Segmentasi Konsumen Produk Kosmetik Berbahan Red Fruit Menggunakan Algoritma K-Means Berbasis Machine Learning dan Visualisasi PCA

Miranti Andhita Scantya*[1], Siwi Putri Andini[2], Lawrence Adi Supriyono[3], Brillyan Aditya Saputra[4], Endraw Denny Hermanto[5], Ganjar Pramudya Wijaya[6], Aqila Zhafira Ridoewan[7], Andhika Ariansyah Nugroho[8]

[1-8] Universitas Jakarta Internasional, Jalan Letjen S.Parman No. 1AA
Slipi, Jakarta barat, 081119167363
e-mail: miranti.scantya@uniji.ac.id*[1], siwi.andini@uniji.ac.id[2], lawrence.supriyono@uniji.ac.id[3], brillyan.saputra@uniji.ac.id[4], endraw.hermanto@uniji.ac.id[5], ganjar.wijaya@uniji.ac.id[6], arid0001@student.uniji.ac.id[7], anug0003@student.uniji.ac.id[8]

Abstrak

Pertumbuhan minat terhadap industri kecantikan meningkat secara signifikan setelah jumlah beauty influencer bertambah dan semakin banyak mempromosikan produk kecantikan. Untuk mengikuti tren tersebut, brand perlu menghasilkan ide-ide baru agar dapat menarik perhatian konsumen; untuk melakukan hal ini, perusahaan perlu memahami karakteristik konsumen sehingga mereka mengetahui produk apa yang akan laku di pasaran dan mana yang tidak. Penelitian ini bertujuan untuk mengklasifikasikan karakteristik konsumen terhadap produk kecantikan baru yang dihasilkan dari buah khas Papua, yaitu buah merah. Penelitian ini menggunakan pendekatan machine learning unsupervised, yaitu algoritma K-Means clustering untuk mengklasifikasikan berbagai karakteristik konsumen terhadap produk kecantikan berbahan dasar buah merah. Data diperoleh dari kuesioner yang dibagikan kepada konsumen, yang terdiri dari pertanyaan terkait persepsi konsumen, minat beli, tingkat kesadaran terhadap buah merah sebagai bahan produk, serta demografi konsumen. Dataset diproses awal (pre-processing) dan dinormalisasi menjadi data numerik untuk diolah menggunakan K-Means clustering. Jumlah klaster (karakteristik) optimal ditentukan menggunakan Metode Elbow. Hasil penelitian menunjukkan bahwa jumlah klaster terbaik adalah 3, yang berarti konsumen dalam penelitian ini terbagi menjadi tiga karakteristik. Tiga klaster utama konsumen dengan karakteristik berbeda tersebut adalah (1) Konsumen dengan Minat Tinggi, (2) Konsumen dengan Minat Sedang, dan (3) Konsumen dengan Minat Rendah. Temuan ini memberikan wawasan berharga bagi strategi pemasaran, penentuan posisi produk, dan identifikasi target audiens untuk produk kosmetik berbahan buah merah.

***Kata kunci**—Pra-pemrosesan data, K-Means Clustering, Karakteristik Konsumen, Machine Learning, Kosmetik Buah Merah.*

Abstract

The growth of interest in beauty industry has raised significantly after the number of beauty influencers increased and advertises beauty products. To keep up with the trend, brands need to come up with new idea so they can attract the consumer's attention; to do this the companies need to understand the characteristic of consumer so they know what will sell in the market and what will not. This study aims to classify the characteristic of consumers to the new beauty product produced from Papua's specialized fruit, red fruit. This study uses unsupervised machine learning approach, the K-Means clustering algorithm to classify variety of consumer characteristics toward red fruit's beauty product. Data were obtained from a questionnaire distributed to consumers, consisting of questions related to consumer perception, purchasing interest, the awareness of red fruit as the ingredients, and consumer demographics. The dataset was pre-processed and normalized into numeric data to be processed with K-Means clustering. The optimal number of clusters (Characters) was determined using the Elbow Method. The results showed the best number of

clusters are 3, it means in this study consumers are divided into three characteristics. Three main consumer clusters with different characteristics are (1) Consumers with High Interest, (2) Consumers with moderate interest, (3) Consumers with Low Interest. These findings provide valuable insights for marketing strategies, product positioning, and target audience identification for red fruit cosmetic products.

Keywords: Data pre-processing, K-Means Clustering, Consumer Characteristics, Machine Learning, Red fruit Cosmetic.

1. PENDAHULUAN

Perkembangan teknologi informasi telah membawa perubahan signifikan dalam cara industri mengambil keputusan strategis, khususnya melalui pemanfaatan data dalam skala besar (data-driven decision making). Industri kecantikan sebagai salah satu sektor dengan pertumbuhan cepat menghasilkan volume data konsumen yang semakin kompleks, mulai dari preferensi produk, persepsi terhadap bahan baku, hingga respons terhadap inovasi produk baru. Kondisi ini menuntut penerapan pendekatan analitik berbasis teknologi informasi untuk memperoleh wawasan yang akurat dan terukur.

