Jurnal JPILKOM, 2 (3) (2024), ISSN: 3025-6887 (Online)
Jurnal JPILKOM (Jurnal Penelitian Ilmu Komputer)
Journal homepage: https://jpilkom.org

Penerapan K-Means untuk Meningkatkan Strategi Pemasaran Melalui Segmentasi Rumah Tangga yang Efektif

M.Dimas Nurwanda[1], Andina Nur Liviasari[2], Dhafin Dhias Pambudi[3]
[1, 2, 3] Fakultas Informatika/Teknik Informatika, Institut Teknologi Telkom Purwokerto, Indonesia

E-Mail : [1]nurwanda@student.telkomuniversity.ac.id, [2]andninayy@student.telkomuniversity.ac.id, [3]dhihaspam@student.telkomuniversity.ac.id

Article Info

Article history:
Received Sept 15, 2024
Revised Sep 20, 2024
Accepted Sep 25, 2024

Kata Kunci:

Data Mining
K-Means
Klastering
Strategi Pemasaran
Segmentasi Pelanggan

Keywords:
Data Mining
K-Means
Clustering
Marketing
Customer Segmentation

ABSTRAK

Di era digital saat ini, strategi pemasaran memiliki peran penting dalam persaingan bisnis. Penelitian ini bertujuan untuk meningkatkan strategi pemasaran melalui segmentasi rumah tangga yang efektif dengan menggunakan algoritma K-Means. Algoritma ini digunakan untuk mengelompokkan pelanggan berdasarkan karakteristik seperti pendapatan, pendidikan, dan pekerjaan. Proses penelitian meliputi tahapan pemilihan data, preprocessing, transformasi data, penerapan algoritma K-Means, dan evaluasi hasil klasterisasi. Hasil penelitian menunjukkan lima klaster pelanggan yang berbeda dengan karakteristik unik masing-masing. Klasterisasi ini membantu perusahaan dalam merancang strategi pemasaran yang lebih spesifik dan tepat sasaran, meningkatkan kepuasan pelanggan, dan mengidentifikasi peluang bisnis baru. Implementasi algoritma K-Means memberikan pemahaman mendalam tentang pola perilaku dan preferensi pelanggan, yang menjadi landasan kuat untuk pengambilan keputusan strategis dalam upaya meningkatkan kinerja bisnis.

ABSTRACT

In today's digital era, marketing strategy has an important role in business competition. This research aims to improve marketing strategies through effective household segmentation using the K-Means algorithm. This algorithm is used to group customers based on characteristics such as income, education, and employment. The research process includes the stages of data selection, preprocessing, data transformation, application of the K-Means algorithm, and evaluation of clustering results. The research results show five different customer clusters with their own unique characteristics. This clustering helps companies design more specific and targeted marketing strategies, increase customer satisfaction, and identify new business opportunities. The implementation of the K-Means algorithm provides a deep understanding of customer behavior patterns and preferences, which becomes a strong basis for making strategic decisions in an effort to improve business performance.

Corresponding Author:
M.Dimas Nurwanda,
Fakultas Informatika/Teknik Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
Jl. D.I. Panjaitan No. 128, Purwokerto Indonesia
Email: nurwanda@student.telkomuniversity.ac.id

1. PENDAHULUAN

Di era digital seperti sekarang, strategi pemasaran memiliki peran penting dalam persaingan bisnis antar perusahaan. Selain mengutamakan strategi pemasaran berorientasi produk, perusahaan juga harus mampu mengutamakan strategi pemasaran berorientasi pelanggan. Hal ini berguna untuk mengelola hubungan yang baik dengan pelanggan sehingga kepuasan pelanggan tercapai dan mendapatkan loyalitas (Adiana et al., 2018). Dalam penyusunan strategi pemasaran berorientasi pelanggan, segmentasi pelanggan dapat dimanfaatkan untuk mengidentifikasi karakteristik dari setiap profil kepala rumah tangga dengan mengelompokkan profil kepala rumah tangga tersebut berdasarkan pendapatan, edukasi, pekerjaan dengan memanfaatkan data mining, data dapat diolah untuk mengetahui pola dan informasi dalam data tersebut (Sembiring Brahmana et al., 2020), (Christy et al., 2021), (Riszky & Sadikin, 2019).

