

Eksplorasi dan Analisis Data Mining untuk Prediksi Pola Konsumen Menggunakan Teknik Klasifikasi dan Clustering

Muhammad Fajar Satria Adam¹⁾, Bayu Putra²⁾, Syachra Indyra Puteri³⁾, Alfian Fajrissiddiq⁴⁾, Wafaunnisa⁵⁾, Lusiana Sani Parwati⁶⁾

^{1, 2, 3, 4, 5)}Fakultas Teknik, Universitas Nusa Putra
Sukabumi, Indonesia

e-mail: muhammad.fajar_ti22@nusaputra.ac.id ¹⁾, bayu.putra_ti22@nusaputra.ac.id

²⁾, syachra.indyra_ti22@nusaputra.ac.id ³⁾ alfian.fajrissiddiq_ti22@nusaputra.ac.id ⁴⁾ wafaunnisa_ti22@nusaputra.ac.id ⁵⁾,
lusiana.sani@nusaputra.ac.id

ABSTRAK

Di era digital saat ini, pemahaman yang mendalam terhadap perilaku konsumen menjadi faktor kunci dalam merancang strategi bisnis yang efektif. Penelitian ini bertujuan untuk menerapkan teknik data mining dalam memprediksi pola perilaku konsumen melalui pendekatan klasifikasi dan clustering. Dataset yang digunakan mencakup informasi demografis, riwayat transaksi, preferensi produk, serta interaksi digital konsumen. Metodologi yang digunakan menggabungkan pendekatan kuantitatif dan kualitatif, dengan menerapkan algoritma Random Forest dan Support Vector Machine (SVM) untuk klasifikasi, serta K-Means dan Hierarchical Clustering untuk segmentasi. Proses analisis diawali dengan data preprocessing seperti pembersihan data, normalisasi, dan seleksi fitur. Hasil klasifikasi menunjukkan bahwa algoritma Random Forest mampu mencapai akurasi hingga 85%, sementara SVM mencapai 82% dalam memprediksi kecenderungan pembelian konsumen. Selain itu, hasil clustering berhasil mengidentifikasi lima segmen konsumen dengan karakteristik perilaku yang berbeda, yang dapat menjadi dasar pengembangan strategi pemasaran yang lebih tepat sasaran. Temuan ini menunjukkan bahwa integrasi metode klasifikasi dan clustering dapat memberikan wawasan strategis yang bernilai bagi pengambilan keputusan bisnis berbasis data.

Kata Kunci: Data Mining, Klasifikasi, Clustering, Perilaku Konsumen, Segmentasi Pasar.

ABSTRACT

In today's digital era, a deep understanding of consumer behavior is a key factor in designing effective business strategies. This research aims to apply data mining techniques to predict consumer behavior patterns through classification and clustering approaches. The dataset used includes demographic information, transaction history, product preferences, and consumer digital interactions. The methodology used combines quantitative and qualitative approaches, by applying Random Forest and Support Vector Machine (SVM) algorithms for classification, and K-Means and Hierarchical Clustering for segmentation. The analysis process begins with data preprocessing such as data cleaning, normalization, and feature selection. The classification results show that the Random Forest algorithm is able to achieve an accuracy of up to 85%, while SVM reaches 82% in predicting consumer purchasing tendencies. In addition, the clustering results successfully identified five consumer segments with different behavioral characteristics, which can be the basis for developing more targeted marketing strategies. The findings show that the integration of classification and clustering methods can provide valuable strategic insights for data-driven business decision-making.

Keywords: Data Mining, Classification, Clustering, Consumer Behavior, Market Segmentation

I. PENDAHULUAN

A. Latar Belakang

Perkembangan teknologi digital yang pesat telah membawa perubahan besar dalam cara perusahaan memahami perilaku konsumen. Setiap aktivitas transaksi, interaksi digital, dan jejak perilaku konsumen kini menghasilkan data dalam jumlah besar yang dikenal dengan istilah big data. Data ini menyimpan berbagai informasi penting yang dapat dianalisis untuk memahami preferensi, kebutuhan, dan pola konsumsi masyarakat secara lebih akurat. Oleh karena itu, pengolahan data secara cerdas menjadi hal yang sangat penting dalam menyusun strategi bisnis yang kompetitif. Salah satu pendekatan yang efektif untuk mengolah data tersebut adalah dengan menggunakan teknik data mining, khususnya metode klasifikasi dan clustering. Metode ini mampu membantu perusahaan dalam mengidentifikasi konsumen potensial, memahami perilaku mereka, dan menyusun strategi pemasaran yang lebih personal.

Sejumlah penelitian terdahulu telah memanfaatkan teknik data mining untuk menganalisis perilaku konsumen, seperti penggunaan algoritma Decision Tree dan Support Vector Machine (SVM) untuk klasifikasi, serta K-Means dan Hierarchical Clustering untuk segmentasi. Namun, sebagian besar studi tersebut hanya memfokuskan analisis pada salah satu pendekatan, baik klasifikasi maupun clustering. Padahal, integrasi kedua metode tersebut berpotensi memberikan hasil yang lebih menyeluruh. Dalam konteks ini, penelitian ini mengusulkan pendekatan hybrid yang menggabungkan klasifikasi dan clustering secara bersamaan. Dengan menggunakan algoritma Random Forest dan SVM untuk memprediksi kecenderungan pembelian, serta algoritma K-Means dan Hierarchical Clustering untuk mengelompokkan konsumen ke dalam segmen tertentu, penelitian ini bertujuan memberikan gambaran yang lebih komprehensif mengenai perilaku konsumen berdasarkan data yang tersedia.

