SEGMENTASI CUSTOMER PERSONALITY MENGGUNAKAN K-MEANS BERDASARKAN INCOME, SPENDING, DAN RECENCY

LATHIIFA ZAHIRA YAHYA

G6401231154

PROGRAM STUDI ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025

I PENDAHULUAN

I.1 Latar Belakang

Dalam dunia bisnis modern, memahami karakteristik dan perilaku pelanggan menjadi kunci utama untuk menyusun strategi pemasaran yang efektif. Tidak semua pelanggan memiliki kebiasaan belanja yang sama. Ada pelanggan yang aktif dan loyal, ada pula yang pasif dan jarang bertransaksi. Oleh karena itu, dibutuhkan pendekatan analitis untuk mengelompokkan pelanggan ke dalam segmen-segmen agar pemasaran lebih terarah, yang pada akhirnya meningkatkan kepuasan dan loyalitas pelanggan. Segmentasi pelanggan adalah strategi dalam membagi pelanggan ke beberapa kelompok yang memiliki karakteristik atau perilaku yang sama (Fadhillah *et al*. 2025).

Dataset *Customer Personality Analysis* dari Kaggle menyediakan informasi demografis dan perilaku konsumen yang dapat dimanfaatkan untuk segmentasi. Metode pengelompokkan berbasis seperti K-Means *Clustering* yang telah terbukti mampu mengklasifikasikan pelanggan secara lebih terperinci dengan memanfaatkan variasi dalam perilaku pelanggan (Rumapea, *et al*. 2024). Melalui metode K-Means, pelanggan dikelompokkan berdasarkan variabel pendapatan tahunan (*Income*), jumlah pengeluaran (*Spending*), dan jumlah hari sejak pembelian terakhir (*Recency*).

Pemrosesan data ini berfokus pada penerapan algoritma K-Means untuk menghasilkan *cluster* pelanggan yang representatif, disertai dengan evaluasi hasil klasterisasi menggunakan metode *Silhouette Coefficient* dan analisis pendukung lainnya.

I.2 Tujuan

Tujuan dari analisis dalam projek ini adalah untuk:

1. Mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan tiga variabel utama, yaitu *Income*, *Total Spending*, dan *Recency*.
2. Mengidentifikasi karakteristik masing-masing *cluster* pelanggan yang diberi label deskriptif seperti *High Spender – Active* atau *Low Spender – Dormant*.
3. Mengevaluasi kualitas hasil segmentasi menggunakan metode *Silhouette Coefficient*, untuk memastikan bahwa pembagian *cluster* sudah cukup representatif dan terpisah dengan baik.

I.3 Manfaat

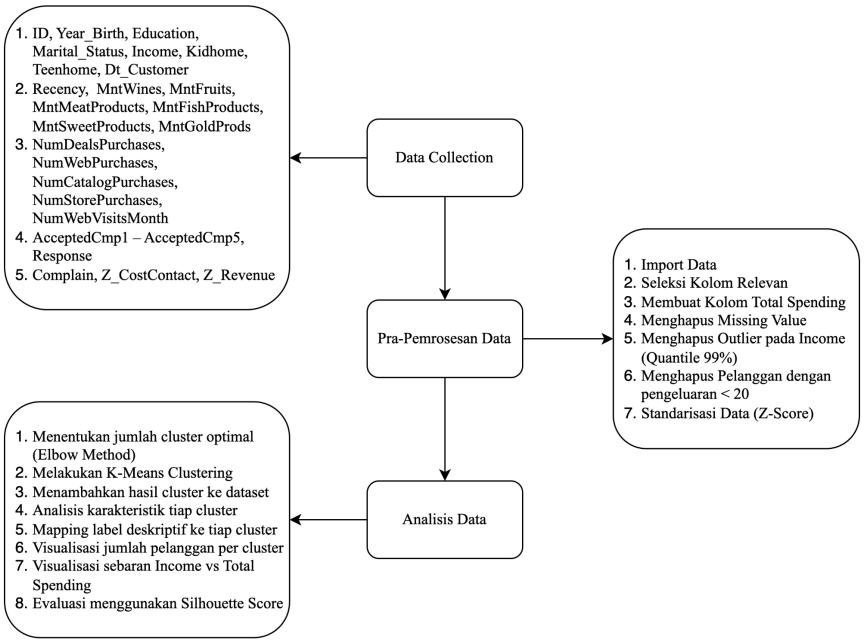
Dengan adanya projek hasil segmentasi pelanggan berdasarkan *income*, *spending*, dan *recency* membantu perusahaan mengenali perilaku pelanggan secara lebih spesifik dengan menyusun strategi pemasaran serta mengoptimalkan penggunaan anggaran promosi.

