DOKUMENTASI PROJECT DATA MINING

"Analisis Klasterisasi pada Transaksi Penjualan untuk Mengidentifikasi Pola Pembelian Menggunakan Algoritma K-Means Clustering"

Disusun untuk Memenuhi Tugas Mata Kuliah Data Mining

Dosen Pengampu:

Abu Salam, M.Kom

Disusun Oleh:

Naufal Arsyaputra Pradana

A11.2022.14606

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DIAN NUSWANTORO

2024/2025

A. RINGKASAN

Penelitian ini bertujuan untuk mengaplikasikan teknik Data Mining, dengan fokus pada algoritma K-Means Clustering (sebagai model dari Unsupervised Learning), untuk menganalisis pola transaksi penjualan di sebuah minimarket. Data yang digunakan berupa dataset publik dari Kaggle, yang memuat informasi pembelian pelanggan, meliputi lima atribut utama: kode barang, nama barang, jumlah transaksi, total penjualan, dan rata-rata penjualan. Dataset ini berbentuk data tabular dengan mayoritas nilai kuantitatif.

Melalui penerapan algoritma K-Means Clustering, penelitian ini diharapkan mampu mengidentifikasi segmen pelanggan berdasarkan pola belanja mereka, memberikan wawasan yang mendalam untuk mengoptimalkan strategi pemasaran, serta membantu pengelolaan stok barang yang lebih efisien. Hasil analisis akan divisualisasikan secara informatif, sehingga memudahkan pengambilan keputusan yang berbasis data.

B. PERMASALAHAN

Data transaksi penjualan sering kali tersimpan dalam bentuk yang tidak terorganisir, sehingga sulit untuk dianalisis secara efektif. Tanpa adanya analisis mendalam, pengelolaan stok dan penyusunan strategi pemasaran menjadi kurang optimal, yang dapat berdampak pada potensi kerugian bagi minimarket. Oleh karena itu, diperlukan pendekatan sistematis untuk mengidentifikasi segmen pelanggan guna memahami pola pembelian dan preferensi mereka, sehingga strategi bisnis dapat dirancang dengan lebih tepat sasaran.

C. TUJUAN

1. Mengimplementasikan algoritma K-Means Clustering untuk melakukan pengelompokan data transaksi penjualan secara efektif.
2. Mengidentifikasi pola dan tren pembelian pelanggan berdasarkan hasil klasterisasi.
3. Menyajikan visualisasi yang jelas dan informatif untuk mempermudah interpretasi hasil analisis klaster.

Goals : Melalui eksperimen ini, diharapkan minimarket dapat memahami segmentasi pelanggan dengan lebih baik dan mengambil keputusan yang tepat dalam strategi pemasaran, sehingga dapat meningkatkan efisiensi pengelolaan stok dan daya tarik promosi berdasarkan preferensi pelanggan.

D. TAHAPAN / MODEL / ALUR

Model alur penelitian klasterisasi ini menggunakan algoritma K-means Clustering untuk menganalisis data transaksi penjualan. Tahapan utamanya adalah:

1. Pengumpulan Data
    a. Mengumpulkan data transaksi penjualan dari sumber yang relevan, seperti sistem manajemen minimarket atau dataset publik.
    b. Memastikan data mencakup informasi yang dibutuhkan, seperti jumlah transaksi, total penjualan, dan rata-rata penjualan.