Di era digital, perilaku konsumen tidak lagi dapat dianalisis secara konvensional. Interaksi melalui media sosial, kampanye digital oleh beauty influencer, serta tren pemasaran berbasis konten menghasilkan pola data yang heterogen dan multidimensional. Oleh karena itu, metode machine learning dan data mining menjadi pendekatan yang relevan dalam mengidentifikasi pola tersembunyi dan segmentasi konsumen secara objektif, terutama dalam mendukung keputusan bisnis berbasis data.

Salah satu tantangan dalam pengembangan produk berbasis inovasi adalah rendahnya tingkat kepastian pasar terhadap bahan baru yang belum dikenal luas oleh konsumen. Buah merah (Pandanus conoideus), sebagai bahan alami lokal dengan kandungan antioksidan dan pigmen alami yang tinggi, memiliki potensi besar untuk dikembangkan menjadi bahan baku produk kecantikan. Namun, keterbatasan pemanfaatan buah merah dalam produk kosmetik menyebabkan minimnya data empiris terkait pengetahuan, persepsi, dan minat konsumen terhadap bahan tersebut. Kondisi ini memerlukan pendekatan analitik yang sistematis untuk mengolah data konsumen dalam jumlah besar guna mendukung pengambilan keputusan produksi dan pemasaran.

Pengolahan data konsumen berskala besar memerlukan tahapan komputasional yang mencakup pembersihan data (data preprocessing), transformasi data, dan analisis pola menggunakan algoritma machine learning. Bahasa pemrograman Python menjadi salah satu alat utama dalam proses tersebut karena dukungannya terhadap pustaka analisis data dan pembelajaran mesin. Dengan pendekatan ini, data hasil survei tidak hanya digunakan sebagai informasi deskriptif, tetapi diolah menjadi pengetahuan (knowledge) yang memiliki nilai strategis.

Pada penelitian ini diterapkan algoritma K-Means clustering sebagai salah satu teknik unsupervised learning untuk melakukan segmentasi konsumen berdasarkan respons mereka terhadap produk kecantikan berbahan buah merah. Algoritma ini dipilih karena kemampuannya dalam mengelompokkan data multidimensional secara efisien dan banyak digunakan dalam analisis perilaku konsumen. Hasil segmentasi konsumen diharapkan mampu memberikan gambaran yang lebih terstruktur mengenai kelompok konsumen potensial, tingkat penerimaan pasar, serta karakteristik masing-masing segmen. Dengan mengintegrasikan pendekatan machine learning dan analisis data konsumen, penelitian ini memberikan kontribusi pada bidang teknologi informasi, khususnya dalam penerapan data mining untuk mendukung keputusan bisnis. Selain itu, penelitian ini juga menunjukkan bagaimana teknologi informasi dapat berperan dalam mendorong inovasi produk berbasis sumber daya lokal melalui analisis data yang objektif dan terukur.

2. METODE PENELITIAN

2.1 *Machine Learning* untuk analitik bisnis

Di era teknologi saat ini, di mana analisis data menjadi penting dan jumlah data yang semakin besar, pegiat bisnis harus mencari metode untuk mengolah data-data untuk mendapatkan informasi yang dapat

digunakan untuk membantu berjalannya bisnis mereka [16]. Karena data-data yang semakin besar dan banyak, *machine learning* dapat digunakan untuk mempermudah proses analisis sehingga membantu industri bisnis dalam mendapatkan informasi penting dan Solusi yang penting dalam menjalankan bisnisnya [2]. Salah satu Business Intelligence (BI) yang dapat dilakukan adalah Customer Relationship Management (CRM). CRM adalah strategi bisnis untuk identifikasi interaksi Perusahaan dengan pelanggan [17]. Model *Machine learning* dapat digunakan untuk membuat Keputusan dengan cara mempelajari data yang diberikan dan menganalisis pola yang terlihat di dalam data tersebut. *Machine learning* sudah banyak digunakan untuk mengambil Keputusan pada bidang bisnis, contohnya pada segmen marketing, *machine learning* digunakan untuk segmentasi pelanggan dan analisis sentimen pelanggan[23]. Hal lainnya yang dapat dilakukan *machine learning* adalah membuat Keputusan dengan akurasi yang lebih baik, analisis prediktif yaitu memprediksi tren di masa depan berdasarkan data-data terdahulu misalnya menganalisis data penjualan dan factor eksternal untuk memprediksi produk apa yang paling diminati [4]. *Machine Learning* dengan Teknik pembelajaran tanpa pengawasan *(unsupervised learning*) seperti *clustering* dapat digunakan untuk membantu melakukan hal-hal tersebut.

2.2 Algoritma *K-Means Clustering*

Clustering adalah salah satu Teknik pada data mining untuk menemukan pola dalam data yang menyediakan informasi dengan mengidentifikasi klasifikasi objek yang mempunyai kemiripan. Clustering dapat digunakan untuk mendapatkan kelompok-kelompok (Kelas) dari objek-objek yang mempunyai karakteristik yang sama [3]. Salah satu metode *clustering* adalah *K-Means. K-Means* adalah salah satu algoritma yang sering dipakai dalam Teknik clustering. K-means dipakai untuk analisis di bidang sains,teknologi dan bisnis. Tujuan dari k-means adalah membagi data-data kedalam beberapa cluster (k) . *K-means* adalah metode clustering non-hirarki untuk mengelompokan data ke dalam beberapa cluster (kelompok). Pengelompokan pada *K-Means* dilakukan dengan partisi yang didasarkan pada sebuah titik pusat. Setiap cluster diwakili oleh titik pusat atau centroid di mana jarak data-data dengan centroid akan dihitung [5]. Algoritma dari *K-means* dijelaskan dengan Langkah-langkah sebagai berikut [6]:

1. Tentukan jumlah cluster (k) pada data set sebagai centroid. Jumlah cluster dapat ditentukan sesuai kebutuhan analisis atau memakai metode penentuan jumlah cluster yang optimal seperti *elbow method*.
2. Untuk Langkah awal, Centroid diinisialisasi secara acak.
3. Setelah centroid ditentukan, hitung jarak data-data dengan centroid menggunakan rumus Euclidean sebagai berikut:

$$D_{(i,j)} = \sqrt{\left(X_{1i} - X_{1j}\right)^2 + \left(X_{2i} - X_{2j}\right)^2 + \cdots + \left(X_{ki} - X_{kj}\right)^2} \quad (1)$$

   Dimana $D_{(i,j)}$adalah jarak antara data i ke centroid j, X_{ki} adalah data ke-i untuk cluster k, X_{kj} adalah centroid untuk cluster k
4. Data-data dikelompokan ke dalam cluster dengan jarak centroid dan data terdekat.
5. Centroid baru akan dilakukan pembaruan setelah data-data dikelompokkan kedalam cluster-cluster terdekat.
6. Pembaruan centroid akan berhenti jika memenuhi syarat yang telah ditentukan atau Ketika centroid sudah tidak berubah lagi.

Dalam penelitian ini penggunaan K-means dapat membantu dalam mengklasifikasikan pelanggan dari data survey yang sudah dikumpulkan kedalam beberapa kelas atau cluster.

2.3 *Data Collection*

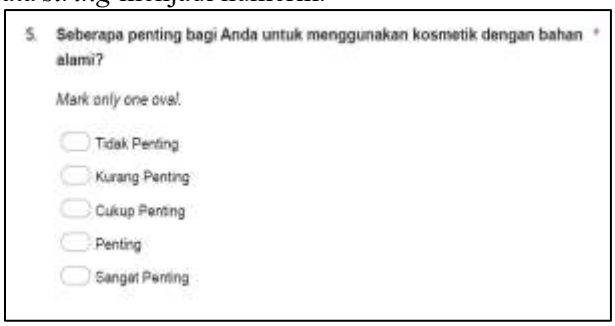
Data collection atau pengumpulan data adalah tahap awal dalam analisis data. Pada tahap ini permasalahan atau tujuan analisis sudah ditentukan untuk mengetahui data apa yang harus dikumpulkan,

dan pada tahap ini tipe data sudah ditentukan. Proses ini dapat mengurangi masalah dan error akibat kurangnya integrasi data [18]. Tahap pengumpulan data dilakukan dengan kuisoner dengan menggunakan google form. Pertanyaan-pertanyaan di dalam kuisoner berjumlah 15 pertanyaan dengan 14 pertanyaan diukur dengan skala likert seperti "Sangat tidak tertarik", "tidak tertarik", "tertarik.", "Sangat tertarik". Pertanyaan-pertanyaan di dalam kuisoner berkaitan dengan pandangan terhadap produk, respon responden terhadap produk (setelah membaca deskripsi singkat tentang *red fruit* di dalam kuisoner), pengetahuan terhadap bahan baku (*red fruit*), dan data skala umur, jenis kelamin, dan pekerjaan. Kuisioner disebarkan selama seminggu dan mendapatkan 3.678 responden.

2.4 *Data Preprocessing*

Pre-processing adalah tahapan awal yang penting dalam pengolahan data untuk menyiapkan dan membersihkan data mentah yang diambil dari pengumpulan data menggunakan kuisoner agar siap digunakan untuk proses selanjutnya yaitu proses *clustering* dengan K-Means [9].

1. *Data Cleaning*: Tahap *pre-processing* pertama yang dilakukan adalah dengan mendeteksi apakah data memiliki noise, data hilang, atau gangguan pada data lainnya. Jika terdeteksi data mempunyai gangguan-gangguan tersebut maka harus dilakukan *data cleaning* yang tepat untuk memperbaiki *noise*, mengisi data hilang, dsb [19].
2. *Data transformation*: transformasi data dilakukan jika data mentah yang ada harus diubah untuk disesuaikan dengan kebutuhan. Pada penelitian ini, data mentah yang dikumpulkan berupa 14 pertanyaan dengan skala likert yang masih berbentuk string dan 1 pertanyaan dengan pilihan jawaban yang dapat dipilih lebih dari satu. Sehingga untuk 14 Pertanyaan dilakukan *encoding* data untuk mengubah tipe data *string* menjadi numerik.



Gambar 1 Contoh pernyataan dalam Kuisioner

Gambar 1 adalah salah satu pertanyaan yang harus direspon oleh responden di mana responden diberikan pertanyaan terkait ketertarikan dalam penggunaan kosmetik berbahan alami. Jawaban yang disediakan berupa sekala dari tidak penting hingga sangat penting yang mempunyai tipe *string*. Untuk mempermudah pemrosesan data maka data yang dihasilkan diubah kedalam tipe numerik.