II DATA

Dataset ini diambil melalui situs *Kaggle* yang berjudul Customer Segmentation: *Clustering*. Dataset ini mengusung topik utama terkait *Customer Personal Analysis* yang berisi informasi demografis dan perilaku pelanggan yang melibatkan pemeriksaan menyeluruh terhadap profil pelanggan suatu perusahaan. Dataset ini digunakan untuk mengelompokkan pelanggan ke dalam segmen-segmen berbeda berdasarkan karakteristiknya yang dapat digunakan oleh bisnis perusahaan untuk menyesuaikan produk guna memenuhi kebutuhan, perilaku, dan perhatian yang berbeda dari berbagai jenis pelanggan.

Tujuan utama dari dataset ini adalah untuk mempraktikkan dan memahami bagaimana algoritma K-Means dapat digunakan untuk segmentasi pelanggan. Oleh karena itu, perusahaan dapat mengidentifikasi segmen yang paling mungkin tertarik pada produk kemudian menyesuaikan strategi pemasaran mereka untuk setiap segmen, meningkatkan kepuasan pelanggan, dan mengoptimalkan penawaran produk atau layanan.

III METODE



Gambar 1 Metodologi yang Digunakan

III.1 Pengumpulan Data

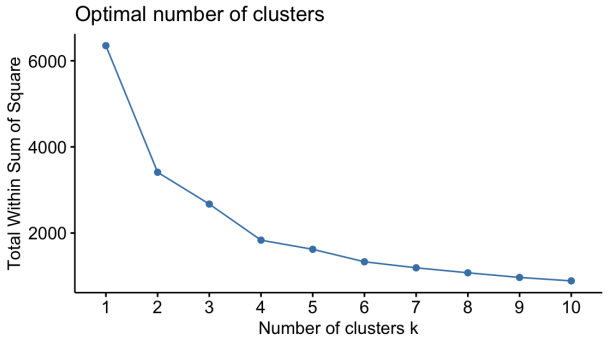
Dataset ini terdiri dari 29 fitur yang mencakup informasi demografis dan perilaku pelanggan, seperti tahun kelahiran, tingkat pendidikan, status pernikahan, pendapatan tahunan, jumlah anak dan remaja di rumah, serta tanggal menjadi pelanggan. Selain itu, dataset ini mencatat pengeluaran pelanggan dalam berbagai kategori produk, frekuensi pembelian melalui berbagai kanal, serta respons terhadap kampanye pemasaran.

III.2 Pra-Pemrosesan Data

1. *Import* Data
   Dataset C*ustomer Personality Analysis* dimasukkan ke dalam R menggunakan fungsi read.csv() untuk diproses dan dilakukan analisis.
2. Seleksi Kolom Relevan
   Beberapa kolom dataset dihapus karena tidak memberikan dampak signifikan terhadap analisis. Hanya kolom yang berhubungan langsung dengan segmentasi berbasis perilaku pembelian yang dipilih, yaitu: *Income*, *Recency*, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, dan MntGoldProds.
3. Membuat Kolom Total *Spending*
   Variabel Total *Spending* dibuat dari total pembelian MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, dan MntGoldProds sebagai indikator utama dalam memahami pola perilaku pelanggan.
4. Menghapus *Missing Value*
   Data yang tersedia tentunya masih belum bersih dari *missing value*, duplikasi, serta data yang inkonsisten (Simanjuntak & Khaira, 2021). Oleh karena itu, kolom income yang memiliki data kosong akan dihapus karena dapat mengganggu analisis.
5. Menghapus *Outlier* Pada *Income*
   Nilai pencilan pada kolom pendapatan dihapus menggunakan pendekatan quantile 99%, yaitu dengan membuang 1% data tertinggi. Hal ini agar data yang terlalu ekstrem tidak memengaruhi hasil segmentasi pelanggan.
6. Menghapus Data dengan Total *Spending* yang Tidak Signifikan
   Pelanggan dengan total pengeluaran di bawah 20 unit dianggap tidak representatif untuk segmentasi. Oleh karena itu, data tersebut dihapus guna mencegah *noise* pada hasil *clustering* yang pengeluarannya terlalu kecil.
7. Standarisasi Data
   Data *Income*, *TotalSpend*, dan *Recency* distandarisasi menggunakan metode Z-score agar memiliki skala yang sama dan tidak memengaruhi hasil *clustering*.