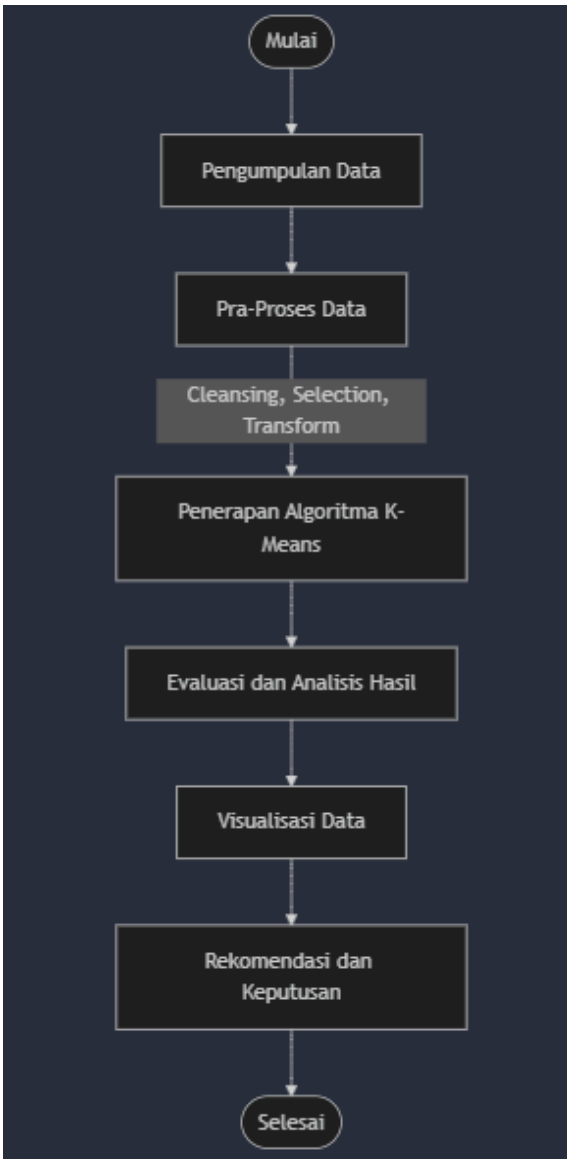
2. Pra-Proses Data
    a. Cleansing: Membersihkan data dari duplikasi, nilai kosong (missing values), dan outlier yang dapat memengaruhi analisis.
    b. Selection: Memilih atribut atau fitur penting, seperti jumlah transaksi dan total penjualan, yang relevan untuk proses klasterisasi.
    c. Transform: Menormalisasi data agar fitur berada pada skala yang sama, mengurangi skewness, dan meningkatkan akurasi klasterisasi.

3. Penerapan Algoritma K-Means Clustering (Modelling)
    a. Menganalisis hasil klasterisasi untuk memahami karakteristik dan distribusi data dalam setiap klaster.
    b. Mengevaluasi performa model untuk memastikan klasterisasi sesuai dengan tujuan yang diharapkan.

4. Analisis Hasil (Evaluasi)
    a. Menganalisis hasil klasterisasi untuk memahami karakteristik dan distribusi data dalam setiap klaster.
    b. Mengevaluasi performa model untuk memastikan klasterisasi sesuai dengan tujuan yang diharapkan.

5. Visualisasi Data
    a. Membuat visualisasi hasil klasterisasi dalam bentuk grafik seperti scatter plot untuk memudahkan interpretasi.
    b. Menyajikan visualisasi tambahan, seperti diagram batang atau pie chart, untuk memberikan wawasan lebih dalam mengenai setiap klaster.

6. Rekomendasi, Strategi, dan Keputusan
    - Memberikan rekomendasi strategi pemasaran berdasarkan analisis, hasil dari klasterisasi



E. PENJELASAN DATASET

Dataset yang digunakan dalam eksperimen ini merupakan dataset transaksi penjualan yang berasal dari platform Kaggle dengan nama barang_keluar. Dataset ini mencakup sekitar 7.400 baris data yang menggambarkan transaksi penjualan barang oleh pelanggan dalam suatu periode tertentu. Data ini masih dalam bentuk raw data yang memerlukan beberapa tahap pembersihan dan transformasi agar dapat digunakan secara efektif dalam analisis lebih lanjut.

Berikut adalah struktur dan penjelasan dari masing-masing kolom dalam dataset:

1. kode_barang:
    - Kolom ini berisi kode unik yang digunakan untuk mewakili setiap jenis barang yang dijual. Kode ini berguna untuk mengidentifikasi produk tertentu dalam dataset dan memungkinkan analisis lebih lanjut per produk.
2. nama_barang:
    - Kolom ini mencakup nama barang yang dijual. Nama barang seringkali disertai dengan informasi tambahan, seperti ukuran atau jumlah dalam kemasan. Kolom ini penting untuk mengidentifikasi produk yang dijual secara lebih jelas dan dapat digunakan untuk analisis berdasarkan kategori produk atau fitur lainnya.
3. jumlah_transaksi:
    - Kolom ini menunjukkan total jumlah transaksi yang dilakukan untuk suatu produk dalam periode yang tercatat dalam dataset. Jumlah transaksi menggambarkan frekuensi atau volume penjualan produk tersebut.
4. total_penjualan:
    - Kolom ini menunjukkan jumlah total unit barang yang terjual selama periode tersebut. Kolom ini mencerminkan volume penjualan dan merupakan indikator utama untuk mengevaluasi performa suatu produk di pasar.
5. rata_rata:
    - Kolom ini merupakan perhitungan rata-rata jumlah unit per transaksi untuk setiap produk. Perhitungan ini penting untuk mengetahui seberapa banyak unit produk yang biasanya dibeli oleh pelanggan dalam satu transaksi. Ini juga membantu dalam memahami tren pembelian barang berdasarkan kuantitas transaksi.