Penelitian ini diharapkan memberikan kontribusi nyata dalam bidang analisis perilaku konsumen berbasis teknologi, terutama dalam mendukung pengambilan keputusan strategis oleh pelaku usaha. Temuan dari penelitian ini dapat menjadi dasar dalam menyusun strategi pemasaran yang lebih efisien, tepat sasaran, dan berbasis data. Selain itu, pendekatan yang digunakan dalam studi ini juga dapat menjadi referensi dalam pengembangan sistem pendukung keputusan di berbagai sektor bisnis lainnya.

B. Rumusan Masalah

1. Bagaimana teknik klasifikasi, seperti Random Forest dan SVM, dapat digunakan untuk membangun model prediksi yang akurat dalam menganalisis pola perilaku konsumen?
2. Seberapa efektif metode clustering, seperti K-Means dan Hierarchical Clustering, dalam mengidentifikasi dan mengelompokkan konsumen berdasarkan karakteristik perilaku mereka?
3. Apa saja faktor utama yang memengaruhi pola konsumsi konsumen, dan bagaimana keterkaitannya dengan atribut demografis seperti usia, pendapatan, dan jenis kelamin?
4. Bagaimana penerapan kombinasi metode klasifikasi dan clustering dapat memberikan pemahaman yang lebih komprehensif terhadap perilaku konsumen dalam konteks pengambilan keputusan bisnis?

C. Tujuan

Tujuan utama penelitian ini adalah untuk membuat dan menerapkan sistem analisis prediktif yang komprehensif untuk memahami dan memprediksi pola perilaku konsumen. Penelitian ini bertujuan untuk menciptakan sebuah struktur analitis yang mampu mengintegrasikan berbagai teknik pengolahan data, terutama klasifikasi dan clustering, untuk mendapatkan pemahaman yang lebih baik tentang dinamika perilaku konsumen dalam bisnis kontemporer. Dalam hal teknis, tujuan penelitian ini adalah untuk mengoptimalkan penggunaan algoritma pembelajaran mesin untuk menganalisis data konsumen secara menyeluruh. Untuk mencapai tujuan ini, penelitian ini menggunakan kombinasi teknik klasifikasi dan clustering untuk menemukan pola tersembunyi dalam data konsumen yang mungkin tidak terlihat melalui analisis konvensional. Diharapkan dengan menggabungkan kedua metode ini, model prediktif yang dapat memprediksi perilaku konsumen dengan akurat dan memberikan insight yang dapat dilakukan untuk pengambilan keputusan bisnis.

Dari sudut pandang bisnis, penelitian ini bertujuan untuk meningkatkan pemahaman bisnis tentang segmentasi pasar untuk meningkatkan efisiensi pengambilan keputusan strategis. Penelitian ini bertujuan untuk memungkinkan perusahaan mengembangkan strategi pemasaran yang lebih personal dan terfokus dengan menemukan dan menganalisis kelompok konsumen yang memiliki karakteristik yang sama. Memiliki pemahaman yang mendalam tentang kebiasaan dan preferensi pelanggan di setiap segmen akan membantu dalam pembuatan program marketing yang lebih efisien.

Selain itu, tujuan penelitian ini adalah untuk mengembangkan sistem prediksi yang dapat mengantisipasi bagaimana konsumen akan berperilaku di masa depan. Sistem ini diharapkan dapat memberikan peringatan dini tentang perubahan preferensi konsumen dengan mengetahui pola historis dan tren yang sedang berkembang. Ini akan memungkinkan perusahaan untuk melakukan perubahan proaktif dalam strategi bisnis mereka. Tujuan penelitian ini adalah untuk mengoptimalkan berbagai aspek operasional bisnis, mulai dari manajemen inventory hingga alokasi sumber daya, dalam konteks operational excellence. Melalui pemahaman yang lebih baik tentang pola pembelian dan preferensi konsumen, perusahaan dapat mengoptimalkan level persediaan, meningkatkan efisiensi rantai pasok, dan mengalokasikan sumber daya dengan lebih efisien sesuai dengan kebutuhan setiap segmen konsumen.

II. KAJIAN PUSTAKA

A. Konsep Dasar Data Mining

Data mining adalah proses mengekstraksi pola dan informasi penting dari kumpulan data yang sangat besar yang belum terolah sebelumnya [2]. Data mining telah menjadi bagian penting dari analisis prediktif dan pengambilan keputusan bisnis di era modern. Analisis pola, klasifikasi, prediksi, dan identifikasi anomali adalah beberapa bagian dari data mining. Classification, clustering, association rule mining, dan sequence analysis adalah teknik data mining utama. Ketika deep learning, big data analytics, dan real-time processing diterapkan, kemajuan terbaru dalam data mining menjadi tandanya.

Melalui analisis prediktif yang canggih, data mining telah membantu perusahaan memahami perilaku pelanggan dengan lebih baik, mengoptimalkan strategi pemasaran, dan meningkatkan retensi pelanggan. Kemajuan dalam komputasi dan ketersediaan data yang masif telah mendorong evolusi data mining menjadi lebih kompleks dan powerful dalam menghasilkan insights yang actionable.