III.3 Analisis Data

1. Menentukan Jumlah *Cluster* Optimal (*Elbow Method*)
   Metode ini untuk menentukan jumlah *cluster* berdasarkan proporsi nilai perbandingan antara jumlah *cluster*. Metode ini menghitung WCSS (*Within Cluster Sum of Squares*) untuk setiap hasil *cluster* (Muningsih & Kiswati, 2018).
2. Melakukan K-Means *Clustering*



Gambar 2 *Clustering* Dataset (*Elbow Method*)

Berdasarkan hasil *Elbow Method*, terlihat bahwa nilai WCSS menurun tajam hingga K = 4, lalu mulai melandai setelahnya. Oleh karena itu, titik belok atau *elbow point* berada pada k = 4, yang dipilih sebagai jumlah *cluster* optimal untuk proses segmentasi pelanggan.

3. Analisis karakteristik tiap *cluster*
   Setiap *cluster* dilakukan perhitungan rata-rata dari *Income*, *TotalSpend*, dan *Recency* untuk memahami karakteristiknya. Hasil dari perhitungan tersebut akan dibuat label perilaku pelanggannya.
4. Mapping label deskriptif ke tiap *cluster*
   Berdasarkan perhitungan karakteristik setiap *cluster*, pelanggan dilabeli guna mempermudah interpretasi sebagai berikut :

Tabel 1 Klasifikasi Label Karakteristik Pelanggan

Cluster	Avg_ Income	Avg_Total Spend	Avg_ Recency	Interpretasi
1	39.744	204	74.6	Middle Spender - Dormant Pendapatan menengah, tingkat belanja rendah, lama tidak berbelanja
2	70.642	1.152	23.0	High Spender - Active Pendapatan tinggi, tingkat belanja tinggi, aktif berbelanja
3	36.338	171	25.1	Low Spender - Active Pendapatan rendah, tingkat belanja rendah, aktif berbelanja

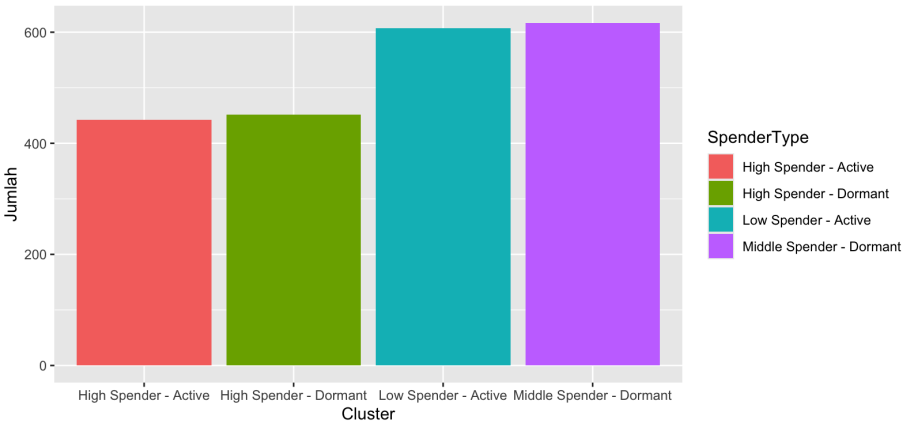
4	72.014	1.278	72.9	High Spender - Dormant Pendapatan tinggi, tingkat belanja tinggi, lama tidak berbelanja