Salah satu metode pada data mining adalah clustering yang menggunakan algoritma k-means yang mana metode ini berguna untuk mengelompokkan data(Khatib Sulaiman et al., n.d.).Teknik clustering dapat digunakan dalam menentukan segmentasi pelanggan dengan menganalisis kelompok data untuk mengetahui karakteristik dari kelompok profil kepala rumah tangga yang terbentuk (Dana et al., 2019). Adapun tahap tahap yang akan digunakan dalam penelitian ini meliputi Selection, Preprocessing, Transformation, Data mining, Evaluation/ Interpretation, Knowledge(Astria et al., 2019).

Adapun penelitian sebelumnya yang berjudul "Analisis segmentasi konsumen menggunakan algoritma K-Means" dilakukan oleh Sulistyowati, Basma Eno Ketherin, Amalia Anjani Arifianti dan Anwar Sodik pada tahun 2018. Proses clustering yang dilakukan dengan K-Means menghasilkan gambaran karakteristik konsumen yang dikelompokkan . Proses pendataan penjualan PT Calista Alba menghasilkan tiga cluster/kelompok konsumen berdasarkan tiga variabel yaitu jenis sepeda motor, jenis pembelian dan pekerjaan. Hasil pengelompokan data konsumen dapat digunakan sebagai informasi pendukung untuk bagian pemasaran dalam menentukan strategi pemasaran. Pengujian perbandingan menggunakan software SPSS menghasilkan jumlah cluster yang sama dengan penelitian yang dilakukan. Namun demikian masih terdapat perbedaan distribusi persentase jumlah anggota cluster, perbedaan rata-rata perbandingan kedua aplikasi adalah 7%. (Sulistyowati dkk., 2018).

Dari penjelasan diatas, maka penulis ingin melakukan penelitian "Penerapan K-Means untuk Meningkatkan Strategi Pemasaran Melalui Segmentasi Rumah Tangga yang Efektif. Salah satu hambatan dalam penelitian ini adalah bagaimana cara nya memahami karakteristik pelanggan untuk dapat meningkatkan strategi pemasaran yang efektif dan efisien. Dan permasalahan yang sering dihadapi adalah bagaimana menggolongkan profil kepala rumah tangga berdasarkan karakteristik dari kelompok profil kepala rumah tangga. Metode yang digunakan penelitian ini adalah K-Means clustering, yang merupakan salah satu teknik data mining yang dapat digunakan untuk mengelompokkan berdasarkan kemiripan karakteristik tertentu dimana data-data yang memiliki kemiripan akan berada pada cluster yang sama(Suharti et al., 2022).

Hasil dari analisis segmentasi pelanggan ini dapat mempermudah perusahaan dalam mengetahui karakteristik pelanggan. Selanjutnya, perusahaan dapat dengan mudah melakukan perencanaan strategi pemasaran yang lebih efektif untuk meningkatkan loyalitas pelanggan sehingga tingkat kesetiaan pelanggan dalam berbelanja pada aplikasi Alfagift dapat meningkat. Tujuan dari penelitian ini adalah untuk mengetahui karakteristik pelanggan untuk meningkatkan strategi pemasaran melalui segmentasi rumah tangga yang efektif. Strategi pemasaran lebih efektif dan tepat sasaran apabila disusun berdasarkan dari tipe dan karakteristik pelanggan dari hasil segmentasi pelanggan yang terbentuk.

2. METODE PENELITIAN

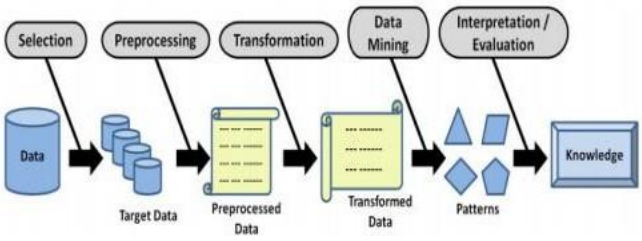
2.1 Data Mining

Data mining adalah proses mengekstrak informasi yang berguna dari data yang besar dan kompleks. Ini mencakup teknik statistik, algoritma, dan proses yang digunakan untuk menemukan pola dan hubungan dalam data. Tujuan dari data mining adalah untuk menemukan informasi yang berguna dan mengubahnya menjadi bentuk yang dapat diterima oleh pengguna. Data mining digunakan dalam berbagai bidang, seperti bisnis, ilmu pengetahuan, teknologi informasi, dan sebagainya untuk menemukan pola dalam data yang dapat digunakan untuk membuat keputusan yang lebih baik (Nahjan et al., 2023).

