DATA SCIENCE
Telkom University Surabaya

Machine Learning

PENERAPAN ALGORITMA K-MEANS DALAM SEGMENTASI PELANGGAN BERDASARKAN PENDAPATAN DAN JUMLAH PEMBELIAN

Ahmad Ihda Falah Annur (1206220015)
Marsha Trista Aulia (1206220007)
Talitha Rahmadewati Witanto (1206220022)

Team Members

DATA SCIENCE
Telkom University Surabaya

Ahmad Ihda Falah Annur

1206220015

Marsha Trista Aulia

1206220007

Talitha Rahmadewati W.

1206220022

Latar Belakang

DATA SCIENCE
Telkom University Surabaya

Segmentasi pelanggan merupakan strategi penting dalam dunia bisnis untuk memahami karakteristik pelanggan dan menyusun pemasaran yang lebih efektif. Dalam tugas ini, kami menggunakan dataset Customer Personality Analysis dari Kaggle yang memuat informasi seperti pendapatan, jumlah pembelian, status pernikahan, dan tingkat pendidikan. Segmentasi dilakukan menggunakan algoritma K-Means Clustering dengan dua variabel utama, yaitu Income dan Purchases, guna mengelompokkan pelanggan ke dalam segmen yang homogen. Tujuannya adalah untuk memberikan wawasan bagi perusahaan dalam merancang strategi pemasaran, produk, dan layanan yang lebih personal, sehingga dapat meningkatkan kepuasan pelanggan serta profitabilitas bisnis.

Informasi Dataset

Dataset digunakan untuk melakukan analisis kepribadian pelanggan, yang bertujuan untuk memahami karakteristik, perilaku, dan preferensi pelanggan secara lebih mendalam. Melalui analisis ini, perusahaan dapat mengembangkan strategi pemasaran dan pengembangan produk yang lebih terarah sesuai dengan segmen pelanggan tertentu.

Nama Kolom	Deskripsi
ID	Identitas unik masing-masing pelanggan
Year_Birth	Tahun kelahiran pelanggan
Education	Tingkat pendidikan terakhir yang dicapai
Marital_Status	Status pernikahan pelanggan
Income	Pendapatan tahunan pelanggan
Kidhome	Jumlah anak kecil dalam rumah tangga
Teenhome	Jumlah remaja dalam rumah tangga
Dt_Customer	Tanggal pertama kali menjadi pelanggan di perusahaan
Recency	Jumlah hari sejak transaksi terakhir
MntWines	Total pengeluaran untuk produk anggur (wine)
MntFruits	Total pengeluaran untuk buah-buahan
MntMeatProducts	Total pengeluaran untuk produk daging
MntFishProducts	Total pengeluaran untuk produk ikan
MntSweetProducts	Total pengeluaran untuk produk manis
MntGoldProds	Total pengeluaran untuk produk emas
NumDealsPurchases	Jumlah pembelian yang dilakukan dengan diskon atau penawaran khusus
NumWebPurchases	Jumlah pembelian melalui situs web perusahaan
NumCatalogPurchases	Jumlah pembelian melalui katalog
NumStorePurchases	Jumlah pembelian di toko fisik
NumWebVisitsMonth	Jumlah kunjungan ke situs web per bulan
AcceptedCmp1 - AcceptedCmp5	Indikator (1 atau 0) apakah pelanggan menerima kampanye pemasaran ke-1 hingga ke-5
Complain	Indikator apakah pelanggan pernah mengajukan keluhan (1 = ya, 0 = tidak)
Z_CostContact	Biaya tetap yang dikeluarkan untuk menghubungi pelanggan
Z_Revenue	Pendapatan tetap dari keberhasilan kampanye
Response	Indikator apakah pelanggan merespons kampanye terakhir (1 = ya, 0 = tidak)

Eksplorasi dan Persiapan Data

DATA SCIENCE
Telkom University Surabaya

Cek Missing Value:

Dataset yang digunakan terdiri dari 2.240 baris dan 29 kolom. Setelah pemeriksaan, hanya kolom “Income” yang memiliki missing value sebanyak 24 data. Untuk mengatasinya, kami menggunakan metode KNN Imputer dengan parameter n_neighbors=5, karena metode ini mengisi nilai hilang berdasarkan rata-rata dari data terdekat yang memiliki kemiripan fitur. Pendekatan ini dianggap lebih akurat dibandingkan pengisian sederhana menggunakan mean atau median keseluruhan.