5. Visualisasi jumlah pelanggan per *cluster*
   Visualisasi data dapat membantu dalam penggalian informasi menggunakan statistik dan membangun visualisasi hingga menjadi bentuk yang dapat dimanfaatkan untuk pengembangan bisnis di masa depan (Amrullah 2023). Visualisasi bar chart dibuat untuk menampilkan distribusi jumlah pelanggan dalam setiap kelompok spender hasil *clustering*. Informasi ini berguna untuk menentukan prioritas strategi pemasaran terhadap tiap segmen.
6. Visualisasi sebaran *Income* vs Total *Spending*
   Visualisasi digunakan untuk mengamati pola hubungan antara pendapatan dan pengeluaran pelanggan berdasarkan *cluster* yang terbentuk. Titik-titik dengan warna berbeda memudahkan dalam mengidentifikasi karakteristik kelompok tertentu.
7. Evaluasi menggunakan *Silhouette Coefficient*
   Evaluasi digunakan untuk mengukur kualitas pemisahan antara *cluster* dan digunakan sebagai dasar pengambilan keputusan. Nilai koefisien *silhouette* yang mendekati 1 menunjukkan hasil pengelompokan yang lebih baik, sedangkan nilai yang kurang dari atau sama dengan 0,25 dianggap sebagai hasil pengelompokan yang buruk (Akbar *et al*. 2023).

IV HASIL DAN PEMBAHASAN

Setelah dataset melalui tahap pra-pemrosesan, dataset kemudian dianalisis menggunakan algoritma K-Means untuk melakukan segmentasi pelanggan dengan jumlah *cluster* optimal yaitu sebanyak empat. Hal ini meningkatkan efektivitas pemasaran dan meningkatkan tingkat konversi dari kampanye promosi (Irawan, *et al*. 2025).

IV.1 Analisis Distribusi Jumlah Pelanggan Per *Cluster*

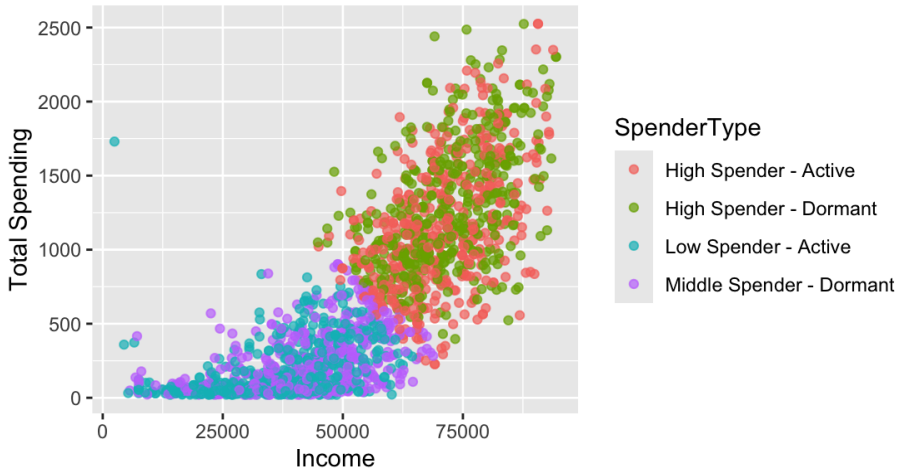


Gambar 3 Jumlah Pelanggan Per *Cluster*

Algoritma K-Means Clustering mampu mengelompokkan data dalam skala besar dan hasil yang mudah diinterpretasikan (Anggraeni, 2025). Gambar berikut memperlihatkan jumlah pelanggan pada masing-masing kelompok hasil segmentasi. Cluster 1 (Middle Spender - Dormant) berjumlah 617 pelanggan, cluster 2 (High Spender - Active) berjumlah 442 pelanggan, pada cluster 3 (Low Spender - Active) berjumlah 607 pelanggan, dan cluster 4 (High Spender - Dormant).

Berdasarkan visualisasi distribusi pelanggan per tipe spender, terlihat bahwa sebagian besar pelanggan berada pada kategori Low Spender. Sementara itu, jumlah pelanggan dalam kategori High Spender - Active relatif sedikit, namun kelompok ini merupakan target utama karena memiliki nilai ekonomi tinggi dan masih aktif bertransaksi. Hal ini menunjukkan bahwa sebagian besar pelanggan dalam dataset ini memiliki daya beli yang rendah meskipun cukup banyak diantara yang masih aktif melakukan pembelian.