Dalam data mining terdapat metode-metode yang dapat digunakan seperti klasifikasi, clustering, regresi, seleksi variabel, dan analisis. Data Mining adalah bidang yang sepenuhnya menggunakan apa yang dihasilkan oleh data warehouse, bersama dengan bidang yang menangani masalah pelaporan dan manajemen data. Suatu pola yang digunakan untuk menemukan hubungan antar item sehingga membentuk pola

pengetahuan yang baru dalam data yang besar dan dapat diselesaikan menggunakan teknik tertentu (Natalia Br Sembiring et al., n.d.).

Adapun untuk menganalisis data dalam penerapan data mining menggunakan proses tahapan Knowledge Discovery in Databases (KDD) yang terdiri dari Selection, Cleaning, Transformation, Data Mining, dan Interpretation/Evaluation (Gustientiedina et al., 2019) seperti pada Gambar 1.



Gambar 1. Tahapan KDD

Adapun penjelasan pada Gambar 1 diatas adalah sebagai berikut :

1. Selection
   Selection digunakan untuk menentukan variabel yang akan diambil agar tidak ada kesamaan dan terjadi perulangan yang tidak diperlukan dalam pengolahan data mining.
2. Preprocessing
   Pada preprocessing terdapat dua tahap, yaitu sebagai berikut :
   a. Data Cleaning Menghilangkan data yang tidak diperlukan seperti menangani missing value, noise data serta menangani data – data yang tidak konsisten dan relevan.
   b. Data Integration Dilakukan terhadap atribut yang mengidentifikasikan entitas yang unik.
3. Transformation
   Merubah data sesuai format ekstension yang sesuai dalam pengolahan data mining karena beberapa metode pada data mining memerlukan format khusus sebelum dapat diproses pada data mining.
4. Datamining
   Proses utama pada metode yang diterapkan untuk mendapatkan pengetahuan baru dari data yang diproses. Pada penelitian ini diterapkan teknik clustering yaitu metode K Means Clustering.
5. Evaluation/Interpretation
   Mengidentifikasi pola – pola yang menarik ke dalam knowledge base yang diidentifikasi. Pada tahap ini, menghasilkan pola – pola khas maupun model prediksi yang dievaluasi untuk menilai kajian yang ada sudah memenuhi target yang diinginkan.
6. Knowledge
   Pola-pola yang dihasilkan akan dipresentasikan kepada pengguna. Pada tahapan ini pengetahuan baru yang dihasilkan bisa dipahami semua orang yang akan dijadikan acuan pengambilan keputusan.

2.2 Clustering

Clustering adalah metode pengelompokan data yang sering digunakan sebagai salah satu metode data mining atau penggalian data. Clustering adalah proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan cluster. Oleh karena itu, metode clustering ini sangat berguna untuk menemukan kelompok yang tidak dikenal dalam data (Prastiwi et al., n.d.).

Pengklusteran tidak digunakan untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari target. Pengklusteran digunakan untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok – kelompok yang memiliki kemiripan, pengklusteran berbeda dengan klasifikasi yang tidak adanya variabel target dalam pengclusteran(Nugraha et al., 2022).

Clustering pada suatu data adalah suatu tahapan untuk menggolongkan himpunan data yang atribut kelasnya belum dideskripsikan, secara konsep clustering adalah untuk memaksimalkan dan meminimalkan kemiripan intra antar kelas. sebagai contoh, ada suatu himpunan objek, proses pertama dapat di klasterisasi menjadi beberapa himpunan kelas selanjutnya menjadi sebuah himpunan beraturan sehingga dapat diturunkan berdasarkan kelompok klasifikasi tertentu. Cluster juga dapat diartikan sebagai kelompok. Maka analisa clustering pada dasarnya akan menghasilkan sejumlah cluster (kelompok)(Muliono & Sembiring, 2019).

2.3 Algoritma K-Means

Dalam algoritma K-Means, obyek atau data yang ada dikelompokkan ke dalam k kelompok atau klaster. Untuk melakukan clustering ini, nilai k harus ditentukan terlebih dahulu. Biasanya user atau pengguna sudah mempunyai informasi awal tentang objek yang sedang dipelajari, termasuk berapa jumlah cluster yang paling tepat. Dalam algoritma K-Mean digunakan ukuran ketidak miripan untuk mengelompokkan objek. Ketidakmiripan ini diterjemahkan dalam konsep jarak. Jika jarak dua objek cukup dekat, maka dua objek tersebut mirip. Semakin dekat berarti semakin tinggi kemiripannya. Semakin tinggi nilai jarak, semakin tinggi pula ketidakmiripannya(Nur Khomarudin, 2003).