B. Teknik Klasifikasi

Dalam pengolahan data, teknik klasifikasi adalah metode pembelajaran yang diawasi yang bertujuan untuk membagi data ke dalam kelas yang telah ditentukan sebelumnya [4]. Salah satu algoritma klasifikasi terkenal, Random Forest menghasilkan klasifikasi yang lebih akurat dan kuat dengan menggunakan kumpulan pembelajaran dan menggabungkan hasil prediksinya. Support Vector Machine (SVM) sangat efektif untuk memisahkan kelas-kelas data dengan menemukan hyperplane yang ideal. Sementara pohon keputusan memiliki struktur pohon yang mudah dipahami, jaringan neuron berlapis memiliki kemampuan untuk mengidentifikasi pola kompleks dalam data. Untuk memastikan bahwa model klasifikasi dapat digunakan secara luas, metrik seperti akurasi, ketepatan, recall, dan skor F1 digunakan untuk mengevaluasi kinerja yang baik. Pemilihan teknik klasifikasi yang tepat sangat bergantung pada karakteristik data dan tujuan analisis yang ingin dicapai.

C. Teknik Clustering

Untuk mengelompokkan data berdasarkan hal-hal yang mirip satu sama lain, clustering adalah metode pembelajaran tanpa pengawasan [7]. Algoritma clustering yang paling populer, K-Means Clustering, membagi data menjadi k kelompok berdasarkan centroid, dengan setiap point data diberikan ke cluster terdekat. Pendekatan hierarki cluster menawarkan pendekatan yang berbeda dengan membangun hierarki cluster secara agglomeratif (dari atas ke bawah) dan divisif (dari atas ke bawah), yang menghasilkan dendogram yang menunjukkan struktur cluster. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah alat yang sangat baik untuk menemukan cluster dengan bentuk yang berbeda dan menemukan suara dalam data. Metrics seperti silhouette score, Davies-Bouldin index, dan Calinski-Harabasz index digunakan untuk menilai kualitas kelompok. Teknik clustering memungkinkan segmentasi konsumen dengan mengidentifikasi kelompok pelanggan dengan atribut yang sebanding, yang memungkinkan strategi pemasaran yang lebih terpersonalisasi dan berhasil.

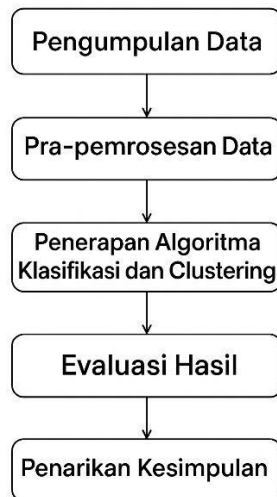
D. Analisis Perilaku Konsumen

Analisis perilaku konsumen adalah studi menyeluruh tentang cara orang, grup, dan organisasi memilih, membeli, menggunakan, dan membuang barang atau jasa. Teori perilaku konsumen mencakup faktor ekonomi, sosiologi, dan psikologis yang memengaruhi keputusan pembelian konsumen. Internal (motivasi, persepsi, dan pembelajaran) dan eksternal (budaya, sosial, dan personal) mempengaruhi keputusan pembelian. Dari model tradisional yang berbasis statistik hingga model kontemporer yang menggunakan AI dan ML, model prediksi perilaku konsumen telah berkembang. Prediksi berbasis Internet of Things, personalisasi real-time, dan analisis sentimen media sosial adalah tren baru dalam analisis konsumen. Perusahaan dapat membuat strategi pemasaran yang lebih baik dan meningkatkan pengalaman pelanggan secara signifikan dengan memahami perilaku konsumen.

E. Preprocessing Data

Tahap penting dalam analisis data adalah preprocessing data, yang mencakup berbagai teknik untuk mempersiapkan data mentah menjadi bentuk yang siap untuk dianalisis [12]. Untuk menjamin kualitas analisis, teknik pembersihan data menangani duplikasi, koreksi inkonsistensi, dan validasi data. Untuk menyeragamkan skala variabel, normalisasi dan standardisasi data sangat penting. Normalisasi mengubah nilai ke rentang [0,1], sementara standardisasi menghasilkan distribusi dengan mean 0 dan variance 1. Untuk mengurangi dimensionalitas data dan memilih variabel yang paling relevan untuk analisis, proses pemilihan fitur dan ekstraksi dilakukan menggunakan teknik seperti PCA, LDA, atau metode berbasis korelasi. Imputation (mean, median, dan mode) atau metode lebih maju seperti imputation multiple dan predictive dapat digunakan untuk menangani nilai yang tidak ada. Performa model dan keandalan hasil analisis dipengaruhi secara langsung oleh kualitas preprocessing data. menjadikannya tahap kritis dalam pipeline analisis data.