Rename Column:

Rename column dilakukan untuk mengganti nama kolom agar lebih mudah dipahami, seperti: kolom “Recency” diubah menjadi “Last_Purchase_Days”, dan lain-lain

Deskripsi Data:

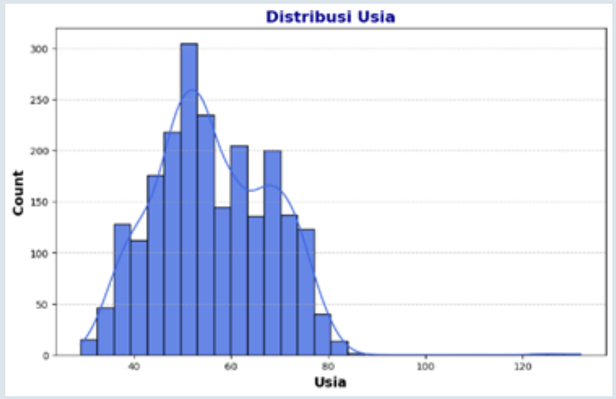
Menggunakan df.describe() untuk mengetahui jumlah data, rata-rata, standar deviasi, nilai minimum, quartil, dan maksimum yang dimiliki oleh setiap fitur yang ada.

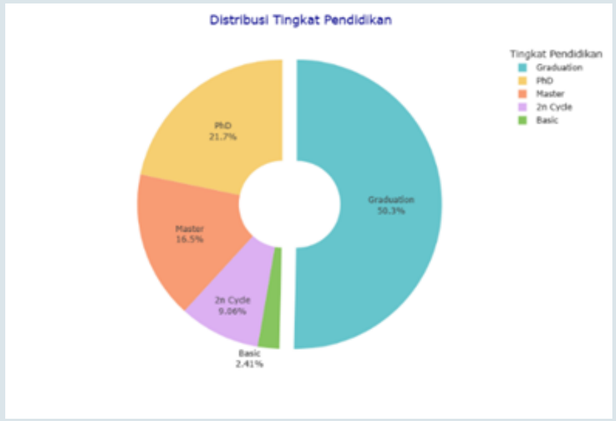
Eksplorasi dan Persiapan Data

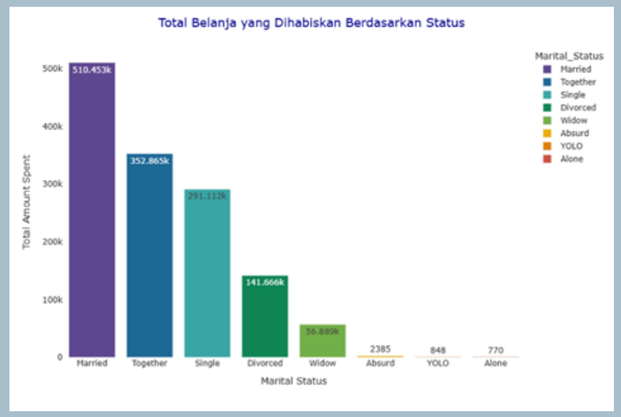
DATA SCIENCE
Telkom University Surabaya

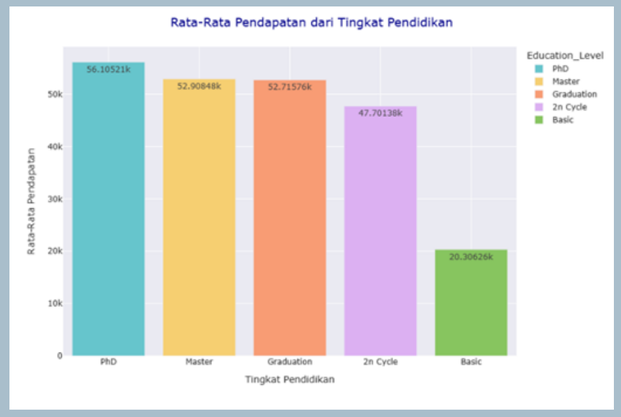
Visualisasi Eksploratif:

Kami juga melakukan visualisasi data untuk memperoleh pemahaman yang lebih baik terhadap fitur-fitur dalam dataset. Visualisasi ini membantu mengidentifikasi pola, distribusi, dan hubungan antar variabel yang mendukung proses segmentasi pelanggan secara lebih efektif.