Langkah-langkah algoritma K-Means adalah sebagai berikut:

a. Pilih secara acak k buah data sebagai pusat cluster.
b. Jarak antara data dan pusat cluster dihitung menggunakan Euclidean Distance. Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

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

dimana:
D(i,j) = Jarak data ke i ke pusat cluster j Xki = Data ke i pada atribut data ke k Xkj = Titik pusat ke j pada atribut ke k

c. Data ditempatkan dalam cluster yang terdekat, dihitung dari tengah cluster.
d. Pusat cluster baru akan ditentukan bila semua data telah ditetapkan dalam cluster terdekat.
e. Proses penentuan pusat cluster dan penempatan data dalam cluster diulangi sampai nilai centroid tidak berubah lagi.

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk menerapkan algoritma K-Means guna melakukan klasterisasi pada dataset pelanggan, yang diharapkan dapat membantu dalam segmentasi pelanggan berdasarkan karakteristik tertentu. Berikut adalah langkah-langkah yang dilakukan dalam proses tersebut:

3.1 Selection

Proses ini dimulai dengan memilih data yang relevan untuk analisis, yaitu data pelanggan yang disimpan dalam file "data_customer.csv". Data ini dipilih karena berisi informasi penting tentang pelanggan yang akan digunakan untuk melakukan klastering. Data dimuat ke dalam DataFrame menggunakan pustaka pandas untuk memudahkan manipulasi dan analisis awal. Pemilihan atribut yang relevan dalam analisis data sangat penting untuk memahami pola dan perilaku pelanggan. Atribut seperti Rentang Pendapatan membantu dalam mengidentifikasi pola pengeluaran, sedangkan Pendidikan mengungkapkan preferensi produk dan kemampuan memahami produk bernilai tinggi. Pekerjaan memberikan wawasan tentang gaya hidup dan kebutuhan pelanggan.

Transformasi data kategorikal menjadi numerik diperlukan agar algoritma K-Means dapat menghitung jarak antar data point dengan tepat. Misalnya, kode numerik digunakan untuk mewakili kategori seperti Jenis Kelamin, Status Pernikahan, dan Pemilik Rumah. Setelah itu, atribut yang dipilih digabungkan menjadi vektor fitur menggunakan VectorAssembler dari pustaka Spark ML. Vektor fitur ini menjadi input utama untuk algoritma K-Means. Keuntungan dari seleksi atribut yang baik termasuk peningkatan akurasi dalam klasterisasi data, pengurangan kompleksitas yang mengarah pada waktu komputasi yang lebih efisien, dan penghindaran overfitting untuk memastikan model dapat diterapkan dengan baik pada data baru. Dengan memilih atribut secara cermat, proses klasterisasi menjadi lebih efektif dan hasilnya lebih dapat diandalkan untuk analisis mendalam dan pengambilan keputusan bisnis.

3.2 Preprocessing Data

Langkah pertama dalam penelitian ini adalah membaca dataset dari file CSV menggunakan pustaka pandas. Dataset yang digunakan berisi informasi tentang pelanggan, termasuk atribut seperti Gender, Marital Status, HomeOwner, Occupation, dan Education yang dapat dilihat pada Tabel 1.

Tabel 1. Data Pelanggan

Atribut	Kategori	Kode
Gender	Laki-laki	0
	Perempuan	1
MaritalStatus	Belum Menikah	0
	Sudah Menikah	1
HomeOwner	Tidak Memiliki Rumah	0
	Memiliki Rumah	1
Occupation	Pekerjaan Manual Tidak Terampil	0
	Pekerjaan Manual Terampil	1
	Administrasi	2
	Penjualan	3
	Manajemen	4
	Profesional	5
Education	SMA	0
	S1	1
	S2	2
	S3	3
	Pascasarjana	4

Tahap pra-pemrosesan bertujuan untuk menyiapkan data agar dapat digunakan dalam model klastering. Langkah pertama adalah mengonversi nilai-nilai kategorikal menjadi numerik. Sebagai contoh, kolom "Gender" diubah menjadi 0 untuk laki-laki dan 1 untuk perempuan. Demikian pula, kolom "MaritalStatus" diubah menjadi 0 untuk belum menikah dan 1 untuk sudah menikah, dan atribut lainnya seperti "HomeOwner", "Occupation", dan "Education" juga diubah menjadi bentuk numerik yang sesuai. Setelah itu, data yang telah diproses diubah menjadi DataFrame Spark untuk memungkinkan pemrosesan yang lebih efisien menggunakan PySpark.