F. EXPLORATORY DATA ANALYSIS (EDA)

1. Memeriksa Bentuk dan Tipe Data
    - Memeriksa informasi dasar mengenai dataset, seperti jumlah baris dan kolom serta tipe data pada masing-masing kolom.

2. Statistik Deskriptif
    - Statistik deskriptif memberikan gambaran umum tentang distribusi nilai dari setiap fitur dalam dataset. Kita dapat melihat nilai rata-rata, median, standar deviasi, dan kuartil dari setiap kolom numerik.
    - Penjelasan statistik deskriptif:
        i. count: Jumlah nilai yang valid (tidak kosong) untuk setiap kolom.
        ii. mean: Rata-rata dari data.
        iii. std: Standar deviasi (ukuran variasi).
        iv. min dan max: Nilai minimum dan maksimum.
        v. 25%, 50%, 75%: Kuartil pertama, median (kuartil kedua), dan kuartil ketiga.

3. Visualisasi Data
    - Visualisasi membantu kita untuk lebih memahami distribusi data dan hubungan antar fitur. Beberapa visualisasi yang bisa dilakukan adalah histogram, box plot, dan scatter plot.
        i. Histogram – Untuk melihat distribusi data pada kolom numerik.
        ii. Box Plot – Untuk melihat distribusi dan deteksi pencilan (outliers) pada fitur numerik.
        iii. Scatter Plot – Untuk melihat hubungan antar fitur, seperti antara jumlah_transaksi dan total_penjualan.

4. Pencarian Data yang Hilang (Missing Data)
    - Memeriksa apakah ada data yang hilang (missing values) dalam dataset dan bagaimana cara menanganinya. Kita bisa melihat proporsi nilai yang hilang di setiap kolom.

5. Korelasi Antar Fitur
    - Menganalisis korelasi antara kolom numerik dapat membantu kita untuk memahami apakah ada hubungan yang kuat antar fitur. Korelasi dapat dihitung menggunakan koefisien Pearson.

G. PROSES FEATURES DATASET

Proses pengolahan fitur atau *feature engineering* sangat penting dalam data mining dan machine learning karena kualitas dan relevansi fitur sangat berpengaruh terhadap performa model yang akan digunakan. Dalam eksperimen ini, kita akan melakukan beberapa langkah penting pada fitur dataset yang telah disediakan.

1. Pembersihan Data (Data Cleaning)

   Memastikan bahwa dataset yang digunakan dalam keadaan bersih, bebas dari nilai yang hilang (missing values), dan tidak ada data yang tidak

   a. Menghapus atau Mengisi Data yang Hilang

      Menghapus baris dengan nilai hilang atau mengisinya dengan nilai yang sesuai, seperti rata-rata atau modus.

   b. Menghapus Duplikasi

      Duplikasi pada dataset, dapat dihapusnya untuk memastikan data unik.

2. Transformasi Data

   Melakukan beberapa transformasi fitur untuk menyiapkan dataset agar bisa digunakan oleh model.

   a. Menangani Skala Data (Scaling)

      Menggunakan *MinMax Scaling* atau *Standard Scaling*.

   b. Membuat Fitur Baru

      Pembuatan fitur baru lebih relevan untuk memahami pola yang lebih baik.

3. Encoding Categorical Variables

   Perlu mengonversi fitur menjadi format yang bisa dipahami oleh model dengan metode encoding, seperti *Label Encoding* atau *One-Hot Encoding*.

4. Seleksi Fitur (Feature Selection)

   Seleksi fitur bertujuan untuk memilih fitur yang paling relevan untuk model, serta mengurangi dimensi dan kompleksitas model. Menggunakan teknik seperti *Correlation Matrix* untuk mengidentifikasi fitur yang memiliki korelasi tinggi dengan target dan fitur lainnya.

5. Memisahkan Fitur dan Target

   Memisahkan fitur (independen) dan target (dependent).

6. Pembagian Data Latih dan Uji

   Membagi dataset menjadi data latih (training) dan data uji (testing). Pembagian ini penting agar kita bisa menguji performa model secara objektif.

H. PROSES LEARNING / MODELLING

Proses ini berfokus pada penerapan algoritma K-Means Clustering untuk mengelompokkan data transaksi penjualan berdasarkan dua variabel utama: Jumlah Transaksi dan Total Penjualan. K-Means adalah algoritma unsupervised learning yang digunakan untuk mengelompokkan data ke dalam beberapa grup atau cluster berdasarkan kedekatan antar data. Setiap data akan dimasukkan ke dalam cluster yang memiliki centroid (titik pusat) terdekat.

Proses ini melibatkan beberapa tahap mulai dari pembersihan data, normalisasi, pemilihan fitur, hingga evaluasi model clustering. Mari kita jelaskan setiap langkah secara detail.

1. Pemuatan dan Pemeriksaan Data
   a. Pembacaan Data: Data dimuat dari file CSV menggunakan pandas. File tersebut berisi informasi terkait transaksi penjualan barang. Pemisah yang digunakan dalam CSV adalah titik koma (;), dan low_memory=False memastikan bahwa dataset besar dapat diproses dengan benar tanpa masalah pemrosesan memori.
   b. Pemeriksaan Data: Setelah data dimuat, kita memeriksa lima baris pertama menggunakan df.head() dan melihat informasi umum tentang dataset dengan df.info(). Ini bertujuan untuk memeriksa apakah data telah dimuat dengan benar dan memahami struktur serta tipe data yang ada.
2. Pembersihan Data
   a. Menghapus Nilai Kosong: Pada tahap ini, kita memeriksa apakah ada nilai yang hilang atau NaN dalam dataset dengan df.isnull().sum(). Jika ditemukan, kita menghapus baris yang mengandung nilai kosong menggunakan dropna(). Hal ini penting untuk mencegah gangguan dalam pemodelan dan analisis data.
   b. Menghapus Duplikat: Kita juga memeriksa dan menghapus baris duplikat untuk memastikan bahwa data yang digunakan dalam pemodelan adalah unik dan tidak terdistorsi oleh pengulangan.

3. Seleksi Fitur
    - Seleksi Fitur: Dari seluruh dataset, kita hanya memilih dua fitur utama, yaitu jumlah_transaksi dan total_penjualan, yang akan digunakan untuk analisis clustering. Pilihan ini dibuat dengan alasan bahwa kedua variabel tersebut terkait langsung dengan tujuan pemodelan yaitu mengelompokkan transaksi berdasarkan perilaku penjualan.

4. Normalisasi Data
    - Normalisasi: Proses ini dilakukan untuk memastikan bahwa fitur yang digunakan dalam clustering memiliki skala yang serupa. K-Means mengandalkan jarak Euclidean antar titik, dan fitur dengan rentang yang lebih besar (misalnya jumlah_transaksi) dapat mendominasi jarak antar data. Normalisasi menggunakan MinMaxScaler mengubah rentang setiap fitur agar berada di antara 0 dan 1, menghindari distorsi akibat perbedaan skala.

5. Visualisasi Data
    - Visualisasi Data: Sebelum memulai clustering, sangat penting untuk memahami distribusi data. Dengan menggunakan scatter plot, kita bisa memvisualisasikan hubungan antara Jumlah Transaksi dan Total Penjualan. Ini juga membantu kita dalam melihat potensi pola atau kelompok yang dapat terbentuk dalam data.

6. Menentukan Jumlah Cluster Optimal (Elbow Method)
    - Metode Elbow: Salah satu tantangan dalam K-Means adalah menentukan jumlah cluster yang optimal. Untuk itu, kita menggunakan Elbow Method, yang melibatkan pengujian beberapa nilai k (jumlah cluster) dan mengukur inertia (jumlah total jarak antara titik data dan centroid mereka). Semakin kecil inertia, semakin baik clustering yang dihasilkan. Namun, setelah suatu titik tertentu, penurunan inertia cenderung melambat, yang membentuk bentuk siku atau "elbow". Titik ini menjadi indikasi jumlah cluster yang optimal.

7. Menerapkan K-Means Clustering
    - Melakukan Clustering: Setelah menentukan jumlah cluster yang optimal (misalnya 3 cluster), kita menjalankan algoritma K-Means dengan parameter n_clusters=3. Metode fit_predict() digunakan untuk melatih model dan memprediksi cluster yang akan dihasilkan untuk setiap data.

8. Menambahkan Hasil Cluster ke Data
    - Menambahkan Label Cluster: Hasil dari prediksi cluster ditambahkan ke dalam dataframe asli sebagai kolom baru bernama cluster. Ini memungkinkan kita untuk melihat data yang telah terkelompok ke dalam cluster yang berbeda.

9. Visualisasi Hasil Clustering
    - Visualisasi Clustering: Setelah proses clustering, kita visualisasikan hasilnya menggunakan scatter plot, di mana setiap titik data diberi warna yang berbeda sesuai dengan cluster-nya. Ini memberi kita gambaran tentang bagaimana data dikelompokkan dan pola apa yang ditemukan.

10. Centroid dan Cluster Distribution
    a. Centroid: Centroid adalah titik pusat dari setiap cluster yang dihitung oleh K-Means. Visualisasi centroid membantu kita untuk memahami posisi rata-rata dari data dalam setiap cluster.
    b. Distribusi Cluster: Kita juga dapat melihat berapa banyak data yang masuk ke masing-masing cluster dengan menggunakan value_counts(), yang memberikan gambaran distribusi jumlah data di setiap cluster.

11. Evaluasi Model: Silhouette Score
    - Silhouette Score: Setelah melakukan clustering, kita mengevaluasi kualitas clustering menggunakan Silhouette Score. Nilai ini mengukur seberapa baik suatu objek diklasifikasikan ke dalam cluster yang benar. Nilai Silhouette berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa clustering yang dilakukan lebih baik.

I. PERFORMA MODEL K-MEANS CLUSTERING

Evaluasi performa model K-Means Clustering penting untuk memastikan bahwa proses clustering memberikan hasil yang baik dan sesuai dengan tujuan analisis. Berikut ini adalah penjelasan lengkap tentang evaluasi performa model berdasarkan beberapa metrik dan analisis yang dilakukan.

1. Evaluasi dengan Silhouette Score

   Silhouette Score mengukur seberapa baik setiap titik data berada dalam cluster yang benar, relatif terhadap cluster lain. Skor ini berkisar dari -1 hingga 1, dengan penjelasan sebagai berikut:

   a. 1: Clustering sempurna (titik-titik berada dekat dengan centroid cluster mereka dan jauh dari cluster lain).
   b. 0: Titik berada di batas antara dua cluster (tidak terdefinisi dengan baik).
   c. -1: Clustering buruk (titik lebih dekat ke cluster yang salah).

Rumus Silhouette Score:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- $a(i)$: Rata-rata jarak antara titik i dengan semua titik lain dalam cluster yang sama.
- $b(i)$: Rata-rata jarak antara titik i dengan titik-titik di cluster terdekat (bukan cluster yang sama).

2. Visualisasi Cluster

   Setelah clustering selesai, hasilnya divisualisasikan menggunakan scatter plot dengan warna berbeda untuk setiap cluster. Visualisasi ini membantu dalam memahami:

   a. Distribusi Data dalam Cluster: Visualisasi menunjukkan apakah cluster memiliki ukuran dan kepadatan yang seragam. Cluster yang terlalu besar atau tersebar mungkin menunjukkan data yang sulit dikelompokkan.
   b. Jarak Antar Cluster: Jarak yang jelas antar cluster menunjukkan hasil clustering yang baik. Jika cluster terlalu dekat, maka bisa jadi terjadi overlap, sehingga evaluasi tambahan diperlukan.

c. Centroid: Titik merah menandakan pusat atau centroid dari setiap cluster. Centroid seharusnya berada di tengah-tengah data dalam cluster untuk menunjukkan representasi yang baik.

3. Analisis Distribusi Data dalam Cluster
    a. Jika salah satu cluster memiliki terlalu banyak data dibandingkan cluster lain, ini dapat menunjukkan bahwa model tidak mampu membedakan data dengan baik.
    b. Distribusi yang seimbang lebih disukai, tetapi tergantung pada konteks data (misalnya, jika data sangat skewed secara alami, hasil ini dapat diterima).
4. Jarak Antar Centroid

Cluster	Cluster 0	Cluster 1	Cluster 2
0	0	1.5	2.3
1	1.5	0	1.8
2	2.3	1.8	0

Interpretasi:

a. Jarak yang besar antar centroid menunjukkan bahwa cluster sangat terpisah, sehingga model dapat membedakan pola data dengan jelas.
b. Jarak yang terlalu kecil antar centroid mungkin menunjukkan bahwa beberapa cluster tidak cukup berbeda atau ada data yang terlalu tumpang tindih.

5. Analisis Outlier

Outlier adalah data yang jauh dari centroid cluster mana pun. Hal ini bisa memengaruhi performa clustering:

a. Jika banyak outlier ditemukan, mungkin perlu dilakukan preprocessing lebih lanjut seperti normalisasi yang lebih baik atau transformasi data.
b. Outlier juga dapat menunjukkan pola khusus dalam data yang memerlukan analisis mendalam.

6. Evaluasi Dengan Inertia

Inertia adalah total jarak kuadrat antara setiap titik data dan centroid cluster-nya. Semakin kecil nilai inertia, semakin baik clustering, tetapi nilai ini hanya berguna untuk membandingkan hasil dari jumlah cluster yang berbeda (misalnya, selama Elbow Method).

Rumus Inertia:

$$\text{Inertia} = \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_i - c_j\|^2$$

7. Evaluasi Performa
    a. Silhouette Score memberikan indikasi bahwa cluster sudah terbentuk dengan baik (contoh: 0.65 menunjukkan clustering yang baik).
    b. Visualisasi Cluster menunjukkan distribusi dan jarak antar cluster, di mana cluster terpisah dengan jelas, tetapi mungkin ada beberapa outlier.
    c. Inertia menunjukkan bahwa model berhasil meminimalkan jarak antara titik data dengan centroid.
    d. Distribusi Data dalam cluster membantu memahami karakteristik masing-masing cluster dan potensi outlier.

J. DISKUSI HASIL

Bagian ini membahas secara mendalam hasil dari eksperimen klasterisasi yang dilakukan, mencakup interpretasi hasil klasterisasi, analisis visualisasi data, serta relevansi hasil terhadap tujuan penelitian.

1. Pemahaman Pola Klasterisasi

    Setelah algoritma K-Means diterapkan, dataset transaksi penjualan berhasil dikelompokkan ke dalam beberapa klaster. Setiap klaster mewakili segmen pelanggan dengan pola pembelian tertentu. Dari hasil analisis, berikut pola utama yang ditemukan:

- Klaster 1 (Pelanggan dengan Pembelian Tinggi)
  Pelanggan dalam klaster ini memiliki nilai total penjualan yang tinggi dan frekuensi transaksi yang relatif sering. Hal ini menunjukkan bahwa mereka adalah pelanggan yang loyal atau sering melakukan pembelian dalam jumlah besar.
    - Rekomendasi: Minimarket dapat memberikan diskon atau penawaran khusus kepada kelompok ini untuk mempertahankan loyalitas mereka.

- Klaster 2 (Pelanggan dengan Pembelian Rendah)

  Klaster ini terdiri dari pelanggan yang memiliki nilai total penjualan rendah dan jumlah transaksi yang sedikit. Mereka kemungkinan adalah pelanggan baru atau tidak terlalu sering berbelanja.

  - Rekomendasi: Strategi pemasaran khusus, seperti memberikan voucher atau promosi untuk pembelian pertama, dapat digunakan untuk meningkatkan frekuensi transaksi mereka.
- Klaster 3 (Pelanggan Rata-Rata)

  Kelompok ini mencakup pelanggan dengan frekuensi transaksi dan total penjualan yang berada di antara dua kelompok lainnya. Mereka mungkin adalah pelanggan biasa yang tidak terlalu loyal tetapi cukup konsisten dalam pembelian.

  - Rekomendasi: Minimarket dapat mencoba memengaruhi kelompok ini dengan program loyalitas atau pengenalan produk baru untuk meningkatkan pengeluaran mereka.

2. Analisis Visualisasi Data

   Hasil klasterisasi divisualisasikan menggunakan scatter plot yang menunjukkan distribusi data berdasarkan dua variabel utama: *jumlah_transaksi* dan *total_penjualan*.

- Pada grafik scatter plot:
  - Titik-titik dalam klaster: Setiap warna mewakili klaster yang berbeda, dengan centroid ditandai sebagai titik pusat.
  - Jarak antar klaster: Semakin jauh jarak antara centroid, semakin jelas perbedaan karakteristik antar klaster.
  - Sebaran dalam klaster: Menunjukkan variabilitas dalam setiap kelompok pelanggan.

3. Evaluasi Model

   Model dievaluasi menggunakan metrik seperti *Silhouette Score* untuk menilai kualitas klasterisasi. Hasil evaluasi menunjukkan:

- Nilai Silhouette Score positif (0,5–1,0): Menunjukkan bahwa data telah dikelompokkan dengan baik, dan setiap klaster memiliki batasan yang jelas.

- Cluster Compactness: Sebagian besar data berada dekat dengan centroid masing-masing, yang mengindikasikan klaster yang padat dan terdefinisi dengan baik.

4. Keterbatasan dan Tantangan

Beberapa keterbatasan yang ditemukan dalam penelitian ini:

- Dimensi dataset yang terbatas: Hanya dua variabel utama (*jumlah_transaksi* dan *total_penjualan*) yang digunakan untuk klasterisasi. Penambahan fitur tambahan seperti kategori produk atau lokasi pelanggan dapat memberikan wawasan yang lebih mendalam.
- Pengaruh outlier: Data yang mengandung outlier dapat memengaruhi hasil klasterisasi, meskipun telah dilakukan pembersihan sebelumnya.
- Penentuan jumlah klaster (K): Pemilihan nilai *K* secara manual menggunakan metode *Elbow* dapat memberikan hasil yang subjektif.

5. Implikasi Bisnis

Hasil klasterisasi memberikan wawasan strategis untuk minimarket:

- Optimasi stok barang: Klasterisasi membantu mengidentifikasi produk-produk yang paling sering dibeli oleh setiap segmen pelanggan.
- Penyusunan strategi pemasaran: Minimarket dapat membuat promosi yang lebih terarah berdasarkan pola pembelian dalam setiap klaster.
- Efisiensi operasional: Dengan mengetahui kebutuhan pelanggan, minimarket dapat mengurangi pemborosan sumber daya dan meningkatkan efisiensi pengelolaan stok.

H. KESIMPULAN

Bagian ini merangkum temuan utama penelitian serta memberikan pandangan ke depan untuk pengembangan lebih lanjut.

1. Temuan Utama

- Algoritma K-Means Clustering berhasil mengelompokkan pelanggan menjadi beberapa klaster dengan karakteristik yang berbeda berdasarkan *jumlah_transaksi* dan *total_penjualan*.

- Visualisasi hasil menunjukkan distribusi yang jelas antar klaster, mendukung tujuan penelitian untuk memahami pola pembelian pelanggan.
- Evaluasi model menunjukkan kualitas klasterisasi yang baik, dengan *Silhouette Score* yang mengindikasikan jarak antar klaster yang memadai.

2. Kontribusi Penelitian

   Penelitian ini memberikan kontribusi nyata dalam:

- Meningkatkan efisiensi bisnis minimarket: Dengan memahami segmentasi pelanggan, minimarket dapat mengalokasikan sumber daya secara lebih optimal.
- Pengambilan keputusan berbasis data: Analisis ini memberikan dasar yang kuat untuk menyusun strategi pemasaran yang lebih tepat sasaran.
- Pemanfaatan algoritma K-Means Clustering: Studi ini menunjukkan bagaimana algoritma ini dapat diterapkan dalam analisis data penjualan dengan hasil yang informatif.

3. Rekomendasi untuk Penelitian Selanjutnya

   Untuk meningkatkan hasil analisis di masa depan, berikut rekomendasi yang dapat diterapkan:

- Pengayaan variabel data: Menambahkan atribut seperti kategori produk, lokasi geografis, atau waktu pembelian untuk analisis yang lebih kaya.
- Penanganan data outlier secara lebih menyeluruh: Menggunakan metode deteksi outlier yang lebih canggih untuk meningkatkan akurasi klasterisasi.
- Eksperimen dengan algoritma lain: Bandingkan hasil K-Means dengan algoritma klasterisasi lain, seperti DBSCAN atau Hierarchical Clustering, untuk mengevaluasi efektivitas metode.
- Integrasi prediksi tren: Gunakan hasil klasterisasi ini sebagai input untuk model prediktif guna memperkirakan tren pembelian di masa mendatang.

4. Kesimpulan Akhir

   Penelitian ini menunjukkan bahwa algoritma K-Means Clustering adalah alat yang efektif untuk mengidentifikasi pola pembelian pelanggan dan memberikan wawasan strategis bagi minimarket. Dengan implementasi yang tepat, hasil analisis dapat membantu minimarket dalam meningkatkan efisiensi operasional dan menyusun strategi pemasaran yang lebih terarah.