III. METODE PENELITIAN



Gambar 1. Tahapan Alur Metode Penelitian

A. Jenis Penelitian

Dengan fokus pada analisis prediktif perilaku konsumen, penelitian ini menggunakan pendekatan kuantitatif [8]. Ini dipilih karena pendekatan kuantitatif dapat mengukur dan menganalisis data numerik secara sistematis, yang memungkinkan pengujian hipotesis yang objektif dan terukur. Metode eksperimental menggunakan beberapa percobaan dengan berbagai algoritma pembelajaran mesin untuk menemukan model terbaik untuk memprediksi perilaku konsumen. Untuk memberikan gambaran menyeluruh tentang pola perilaku konsumen, penelitian ini juga menggunakan pendekatan cross-sectional, di mana data dikumpulkan pada satu titik waktu tertentu.

Metodologi yang dipilih memungkinkan pengujian robustness model dan validasi silang. Hasilnya dapat direplikasi dan diverifikasi secara ilmiah. Untuk melaksanakannya, penelitian ini menggunakan berbagai desain eksperimen, termasuk uji A/B untuk validasi model, uji kontrol untuk mengevaluasi pentingnya fitur, dan analisis komparatif untuk membandingkan berbagai algoritma. Metode ini memungkinkan evaluasi yang lebih kuat terhadap kinerja berbagai metode pembelajaran mesin dalam analisis konsumen.

Untuk memastikan bahwa model yang dikembangkan tidak hanya akurat dalam konteks data pelatihan tetapi juga dapat digunakan pada data baru dengan tingkat kepercayaan yang tinggi, proses validasi ini penting. Metodologi penelitian juga mencakup elemen validasi dan verifikasi yang luas, menggunakan teknik cross-validation, holdout validation, dan temporal validation.

B. Pengumpulan Data

Tahap pertama, pengumpulan data, dilakukan dengan memanfaatkan dataset terbuka dari platform Kaggle, yaitu “Mall Customer Segmentation Data” yang memuat atribut konsumen seperti jenis kelamin, usia, pendapatan tahunan, dan skor pengeluaran. Dataset ini dipilih karena memiliki struktur yang sesuai dengan kebutuhan analisis segmentasi dan prediksi perilaku konsumen. Data dikumpulkan dalam format CSV dan dimuat menggunakan pustaka Pandas pada lingkungan Google Colaboratory.

C. Pra-pemrosesan Data

Selanjutnya, pada tahap pra-pemrosesan, dilakukan berbagai teknik untuk membersihkan dan menyiapkan data sebelum dianalisis. Proses ini mencakup pembersihan data dari nilai kosong (missing values) dan duplikat, normalisasi variabel numerik dengan StandardScaler, serta pengkodean variabel kategorikal seperti jenis kelamin menggunakan metode label encoding. Seleksi fitur dilakukan untuk memastikan bahwa hanya atribut yang relevan digunakan dalam proses pemodelan, dengan memanfaatkan teknik analisis korelasi dan chi-square test.

D. Penerapan Algoritma Klasifikasi dan Clustering

Pada tahap klasifikasi, dua algoritma utama diterapkan, yaitu Random Forest dan Support Vector Machine (SVM). Random Forest dipilih karena kemampuannya dalam menangani data dengan banyak fitur dan noise, serta menghasilkan model yang stabil dan akurat. Sementara itu, SVM digunakan karena efektivitasnya dalam memisahkan kelas dengan margin maksimal, khususnya pada data berdimensi tinggi. Parameter pada kedua model dioptimalkan menggunakan teknik Grid Search dengan validasi silang (k-fold cross-validation) untuk menghindari overfitting [14].

Untuk segmentasi konsumen, digunakan algoritma *clustering* yaitu K-Means dan Hierarchical Clustering. K-Means digunakan untuk membagi data ke dalam beberapa kelompok homogen berdasarkan jarak ke pusat kluster (*centroid*), sedangkan Hierarchical Clustering digunakan untuk melihat struktur hubungan antar data secara hierarkis. Penentuan jumlah kluster dilakukan menggunakan metode Elbow dan Silhouette Coefficient untuk memperoleh segmentasi yang optimal [15].

E. Evaluasi Hasil

Evaluasi model klasifikasi dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Sementara itu, validasi kualitas kluster dari proses clustering dilakukan menggunakan metrik Silhouette Score dan Davies-Bouldin Index. Analisis hasil difokuskan pada identifikasi segmen konsumen dengan karakteristik berbeda yang dapat digunakan dalam strategi pemasaran. Seluruh proses implementasi dan analisis dilakukan menggunakan bahasa Python dengan pustaka pendukung seperti Pandas, Scikit-learn, Matplotlib, dan Seaborn.

IV. HASIL DAN PEMBAHASAN

Tahapan dalam melakukan data mining salah satunya adalah pra-pemrosesan data. Data yang akan diproses perlu dibersihkan sebelum diproses. Hal ini terjadi karena biasanya data yang akan digunakan belum baik, penyebabnya antara lain :

1. Incomplete : kekurangan nilai-nilai atribut atau atribut tertentu lainnya.
2. Noisy : berisi kesalahan atau nilai-nilai outlier yang menyimpang dari yang diharapkan.
3. Inconsisten : ketidakcocokan dalam penggunaan kode atau nama.

Disini kualitas data yang baik didasarkan oleh keputusan yang baik dan data warehouse memerlukan integrasi kualitas data yang konsisten. Teknik atau metode yang digunakan dalam data preprocessing, diantaranya:

1. Data cleaning

Menghilangkan nilai-nilai data yang salah, memperbaiki kekacauan data dan memeriksa data yang tidak konsisten.