3.3 Transformation

Transformasi data merupakan tahap krusial yang bertujuan untuk mempersiapkan data agar sesuai dengan format yang diperlukan oleh algoritma klastering. Pada langkah ini, VectorAssembler dari PySpark digunakan untuk menggabungkan beberapa kolom menjadi satu kolom fitur yang akan digunakan oleh model KMeans. Secara khusus, kolom-kolom "IncomeRange", "Education", dan "Occupation" dipilih sebagai atribut penting yang akan dikombinasikan menjadi vektor fitur tunggal.

Tabel 2. Model Tranform

Cluster	Count
0	2608
1	2009
2	1974
3	1964
4	1989

VectorAssembler bekerja dengan menggabungkan nilai-nilai dari kolom-kolom tersebut ke dalam satu kolom baru bernama "features". Hal ini diperlukan karena algoritma KMeans bekerja dengan data dalam bentuk vektor fitur, bukan nilai-nilai individual dari berbagai kolom. Setelah vektor fitur ini dibuat, dataset yang telah diubah ini dipisahkan ke dalam dua kolom utama: "CustomerName" dan "features". Kolom "CustomerName" tetap ada untuk referensi identifikasi, sementara kolom "features" digunakan dalam proses pelatihan model KMeans. Dengan demikian, data siap digunakan untuk proses klastering, di mana setiap entri dalam dataset sekarang direpresentasikan sebagai vektor dalam ruang fitur multi-dimensi.

Tabel 4. Dataset

Costumername	Cluster
UsmanSetiawan	3
Fitri Anggraini	2
Agus Susanto	2

Costumername	Cluster
Taufik Haryanto	0
Indra Hidayat	0
Rizki Mahendra	1
Fina Fauzan	3
Nina Setiawan	4
Indra Purnomo	2
Wati Susanto	4

Sebelum menjalankan algoritma K-Means, terlebih dahulu atribut yang paling relevan telah dipilih. Misalnya, atribut IncomeRange, Education, dan Occupation dipilih karena masing-masing memberikan wawasan yang signifikan terhadap preferensi dan gaya hidup pelanggan. Selanjutnya, data kategorikal telah diubah menjadi representasi numerik menggunakan kode-kode tertentu, seperti jenis kelamin, status pernikahan, dan kepemilikan rumah. Hal ini diperlukan agar algoritma K-Means dapat mengoperasikan data dengan benar, menghitung jarak antar data point berdasarkan vektor fitur yang telah dibentuk dengan 'VectorAssembler'.

Proses selanjutnya adalah menentukan jumlah klaster (K) yang optimal. Dalam kasus ini, jumlah klaster sebanyak 5 dipilih setelah pertimbangan matang terhadap struktur data dan tujuan analisis yang diinginkan. Dengan memilih jumlah klaster yang tepat, klasterisasi menjadi lebih representatif dan dapat memberikan pemahaman yang lebih dalam terhadap kelompok pelanggan yang berbeda.

3.4 Data Mining

Setelah atribut yang relevan dipilih dan data kategorikal diubah menjadi representasi numerik, langkah selanjutnya dalam proyek ini adalah penerapan algoritma K-Means. Pertama, atribut yang telah dipilih seperti Rentang Pendapatan, Pendidikan, dan Pekerjaan digabungkan menjadi satu vektor fitur menggunakan `VectorAssembler` dari pustaka Spark ML. Vektor fitur ini digunakan sebagai input utama untuk algoritma K- Means.

Cluster 0 Description:

	Age	MaritalStatus	IncomeRange	Gender	TotalChildren
count	2068.000000	2068.000000	2068.000000	2068.000000	2068.000000
mean	43.338008	0.487427	33567.519342	0.518375	2.531431
std	15.134436	0.499963	7777.234089	0.499783	1.718022
min	18.000000	0.000000	20005.000000	0.000000	0.000000
25%	30.000000	0.000000	26862.000000	0.000000	1.000000
50%	43.000000	0.000000	33591.500000	1.000000	3.000000
75%	56.250000	1.000000	40374.000000	1.000000	4.000000
max	69.000000	1.000000	46912.000000	1.000000	5.000000

	ChildrenAtHome	Education	Occupation	HomeOwner	Cars \
count	2068.000000	2068.000000	2068.000000	2068.000000	2068.000000
mean	1.284333	1.972921	2.500967	0.504836	1.494197
std	1.414353	1.417372	1.722528	0.500098	1.132903
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	1.000000	0.000000	0.000000
50%	1.000000	2.000000	2.000000	1.000000	1.000000
75%	2.000000	3.000000	4.000000	1.000000	3.000000
max	5.000000	4.000000	5.000000	1.000000	3.000000