Q1	Q2	Q3	Q4	Q5
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.)	Sering (6–8 kali)	Cukup Penting
<18	Perempuan	Media sosial (TikTok, Instagram, dll.), Influenc	Jarang (1-2 kali)	Sangat Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Sangat Penting
25-34	Laki-laki	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Kurang Penting
<18	Perempuan	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Cukup Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.), Influenc	Jarang (1-2 kali)	Sangat Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.), Influenc	Jarang (1-2 kali)	Sangat Penting
25-34	Perempuan	Media sosial (TikTok, Instagram, dll.)	Cukup sering (3–5 kali)	Sangat Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.), Influenc	Jarang (1-2 kali)	Penting
25-34	Perempuan	Media sosial (TikTok, Instagram, dll.)	Tidak Pernah	Cukup Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.)	Cukup sering (3–5 kali)	Cukup Penting
<18	Perempuan	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Sangat Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Sangat Penting
<18	Perempuan	Media sosial (TikTok, Instagram, dll.)	Cukup sering (3–5 kali)	Kurang Penting
25-34	Perempuan	Media sosial (TikTok, Instagram, dll.)	Sering (6–8 kali)	Sangat Penting
<18	Perempuan	Media sosial (TikTok, Instagram, dll.), Teman/K	Cukup sering (3–5 kali)	Penting
18-24	Perempuan	Media sosial (TikTok, Instagram, dll.)	Jarang (1-2 kali)	Cukup Penting

Gambar 2 Hasil Kuisioner

Gambar 2 adalah hasil kuisioner 5 pertanyaan dari 15 pertanyaan yang diajukan kepada responden. Pertanyaan-pertanyaan diganti dengan variabel Q1-Q16 untuk memperkecil ukuran kolom. Untuk pertanyaan Q1,Q2,Q3 dan Q5 respon dengan tipe string akan diubah kedalam bentuk skala likert berbentuk angka. Sehingga hasilnya akan seperti yang ditunjukan dalam Gambar 3

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
2	0	4.6	4	3	2	3	3	3	3	3	3	3	3	3
1	0	4.1	2	5	1	5	4	4	4	1	5	5	5	5
2	0	4.6	2	5	1	3	5	4	4	2	4	4	4	4
3	1	4.6	2	2	4	5	5	4	4	1	3	3	4	4
1	0	4.6	2	3	1	2	3	3	3	1	3	2	3	3
2	0	4.8	2	5	1	4	4	4	3	2	4	5	5	5
2	0	4.8	2	5	1	4	4	4	3	2	4	5	5	5
3	0	4.6	3	5	1	3	4	3	4	2	5	5	5	5
2	0	4.1	2	4	3	3	5	5	4	2	3	3	5	5
3	0	4.6	1	3	4	4	4	3	4	1	3	2	5	2
2	0	4.6	3	3	2	3	3	3	3	1	3	3	4	3
1	0	4.6	2	5	1	3	4	5	4	1	5	2	5	3
2	0	4.6	2	5	3	3	4	3	4	2	3	3	3	4
1	0	4.6	3	2	3	4	4	5	5	5	3	4	1	1
3	0	4.6	4	5	2	4	5	5	5	2	4	4	4	3
1	0	2.8	3	4	4	4	4	4	4	2	5	4	5	4
2	0	4.6	2	3	1	3	5	3	4	3	3	4	5	5
2	0	4.6	3	3	3	3	3	3	4	1	3	3	3	3

Gambar 3 *Encoding* Hasil Kuisioner

Pertanyaan Q3 membutuhkan lebih dari satu jawaban sehingga untuk mengubah tipe respon menjadi skala likert diterapkan metode lain yaitu *House of Quality. House of Quality* adalah sebuah matriks untuk mengkonversi kebutuhan pelanggan kedalam Solusi. Ada 5 jawaban yang dapat dipilih untuk pertanyaan Q3 yaitu *media social* (TikTok, Instagram,dll.), Teman/keluarga, *Influencer/beauty*, Iklan, dan Toko kosmetik/*Beauty Advisor (BA)*. Sebelum proses konversi jawaban yang dipilih responden diubah kedalam angka biner, 1 untuk jawaban yang dipilih dan 0 untuk jawaban yang tidak dipilih. Setelah itu setiap jawaban dijumlahkan untuk dihitung frekuensinya, untuk melihat seberapa sering jawaban dipilih dilihat dari bobotnya. Kemudian tentukan nilai hubungan antara setiap jawaban (sumber informasi) dengan persyaratan teknis seperti jangkauan informasi, kredibilitas sumber, pengaruh terhadap kepercayaan, dan visual/testimoni. Setelah mendapatkan nilai hubungan tersebut dilakukan perhitungan prioritas produk dengan mengalikan bobot dan nilai hubungan untuk mendapatkan skala likert untuk setiap jawaban.