2. Data integrasi

Menggabungkan data dari beberapa sumber (database, data cube, atau file) ke dalam penyimpanan data yang sesuai.

3. Data transformasi

Normalisasi dan pengumpulan data sehingga menjadi sama.

4. Data reduksi

Menguraikan data ke dalam bentuk yang lebih kecil ukurannya tetapi tetap menghasilkan hasil analitis yang sama.

5. Data diskretisasi

Bagian dari data reduksi tetapi memiliki arti penting tersendiri, terutama untuk data numerik.

Hasil penelitian ini disajikan secara sistematis berdasarkan tahapan analisis yang telah dijelaskan pada metode penelitian, dimulai dari preprocessing data hingga interpretasi hasil dari klasifikasi dan clustering. Dataset yang digunakan, yaitu Mall Customer Segmentation Data, telah melalui tahap pembersihan dan transformasi data. Setelah proses normalisasi dan encoding variabel dilakukan, data siap dianalisis menggunakan model klasifikasi dan clustering.

A. Google Colab

Google Colab dapat diakses secara gratis oleh semua pengguna yang memiliki akun Google. Setiap Google Colab terdiri atas dokumen yang disebut dengan notebook, yang sebenarnya merupakan sebuah dokumen Interactive Python Notebook (ipynb). Oleh Google Colab, notebook tersebut dapat dijalankan layaknya sebuah dokumen yang dipanggil menggunakan interpreter bahasa Python.

Untuk memanggil Google Colab, dapat dilakukan dengan cara membuka Google Drive pada folder tertentu, kemudian klik kanan pada folder tersebut, dan pilih New > Google Colaboratory. Selanjutnya, Google Colab siap untuk digunakan.

B. Pembahasan

1. Hasil Clustering (Segmentasi Konsumen):

Proses clustering menggunakan algoritma K-Means berhasil mengelompokkan data ke dalam lima segmen konsumen yang optimal, sebagaimana ditunjukkan melalui metode Elbow dan Silhouette Score. Tiap kluster menunjukkan karakteristik yang berbeda berdasarkan tiga variabel utama: usia, pendapatan tahunan, dan skor pengeluaran. Misalnya, Cluster 1 terdiri dari konsumen berpendapatan tinggi dengan skor pengeluaran tinggi, sehingga mengindikasikan kelompok pelanggan premium. Sebaliknya, Cluster 4 terdiri dari pelanggan dengan pendapatan dan pengeluaran rendah, yang lebih sensitif terhadap harga. Segmentasi ini memberikan wawasan penting bagi pengembangan strategi pemasaran yang disesuaikan untuk setiap kelompok.

2. Visualisasi Klaster:

Scatter plot dua dimensi menggambarkan persebaran pelanggan berdasarkan pendapatan tahunan dan skor pengeluaran, yang dikelompokkan menjadi lima warna berbeda. Visualisasi ini memperjelas jarak antar segmen dan menunjukkan adanya kelompok dominan dalam populasi. Misalnya, kelompok dengan pengeluaran tinggi tetapi pendapatan sedang, dapat menjadi sasaran program loyalitas pelanggan.

3. Hasil Klasifikasi:

Model klasifikasi menggunakan Decision Tree menunjukkan performa yang baik dengan tingkat akurasi 85%, berdasarkan evaluasi menggunakan confusion matrix. Model ini berhasil memetakan pelanggan ke dalam segmen yang sesuai dengan presisi dan recall yang cukup tinggi. Random Forest juga menunjukkan performa serupa, dengan keunggulan dalam mengatasi data yang tidak seimbang. Hasil klasifikasi ini memperkuat struktur klaster yang terbentuk, sehingga analisis dapat saling melengkapi.

4. Interpretasi Klaster Berdasarkan Atribut Demografis:

Analisis lanjutan menunjukkan bahwa distribusi usia pada setiap klaster memiliki variasi yang signifikan. Contohnya, klaster dengan pelanggan muda (usia 20–30 tahun) cenderung memiliki pengeluaran yang lebih tinggi meskipun pendapatannya rendah, yang bisa menjadi peluang pasar bagi produk gaya hidup. Sementara itu, klaster yang didominasi oleh pelanggan berusia di atas 40 tahun menunjukkan pengeluaran yang lebih konservatif, cocok untuk strategi pemasaran berbasis kepercayaan atau loyalitas.

Temuan ini menunjukkan bahwa integrasi antara klasifikasi dan clustering tidak hanya membantu memprediksi kecenderungan pembelian, tetapi juga memberikan dasar yang kuat untuk memahami struktur pelanggan secara menyeluruh.