Gambar 2. Cluster Prediction 0

Selanjutnya, algoritma K-Means diterapkan pada data yang telah dipersiapkan dengan tujuan membentuk 5 klaster yang mewakili segmen-segmen berbeda dari populasi pelanggan. Proses ini melibatkan penempatan titik-titik pusat klaster awal secara acak, diikuti dengan iterasi untuk mengoptimalkan lokasi pusat klaster berdasarkan jarak dari titik data. Setelah konvergensi, model K-Means dilatih menggunakan data tersebut untuk menghasilkan klaster yang optimal.

```
Cluster 4 Description:
                Age  MaritalStatus   IncomeRange       Gender  TotalChildren  \
count  1989.000000    1989.000000   1989.000000  1989.000000    1989.000000
mean     43.589744       0.524887  60377.191051     0.496229       2.480644
std      14.982673       0.499506   7602.168869     0.500112       1.713437
min      18.000000       0.000000  46957.000000     0.000000       0.000000
25%      31.000000       0.000000  53836.000000     0.000000       1.000000
50%      44.000000       1.000000  60260.000000     0.000000       2.000000
75%      57.000000       1.000000  67038.000000     1.000000       4.000000
max      69.000000       1.000000  73318.000000     1.000000       5.000000

       ChildrenAtHome    Education   Occupation    HomeOwner         Cars  \
count     1989.000000  1989.000000  1989.000000  1989.000000  1989.000000
mean         1.258924     1.945199     2.508296     0.503771     1.485168
std          1.399492     1.400098     1.707056     0.500112     1.111222
min          0.000000     0.000000     0.000000     0.000000     0.000000
25%          0.000000     1.000000     1.000000     0.000000     0.000000
50%          1.000000     2.000000     3.000000     1.000000     2.000000
75%          2.000000     3.000000     4.000000     1.000000     2.000000
max          5.000000     4.000000     5.000000     1.000000     3.000000
```

Gambar 3. Cluster Prediction 4

Penerapan K-Means dalam konteks ini memberikan pemahaman yang lebih mendalam tentang pola perilaku dan preferensi pelanggan dalam kelompok-kelompok yang berbeda. Dengan membagi pelanggan ke dalam klaster yang berarti, perusahaan dapat mengidentifikasi peluang-peluang baru, mengkustomisasi strategi pemasaran, dan meningkatkan pengalaman pelanggan secara keseluruhan. Hasil analisis dari model ini akan menjadi landasan yang kuat untuk pengambilan keputusan strategis dalam upaya meningkatkan kinerja bisnis dan kepuasan pelanggan.

3.5 Evaluation

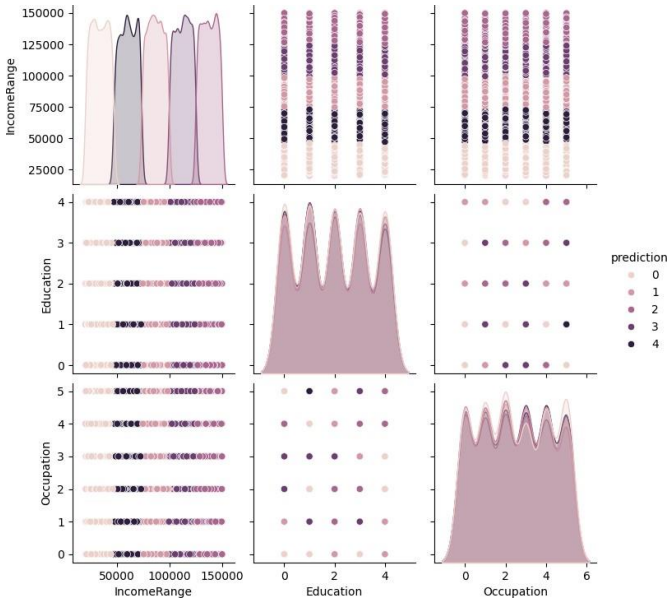
Setelah model K-Means dilatih menggunakan data pelanggan yang telah diproses, langkah selanjutnya adalah mengevaluasi dan menginterpretasi hasil klasterisasi untuk memahami karakteristik setiap klaster. Proses ini penting karena akan memberikan wawasan mendalam tentang berbagai kelompok pelanggan yang ada. Dengan memahami perbedaan dalam atribut seperti Rentang Pendapatan, Pendidikan, dan Pekerjaan di setiap klaster, perusahaan dapat merancang strategi pemasaran yang lebih tepat sasaran. Hasil dari analisis ini juga dapat digunakan untuk menyesuaikan produk atau layanan, meningkatkan pengalaman pelanggan, dan mengidentifikasi peluang bisnis baru yang dapat dioptimalkan untuk setiap segmen. Melalui integrasi kode-kode Python dan penggunaan alat-alat analisis data seperti Spark ML, proses ini tidak hanya efisien tetapi juga dapat memberikan solusi yang komprehensif dalam pengambilan keputusan strategis. Dengan demikian, perusahaan dapat meningkatkan kinerja bisnis mereka dengan pendekatan yang didasarkan pada pemahaman yang mendalam tentang preferensi dan perilaku pelanggan.

```
Predictions made successfully.
Silhouette Score: 0.7237541215009036
```