2.6 Clustering with K-Means

Cluster adalah istilah untuk sekelompok object yang mempunyai kemiripan, Cluster dapat digunakan untuk segmentasi pelanggan atau calon pelanggan untuk analisis data [11]. Pada penelitian ini segmentasi calon pelanggan dilakukan untuk mengetahui apakah ada ketertarikan pelanggan terhadap kosmetik dengan bahan buah merah atau *red fruit.* Segmentasi calon pelanggan dilakukan menggunakan K-means untuk mengelompokan calon pelanggan berdasarkan ketertarikan, dan pengetahuan pelanggan terhadap produk kosmetik terutama kosmetik dengna bahan utam red Fruit [8]. Setelah data siap untuk digunakan maka pemrosesan data dilanjutkan kedalam tahap c*lustering*. Banyaknya titik pusat atau centroid dapat ditentukan sesuai dengan kebutuhan klasifikasi, seperti berapa klasifikasi yang digunakan untuk membagi data. Selain itu, ada dua metode yang dapat digunakan untuk menentukan jumlah k yaitu metode *elbow* dan metode skor *Silhouette*.

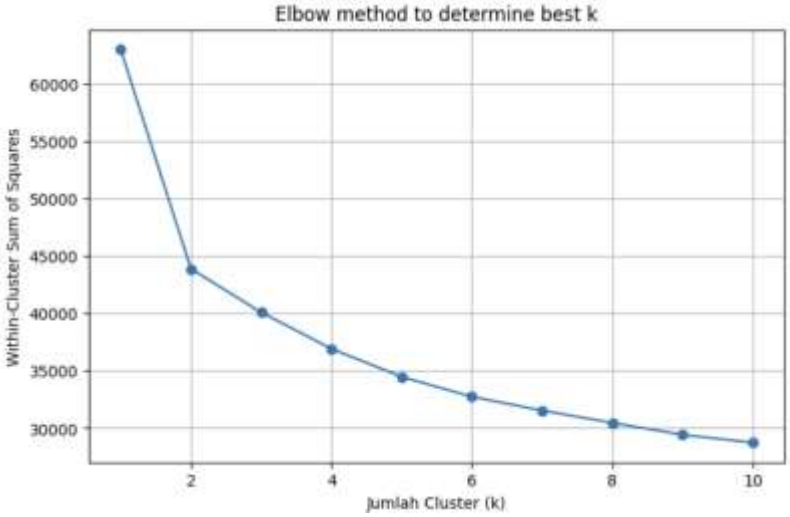
a. Metode elbow: Teknik dengan menggunakan representasi visual untuk menentukan jumlah *cluster* yang optimal dengan menghitung distorsi atau jarak setiap titik data dan centroid (SSE). Ketika jumlah k bertambah nilai SSE akan menurun, dan di jumlah k tertentu nilai SSE akan menurun drastik dari k sebelumnya sehingga membentuk siku. Hal ini menandakan bahwa k tersebut adalah jumlah k terbaik [7].
b. Metode Skor *Silhouette*: Teknik ini mengukur seberapa dekat data-data ke centroid terdekat. Teknik ini juga mengukur jarak antar *cluster*. Semakin tinggi nilai semakin optimal pengelompokan yang dilakukan [20].

Dalam penelitian ini metode *elbow* dan *Silhouette* digunakan untuk menentukan jumlah k yang terbaik untuk dianilisis perbandingan hasil dari kedua metode tersebut.

3. HASIL DAN PEMBAHASAN

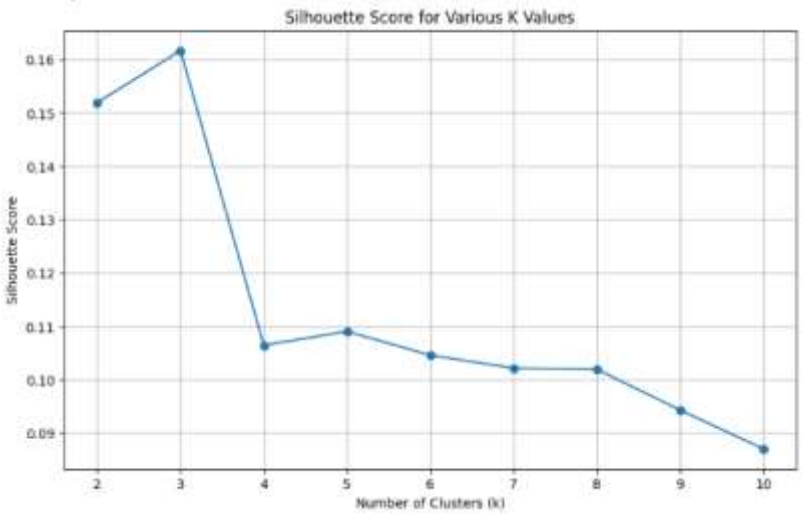
3.1 Pemilihan jumlah *Cluster*

Untuk mengklasifikasikan data ke dalam beberapa *cluster* (k), jumlah k yang optimal harus ditentukan. Jumlah k biasanya adalah sebuah nilai acak yang ditentukan sebelum proses *clustering* menggunakan *k-means*, setelah itu proses klasifikasi data dilakukan dengan menghitung jarak antara data-data dan k titik yang jumlahnya sesuai dengan jumlah yang sudah ditentukan. Dalam penelitian ini, kami menggunakan metode *elbow* dan *silhouette* untuk menentukan berapa banyak k yang paling sesuai untuk penelitian.