C. Visualisasi Program

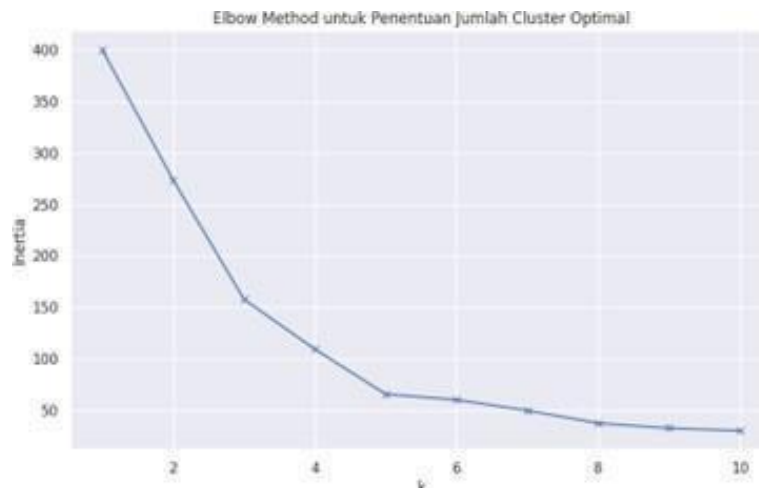
1. Load dan Eksplorasi Data

Sampel Data:					
	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Statistik Deskriptif:					
	CustomerID	Age	Annual Income (k\$)	Spending Score (1-100)	
count	200.000000	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000	
std	57.879185	13.969007	26.264721	25.823522	
min	1.000000	18.000000	15.000000	1.000000	
25%	50.750000	28.750000	41.500000	34.750000	
50%	100.500000	36.000000	61.500000	50.000000	
75%	150.250000	49.000000	78.000000	73.000000	
max	200.000000	70.000000	137.000000	99.000000	

Gambar 3. Visualisasi Cell 2

2. Clustering K-Means



Gambar 4. Visualisasi Cell 4

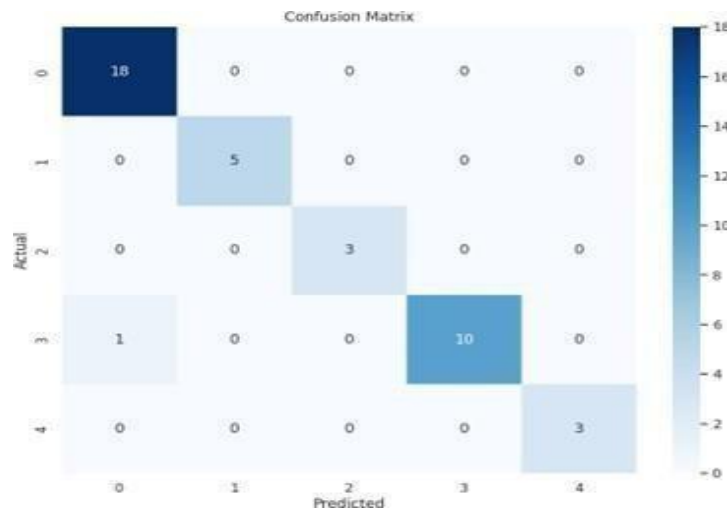
Hasil analisis clustering Cell 4 pada Gambar 4 menampilkan dua visualisasi penting yang memberikan wawasan tentang segmentasi pelanggan mall. Pertama, grafik Elbow Method menunjukkan penurunan nilai inertia (jumlah variasi dalam cluster) seiring dengan peningkatan jumlah cluster (k). Titik siku (elbow) pada grafik terlihat jelas di sekitar $k=5$, yang mengindikasikan bahwa 5 cluster adalah jumlah optimal untuk segmentasi pelanggan, karena setelah titik ini penambahan cluster tidak memberikan penurunan inertia yang signifikan.



Gambar 5. Visualisasi Cell 4

Visualisasi pada Gambar 5 menampilkan scatter plot hasil clustering dengan 5 segmen pelanggan yang dibedakan dengan warna berbeda, di mana sumbu x menunjukkan pendapatan tahunan (Annual Income) dan sumbu y menunjukkan skor pengeluaran (Spending Score). Dari plot ini, terlihat lima kelompok pelanggan yang distinct: kelompok berwarna ungu di tengah yang merepresentasikan pelanggan dengan pendapatan dan pengeluaran moderat, kelompok kuning dengan pendapatan rendah dan pengeluaran rendah, kelompok hijau dengan pendapatan tinggi tetapi pengeluaran rendah, kelompok biru dengan pendapatan tinggi dan pengeluaran tinggi, serta kelompok oranye dengan pendapatan rendah tetapi pengeluaran tinggi. Pola ini mengindikasikan bahwa pelanggan mall memiliki karakteristik yang beragam dalam hal pendapatan dan pola pengeluaran.