IV.2 Hubungan *Income* dengan total *spender* Per *Cluster*

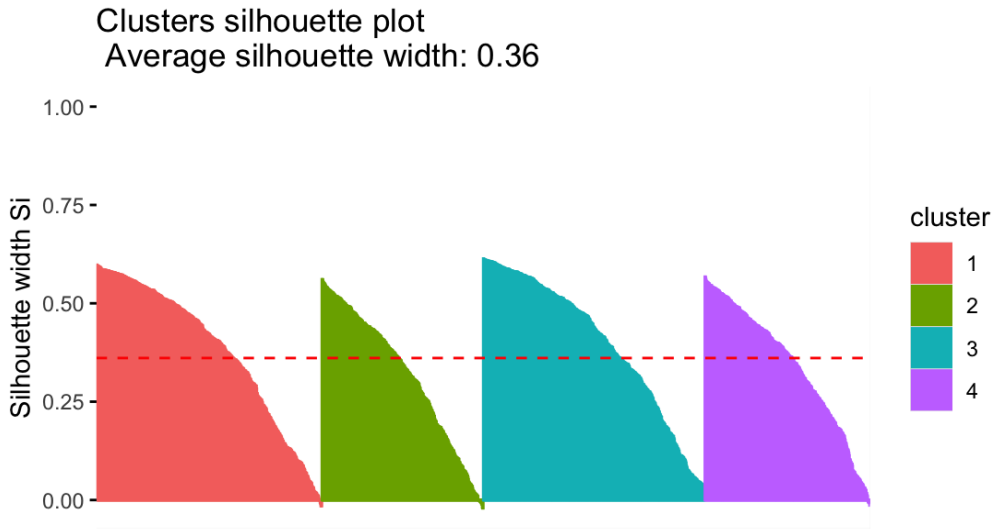


Gambar 4 Perbandingan Income dan Total Spending Per *Cluster*

Scatter plot ini memperlihatkan adanya korelasi positif antara pendapatan dan total pengeluaran pelanggan. Cluster *High Spender* terlihat dominan di wilayah kanan atas grafik, menandakan pelanggan dengan daya beli tinggi. Di sisi lain, cluster *Low Spender* lebih terkonsentrasi di bagian kiri bawah grafik. Seseorang dengan tingkat penghasilan yang tinggi akan menimbulkan perilaku konsumtif yang tinggi, begitu pula sebaliknya (Attan & Natsir, 2023).

Dari grafik ini dapat diartikan semakin tinggi *income*, semakin tinggi *spending* yang terjadi. Konsumen high income menunjukkan pola pengeluaran belanja yang fluktuatif, sering kali melesat dari perencanaan keuangan yang telah dibuat (Japarianto & Sugiharto 2012). Terlihat ada korelasi positif secara umum pelanggan dengan pendapatan tinggi lebih mungkin untuk membelanjakan lebih banyak. Kelompok *Active* memiliki nilai spending yang lebih tinggi dari *Dormant* meskipun income-nya mirip. Ini menunjukkan bahwa recency bisa menjadi faktor penting dalam memaksimalkan nilai pelanggan, tidak hanya dari pendapatan.

IV.3 Evaluasi Kualitas Segmentasi menggunakan *Silhouette Coefficient*



Gambar 5 Visualisasi Silhouette Plot

Evaluasi hasil klasterisasi model K-Means dengan k = 4 menggunakan *Elbow Method* menghasilkan rata-rata *silhouette coefficient* sebesar 0.36. Dalam menentukan tingkat keakuratan nilai *Silhouette Coefficien*t jika nilainya ≥0 maka hasilnya dinyatakan cukup signifikan (Zurfani, *et al.* 2024). Ini menunjukkan bahwa hasil *clustering* berada dalam kategori cukup baik dan signifikan. Nilai ini mengindikasikan bahwa sebagian besar pelanggan sesuai dengan cluster-nya, meskipun ada beberapa data yang mungkin berada di batas antar cluster. Hal ini dapat terjadi karena karakteristik pelanggan dapat serupa antar satu *cluster* dengan yang lain, terutama pada kelompok dengan pendapatan menengah. Secara keseluruhan, hasil ini menunjukkan bahwa

segmentasi pelanggan ke dalam empat kelompok dapat diterima dan relevan untuk pengambilan keputusan bisnis.

V SIMPULAN DAN SARAN

V.1 Simpulan

Dari hasil segmentasi yang ditentukan oleh K-Means menggunakan *elbow method*, jumlah cluster yang optimal dari dataset ini adalah 4 *cluster*. Segmentasi tersebut terdiri dari: Cluster 1 (Middle Spender - Dormant), cluster 2 (High Spender - Active), cluster 3 (Low Spender - Active), dan cluster 4 (High Spender - Dormant).Setiap *cluster* berisi dari perhitungan rata-rata *income*, *TotalSpend*, dan *recency*.

Setiap *cluster* menyajikan informasi yang sangat jelas tentang *customer personality*. Oleh karena itu, dari analisis dataset ini, perusahaan dapat mengatur dan mengembangkan strategi pemasarannya mengikuti karakteristik pelanggannya. Seperti, pelanggan aktif dengan nilai belanja tinggi dapat difokuskan pada program loyalitas mulai dari penawaran eksklusif, *reward* poin, akses tertentu hingga undangan *event* produk. Pelanggan yang memiliki potensial tidak aktif dapat dijangkau melalui kampanye promosi berupa diskon khusus atau bundling produk dengan harga terjangkau.

V.2 Saran

Saran untuk analisis berikutnya adalah dengan mempertimbangkan eksplorasi segmentasi terhadap jumlah cluster untuk mendapatkan *silhouette coefficient* yang lebih optimal. Kondisi tersebut dapat dioptimalkan dengan cara mengganti metode *clustering* yang lebih baik, menggunakan PCA dan t-SNE untuk visualisasi hasil *cluster* untuk mengurangi reduksi dimensi.

DAFTAR PUSTAKA

Akbar T, Tinungki GM, Siswanto. 2023. Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application with Noise Using Silhouette Coefficient Test. *J Mathematics and Its Applications*. 17(3): 1605-1616. DOI: 10.30598/barekengvol17iss3pp1605-1616.

Amrullah DS. 2023. Penerapan Visualisasi Data pada PD. Fokus Bandung. *J Penelitian Mahasiswa Teknik Dan Ilmu Komputer* (JUPITER). 3(1): 44-52.

Anggraeni SR. 2025. Integrasi Data Analytics dalam Kajian Perilaku Pengguna untuk Pengembangan Layanan Informasi. *J Informatika dan Sains Teknologi* (MODEM). 3(2): 1-12. DOI: 10.62951/modem.v3i2.364

Attan MN, Natsir K. 2023. Studi Tentang Faktor-Faktor yang Memengaruhi Consumptive Behavior pada Kolektor Merchandise K-Pop. *J Muara Ilmu Ekonomi dan Bisnis.* 7(1): 187-201. DOI: 10.24912/jmieb.v7i1.22937 187.

Fadhillah MF, Suyoso ALA, Puspitasari I. 2025. Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM). *J Machine Learning and Computer Science* (MALCOM). 5(1): 48-56. DOI: 10.57152/malcom.v5v1.1491.

Irawan D, Wijaya G, Warisaji TT. 2025. Penerapan Algoritma K-Means Clustering untuk Segmentasi Nasabah Bank. *J Teknologi Informasi dan Rekayasa Komputer* (BIOS). 6(1): 47-53. DOI: 10.37148/bios.v6i1.162.

Japarianto E, Sugiharto S. 2012. Pengaruh Shopping Lifestyle dan Fashion Involvement Terhadap Impulsive Buying Behavior Masyarakat High Income Surabaya. *J Manajemen Pemasaran*. 6(1): 32-41. DOI: 10.9744/pemasaran.6.1.32-41.

Muningsih E, Kiswati S. 2018. Sistem aplikasi berbasis optimasi metode elbow untuk penentuan clustering pelanggan. Joutica, 3(1), 117.

Rumapea AYN, Pratiwi D, Sari S. 2024. Analisis Segmentasi Pelanggan Ritel Online Menggunakan K-Means Clustering Berdasarkan Model Recency, Frequency, Monetary (RFM). *J Sains dan Teknologi*. 6(3): 292-299. DOI: 10.55338/saintek.v6i3.4607.

Simanjuntak KP, Khaira U. 2021. Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering. *Machine Learning and Computer Science* (MALCOM). 1(1): 7-16.

Zurfani FA, Sawaluddin, Mardiningsih, Syahputra MR. 2024. Analisis Metode Clustering K-Means pada Zonasi Daerah Terdampak Banjir di Kota Medan dengan Evaluasi Silhouette Coefficient. *J Matematika, Ilmu pengetahuan Alam, Kebumian dan Angkasa.* 2(6): 170-181. DOI: 10.62383/algoritma.v2i6.270.