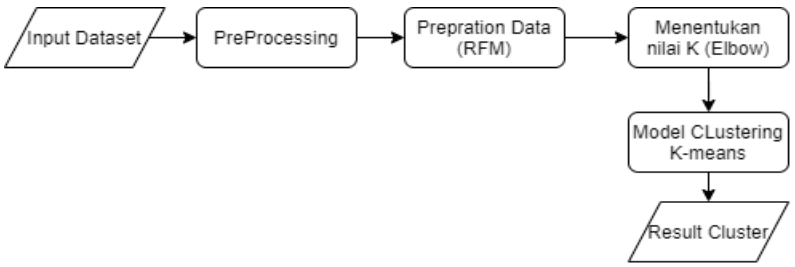
Dataset yang akan diproses menggunakan algoritma akan di preprocess terlebih dahulu. Tahapan nya yaitu menghilangkan *missing value* dengan menghapus baris yang mengandung nilai kosong atau *null*. Selanjutnya mengubah tipe data *CustomerID* dari *integer* menjadi *string*

6. Data Preparation

Data preparation ini merupakan proses mempersiapkan data-data yang dibutuhkan dalam proses algoritma. Adapun proses yang dilakukan pada tahap ini yaitu dengan membuat tiga atribut baru yaitu [10]:

- **Monetary**, merupakan jumlah uang yang dibelanjakan oleh customer selama periode waktu tertentu, dimana semakin tinggi nilainya , maka semakin banyak keuntungan yang akan dihasilkan oleh perusahaan.
- **Frequency**, merupakan periode antara dua pembelian pelanggan berikutnya. maka dapat disimpulkan jika nilai frekuensi tinggi, maka semakin banyak juga kunjungan pelanggan ke perusahaan.
- **Recency**, recency mengacu kepada jumlah hari sebelum tanggal frekuensi ketika pelanggan melakukan pembelian terakhir, maka jika semakin rendah nilai recency, akan semakin tinggi kunjungan pelanggan ke perusahaan.

Implementasi Algoritma k-means untuk menghasilkan beberapa clustering. Gambar 1 merupakan gambaran sistem yang akan dibangun.



Gambar 1 Flowchart Sistem

7. Implementasi

Pada tahap implementasi dilakukan dengan menggunakan bahasa pemrograman phyton dan algoritma k-means yang akan menghasilkan beberapa cluster.

8. Algoritma *K-Means*

Merupakan salah satu metode pengelompokan data non hirarki (sekatan) yang berusaha membagi data ke dalam bentuk dua atau lebih kelompok. Metode ini membagi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berbeda karakteristik dikelompokkan ke dalam kelompok yang lain [11]. Pengelompokan data dengan metode K- Means secara umum dilakukan dengan algoritma sebagai berikut [11].

- Menentukan berapa banyaknya k kelompok
- Membagi data ke dalam k kelompok
- Menghitung pusat kelompok (sentroid) dari data yang ada di masing-masing kelompok dan dinyatakan dalam bentuk persamaan(2)
- Dimana C adalah sentroid, M adalah banyak data, j adalah banyak kelompok.
- Masing-masing data dialokasikan ke sentroid terdekat. Menghitung jarak data ke setiap centroid menggunakan jarak Euclidean dan dinyatakan dalam bentuk persamaan berikut:

$$D(X_I, C_i = \sqrt{\sum_{j=1}^{q}(x_{ij} - C_{IJ})^2}$$

3 Hasil dan Pembahasan

Penelitian ini menggunakan algoritma *K-Means* untuk membuat *cluster* berdasarkan *Recency Frequency Monetary (RFM).* Pada tahap awal dilakukan input dataset retail yang didapat melalui kaggle. Dataset ini memiliki 8 variabel seperti pada Gambar 3.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850.0	United Kingdom

Gambar 2 Dataset Retail Online

Dataset diatas kemudian masuk kedalam preprocessing terlebih dahulu, yaitu dengan menghapus data yang memiliki nilai null atau tidak memiliki arti, selain itu dilakukan juga pengubahan tipe data *customer* id dari type int menjadi tipe string. setelah data dirasa telah siap untuk diolah, maka tahap selanjutnya yaitu preparation, dimana pada proses ini dilakukan pembuatan atribut baru yaitu *Recency,Frequency* dan *Monetary* (RFM).

1. Pembuatan atribut baru *Monetary*

Atribut monetary dibuat dengan mengalikan atribut *quantity* dan *Unitprice*, maka hasilnya seperti pada Gambar 3.

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

	CustomerID	Amount
0	12346.0	0.00
1	12347.0	4310.00
2	12348.0	1797.24
3	12349.0	1757.55
4	12350.0	334.40

Gambar 3 Hasil pembuatan atribut baru (Monetary)

2. Pembuatan atribut baru *Frequency*

Atribut *Frequency* dibuat berdasarkan banyak nya data *invoice* per *customer*, hasilnya seperti pada Gambar 4.