Gambar 4. Sillhoute Score

3.6 Knowledge

Tahap akhir adalah ekstraksi pengetahuan dari model yang telah dibuat. Analisis deskriptif dilakukan pada setiap klaster untuk memahami karakteristik unik masing-masing klaster. Misalnya, statistik deskriptif seperti rata-rata dan standar deviasi dari atribut dalam setiap klaster dihitung dan dianalisis. Selain itu, visualisasi klaster dilakukan menggunakan pairplot dari seaborn untuk memperlihatkan distribusi data dalam setiap klaster berdasarkan atribut yang digunakan dalam klastering. Visualisasi ini membantu dalam memahami pola dan hubungan antar atribut dalam setiap klaster. Secara keseluruhan, proses ini memberikan panduan lengkap mulai dari pemilihan data hingga ekstraksi pengetahuan dengan menggunakan teknik data mining, khususnya klastering pelanggan menggunakan algoritma KMeans di PySpark.



Gambar 5. Visualisasi Data

Setelah menerapkan algoritma K-Means pada dataset pelanggan, terbentuk lima klaster yang membagi pelanggan berdasarkan karakteristik seperti pendapatan, pendidikan, dan pekerjaan. Setiap klaster memiliki centroid yang mewakili rata-rata atribut yang dipilih. Distribusi pelanggan dalam klaster dapat divisualisasikan untuk memahami proporsi masing-masing. Analisis lebih lanjut mengidentifikasi karakteristik unik dari setiap klaster, seperti pelanggan dengan pendapatan tinggi dan pendidikan tinggi di Klaster 1, atau pendidikan SMA dan pekerjaan manual di Klaster 3. Hasil klasterisasi disajikan dalam visualisasi scatter plot untuk memperlihatkan sebaran data. Pemilihan atribut yang tepat seperti IncomeRange, Education, dan Occupation memungkinkan perancangan strategi pemasaran yang lebih spesifik, meningkatkan kepuasan pelanggan, dan memahami preferensi konsumen lebih baik.

4. KESIMPULAN

Penelitian ini bertujuan untuk meningkatkan strategi pemasaran melalui segmentasi rumah tangga yang efektif menggunakan algoritma K-Means. Dengan menerapkan metode data mining ini, kami mengelompokkan pelanggan berdasarkan karakteristik seperti pendapatan, pendidikan, dan pekerjaan. Hasilnya adalah lima klaster yang mewakili segmen berbeda dari populasi pelanggan, dengan setiap klaster memiliki centroid yang merepresentasikan rata-rata atribut yang dipilih.

Proses penelitian ini meliputi beberapa tahapan penting: pemilihan atribut yang relevan seperti Rentang Pendapatan, Pendidikan, dan Pekerjaan, transformasi data kategorikal menjadi representasi numerik, dan penerapan algoritma K-Means untuk membentuk klaster. Penentuan jumlah klaster (K) ditetapkan sebagai lima untuk memberikan segmentasi yang lebih mendetail.

Hasil klasterisasi dievaluasi dan diinterpretasikan untuk memahami karakteristik unik dari masing-masing klaster. Visualisasi scatter plot digunakan untuk memperlihatkan distribusi pelanggan di dalam klaster-kelaster tersebut. Dengan memahami karakteristik ini, perusahaan dapat merancang strategi pemasaran yang lebih spesifik dan tepat sasaran, meningkatkan kepuasan pelanggan, serta mengidentifikasi peluang-peluang baru dalam bisnis. Penelitian ini memberikan landasan kuat untuk pengambilan keputusan strategis dalam meningkatkan kinerja bisnis, dengan fokus pada pemahaman yang mendalam terhadap preferensi dan perilaku pelanggan.