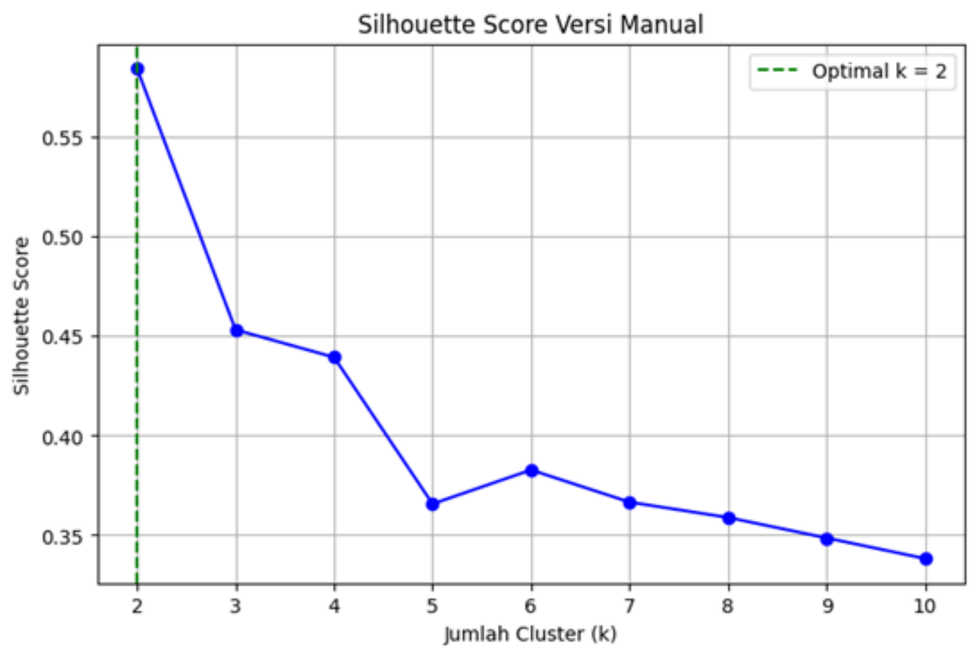
Pemodelan

DATA SCIENCE
Telkom University Surabaya

1. Menentukan Jumlah Cluster (k)

- **Silhouette Score**

  Menilai seberapa baik data dikelompokkan dalam klaster-nya sendiri.



Uji Silhouette Score menunjukkan bahwa nilai silhouette score tertinggi ada di k = 2 dengan nilai 0.5890. Penurunan drastis nilai silhouette score terjadi setelah k = 2, yang mengindikasikan kualitas klaster semakin rendah