	CustomerID	Frequency
0	12346.0	2
1	12347.0	182
2	12348.0	31
3	12349.0	73
4	12350.0	17

Gambar 4 Hasil pembuatan atribut baru (Frequency)

Setelah kedua atribut baru tersebut dibuat, maka tahap selanjutnya yaitu menyatukan kedua atribut tersebut kedalam satu frame. Gambar 5 merupakan hasil dari penggabungan kedua atribut baru tersebut

	CustomerID	Amount	Frequency
0	12346.0	0.00	2
1	12347.0	4310.00	182
2	12348.0	1797.24	31
3	12349.0	1757.55	73
4	12350.0	334.40	17

Gambar 5 Hasil penggabungan dua atribut baru

3. Pembuatan atribut baru *Recency*

Untuk membuat atribut *Recency* perlu meng convert terlebih dahulu tipe data pada *invoiceDate* menjadi format (Day,Month,year, hour:minute), kemudian menghitung tanggal maksimum untuk mengetahui tanggal terakhir transaksi. Maka hasil setelah data digabungkan ke dalam dataframe, hasil nya seperti yang terlihat pada Gambar 6.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Amount	Diff
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	373 days 04:24:00
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	373 days 04:24:00
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00	373 days 04:24:00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	373 days 04:24:00
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	373 days 04:24:00

Gambar 6 Hasil penggabungan atribut baru

Selanjutnya yaitu menghitung tanggal transaksi terakhir untuk mendapatkan *recency* atau kebaruan pelanggan, Hasilnya seperti yang terlihat pada Gambar 7.

	CustomerID	Diff
0	12346.0	325 days 02:33:00
1	12347.0	1 days 20:58:00
2	12348.0	74 days 23:37:00
3	12349.0	18 days 02:59:00
4	12350.0	309 days 20:49:00

Gambar 7 Hasil dari recency pelanggan

Step terakhir, menggabungkan hasil dari recency berupa data hari dengan semua atribut baru yang telah dibuat. maka hasilnya seperti pada Gambar 8.

	CustomerID	Amount	Frequency	Recency
0	12346.0	0.00	2	325
1	12347.0	4310.00	182	1
2	12348.0	1797.24	31	74
3	12349.0	1757.55	73	18
4	12350.0	334.40	17	309

Gambar 8 Penggabungan tiga atribut baru

Setelah atribut baru terbentuk, kami memvisualisasikan dulu datanya untuk mengetahui apakah ada data *outlier* atau tidak. Selanjutnya menghapus data outlier yang terdapat pada setiap atribut. Selanjutnya mengubah skala variabel sehingga memiliki skala yang sama dengan melakukan standarisasi (mean-0, sigma-1), maka hasilnya dapat dilihat pada Gambar 9.

	Amount	Frequency	Recency
0	-0.723738	-0.752888	2.301611
1	1.731617	1.042467	-0.906466
2	0.300128	-0.463636	-0.183658
3	0.277517	-0.044720	-0.738141
4	-0.533235	-0.603275	2.143188

Gambar 9 hasil normalisasi variabel

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

Tahap selanjutnya yaitu penentuan jumlah “k” atau jumlah cluster yang akan dibuat. Disini kami menggunakan metode *Silhouette*. Sehingga menghasilkan hasil seperti Gambar 10.

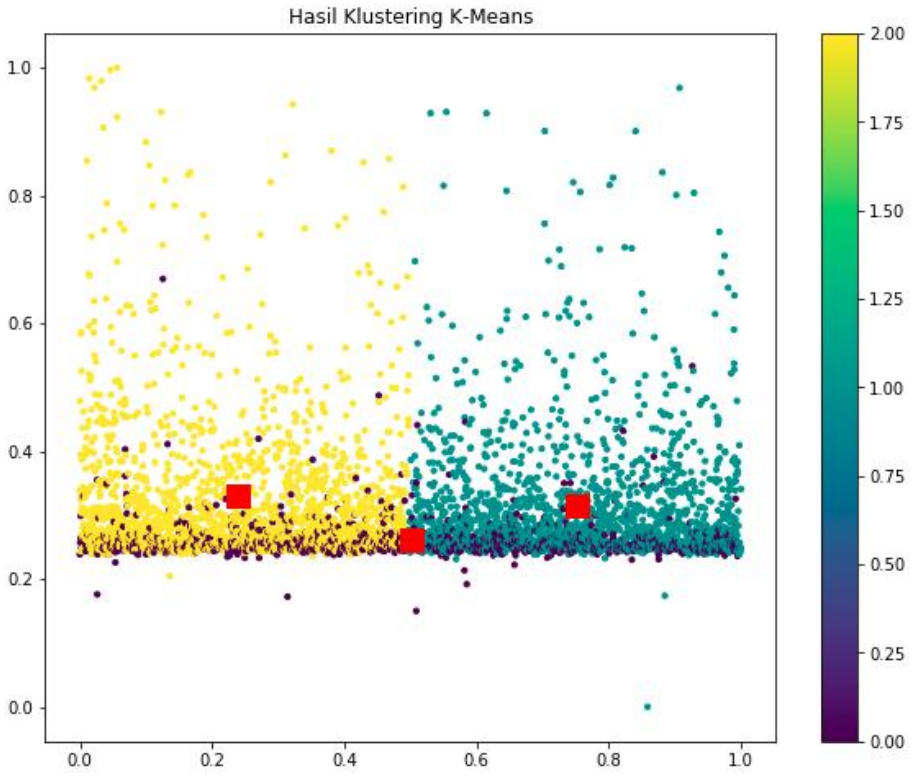
```
For n_clusters=2, the silhouette score is 0.33514448734200425
For n_clusters=3, the silhouette score is 0.39705628252134517
For n_clusters=4, the silhouette score is 0.3868714505806646
For n_clusters=5, the silhouette score is 0.3969730854409133
For n_clusters=6, the silhouette score is 0.34114509558953343
For n_clusters=7, the silhouette score is 0.34544219626269174
For n_clusters=8, the silhouette score is 0.33949230629942745
```