Memberikan pernyataan bahwa apa yang diharapkan seperti yang tertera pada bab “Pendahuluan” pada akhirnya dapat menghasilkan bab “Hasil dan Pembahasan”, sehingga ada kesesuaian. Selain itu juga dapat ditambahkan prospek pengembangan hasil penelitian dan prospek penerapan studi lanjutan ke depan (berdasarkan hasil dan pembahasan).

ACKNOWLEDGEMENTS

Kami mengucapkan terima kasih kepada seluruh pihak yang telah berkontribusi dalam pelaksanaan penelitian ini. Terima kasih kepada para pembimbing, rekan-rekan, dan keluarga yang telah memberikan dukungan dan motivasi. Kami juga berterima kasih kepada institusi pendidikan kami yang telah menyediakan fasilitas dan sumber daya yang diperlukan. Penghargaan khusus kami sampaikan kepada para responden yang telah bersedia memberikan data yang sangat berharga bagi penelitian ini. Penelitian ini tidak akan berhasil tanpa kontribusi dan dukungan dari semua pihak yang terlibat.

REFERENCES

Adiana, B. E., Soesanti, I., & Permanasari, A. E. (2018). Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering. 2. https://doi.org/10.21460/jutei.2017.21.76

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data.

Physics Procedia, 78, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Astria, C., Windarto, A. P., & Hartama, D. (2019). PENERAPAN K-MEDOID PADA RUMAH TANGGA YANG MEMILIKI SUMBER PENERANGAN LISTRIK PLN BERDASARKAN PROVINSI.

KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer), 3(1). https://doi.org/10.30865/komik.v3i1.1667

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences, 33(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

Dana, R. D., Rohmat, C. L., & Rinaldi, A. R. (2019). Strategi Marketing Penerimaan Mahasiswa Baru Menggunakan Machine Learning dengan Teknik Clustering. Jurnal Informatika: Jurnal Pengembangan IT, 4(2–2), 201–204. https://doi.org/10.30591/jpit.v4i2-2.1879

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. Jurnal Nasional Teknologi Dan Sistem Informasi, 5(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24

Khatib Sulaiman, J., Meiriza, A., Ali, E., & AMIK Riau Pekanbaru, S. (n.d.). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Program BPJS Ketenagakerjaan. Indonesian Journal of Computer Science.

Muliono, R., & Sembiring, Z. (2019). DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K- MEANS UNTUK KLASTERISASI TINGKAT TRIDARMA PENGAJARAN DOSEN (Vol. 4, Issue 2).

Nahjan, M. R., Heryana, N., & Voutama, A. (2023). IMPLEMENTASI RAPIDMINER DENGAN METODE CLUSTERING K-MEANS UNTUK ANALISA PENJUALAN PADA TOKO OJ CELL. In Jurnal Mahasiswa Teknik Informatika (Vol. 7, Issue 1).

Natalia Br Sembiring, S., Winata, H., Kusnasari, S., Informasi, S., & Triguna Dharma, S. (n.d.).

Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means.

Nugraha, A., Nurdiawan, O., & Dwilestari, G. (2022). PENERAPAN DATA MINING METODE K-MEANS CLUSTERING UNTUK ANALISA PENJUALAN PADA TOKO YANA SPORT. In Jurnal

Mahasiswa Teknik Informatika) (Vol. 6, Issue 2).

Nur Khomarudin, A. (2003). Teknik Data Mining : Algoritma K-Means Clustering. https://agusnkhom.wordpress.com

Prastiwi, H., Pricilia, J., & Raswir, E. (n.d.). Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM) Implementasi Data Mining Untuk Menentuksn Persediaan Stok Barang Di Mini Market Menggunakan Metode K-Means Clustering.

Riszky, A. R., & Sadikin, M. (2019). Data Mining Menggunakan Algoritma Apriori untuk Rekomendasi Produk bagi Pelanggan. Jurnal Teknologi Dan Sistem Komputer, 7(3), 103–108. https://doi.org/10.14710/jtsiskom.7.3.2019.103-108

Sembiring Brahmana, R. W., Mohammed, F. A., & Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 11(1), 32. https://doi.org/10.24843/lkjiti.2020.v11.i01.p04

Suharti, P. H., Suryandari, A. S., & Amalia, R. N. (2022). ANALISIS KINERJA MODUL PENGENDALI TEKANAN UDARA PCT-14 BERBASIS PLC DENGAN BERBAGAI METODA TUNING. Sebatik,

26(2), 420–427. https://doi.org/10.46984/sebatik.v26i2.2134