Gambar 4 Visualisasi Metode Elbow

Metode elbow digunakan untuk menentukan jumlah optimal dari cluster dengan cara mengidentifikasi titik data Dimana peningkatan jumlah cluster, menurunkan jumlah varians dari data [10]. data Dari visualisasi metode elbow dalam gambar 4 terlihat bahwa nilai SSE menurun di jumlah *cluster* k=2, penurunan ini membentuk siku sehingga berdasarkan pengertiannya, k optimal menurut metode *elbow* adalah titik k yang membentuk siku yaitu k=2.



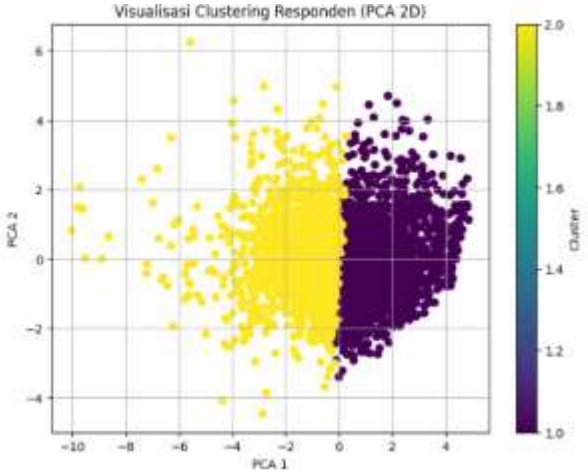
Gambar 5 Visualisasi metode *Silhouette*

Metode *Silhouette* adalah metode untuk menghitung k optimal dengan mengambil nilai k yang mempunyai Skor tertinggi. Pada gambar 5 terlihat bahwa k dengan skort tertinggi adalah k=3. Karena dua metode yang digunakan memberikan hasil yang berbeda maka pada penelitian ini akan dilakukan klasifikasi

data kedalam 2 dan 3 kelas (*cluster*) dan akan dianalisis hasil mana yang mempunyai klasifikasi terbaik dilihat dari visualisasinya.

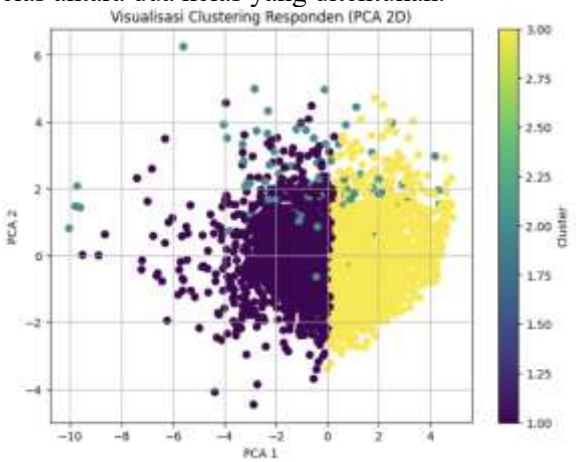
3.2 Hasil *Clustering*

Hasil *clustering* divisualisasikan dengan menggunakan metode PCA. PCA membantu klasifikasi dengan cara membagi 16 variabel dari 16 pertanyaan yang diajukan menjadi dua variable sehingga dapat divisualisasikan dengan grafik dua dimensi.



Gambar 6 Visualisasi Clustering k=2

Berdasarkan metode *elbow* jumlah k yang paling optimal adalah k=2, sehingga dilakukan *clustering* menggunakan k-means dengan membagi data kedalam dua kelas (*cluster*). Kelas 1 ditunjukan dengan titik berwarna ungu tua, kelas ini adalah kelas 1 (*cluster* 1) yang merepresentasikan responden yang TIDAK TERTARIK dengan produk berbahan *red fruit*. Sementara kelas 2 ditunjukan dengan titik kuning yang merepresentasikan responden yang TERTARIK dengan produk kecantikan berbahan *red fruit*. Terlihat dari gambar 6 klasifikasi terlihat jelas antara dua kelas yang ditentukan.



Gambar 7 Visualisasi Clustering k=3

Sementara itu, hasil dari metode *silhouette* menyatakan bahwa klasifikasi kedalam tiga akelas adalah jumlah *cluster* yang paling optimal. Dari gambar 4.4 Kelas 1 atau kelompok responden yang TIDAK TERTARIK direpresentasikan dengan titik berwarna ungu tua. *Cluster* 2 yang merepresentasikan kelompok responden yang mempunyai reaksi normal terhadap produk dengan bahan utama *red fruit*, ditunjukan dengan titik berwarna hijau. Sementara kelas 3 yaitu responden yang TERTARIK ditunjukan dengan titik kuning. Visualisasi pada gambar 7 menunjukan bahwa *Cluster* 1 (ungu) ada di sisi kiri PCA 1, *Cluster* 2

(hijau) berada di Tengah di antara dua kelas (*Cluster* 1 dan *Cluster* 3), dan *Cluster* 3 (Kuning) berada di sisi kanan PCA 1. Pengelompokan ini menunjukan PCA 1 adalah komponen yang sangat mempengaruhi pembagian *cluster*.

```
--- Step 7: Cluster Summary ---
                Q1   Q2        Q3        Q4        Q5        Q6        Q7  \
Cluster
1         2.002593  0.0  4.332227  2.252075  3.718880  1.644710  3.047718
2         2.242718  1.0  4.058576  2.242718  3.737864  1.970874  3.407767
3         2.092896  0.0  4.370998  2.423194  4.394657  1.983607  3.867638

                Q8        Q9       Q10       Q11       Q12       Q13       Q14  \
Cluster
1         3.362552  3.206950  3.525415  1.875000  3.245851  3.252593  3.754149
2         3.747573  3.592233  3.737864  2.271845  3.349515  3.466019  3.825243
3         4.238009  4.131755  4.296903  1.880996  4.276867  4.251366  4.658774

                Q15      PCA1      PCA2
Cluster
1         3.394191 -1.411699 -0.006066
2         3.631068 -0.762760  1.935973
3         4.343655  1.700254 -0.113970
```