Gambar 10 Hasil pencarian K menggunakan Silhouette

Berdasarkan hasil analisis silhouette diatas, kami menggunakan nilai k=3 untuk banyaknya jumlah cluster yang dibuat. Selanjutnya dataset 3 variabel RFM tersebut diclustering dengan model yang sudah dibuat, sehingga hasilnya seperti yang terlihat pada Gambar 11 dan 12.

	CustomerID	Amount	Frequency	Recency	Cluster_Id
0	12346.0	0.00	2	325	1
1	12347.0	4310.00	182	1	2
2	12348.0	1797.24	31	74	0
3	12349.0	1757.55	73	18	0
4	12350.0	334.40	17	309	1

Gambar 11 Hasil penetapan label



Gambar 12 Visualisasi hasil clustering

4 Simpulan

Berdasarkan percobaan kami mengenai implementasi algoritma mean shift untuk clustering. Dapat dilihat bahwa algoritma meanshift clustering untuk dataset ini dapat melakukan klasterisasi penjualan di sebuah toko online dengan baik, hal ini dapat dilihat dari hasil visualisasi data yang menunjukan tiga klaster yang terbentuk dengan masing - masing cluster_id 0, 1, dan 2. Dimana perbandingan cluster_id dan amount menggunakan boxplot model menunjukkan cluster tertinggi yakni cluster_id 2 dengan warna boxplot hijau, untuk perbandingan cluster_id dan frequency didapatkan cluster dengan frequeny transaksi tertinngi yakni cluster_id 1 dengan warna bloxplot orange, dan untuk hasil perbandingan cluster_id dengan recency didapatkanhasil cluster dengan recency tertinggi yakni cluster 0 dengan warna boxplot biru.

Referensi

[1] V. No and N. Mona, "Konsep Isolasi Dalam Jaringan Sosial Untuk Meminimalisasi Efek Contagious (Kasus Penyebaran Virus Corona Di Indonesia)," *J. Sos. Hum. Terap.*, vol. 2, no. 2, pp. 117–125, 2020, doi: 10.7454/jsht.v2i2.86.

[2] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 151, 2018, doi: 10.24076/citec.2017v4i2.106.

[3] T. Taufik and E. A. Ayuningtyas, "Dampak Pandemi Covid 19 Terhadap Bisnis Dan Eksistensi Platform Online," *J. Pengemb. Wiraswasta*, vol. 22, no. 01, p. 21, 2020, doi: 10.33370/jpw.v22i01.389.

[4] B. Santosa, "Data mining:Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu- Bisnis.Edisi Pertama.," *Data miningTeknik Pemanfaat. Data untuk Keperluan Bisnis. Yogyakarta Graha Ilmu- Bisnis.Edisi Pertama.*, vol. 33, no. 4, pp. 365–373, 2007.

[5] mohamad jajuli nurul rohmawati, sofi defiyanti, "Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa," *Jitter 2015*, vol. I, no. 2, pp. 62–68, 2015.

[6] F. Nur, M. Zarlis, and B. B. Nasution, "Penerapan Algoritma K-Means Pada Siswa Baru Sekolahmenengah Kejuruan Untuk Clustering Jurusan," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 1, no. 2, pp. 100–105, 2017, doi: 10.30743/infotekjar.v1i2.70.

[7] A. K. Wardhani, "Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan," *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.

[8] R. A. Asroni, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *Ilm. Semesta Tek.*, vol. 18, no. 1, pp. 76–82, 2015.

[9] A. Bastian, H. Sujadi, and G. Febrianto, "Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka)," no. 1, pp. 26–32.

[10] T. Hermanto, "Implementasi Algoritma Association Rule Dan K-Means Sebagai Sistem Rekomendasi Produk Pada Website Penjualan Online," *Stt-Wastukancana.Ac.Id*, pp. 70–73.

[11] B. S. Ashari, S. C. Otniel, and Rianto, "Perbandingan Kinerja K-Means Dengan DSCAN Untuk Metode Clustering Data Penjualan Online Retail," *J. Siliwangi*, vol. 5, no. 2, pp. 72–77, 2019.