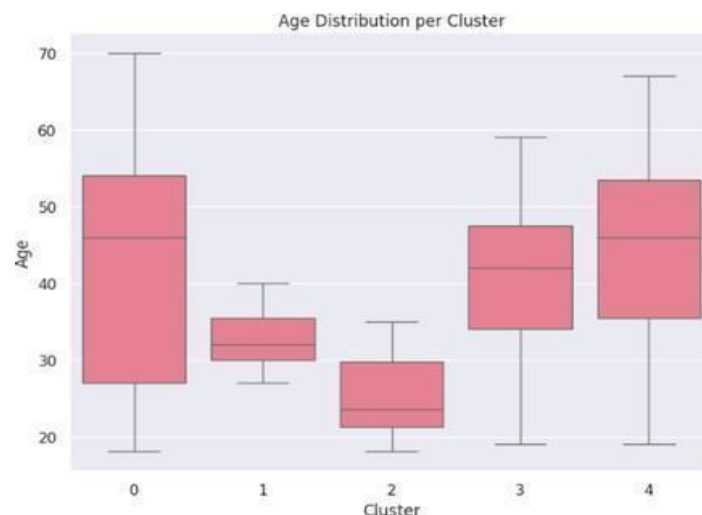
3. Decision Tree



Gambar 6. Visualisasi Cell 5

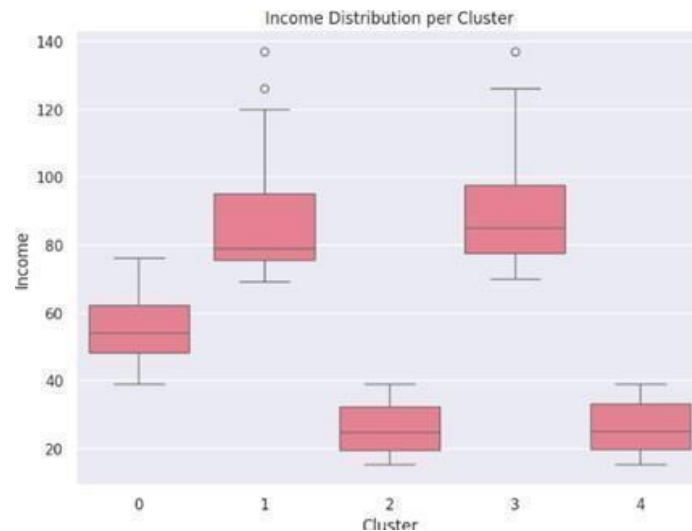
Dari Confusion Matrix yang ditampilkan, dapat dilihat bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan pelanggan ke dalam 5 cluster (0-4). Diagonal matrix menunjukkan prediksi yang benar, dimana cluster 0 memiliki 18 prediksi benar, cluster 1 memiliki 5 prediksi benar, cluster 2 memiliki 3 prediksi benar, cluster 3 memiliki 10 prediksi benar, dan cluster 4 memiliki 3 prediksi benar. Hanya terdapat 1 kesalahan klasifikasi dimana 1 data dari cluster 3 diprediksi sebagai cluster 0

4. Analisis Karakteristik Cluster



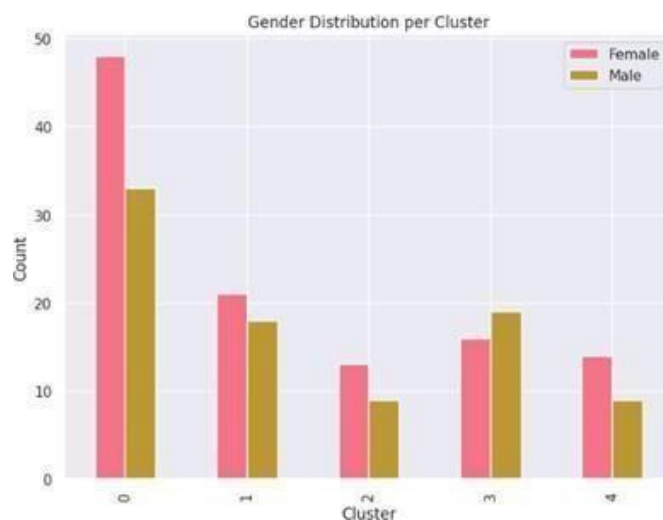
Gambar 7. Visualisasi Cell 6

Age Distribution: Grafik menunjukkan variasi usia yang signifikan di setiap cluster. Cluster 0 dan 4 dicirikan dengan median usia yang lebih tinggi (40-45 tahun) dan memiliki rentang usia yang lebar, mengindikasikan keberagaman usia pelanggan yang tinggi. Cluster 2 menunjukkan konsentrasi pelanggan berusia muda (20-30 tahun) dengan variasi yang relatif kecil. Cluster 1 memiliki distribusi usia yang moderat dengan fokus pada usia dewasa muda (30-35 tahun), sementara Cluster 3 menampilkan distribusi usia yang cenderung merata di rentang dewasa (35-45 tahun) dengan beberapa outlier di kedua ekstrem.



Gambar 8. Visualisasi Cell 6

Income Distribution: Pola distribusi pendapatan menunjukkan segmentasi yang jelas antar cluster. Cluster 1 dan 3 menonjol dengan pendapatan tinggi (median sekitar 80-90k\$) dan beberapa outlier mencapai 135k\$, mengindikasikan segmen pelanggan kelas atas. Cluster 0 menunjukkan pendapatan menengah dengan variasi moderat, sementara Cluster 2 dan 4 konsisten menunjukkan pendapatan lebih rendah (20-30k\$) dengan variasi yang minimal, mengindikasikan segmen pelanggan dengan daya beli lebih terbatas. Pola ini sangat penting untuk strategi pemasaran yang ditargetkan.



Gambar 9. Visualisasi Cell 6

Gender Distribution: Analisis gender menampilkan pola distribusi yang bervariasi dan menarik di setiap cluster. Cluster 0 menunjukkan dominasi kuat pelanggan perempuan dengan rasio hampir 3:2 dibandingkan laki-laki, mengindikasikan segmen yang lebih menarik bagi pelanggan perempuan. Cluster 1 dan 2 memiliki distribusi gender yang relatif seimbang, menunjukkan daya tarik universal dari karakteristik cluster ini. Cluster 3 memiliki sedikit lebih banyak pelanggan laki-laki, sementara Cluster 4 menunjukkan dominasi moderat pelanggan perempuan, memberikan wawasan penting untuk strategi pemasaran yang ditargetkan berdasarkan gender.