Gambar 8 hasil perhitungan PCA

Gambar 8 menunjukan skor rata-rata untuk setiap pertanyaan di dalam 3 *cluster.* Data dari Q1-Q15 pada *Cluster* 1 mempunyai rata-rata yang rendah, seperti pada Q5 dengan nilai rata-rata 3.72, dan Q6 = 1.64. hasil PCA1 untuk *cluster*1 adalah –1.41 nilai ini merepresentasikan grup paling kiri (bernilai minus) yang kemungkinan besar merepresentasikan responden yang tidak tertarik kedalam Produk kecantikan *red fruit*. Nilai rata-rata setiap pertanyaan di Cluster 2 mempunyai nilai yang berada ditengah-tengah nilai rata-rata *cluster* 1 dan *cluster* 3. Nilai PCA 1 untuk cluster ini adalah sebesar 0.76 sehingga letak titik-titik (data) untuk *cluster* 2 ada di Tengah grafik. Responden dalam kelompok ini memiliki tingkat ketertarikan yang sedang atau rata-rata, kemungkinan besar memiliki minat ringan terhadap Kosmetik *red fruit*. Sementara untuk Cluster 3 mempunyai nilai rata-rata yang paling tinggi untuk pertanyaan-pertanyaan Q1 hingga Q15 dibandingkan *cluster* lain. Nilai PCA 1 Untuk *cluster* ini adalah 1.70 sehingga titik-titik data berada di sisi kanan grafik. Kelompok ini cenderung memberikan tanggapan positif yang lebih kuat dari kelompok lainnya dan kemungkinan besar Sangat Tertarik pada produk kecantikan *Red Fruit*.

4. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan machine learning untuk melakukan segmentasi konsumen terhadap produk kosmetik berbahan dasar buah merah (red fruit) berdasarkan data kuesioner yang telah melalui proses data preprocessing dan encoding numerik. Algoritma K-Means digunakan sebagai metode utama dalam pengelompokan konsumen, sementara metode Elbow dan Silhouette Coefficient digunakan untuk menentukan jumlah cluster yang optimal.

Hasil evaluasi menunjukkan bahwa metode Elbow menghasilkan jumlah cluster optimal sebesar k = 2, sedangkan metode Silhouette menunjukkan hasil optimal pada k = 3. Perbandingan kedua skenario tersebut memperlihatkan bahwa segmentasi dengan tiga cluster (k = 3) memberikan hasil yang lebih informatif dan mudah diinterpretasikan, karena mampu merepresentasikan variasi tingkat ketertarikan konsumen secara lebih rinci, yaitu konsumen dengan minat rendah, sedang, dan tinggi terhadap produk kosmetik berbahan dasar buah merah.

Visualisasi menggunakan Principal Component Analysis (PCA) menunjukkan bahwa komponen utama pertama memiliki peran dominan dalam memisahkan kelompok konsumen berdasarkan tingkat ketertarikan mereka. Hal ini mengindikasikan bahwa karakteristik responden yang terkait dengan persepsi, minat, dan penerimaan terhadap produk memiliki kontribusi signifikan dalam proses segmentasi. Dari visualisasi PCA terlihat bahwa k=2 membagi konsumen menjadi dua kelompok Tertarik dan Tidak Tertarik, sedangkan k=3 membagi responden kedalam 3 kelompok yaitu:

1. Konsumen dengan ketertarikan rendah terhadap produk kecantikan *red fruit* (Tidak Tertarik).

2. Konsumen dengan minat sedang terhadap produk kecantikan *red fruit.*
3. Konsumen dengan ketertarikan tinggi terhadap produk kecantikan *red fruit* (Tertarik)

Secara keseluruhan, penelitian ini membuktikan bahwa penerapan teknik data mining dan machine learning dapat menjadi alat bantu yang efektif dalam pengambilan keputusan bisnis, khususnya dalam memahami perilaku dan preferensi konsumen. Hasil segmentasi yang diperoleh dapat dimanfaatkan oleh produsen kosmetik sebagai dasar dalam menentukan strategi pemasaran, pengembangan produk, serta edukasi konsumen secara lebih tepat sasaran. Selain itu, penelitian ini juga memberikan kontribusi pada pemanfaatan teknologi informasi dalam analisis data konsumen berbasis survei, terutama untuk produk inovatif berbahan alami yang masih memiliki tingkat pengetahuan konsumen yang relatif rendah.

DAFTAR PUSTAKA

[1] G. S. Mozes, K. P. A. Nugroho, and D. Puspita, "Pemanfaatan buah merah (*Pandanus conoideus*) sebagai bahan baku dalam pembuatan saus dan potensinya sebagai bahan tambahan pangan," *Prosiding Seminar Nasional Mahasiswa Unimus*, vol. 1, pp. 218-226, 2018.

[2] T. V. Iyelolu and P. O. Paul, "Implementing machine learning models in business analytics: Challenges, solutions, and impact on decision-making," *World Journal of Advanced Research and Reviews*, vol. 22, no. 3, pp. 1906-1916, 2024, doi: 10.30574/wjarr.2024.22.3.1959.

[3] S. Setyaningtyas, B. I. Nugroho, and Z. Arif, "Tinjauan pustaka sistematis pada data mining: Studi kasus algoritma K-Means clustering," Jurnal Teknoif Teknik Informatika Institut Teknologi Padang, vol. 10, no. 2, pp. 52-61, Oct. 2022, doi: 10.21063/jtif.2022.V10.2.52-61.

[4] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas, "Prescriptive analytics: Literature review and research challenges," *International Journal of Information Management*, vol. 50, pp. 57-70, 2020.

[5] W. Sudrajat, I. Cholid, and J. Petrus, "Penerapan algoritma K-Means clustering untuk pengelompokan UMKM menggunakan RapidMiner," 2022.

[6] T. Tendean and W. Purba, "Analisis cluster provinsi Indonesia berdasarkan produksi bahan pangan menggunakan algoritma K-Means," *Jurnal Sains dan Teknologi*, vol. 1, no. 2, pp. 5-11, Mar. 2020.

[7] E. Erzenemine, "Elbow method: Finding the optimal number of clusters in K-means," *Medium*, 2025. [Online]. Available: https://medium.com/@erzeneminegul/elbow-method-finding-the-optimal-number-of-clusters-in-k-means-9bc652a403ef

[8] P. Utomo, M. Tampi, D. Claudia, and J. Haikal, "K-Means clustering segmentation based on consumer interest using SPSS program at XYZ Indonesia customers," *Dinasti International Journal of Digital Business Management*, vol. 4, no. 3, pp. 451-460, 2023.

[9] D. A. Awaliyah, B. Prasetiyo, R. Muzayanah, and A. D. Lestari, "Optimising customer segmentation in online retail transactions through the implementation of the K-Means clustering algorithm," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 135-146, 2024.

[10] N. Kumar, "Intelligent customer segmentation: Unveiling consumer patterns with machine learning," *Journal of Umm Al-Qura University for Engineering and Architecture*, 2025. [Online]. Available: https://doi.org/10.1007/s43995-025-00180-7

[11] P. Anitha, "RFM model for customer purchase behavior segmentation using K-Means," *International Journal of Research in Engineering and Science*, 2022.

[12] B. N. Yulisasih, "K-Means clustering method for customer segmentation: FMCG purchasing potential," *Jurnal Teknologi Informasi*, 2024.
[13] N. Migau, "The antioxidant activity of red fruit extract (*Pandanus conoideus* L.) from Nabire Papua," *Indonesian Biomedical Journal*, vol. 5, no. 1, pp. 45-52, 2023.
[14] Heriyanto, I. A. Gunawan, R. Fujii, T. Maoka, Y. Shioi, K. M. B. Kameubun, L. Limantara, and T. H. P. Brotosudarmo, "Carotenoid composition in buah merah (*Pandanus conoideus* Lam.), an indigenous red fruit of the Papua Islands," *Journal of Food Composition and Analysis*, vol. 96, 2021, doi: 10.1016/j.jfca.2020.103722.
[15] C. H. Dumaria, A. A. G. P. Wiraguna, and W. Pangkahila, "Krim ekstrak buah merah (*Pandanus conoideus*) 10% sama efektifnya dengan krim hidrokuinon 4% dalam mencegah peningkatan jumlah melanin kulit marmut (*Cavia porcellus*) yang dipapar sinar ultraviolet B," *Jurnal Biomedik (JBM)*, vol. 10, no. 2, 2018.
[16] G. Bohanec, M. Robnik-Šikonja, and M. Kljajić Borštnar, "Decision-making framework with machine learning and data mining for business intelligence," *International Journal of Information Management*, vol. 43, pp. 364-379, 2018.
[17] X. Chen, Z. Xu, and Y. Zhang, "Customer segmentation using K-means clustering and decision tree models," *Procedia Computer Science*, vol. 122, pp. 240-247, 2017.
[18] H. Taherdoost, "Data collection methods and tools for research: A step-by-step guide to choose data collection technique for academic and business research projects," *International Journal of Academic Research in Management*, 2021.
[19] A. A. A. Daniswara and I. K. D. Nuryana, "Data preprocessing pola pada penilaian mahasiswa program profesi guru," *JINACS: Journal of Informatics and Computer Science*, vol. 5, no. 1, pp. 97-100, 2023.
[20] K. Kertanah, W. P. Nurmayanti, S. R. Aini, L. M. Amrullah, and M. Sya'roni, "Comparison of algorithms K-Means and DBSCAN for clustering student cognitive learning outcomes in physics subject," *Kappa Journal*, vol. 7, no. 1, pp. 251-255, 2023.
[21] A. S. Aldila and L. A. Supriyono, "Predictive modeling of Covid-19 spread with machine learning: A focus on decision tree accuracy," Jurnal Teknik Informatika Unika St. Thomas (JTIUST), vol. 9, no. 2, pp. 223-230, Dec. 2024.