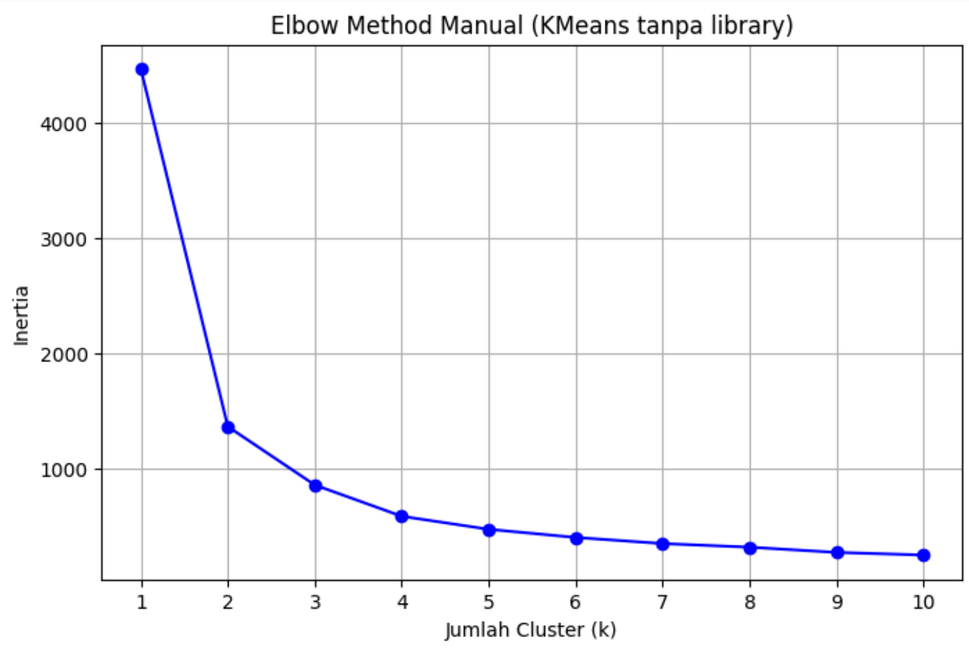
Pemodelan

DATA SCIENCE
Telkom University Surabaya

1. Menentukan Jumlah Cluster (k)

- **Elbow Method**

  Mengukur nilai inertia untuk k dari 1 sampai 10 dan melihat titik siku grafik



Uji Elbow Method menunjukkan bahwa terjadi penurunan inertia yang sangat signifikan dari k = 1 ke k = 2. Penurunan mulai melambat (membentuk “siku” atau elbow) terutama di k = 2 hingga k = 4

Pemodelan

DATA SCIENCE
Telkom University Surabaya

2. Penerapan K-Means Clustering

Langkah-langkah Implementasi K-Means Manual

- **Pengambilan Fitur**

  Fitur yang digunakan: Annual Income dan Purchases.

- **Menentukan Parameter Awal**

  Jumlah kluster = 2, sesuai dengan uji yang dilakukan.

- **Inisialisasi Centroid Awal**

  Menggunakan np.random.choice

- **Proses Iteratif K-Means**

  a. Menghitung Jarak Euclidean $\longrightarrow d = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$

  b. Menentukan Klaster Terdekat

  c. Memperbarui Posisi Centroid

  d. Cek Konvergensi

  e. Update Centroid

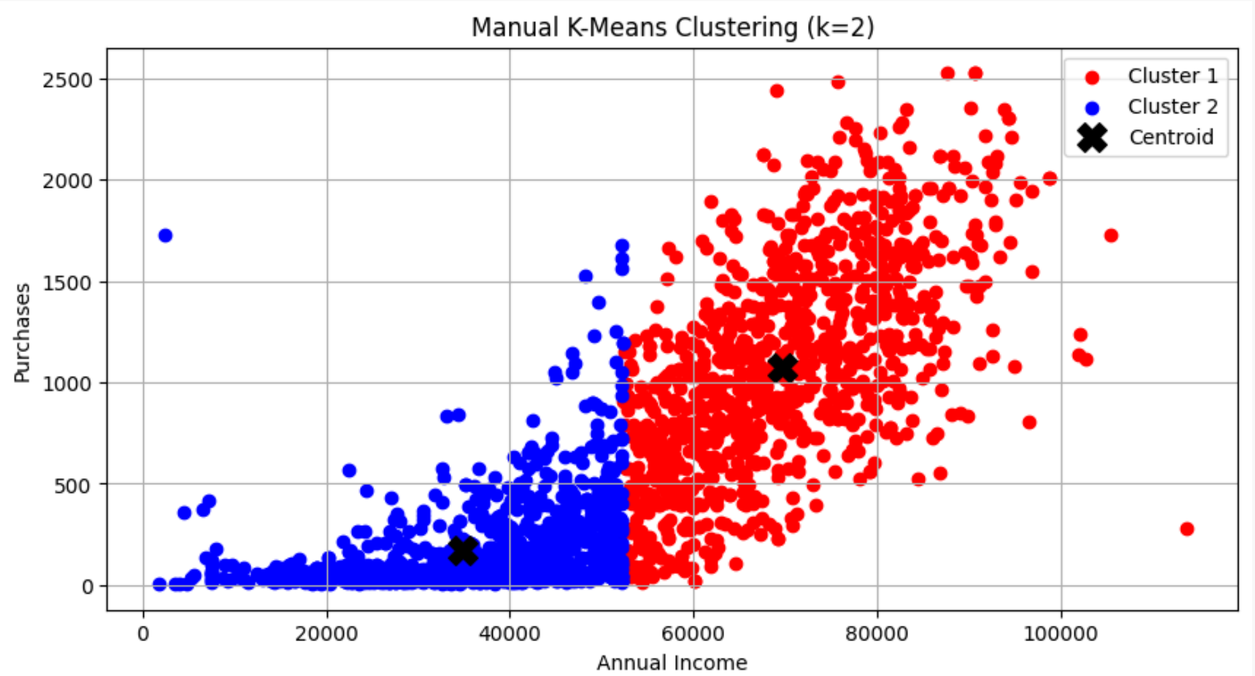
- **Visualisasi Hasil Clustering**

Pemodelan

DATA SCIENCE
Telkom University Surabaya

Hasil Clustering

Setelah proses K-Means selesai, hasil klasterisasi divisualisasikan dalam bentuk scatter plot. Titik-titik data akan diberi warna sesuai dengan klasternya, dan posisi centroid ditandai dengan simbol X berwarna hitam.



Visualisasi menunjukkan bahwa data berhasil dikelompokkan menjadi dua klaster yang jelas:

- Cluster 1 (Merah): Pelanggan dengan pendapatan tinggi dan pembelian tinggi (high-value customers).
- Cluster 2 (Biru): Pelanggan dengan pendapatan rendah dan pembelian rendah.

Evaluasi

DATA SCIENCE
Telkom University Surabaya

Untuk mengevaluasi performa model **K-Means Clustering** yang dibangun secara manual, kami menggunakan dua metode utama:

1. **Silhouette Score,** yang menghasilkan nilai 0.5890 untuk jumlah klaster k = 2, menunjukkan pemisahan klaster yang cukup baik dan stabil.
2. **Visualisasi Scatter Plot** berdasarkan fitur Annual Income dan Purchases, menunjukkan hasil klaster yang jelas terpisah: Cluster 1: pelanggan dengan pendapatan dan pembelian tinggi, Cluster 2: pelanggan dengan pendapatan dan pembelian rendah

Evaluasi ini membuktikan bahwa model mampu mengelompokkan pelanggan secara efektif sesuai karakteristik utamanya.

Evaluasi

DATA SCIENCE
Telkom University Surabaya

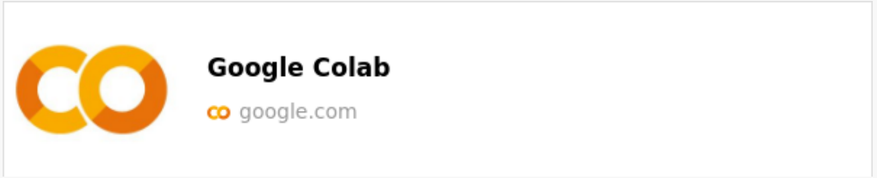
Justifikasi Pemilihan Metode Evaluasi

- Silhouette Score = Digunakan untuk menilai seberapa baik data dikelompokkan dalam clusternya. Nilai mendekati 1 menunjukkan pemisahan antar cluster yang baik dan segmentasi yang jelas. Cocok untuk memastikan segmen pelanggan homogen dan terpisah dengan baik.
- Elbow Method = Membantu menentukan jumlah cluster (k) yang optimal dengan melihat titik perubahan signifikan pada grafik WCSS. Metode ini memastikan keseimbangan antara jumlah cluster yang efisien dan kemiripan dalam tiap cluster.

Eksperimen

DATA SCIENCE
Telkom University Surabaya

Terdapat beberapa eksperimen yang telah kami lakukan. Penjelasan lebih lengkap dapat dilihat langsung melalui kode program dalam file notebook (ipynb) yang tersedia di Google Colab.



Kesimpulan

DATA SCIENCE
Telkom University Surabaya

Kami berhasil menerapkan K-Means Clustering untuk segmentasi pelanggan berdasarkan Annual Income dan Purchases, dengan jumlah klaster optimal k = 2, yang menunjukkan pemisahan jelas antara pelanggan bernilai rendah dan tinggi (high-value customers). Model K-Means manual berhasil konvergen dalam 8 iterasi dan visualisasi menunjukkan korelasi positif antara pendapatan dan pembelian. Pada eksperimen lanjutan dengan fitur demografi dan perilaku, jumlah klaster optimal adalah k = 9, dengan konvergensi dalam 27 iterasi, menghasilkan segmentasi pelanggan yang lebih detail dan spesifik.

DATA SCIENCE
Telkom University Surabaya

Thank you