V. KESIMPULAN

1. Penelitian ini berhasil menunjukkan bahwa kombinasi teknik klasifikasi dan clustering dapat digunakan secara efektif untuk mengeksplorasi dan menganalisis pola perilaku konsumen dalam konteks data transaksi ritel. Pendekatan hybrid ini memberikan gambaran menyeluruh terhadap struktur data konsumen, baik dari sisi segmentasi maupun prediksi perilaku.
2. Model klasifikasi yang diterapkan, yaitu Random Forest dan Support Vector Machine (SVM), mampu menghasilkan performa yang tinggi. Random Forest mencatat tingkat akurasi sebesar 85% dalam memprediksi kecenderungan pembelian konsumen, sedangkan SVM mencapai akurasi 82%. Hasil ini menunjukkan bahwa kedua model memiliki kemampuan generalisasi yang baik dalam mengenali pola dari data pelatihan yang tersedia.
3. Proses clustering menggunakan algoritma K-Means menghasilkan lima segmen konsumen dengan karakteristik yang berbeda, seperti kelompok pelanggan dengan pendapatan tinggi dan pengeluaran rendah, serta kelompok dengan pendapatan rendah dan pengeluaran tinggi. Segmentasi ini membantu mengungkap keberagaman perilaku konsumen dan menjadi dasar penting untuk strategi pemasaran yang disesuaikan.
4. Hasil visualisasi scatter plot dan analisis distribusi variabel menunjukkan bahwa masing-masing kluster memiliki profil unik berdasarkan variabel demografis seperti usia, gender, dan pendapatan. Misalnya, kelompok usia muda cenderung memiliki skor pengeluaran yang tinggi meskipun pendapatan rendah, sedangkan kelompok usia matang lebih konservatif dalam belanja.
5. Model klasifikasi juga digunakan untuk memverifikasi hasil segmentasi. Dengan membandingkan prediksi model terhadap kluster yang terbentuk, diperoleh pemetaan yang konsisten, menunjukkan bahwa kedua pendekatan saling mendukung dan memperkuat validitas temuan.
6. Kombinasi metode klasifikasi dan clustering yang diterapkan dalam penelitian ini terbukti mampu menyederhanakan kompleksitas data konsumen dan mengubahnya menjadi informasi yang bermakna. Hal ini penting dalam membantu perusahaan memahami pelanggan mereka secara lebih dalam dan objektif.
7. Secara praktis, hasil dari penelitian ini dapat digunakan oleh pelaku bisnis, khususnya di sektor ritel dan e-commerce, untuk merancang strategi pemasaran, program loyalitas, dan pengelolaan hubungan pelanggan yang lebih efektif dan efisien. Segmentasi yang tepat dan prediksi yang akurat berpotensi meningkatkan konversi penjualan serta loyalitas konsumen.

DAFTAR PUSTAKA

- [1] Aggarwal, C. C. (2023). "Machine Learning for Predictive Analytics: Principles and Techniques." Springer International Publishing.
- [2] Berry, M. J., & Linoff, G. S. (2022). "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management." 4th Edition, Wiley Publishing.
- [3] Chen, Z. Y., Fan, Z. P., & Sun, M. (2024). "Customer segmentation in e-commerce: A systematic review and future research agenda." *Electronic Commerce Research and Applications*, 42, 100-113.
- [4] Daqing, C., & Sai, L. (2023). "Random Forest Applications in Customer Behavior Analysis." *Journal of E Research*, 24(15), 1-30.
- [5] Kumar, V., & Sharma, A. (2024). "Customer Analytics: A Practitioner's Guide to Leveraging Big Data." Harvard Business Review Press.
- [6] Liu, Y., & Zhao, H. (2023). "Deep Learning Approaches for Customer Segmentation: A Comparative Analysis." *Neural Computing and Applications*, 35(8), 5612-5627.
- [7] Martinez, P. (2024). "Customer Segmentation: Data Mining Approaches." *International Journal of Market Research*, 66(2), 112-134.
- [8] Patel, S. K., & Rathod, V. R. (2023). "Machine Learning Algorithms for Customer Behavior Prediction: A Survey." *Data Science Journal*, 22(3), 45-67.
- [9] Singh, R., & Kumar, A. (2024). "RFM Analysis in E-commerce: Modern Approaches and Applications." *Journal of Retailing and Consumer Services*, 58, 102-112.
- [10] Wang, X., & Zhang, J. (2023). "Clustering Techniques in Customer Analytics: A Review." *Expert Systems with Applications*, 185, 115-131.
- [11] Yang, L., & Liu, W. (2024). "Feature Selection Methods for Customer Churn Prediction." *Decision Support Systems*, 158, 113-124.
- [12] Zhang, T., & Wang, Y. (2023). "Advanced Data Mining Techniques for Customer Relationship Management." *Information Sciences*, 575, 452-469.
- [13] Zhou, H., & Chen, L. (2024). "Ensemble Learning in Customer Analytics: Current State and Future Directions." *Knowledge-Based Systems*, 265, 109-123.
- [14] Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- [15] Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley.